

Noise-Consistent Siamese-Diffusion for Medical Image Synthesis and Segmentation

Kunpeng Qiu^{1,2†}, Zhiqiang Gao^{3*}, Zhiying Zhou^{1,2†}, Mingjie Sun^{4*}, Yongxin Guo^{1,2,5*}

¹National University of Singapore, ²National University of Singapore Suzhou Research Institute, ³Wenzhou-Kean University

⁴Soochow University, ⁵City University of Hong Kong

kunpeng-qiu@u.nus.edu, zgao@kean.edu, elezzy@nus.edu.sg,

mjsun@suda.edu.cn, yongxin.guo@cityu.edu.hk

Abstract

Deep learning has revolutionized medical image segmentation, yet its full potential remains constrained by the paucity of annotated datasets. While diffusion models have emerged as a promising approach for generating synthetic image-mask pairs to augment these datasets, they paradoxically suffer from the same data scarcity challenges they aim to mitigate. Traditional mask-only models frequently yield low-fidelity images due to their inability to adequately capture morphological intricacies, which can critically compromise the robustness and reliability of segmentation models. To alleviate this limitation, we introduce Siamese-Diffusion, a novel dual-component model comprising Mask-Diffusion and Image-Diffusion. During training, a Noise Consistency Loss is introduced between these components to enhance the morphological fidelity of Mask-Diffusion in the parameter space. During sampling, only Mask-Diffusion is used, ensuring diversity and scalability. Comprehensive experiments demonstrate the superiority of our method. Siamese-Diffusion boosts SANet’s mDice and mIoU by 3.6% and 4.4% on the Polyps, while UNet improves by 1.52% and 1.64% on the ISIC2018.

1. Introduction

Recent advances in deep learning have led to unprecedented success in medical image analysis [1, 35, 52, 54]. However, the full potential of segmentation remains untapped due to limited annotated datasets [53]. Unfortunately, medical data annotation is often costly, and acquiring images can be difficult or impossible due to privacy and copyright concerns. To overcome these challenges, diffusion models [21, 36, 45] have emerged as effective tools to generate image mask pairs, widely used to expand data sets for medical image segmentation tasks [14, 33, 40].

*Corresponding Authors. † Equal Contribution.

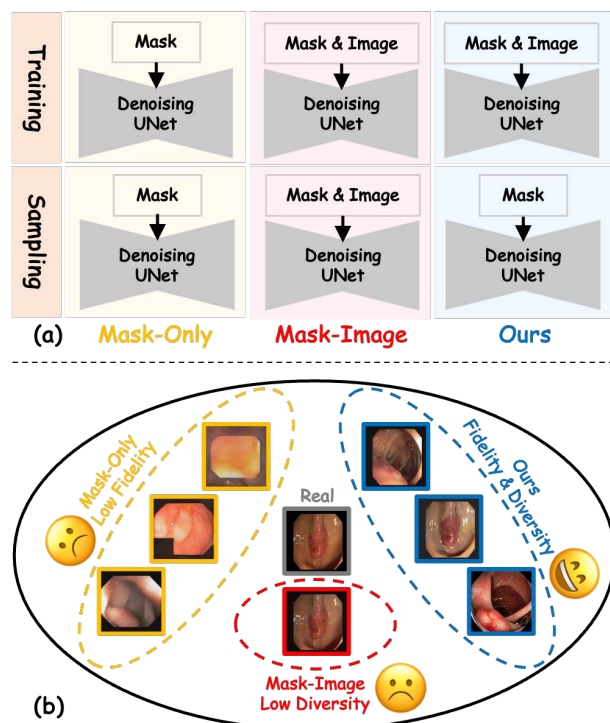


Figure 1. (a) Workflow comparisons between our method and existing ones during the training and sampling phases. (b) Differences in synthesized images across methods. Mask-only methods (yellow) lack morphological characteristics (e.g., surface texture), resulting in low fidelity. Mask-image methods (red) produce high-fidelity images; however, their reliance on extra image prior control results in low diversity and scalability. Our method (blue) enhances morphological fidelity while preserving diversity.

Typically, generating medical images involves using masks to indicate the lesion areas as prior controls injected into the diffusion model [31, 59]. However, obtaining high-fidelity synthetic medical images relies on large-scale train-

ing datasets, creating a paradox: while generative models aim to alleviate data scarcity for segmentation models, they themselves face the same challenge.

As shown in Fig. 1(a), mask-only models primarily focus on aligning the lesion shape to the mask [14, 27], often getting trapped in local minima with low morphological fidelity. Consequently, the medical images synthesized, highlighted by yellow boxes in Fig. 1(b), frequently neglect the indispensable morphological characteristics (*e.g.*, surface texture) [48, 49]. This neglect obscures the discriminative features of the lesions learned by the segmentation models, rendering them unexplainable and elusive, which significantly undermines the reliability of the enhanced segmentation models.

Inspired by [9, 20], incorporating additional prior guidance through the combination of images and masks significantly enhances the morphological fidelity of the synthetic images. A picture is worth a thousand words: these images contain intricate morphological information that even experienced clinicians often struggle to describe accurately. While the joint prior control substantially enhances the fidelity of synthesized morphological characteristics (marked with the red box in Fig. 1(b)), its close resemblance to real data (marked with the gray box in Fig. 1(b)) leads to a disastrous reduction in diversity. Furthermore, the scalability of the synthesized images is constrained by the scarcity of sample pairs, ultimately failing to genuinely address the data scarcity issue for downstream segmentation models.

To overcome the limitations of the aforementioned paradigms and capitalize on their strengths, we propose the resource-efficient *Siamese-Diffusion* model, which trains the same diffusion model under varying prior controls. When guided by a mask alone, this process is termed *Mask-Diffusion*, and when guided by both the image and its corresponding mask, it is called *Image-Diffusion*. To alleviate the inherent issues of *Image-Diffusion*, we introduce a *Noise Consistency Loss*, enabling the noise predicted by *Image-Diffusion* to act as an additional anchor, steering the convergence trajectory of *Mask-Diffusion* toward a local minimum with higher morphological fidelity in the parameter space. Ultimately, only *Mask-Diffusion* is employed during the sampling phase, where the diversity limitation is mitigated by the flexibility of using arbitrary masks. Our contributions are summarized as follows:

- We identify that both current generative models and medical image segmentation models encounter the same data scarcity issue, leading to low morphological fidelity. We argue that only synthesized medical images containing essential morphological characteristics can ensure the reliability of enhanced segmentation models.
- We propose the *Siamese-Diffusion* model, which leverages *Noise Consistency Loss* and *Image-Diffusion* to guide *Mask-Diffusion* towards a local minimum with

high morphological fidelity. During sampling, only *Mask-Diffusion* is used to synthesize medical images with realistic morphological characteristics.

- Extensive experiments demonstrate that our method outperforms existing approaches in both image quality and segmentation performance. SANet improves mDice and mIoU by 3.6% and 4.4% on the Polyps dataset, and UNet achieves gains of 1.52% and 1.64% on the ISIC2018 dataset, highlighting the superiority of our method.

2. Related Work

2.1. Controllable Diffusion Models

In recent years, Diffusion Probabilistic Models [21, 43, 45] have become dominant in high-quality image synthesis. Latent Diffusion Models [36], which later evolved into Stable Diffusion, significantly enhance both the speed and quality of image generation by mapping the diffusion process into a latent feature space [16]. Furthermore, controllable conditions, such as classifier guidance [9, 39], text guidance [20], and structure-image guidance [31, 59], have introduced novel approaches for advancing foundational model research. These techniques have found widespread applications in data-scarce fields, such as medical diagnosis [14, 48] and industrial anomaly detection [12, 13, 23].

2.2. ReferenceNet

Prompts often struggle to precisely control fine details and low-level semantics in generated images [36]. To address this challenge, several models [6, 22, 56, 57] replace complex prompts with exemplar images. These models leverage external hypernetworks [17] to extract rich semantic information, enabling finer control over image generation. However, they prioritize consistency over diversity, limiting their applicability in tasks like medical image segmentation, which demands both morphological diversity and fidelity. Additionally, while some methods synthesize abnormal samples from normal images [4, 41], they still raise ethical concerns. Meanwhile, adopting the IP-Adapter paradigm [58] requires a correlation between the image prior control and noisy images (*e.g.*, color or elements) during training, which is impractical for scarce RGB datasets. In contrast, the mask-only method offers a more accessible and streamlined framework, making it a compelling alternative for broader adoption [11, 14, 40].

3. Method

3.1. Preliminary

Diffusion models [21, 44] comprise a diffusion process and a denoising process. The diffusion process $q(x_{1:T}|x_0)$ constitutes a fixed T -timestep Markov chain devoid of trainable parameters. The original input x_0 is iteratively perturbed by

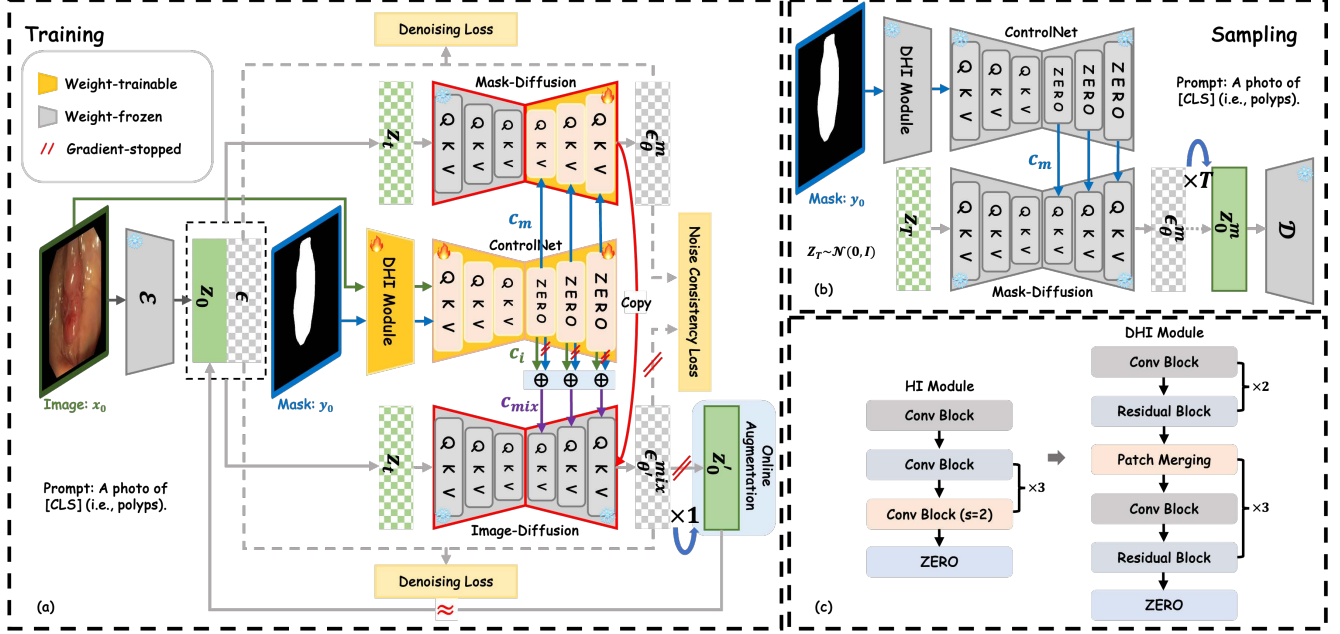


Figure 2. (a) Illustration of our method during the training phase. The noisy image input z_t in the latent space is processed through the same diffusion model (i.e., “Copy”) under the mask and image-mask conditions, generating the noise predictions ϵ_θ^m and ϵ_θ^{mix} , respectively. These two processes are referred to as Mask-Diffusion and Image-Mask-Diffusion. The entire framework is optimized using the denoising losses from both cases and the proposed *Noise Consistency Loss* between ϵ_θ^m and ϵ_θ^{mix} . The *Online-Augmentation* module employs single-step sampling to obtain the denoised z_0' , which is then recombined with the mask y_0 to train the Mask-Diffusion. (b) During the sampling phase, only Mask-Diffusion is utilized to generate high-fidelity and diverse synthetic images. (c) Replacement of the Hint Input (HI) module with the proposed *Dense Hint Input* (DHI) module enhances the extraction of prior guidance from the image.

Gaussian noise according to a variance schedule $\{\beta_t\}_{t=1:T}$:

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, \quad \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad (1)$$

where ϵ denotes random noise sampled from a Gaussian distribution, $\alpha_t = 1 - \beta_t$, and $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$.

Conversely, the denoising process $p_\theta(x_{t-1}|x_t)$ is trained to reverse the parameterized Markov noise process, transforming the noised Gaussian $p(x_T)$ at timestep T into a reconstruction of the original data x_0 , as follows:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)), \quad (2)$$

where $\mu_\theta(x_t, t)$ represents the network-predicted mean of the denoised signal at timestep t , $\Sigma_\theta(x_t, t)$ denotes the variance of the Gaussian distribution, which can be fixed to a constant. For simplicity, Ho et al. [21] parameterized this model as the function $\epsilon_\theta(x_t, t)$, predicting the noise component of a noisy sample x_t at any timestep t . The training objective is to minimize the mean squared error loss between the actual noise ϵ and the predicted noise $\epsilon_\theta(x_t, t)$:

$$\mathcal{L} = \mathbb{E}_{x_t, t, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} [\|\epsilon_\theta(x_t, t) - \epsilon\|^2]. \quad (3)$$

Subsequently, Stable Diffusion [36] employs a pre-trained VQ-VAE [50] to encode images into the latent

space, conducting training on the latent representation z_0 . In the context of controlled generation, given a text prompt c_t and task-specific conditions c_f , the diffusion training loss at time step t can be reformulated as:

$$\mathcal{L} = \mathbb{E}_{z_t, t, c_t, c_f, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} [\|\epsilon_\theta(z_t, t, c_t, c_f) - \epsilon\|^2]. \quad (4)$$

3.2. Architecture of Siamese-Diffusion

ControlNet [59] and the pre-trained Stable Diffusion [36] serve as the foundational framework for our method. Following the approach outlined in [59], both the VQ-VAE module and the denoising U-Net encoder module within the Stable Diffusion framework are kept frozen.

During the training phase, as illustrated in Fig. 2(a), the image x_0 is compressed into the latent space z_0 using the VQ-VAE encoder $\mathcal{E}$. The proposed *Dense Hint Input* (DHI) and ControlNet are combined in series to form a feature extraction network. This network encodes the input image x_0 and its corresponding mask y_0 into c_i and c_m . Different prior controls, c_m and c_{mix} , are then injected into the denoising U-Net decoder of the same diffusion model (i.e., “Copy” in Fig. 2(a)) to predict the noise ϵ_θ^m and ϵ_θ^{mix} , respectively. Here, c_{mix} is the mixed features of c_i and c_m :

$$c_{mix} = w_i \cdot c_i + w_m \cdot \text{sg}[c_m], \quad (5)$$

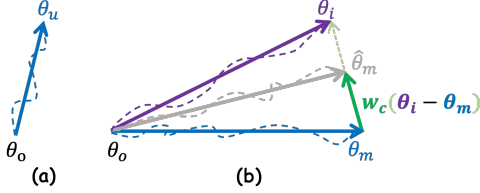


Figure 3. (a) Parameter update direction. (b) *Mask-Diffusion* parameter update direction, scaled by the Noise Consistency Loss.

where $w_i = \frac{k}{N_{iter}}$ and w_m denote the weights for the image and mask prior controls. N_{iter} is the total number of training iterations, and k is the current iteration. For brevity, processes utilizing mask-only (i.e., c_m) and mask-image (i.e., c_{mix}) are referred to as *Mask-Diffusion* and *Image-Diffusion*. In Eq. (5), c_m is the same as that used in *Mask-Diffusion* with truncated gradients. This Siamese architecture distinguishes it from knowledge distillation models [19, 30], which are resource-intensive due to the need for training both a teacher and a student network. Ultimately, the loss employed to train *Siamese-Diffusion* is defined as:

$$\mathcal{L} = \mathcal{L}_m + \mathcal{L}_i + \mathcal{L}_c + \mathcal{L}_{m'}, \quad (6)$$

where $\mathcal{L}_c$ is the proposed Noise Consistency Loss, and $\mathcal{L}_{m'}$ arises from *Online-Augmentation*, both of which will be elaborated upon subsequently. $\mathcal{L}_m$ and $\mathcal{L}_i$ are the Denoising Losses for *Mask-Diffusion* and *Image-Diffusion*:

$$\mathcal{L}_m = \mathbb{E} [||\epsilon_\theta^m - \epsilon||^2], \epsilon_\theta^m = \epsilon_\theta(z_t, t, c_t, c_m), \quad (7)$$

$$\mathcal{L}_i = \mathbb{E} [||\epsilon_{\theta'}^{mix} - \epsilon||^2], \epsilon_{\theta'}^{mix} = \epsilon_{\theta'}(z_t, t, c_t, c_{mix}), \quad (8)$$

where θ is the *Mask-Diffusion* parameter, and θ' is a Deep-Copy of θ . Eq. (7) and Eq. (8) share the same $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

During the sampling phase, as illustrated in Fig. 2(b), only *Mask-Diffusion* is utilized with arbitrary mask prior control to synthesize medical images for segmentation.

3.3. Noise Consistency Loss

As mentioned above, in data-scarce scenarios, *Mask-Diffusion* tends to get trapped in low-fidelity local minima, yielding synthesized medical images that lack indispensable morphological characteristics, such as surface texture. In contrast, *Image-Diffusion*, aided by the additional image prior control x_0 , more effortlessly converges to high-fidelity local optima. However, during the sampling phase, the strong prior control from the image x_0 , which leads to synthesized images closely resembling real images as shown in Fig. 1(b), restricts the diversity of morphological characteristics and limits the scalability due to the scarcity of paired samples. As illustrated in Tab. 2, both cases lead to catastrophic degradation of segmentation performance.

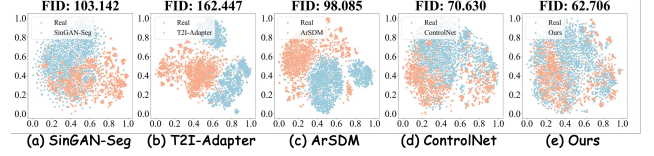


Figure 4. t-SNE visualization of data distribution. (a)–(e) illustrate the distribution differences between real polyp images and those synthesized by each respective mask-only method. The distribution of polyp images generated by our method nearly overlaps with the real data, underscoring its exceptional ability to produce highly realistic polyp images.

To help *Mask-Diffusion* escape from low-fidelity local minima and move toward higher-fidelity regions in the parameter space, we introduce the Noise Consistency Loss:

$$\mathcal{L}_c = w_c \cdot \mathbb{E} [||\epsilon_\theta^m - \text{sg}[\epsilon_{\theta'}^{mix}]]^2], \quad (9)$$

where $\epsilon_{\theta'}^{mix}$, the predicted noise from *Image-Diffusion*, is more accurate than ϵ_θ^m predicted by *Mask-Diffusion*, owing to the additional image prior control. The stop-gradient operation [5] sg enables the more accurate $\epsilon_{\theta'}^{mix}$ to serve as an anchor, guiding the convergence trajectory of *Mask-Diffusion* towards higher-fidelity local minima in the parameter space. The “Copy” operation in Fig. 2(a) and the identical c_m in Eq. (5), jointly ensure that the strong prior control provided by the additional image is successfully propagated to *Mask-Diffusion*. Meanwhile, $w_i = \frac{k}{N_{iter}}$ gradually increases, with Denoising Loss dominating the early stages to ensure stable convergence. w_c controls the steering strength, as discussed in Sec. 5.3.2.

Fig. 3 visually illustrate the above operation in parameter space. Using arithmetic operations [24, 34], the update direction of the parameters can be depicted by the difference between the original model parameters θ_o , and the updated model parameters θ_u . Therefore, the final update direction for the *Mask-Diffusion* parameters can be approximated as:

$$\hat{\theta}_m = \theta_m + w_c \cdot (\theta_i - \theta_m), \quad (10)$$

where θ_m and θ_i are the parameters of *Mask-Diffusion* and *Image-Diffusion*, and $\hat{\theta}_m$ is the refined *Mask-Diffusion* parameter. Throughout this process, collaborative updates to the feature extraction network and the diffusion model maintain the alignment of mask features c_m with noisy image features z_t in the latent space, ensuring the fused latent features reside within a suitable manifold. Consequently, *Mask-Diffusion* can operate independently during sampling. As shown in Fig. 1(b), the morphological characteristics of the synthesized images are enhanced by the refined *Mask-Diffusion*, making these images competitive with those synthesized by *Image-Diffusion*. Segmentation models enhanced with these images ensure interpretability and reliability. Unlike the classifier-free guid-

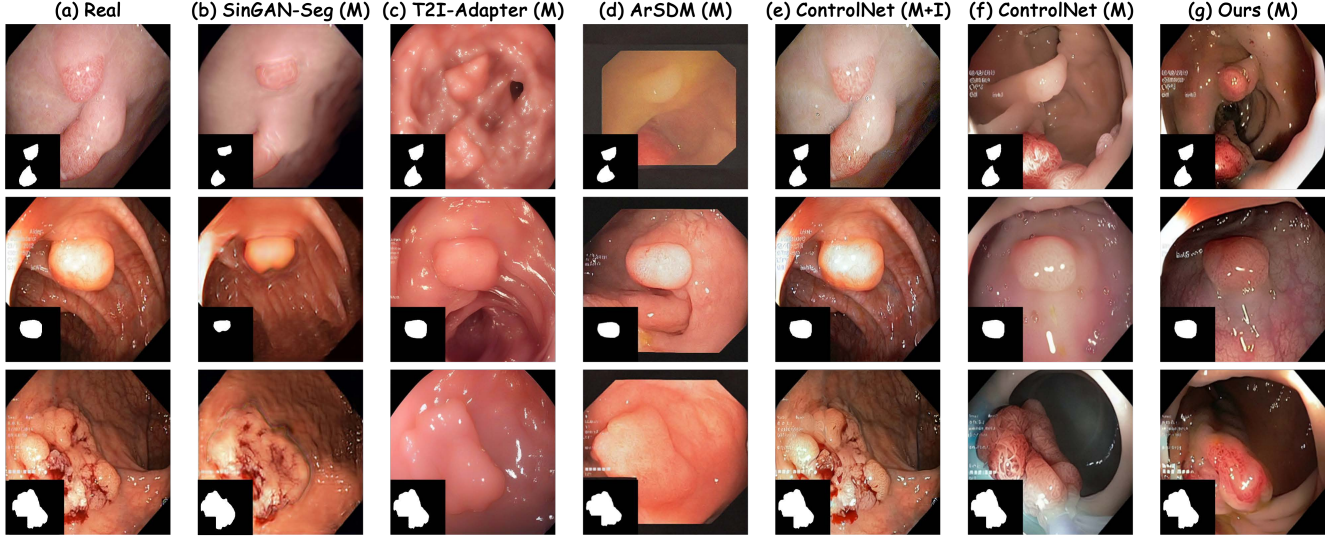


Figure 5. (a) Examples of real polyp images. (b)–(g) Examples of synthetic polyp images generated by each respective method. “M” denotes that mask-only prior control, while “M+I” denotes mask-image joint prior control. The synthetic polyp images generated by our method achieve competitive morphological fidelity while also exhibiting morphological diversity (Zoom in for better visualization).

ance (CFG) [20], which improves fidelity by adjusting guidance strength in the sample space, our proposed *Siamese-Diffusion* enhances morphological fidelity within the parameter space. This distinction eliminates the limitations imposed by the reliance on paired controls during sampling.

3.4. Dense Hint Input Module

The prior controls used in [59] are typically low-density semantic images, such as segmentation masks, depth maps, and sketches, from which features can be effortlessly extracted using a sparse Hint Input (HI) module. However, when handling high-density semantic images, such as medical images, the original sparse HI module is insufficient for capturing nuanced details, including texture and color. Thus, a *Dense Hint Input* (DHI) module is introduced. As illustrated in Fig. 2(c), the DHI module incorporates denser residual blocks with channel sizes of 16, 32, 64, 128 and 256, along with a more advanced patch merging module [28]. This design enable the DHI module effectively accommodates both image and mask prior controls [56].

3.5. Online-Augmentation

Leveraging the advantages of the *Siamese-Diffusion* training paradigm, *Online-Augmentation* is introduced to expand the *Mask-Diffusion* training set, as shown in Fig. 2(a). The additional image prior control x_0 improves the accuracy of noise predicted by *Image-Diffusion*, enabling single-step sampling [27, 44] to generate z'_0 online. In the first step, z'_0 is obtained through the reverse process of Eq. (1), with x_t replaced by z_t , as follows:

Table 1. Comparison of synthetic polyp image quality generated by each respective mask-only method, evaluated using FID [18], KID [3], CLIP-I [38], LPIPS [60], CMMD [25], and MOS metrics.

Methods	FID (↓)	KID (↓)	CLIP-I (↑)	LPIPS (↓)	CMMD (↓)	MOS (Confidence) (↑)
SinGAN-Seg [47]	103.142	0.0898	0.851	0.639	0.929	0.233 (8.25)
T2I-Adapter [31]	162.447	0.1517	0.875	0.672	0.797	-
ArSDM [14]	98.085	0.0927	0.845	0.727	0.811	-
ControlNet [59]	70.630	0.0509	0.887	0.612	0.543	0.487 (5.94)
Ours	62.706	0.0395	0.892	0.586	0.515	0.587 (6.04)

$$z_0 \approx z'_0 = \frac{z_t - \sqrt{1 - \bar{\alpha}_t} \text{sg}[\epsilon_{\theta'}^{mix}]}{\sqrt{\bar{\alpha}_t}}, \quad (11)$$

where $\text{sg}[\epsilon_{\theta'}^{mix}]$ is the noise predicted by *Image-Diffusion* with truncated gradients. In the second step, z'_0 approximates z_0 , recombining with c_m to train *Mask-Diffusion*:

$$\mathcal{L}_{m'} = w_a \cdot \mathbb{E} \left[\|\epsilon_{\theta'}^{m'} - \epsilon\|^2 \right], \epsilon_{\theta'}^{m'} = \epsilon_{\theta}(z'_t, t, c_t, c_m), \quad (12)$$

where ϵ is as in Eq. (7), and w_a controls alignment between z'_0 and c_m , following [27]:

$$w_a = \begin{cases} 1, & \text{if } k > K_{\tau} \text{ and } t < T_{\tau}, \\ 0, & \text{otherwise,} \end{cases} \quad (13)$$

where k is the current iteration, K_{τ} is the iteration threshold, t is the timestep, and T_{τ} is the timestep threshold.

Table 2. Comparison of SANet [52], Polyp-PVT [10], and CTNet [54] performance, evaluated using mDice (%) and mIoU (%) metrics. “+” denotes that the training dataset consists of both real polyp images and synthetic polyp images generated by each respective method. “M” denotes that mask-only prior control, while “M+I” denotes mask-image joint prior control.

Methods	EndoScene [51]		ClinicDB [2]		Kvasir [26]		ColonDB [46]		ETIS [42]		Overall	
	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU
SANet [52]	88.8	81.5	91.6	85.9	90.4	84.7	75.3	67.0	75.0	65.4	79.4	71.4
+Copy-Paste	89.7	83.0	90.2	85.1	90.3	84.8	77.7	70.0	77.4	68.8	81.1	73.7
+SinGAN-Seg [47] (M)	88.3	81.6	90.9	85.3	91.0	85.8	77.3	69.4	73.7	65.4	80.0	72.6
+T2I-Adapter [31] (M)	86.5	78.9	90.6	85.0	89.9	84.1	77.7	69.4	78.6	69.7	81.1	73.2
+ArSDM [14] (M)	90.2	83.2	91.4	86.1	91.1	85.6	77.7	70.0	78.0	69.5	81.5	74.1
+ControlNet [59] (M)	90.2	83.7	91.6	85.4	90.6	85.0	77.0	69.0	75.8	66.5	80.5	72.8
+ControlNet [59] (M+I)	89.9	83.3	91.4	86.2	91.2	85.5	78.2	70.5	75.4	66.6	81.0	73.6
+Ours (M)	90.4	83.9	93.0	88.1	91.2	85.5	79.1	71.3	81.1	73.4	83.0	75.8
Polyp-PVT [10]	90.0	83.3	93.7	88.9	91.7	86.4	80.8	72.7	78.7	70.6	83.3	76.0
+Copy-Paste	88.0	80.9	93.4	88.7	91.7	87.1	79.8	71.8	79.2	71.3	82.8	75.6
+SinGAN-Seg [47] (M)	87.0	79.7	91.7	87.0	92.8	88.1	76.9	69.0	74.2	66.7	80.1	73.0
+T2I-Adapter [31] (M)	88.7	81.6	93.1	88.1	91.3	86.0	80.6	72.4	79.0	71.0	83.1	75.7
+ArSDM [14] (M)	88.2	81.2	92.2	87.5	91.5	86.3	81.7	73.8	80.6	72.9	84.0	76.7
+ControlNet [59] (M)	88.0	81.1	93.7	89.0	91.5	86.3	80.8	72.5	76.1	68.1	82.5	75.1
+ControlNet [59] (M+I)	90.3	83.5	92.2	87.7	91.1	86.2	81.6	73.7	77.2	69.1	83.2	76.0
+Ours (M)	90.4	83.8	93.9	89.3	91.1	85.9	81.7	74.0	81.5	73.4	84.4	77.3
CTNet [54]	90.8	84.4	93.6	88.7	91.7	86.3	81.3	73.4	81.0	73.4	84.2	77.0
+Copy-Paste	91.6	85.5	92.3	87.1	91.5	86.3	82.5	74.7	78.1	70.6	84.0	76.9
+SinGAN-Seg [47] (M)	89.4	82.3	93.6	88.8	92.9	88.6	78.0	70.4	76.4	68.7	81.5	74.6
+T2I-Adapter [31] (M)	90.8	84.3	93.7	88.7	91.9	86.9	81.7	74.1	80.4	72.5	84.3	77.2
+ArSDM [14] (M)	89.8	82.9	92.9	87.8	90.5	85.2	81.9	74.0	80.9	73.5	84.2	77.0
+ControlNet [59] (M)	90.2	84.1	92.3	87.5	91.7	86.9	81.4	73.5	79.5	71.7	83.7	76.6
+ControlNet [59] (M+I)	90.7	84.2	93.2	88.6	92.0	87.2	81.1	73.6	80.9	73.1	84.1	77.1
+Ours (M)	90.1	83.6	93.8	88.9	92.9	89.1	83.8	75.8	79.5	72.0	85.1	78.1

4. Experimental

4.1. Dataset and Metrics

We conduct experiments on three public medical datasets—Polyps [2, 26], ISIC2016 [8], and ISIC2018 [7]—and two internal natural datasets, Stain and Faeces.

Training Set for Generative Models: Following [14], the Polyps dataset consists of 1,450 samples, comprising 900 samples from Kvasir [26] and 550 samples from CVC-ClinicDB [2]. The ISIC2016 [8] and ISIC2018 [7] datasets contain 900 and 2,594 samples, respectively. The Stain (500 samples) and Faeces (458 samples) datasets are partitioned into training, validation, and test sets in a 3:1:1 ratio, yielding training sets of 300 and 275 samples, respectively.

Training Set for Segmentation Models: The medical datasets leverage synthetic images generated using masks derived from their original training sets, which are then amalgamated with the original images to create new training sets. For the Stain and Faeces datasets, the original masks are augmented through various transformations (*e.g.*, scaling) to produce 1,000 samples for training.

Evaluation Metrics: We use several image quality evaluation metrics, including FID [18], KID [3], CLIP-I [38], LPIPS [60], CMMD [25], and the Mean Opinion Score (MOS), which is assessed by 3 experienced clinicians (see Appendix for details). We use mDice and mIoU to evaluate segmentation performance with the default settings of CNN and Transformer-based models.

Table 3. Comparison of UNet [37] and SegFormer [55] performance, evaluated using mDice (%) and mIoU (%). “+” denotes that the training dataset consists of both real and synthetic skin lesion images generated by each respective mask-only method. “Real” denotes that masks are derived from real data, while “Random” indicates transformed versions of real masks (*e.g.*, scaling).

Methods	ISIC2016 [8]				ISIC2018 [7]			
	UNet [37]		SegFormer [55]		UNet [37]		SegFormer [55]	
	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU
Real Dataset	89.58	86.62	93.57	89.06	82.19	78.45	91.35	86.10
+Copy-Paste	89.43	86.48	93.53	88.85	81.00	77.16	91.25	86.09
+T2I-Adapter [31]	89.48	86.47	93.71	89.24	81.16	77.04	91.19	85.73
+ControlNet [59]	89.42	86.51	93.54	88.83	81.76	77.94	91.52	86.19
+Ours (Real)	89.91	87.01	94.14	89.76	82.81	79.04	91.80	86.67
+Ours (Random)	90.06	87.25	94.31	90.01	83.71	80.09	91.93	87.12

Table 4. Comparison of SegFormer [55] performance, evaluated on mDice (%) and mIoU (%). Training dataset includes 1,000 synthetic Stain/Faeces images generated by each mask-only method.

Methods	Stain		Faeces	
	SegFormer [55]		SegFormer [55]	
	mDice	mIoU	mDice	mIoU
Real Dataset	93.46	89.78	90.50	87.78
DFMGAN [15]	49.59	49.19	49.47	48.95
AnomalyDiffusion [23]	70.84	64.47	81.59	73.39
T2I-Adapter [31]	74.75	70.57	88.02	83.88
ControlNet [59]	76.65	70.48	88.50	83.58
Ours	87.96	83.37	94.11	90.26

4.2. Implementation Details

Pre-trained Stable Diffusion V1.5 [36] forms the foundation of *Siamese-Diffusion*. We fine-tune Stable Diffusion across all five datasets using the AdamW [29] optimizer, with a learning rate of 1×10^{-5} and a weight decay of 1×10^{-2} . The maximum number of training iterations (*i.e.*, N_{iter}) is set to 3,000 for the Polyps [2, 26], ISIC2016 [8], and ISIC2018 [7] datasets. It is set to 1,500 for the Stain and Faeces datasets. In Eq. (5), the weight w_m is set to 1.0. In Eq. (13), the iteration threshold $K_\tau = \frac{N_{iter}}{3}$ is empirically chosen and the timestep threshold $T_\tau = 200$ follows [27]. We use an optimal value of $w_c = 1.0$, as elaborated in Sec. 5.3.2. The images of the five datasets are resized to 384×384 , and 5% probability is applied with no prompt [20, 59]. We execute training on 8 NVIDIA 4090 GPUs with a batch size of 6, culminating in a total batch size of 48. During sampling, we utilize DDIM [44] ($\eta = 0$) with 50 steps and a guidance scale of $\lambda = 9$, in accordance with the methodology in [59].

Table 5. Comparison of the impact of different components on polyp image quality and segmentation performance. Image quality is evaluated using FID [18], KID [3], and LPIPS [60] metrics. Segmentation performance is evaluated using weighted mDice (%) and mIoU (%) metrics across five public test datasets. The training dataset consists of real polyp images and images synthesized by our proposed *Siamese-Diffusion* method for each setting.

Settings	DHI	Online-Aug	$\mathcal{L}_s$	FID ($\downarrow$) KID ($\downarrow$) LPIPS ($\downarrow$)			SAnet [52]		Polyp-PVT [10]		CTNet [54]	
				mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU	
1				70.630	0.0509	0.612	80.5	72.8	82.5	75.1	83.7	76.6
2	✓			70.479	0.0496	0.610	80.8	73.3	83.1	75.9	83.8	76.8
3		✓		65.927	0.0459	0.605	80.9	73.4	82.7	75.5	84.0	76.9
4			✓	67.508	0.0476	0.602	81.8	74.6	82.8	75.6	84.3	77.1
5		✓	✓	65.266	0.0458	0.601	82.0	74.4	83.7	76.5	84.4	77.3
6	✓	✓		62.706	0.0385	0.598	81.7	74.2	83.5	76.4	84.1	77.1
7	✓		✓	63.059	0.0405	0.590	82.7	75.2	84.2	76.8	84.7	77.7
Ours	✓	✓	✓	62.706	0.0395	0.586	83.0	75.8	84.4	77.3	85.1	78.1

5. Evaluations and results

5.1. Image Quality Comparison

We evaluate synthetic images from prior mask-only models quantitatively and qualitatively on the Polyps dataset, with results from other datasets provided in the Appendix.

Quantitative Evaluation: Our method outperforms others across six metrics, achieving a 35-point reduction in FID compared to the previous SOTA ArSDM [14], as shown in Tab. 1. The t-SNE visualizations in Fig. 4 further substantiate the realism of our synthetic images. Additionally, our method achieves a KID score of 0.0395 and a CLIP-I rating of 0.892, reflecting superior image quality. Meanwhile, the best LPIPS score of 0.586 and CMMD score of 0.515 demonstrate a closer alignment with human perception, while the top MOS score of 0.587 supports this conclusion, highlighting enhanced morphological characteristics.

Qualitative Evaluation: Fig. 5 presents polyp images generated by various methods. SinGAN-Seg [47] often introduces artifacts and lacks diversity due to its “editing-like” approach. T2I-Adapter [31] produces a “glossy” appearance, leading to unrealistic images. ArSDM [14] fails to synthesize key morphological characteristics (e.g., surface texture), resulting in low fidelity. Our method, based on ControlNet [59], outperforms the mask-only (i.e., “M”) ControlNet [59] shown in Fig. 5(f) in three critical areas: mask alignment, morphological texture, and color, demonstrating its superiority. Compared to the mask-image (i.e., “M+I”) ControlNet [59] shown in Fig. 5(e), our method exhibits competitive fidelity while preserving diversity.

5.2. Segmentation Performance Comparison

Polyp segmentation performance comparisons are provided in Tab. 2. We retrain models on a duplicated dataset (i.e., “Copy-Paste”) to establish a new baseline. Our method surpasses the previous SOTA, ArSDM [14], significantly improving mDice and mIoU for SANet [52] by 3.6% and

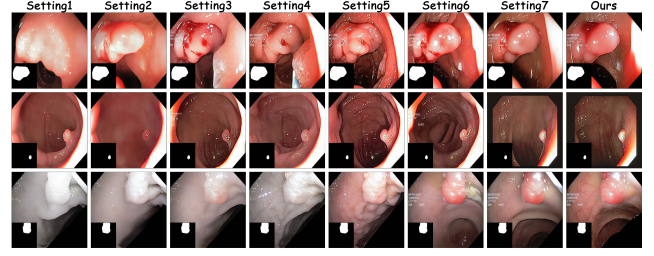


Figure 6. Visualizing the impact of different components on the synthesis of polyp images (Zoom in for better visualization).

4.4%. For more advanced architectures, such as Polyp-PVT [10] and CTNet [54], mDice scores increase by 1.1% and 0.9%, and mIoU increases by 1.3% and 1.1%. Only minor declines are observed on Kvasir with Polyp-PVT and on EndoScene and ETIS with CTNet, highlighting the robustness of our method across diverse models and test sets.

Intriguingly, despite ControlNet’s [59] superior morphological quality with mask-image (i.e., “M+I”) joint prior controls, the segmentation performance of Polyp-PVT and CTNet declines, similar to Copy-Paste. Conversely, ArSDM [14], despite lower morphological fidelity, improves segmentation performance. These observations reveal a crucial insight: image quality alone does not solely determine segmentation performance; morphological diversity is equally essential. However, we emphasize that, in medical imaging applications, morphological diversity must be paired with high morphological fidelity to ensure interpretable and reliable segmentation.

Similar phenomena can be observed in Tab. 3. Interestingly, both UNet [37] and SegFormer [55] perform better with skin lesion images synthesized using random masks (i.e., “Random”) compared to those derived from real data (i.e., “Real”). UNet achieves a remarkable increase of 1.52% in mDice and 1.64% in mIoU on ISIC2018, underscoring the sensitivity of segmentation to morphological diversity. Tab. 4 presents segmentation performance comparisons using 1,000 Stain or Faeces samples synthesized with transformed masks (e.g., scaling), without real samples. Our method significantly outperforms AnomalyDiffusion [23], even surpassing results achieved with real faeces samples, highlighting its superior scalability.

5.3. Ablation Studies

5.3.1. Contribution of Main Components

We conduct the first ablation study to evaluate the contribution of each main component.

Quantitative Analysis: As illustrated in Tab. 5, Setting 1 corresponds to ControlNet [59]. Setting 2 introduces DHI, which improves segmentation performance by refining mask alignment but does not enhance morphological fidelity. Setting 3 employs Online-Augment to expand the

Table 6. Comparison of the impact of different values of w_c on medical image segmentation performance, evaluated using weighted mDice (%) and mIoU (%) metrics across three datasets. The training dataset consists of real images and images synthesized by our proposed *Siamese-Diffusion* method for each w_c .

	Polyps						ISIC2016				ISIC2018			
w_c	SANet		Polyp-PVT		CTNet		UNet		SegFormer		UNet		SegFormer	
	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU
0.0	80.5	72.8	82.5	75.1	83.7	76.6	89.42	86.51	93.54	88.83	81.76	77.94	91.52	86.19
0.5	82.0	74.4	84.0	76.7	84.2	77.0	89.88	86.98	94.02	89.72	82.24	78.52	91.75	86.65
1.0	83.0	75.8	84.4	77.3	85.1	78.1	89.91	87.01	94.14	89.76	82.81	79.04	91.80	86.67
1.5	82.4	75.1	84.0	76.7	84.6	77.7	89.96	87.16	94.07	89.71	82.41	78.56	92.05	87.19
2.0	81.4	73.7	83.8	76.4	83.8	76.5	89.77	86.82	94.11	89.74	82.27	78.46	91.54	86.30

training data, thereby enhancing both image quality and segmentation performance. Setting 4 integrates $\mathcal{L}_c$, boosting the performance of SANet [52], a texture-sensitive CNN model [32], and surpassing the previous SOTA model, ArSDM [14]. Settings 5 through 7 demonstrate the synergistic effects of combining components. DHI acts as a catalyst for Online-Augment and $\mathcal{L}_c$, substantially improving image quality in Setting 6 and segmentation performance in Setting 7, thoroughly surpassing ArSDM. The final model, which integrates all components, achieves optimal results in both image fidelity and segmentation performance.

As discussed in Sec. 5.2, statistical quality and segmentation performance are not always directly correlated. Metrics like FID capture high-dimensional data distributions but tend to overfit with scarce data [23] and miss fine-grained morphological details [25], crucial for segmentation. Consequently, while $\mathcal{L}_c$ significantly improves segmentation performance by refining morphological characteristics, such as surface textures and color representation in Settings 4 and 7, its quality assessment results are lower than those of Online-Augment in Settings 3 and 6.

Qualitative Analysis: Fig. 6 depicts the impact of each component. Settings 2 through 4 provided marginal improvements when applied individually. Setting 2 improves lesion-mask alignment, as shown in the first two columns. Setting 3 enhances color representation, while Setting 4 refines morphological texture features. In contrast, Settings 5 through 7, which combine components, significantly enhance morphological details. Our *Siamese-Diffusion*, integrating all components, generates the most realistic images with intricate details, demonstrating its efficacy.

5.3.2. Weight Adjustment of Noise Consistency Loss

Similar to CFG [20], our method is akin to Image-Free Guidance (IFG) in the parameter space, with w_c serving as the guidance strength. In this section, we investigate the impact of w_c for $\mathcal{L}_c$ on the performance of various segmentation models, as shown in Tab. 6. We evaluate weights of 0.0, 0.5, 1.0, 1.5, and 2.0. Setting $w_c = 0.0$ corresponds to ControlNet [59]. When w_c is set to 0.5, 1.0, or 1.5, the mDice and mIoU scores of the segmentation models are

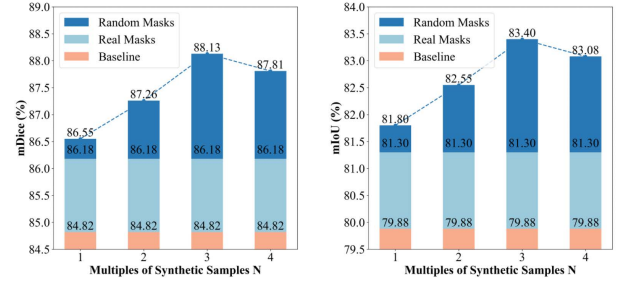


Figure 7. Visualizing the trends of mDice (%) and mIoU (%) metrics for SegFormer [55]. The training dataset consists of real polyp images and synthetic polyp images generated at different multiples by our proposed *Siamese-Diffusion*.

consistently on par with or exceed those of the SOTA models. These results validate the effectiveness of the proposed *Noise Consistency Loss* $\mathcal{L}_c$ in enhancing segmentation performance and demonstrate the robustness of our approach. Finally, we select $w_c = 1.0$ as the optimal stable default.

5.4. Discussion

While the advantages of the mask-only sampling paradigm in terms of morphological diversity and scalability are preliminarily demonstrated in Tab. 3 and Tab. 4, we further investigate the impact of varying data volumes on segmentation performance using SegFormer [55]. As shown in Fig. 7, when the synthetic data volume equals one fold, images generated with transformed masks (e.g., scaling) show more pronounced improvements compared to those generated with real masks. This advantage continues to escalate as the data volume increases, peaking at threefold, where the mDice and mIoU metrics reach impressive values of 88.13% and 83.40%. These results demonstrate that both scalability and morphological diversity are critical factors for segmentation. However, when the augmented data volume increases to fourfold, a decline in performance gains is observed. This intriguing phenomenon may be attributed to a unique “long-tail problem” arising from the complex distribution within the Polyps training and testing datasets, which we plan to explore further in future work.

6. Conclusion

In this paper, we introduce a novel image synthesis method, *Siamese-Diffusion*, which enhances the morphological fidelity of medical images while preserving the diversity inherent in the mask-only sampling paradigm. By employing *Noise Consistency Loss*, we generate high-fidelity medical images rich in morphological characteristics using arbitrary masks during the sampling phase. Extensive experiments across five datasets demonstrate the effectiveness and superiority of our method.

Acknowledgement

This work was supported in part by the Startup Grant for Professors (SGP) — CityU SGP, City University of Hong Kong, under Grant 9380170; partially by the Young Scientists Fund of the National Natural Science Foundation of China (Grant No. 62302328) and the Jiangsu Province Foundation for Young Scientists (Grant No. BK20230482); and partially by the WKU 2025 International Collaborative Research Program (Grant No. ICRPSP2025001).

References

- [1] Reza Azad, Ehsan Khodapanah Aghdam, Amelie Rauland, Yiwei Jia, Atlas Haddadi Avval, Afshin Bozorgpour, Sanaz Karimijafarbigloo, Joseph Paul Cohen, Ehsan Adeli, and Dorit Merhof. Medical image segmentation review: The success of u-net. *IEEE TPAMI*, pages 1–20, 2024. 1
- [2] 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. *COMPUT MED IMAG GRAP*, 43:99–111, 2015. 6
- [3] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. In *ICLR*, 2018. 5, 6, 7
- [4] Qi Chen, Xiaoxi Chen, Haorui Song, Zhiwei Xiong, Alan Yuille, Chen Wei, and Zongwei Zhou. Towards generalizable tumor synthesis. In *CVPR*, 2024. 2
- [5] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *CVPR*, 2021. 4
- [6] Yisol Choi, Sangkyung Kwak, Kyungmin Lee, Hyungwon Choi, and Jinwoo Shin. Improving diffusion models for virtual try-on. In *ECCV*, 2024. 2
- [7] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*, 2019. 6
- [8] Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *ISBI*, 2018. 6
- [9] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, 2021. 2
- [10] Bo Dong, Wenhai Wang, Deng-Ping Fan, Jinpeng Li, Huazhu Fu, and Ling Shao. Polyp-pvt: Polyp segmentation with pyramid vision transformers. *CAAI AIR*, 2:9150015, 2021. 6, 7
- [11] Zolnamar Dorjsembe, Hsing-Kuo Pao, and Furen Xiao. Polyp-ddpm: Diffusion-based semantic polyp synthesis for enhanced segmentation. In *EMBC*, 2024. 2
- [12] Xuefeng Du, Zhaoning Wang, Mu Cai, and Yixuan Li. Vos: Learning what you don’t know by virtual outlier synthesis. In *ICLR*, 2022. 2
- [13] Xuefeng Du, Yiyu Sun, Jerry Zhu, and Yixuan Li. Dream the impossible: Outlier imagination with diffusion models. In *NeurIPS*, 2024. 2
- [14] Yuhao Du, Yuncheng Jiang, Shuangyi Tan, Xusheng Wu, Qi Dou, Zhen Li, Guanbin Li, and Xiang Wan. Arsdm: colonoscopy images synthesis with adaptive refinement semantic diffusion models. In *MICCAI*, 2023. 1, 2, 5, 6, 7, 8
- [15] Yuxuan Duan, Yan Hong, Li Niu, and Liqing Zhang. Few-shot defect image generation via defect-aware feature manipulation. In *AAAI*, 2023. 6
- [16] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, 2021. 2
- [17] David Ha, Andrew Dai, and Quoc V Le. Hypernetworks. In *ICLR*, 2017. 2
- [18] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017. 5, 6, 7
- [19] Geoffrey Hinton. Distilling the knowledge in a neural network. In *NeurIPS Workshop*, 2015. 4
- [20] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS Workshop*, 2022. 2, 5, 6, 8
- [21] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 1, 2, 3
- [22] Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *CVPR*, 2024. 2
- [23] Teng Hu, Jiangning Zhang, Ran Yi, Yuzhen Du, Xu Chen, Liang Liu, Yabiao Wang, and Chengjie Wang. Anomalydiffusion: Few-shot anomaly image generation with diffusion model. In *AAAI*, 2024. 2, 6, 7, 8
- [24] Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *ICLR*, 2022. 4
- [25] Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. Rethinking fid: Towards a better evaluation metric for image generation. In *CVPR*, 2024. 5, 6, 8
- [26] 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 *MMM*, 2020. 6
- [27] Ming Li, Taojiannan Yang, Huafeng Kuang, Jie Wu, Zhaoning Wang, Xuefeng Xiao, and Chen Chen. Controlnet-plus-plus: Improving conditional controls with efficient consistency feedback. In *ECCV*, 2024. 2, 5, 6
- [28] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 5

- [29] I Loshchilov. Decoupled weight decay regularization. In *ICLR*, 2017. 6
- [30] Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. In *CVPR*, 2023. 4
- [31] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *AAAI*, 2024. 1, 2, 5, 6, 7
- [32] Muhammad Muzammal Naseer, Kanchana Ranasinghe, Salman H Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Intriguing properties of vision transformers. In *NeurIPS*, 2021. 8
- [33] Quang Nguyen, Truong Vu, Anh Tran, and Khoi Nguyen. Dataset diffusion: Diffusion-based synthetic data generation for pixel-level semantic segmentation. In *NeurIPS*, 2024. 1
- [34] Mang Ning, Enver Sangineto, Angelo Porrello, Simone Calderara, and Rita Cucchiara. Input perturbation reduces exposure bias in diffusion models. In *ICML*, 2023. 4
- [35] Kunpeng Qiu, Zhiying Zhou, and Yongxin Guo. Learn from zoom: Decoupled supervised contrastive learning for wce image classification. In *ICASSP*, 2024. 1
- [36] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1, 2, 3, 6
- [37] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 6, 7
- [38] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, 2023. 5, 6
- [39] Vikash Sehwal, Caner Hazirbas, Albert Gordo, Firat Ozgenel, and Cristian Canton. Generating high fidelity data from low-density regions using diffusion models. In *CVPR*, 2022. 2
- [40] Shitong Shao, Xiaohan Yuan, Zhen Huang, Ziming Qiu, Shuai Wang, and Kevin Zhou. Diffuseexpand: Expanding dataset for 2d medical image segmentation using diffusion models. In *IJCAI Workshop*, 2024. 1, 2
- [41] Vanshali Sharma, Abhishek Kumar, Debesh Jha, Manas Kamal Bhuyan, Pradip K Das, and Ulas Bagci. Controlpolypnet: towards controlled colon polyp synthesis for improved polyp segmentation. In *CVPR Workshop*, 2024. 2
- [42] Juan Silva, Aymeric Histace, Olivier Romain, Xavier Dray, and Bertrand Granado. Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. *INT J COMPUT ASS RAD*, 9:283–293, 2014. 6
- [43] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015. 2
- [44] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2020. 2, 5, 6
- [45] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2020. 1, 2
- [46] Nima Tajbakhsh, Suryakanth R Gurudu, and Jianming Liang. Automated polyp detection in colonoscopy videos using shape and context information. *IEEE TMI*, 35(2):630–644, 2015. 6
- [47] Vajira Thambawita, Pegah Salehi, Sajad Amouei Sheshkal, Steven A Hicks, Hugo L Hammer, Sravanthi Parasa, Thomas de Lange, Pål Halvorsen, and Michael A Riegler. Singan-seg: Synthetic training data generation for medical image segmentation. *PloS one*, 17(5), 2022. 5, 6, 7
- [48] Soobin Um and Jong Chul Ye. Don’t play favorites: Minority guidance for diffusion models. In *ICLR*, 2023. 2
- [49] Soobin Um and Jong Chul Ye. Self-guided generation of minority samples using diffusion models. In *ECCV*, 2024. 2
- [50] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. In *NeurIPS*, 2017. 3
- [51] David Vázquez, Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Antonio M López, Adriana Romero, Michal Drozdal, and Aaron Courville. A benchmark for endoluminal scene segmentation of colonoscopy images. *J HEALTHC ENG*, 2017(1):4037190, 2017. 6
- [52] Jun Wei, Yiwen Hu, Ruimao Zhang, Zhen Li, S.Kevin Zhou, and Shuguang Cui. Shallow attention network for polyp segmentation. In *MICCAI*, 2021. 1, 6, 7, 8
- [53] Weijia Wu, Yuzhong Zhao, Mike Zheng Shou, Hong Zhou, and Chunhua Shen. Diffumask: Synthesizing images with pixel-level annotations for semantic segmentation using diffusion models. In *ICCV*, 2023. 1
- [54] Bin Xiao, Jinwu Hu, Weisheng Li, Chi-Man Pun, and Xiuli Bi. Ctnet: Contrastive transformer network for polyp segmentation. *IEEE TCYB*, 54(9):5040–5053, 2024. 1, 6, 7
- [55] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *NeurIPS*, 2021. 6, 7, 8
- [56] Xingqian Xu, Jiayi Guo, Zhangyang Wang, Gao Huang, Irfan Essa, and Humphrey Shi. Prompt-free diffusion: Taking” text” out of text-to-image diffusion models. In *CVPR*, 2024. 2, 5
- [57] Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. Magicanimate: Temporally consistent human image animation using diffusion model. In *CVPR*, 2024. 2
- [58] Hu Ye, Jun Zhang, Sibio Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 2
- [59] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 1, 2, 3, 5, 6, 7, 8
- [60] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 5, 6, 7