Semantic Interpolative Diffusion Model: Bridging the Interpolation to Masks and Colonoscopy Image Synthesis for Robust Generalization

Chanyeong Heo[0009−0008−5884−4720] and Jaehee Jung[0000−0002−0932−3039]

Department of Information and Communication Engineering, Myongji University, South Korea
{hcn98,jhjung}@mju.ac.kr

Abstract. Polyp segmentation is a representative task in computer-aided clinical diagnosis in colonoscopy analysis. However, strict regulations limit the availability of large, high-quality image-mask paired datasets for segmentation. As a result, recent studies have focused on models that generate images conditioned on masks. However, due to rigid annotation constraints and a high reliance on fixed masks, the synthesized images often exhibit limited variation, leading to a lack of generalization in downstream tasks. This study introduces the Semantic Interpolative Diffusion Model (SIDM), which applies interpolation to both the given masks and the colonoscopy images to generate pairs of interpolated masks and images. First, a background semantic label was devised by labeling background regions based on the colonoscopy imaging environment. Both the masks and the background semantic labels are applied as multi-conditions to the diffusion model for colonoscopy image generation. After training, interpolation on both the masks and background semantic labels is performed at a chosen ratio. Applying the interpolated masks and labels to the model generates an intermediate perspective of colonoscopy images that partially incorporates features from each condition. By augmenting the dataset with these pairs of interpolated masks and generated images with interpolated conditions, segmentation models can extend the coverage of possible colonoscopy scenarios and mitigate the limitations of fixed masks, leading to robust generalization. Experimental comparisons against existing generative models, using the same test data across different segmentation models and different test datasets with the same model, demonstrate the effective generalization of the proposed model. The code is available at https://github.com/DSLab-MJU/SIDM.

Keywords: Diffusion model · Semantic Interpolation · Colonoscopy Image Synthesis · Data Augmentation · Polyp Segmentation.

1 Introduction

Colorectal cancer (CRC) is a serious threat to human health and is the second leading cause of cancer-related deaths worldwide. In particular, CRC remains

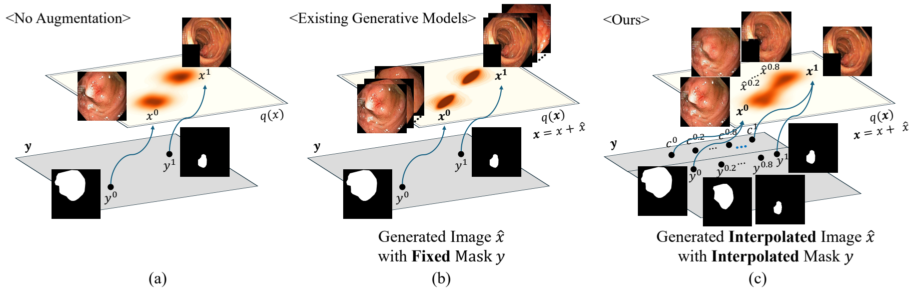
<No Augmentation>
<Existing Generative Models>
<Ours>
x^1
x^0
$q(x)$
y
y^1
y^0
$q(\boldsymbol{x})$
$\boldsymbol{x} = x + \hat{x}$
$\hat{x}^{0.2}$
$\hat{x}^{0.8}$
c^0
$c^{0.2}$
$c^{0.8}$
c^1
$y^{0.2}$
$y^{0.8}$
Generated Image $\hat{x}$ with **Fixed** Mask y
Generated **Interpolated** Image $\hat{x}$ with **Interpolated** Mask y
(a)
(b)
(c)

Fig. 1. A schematic representation of the data augmentation process using a generative model. (a) represents the data before augmentation, (b) illustrates augmentation using existing generative models, and (c) shows augmentation using our proposed method.

one of the leading causes of cancer mortality in the United States, with over 150,000 new cases and 50,000 deaths expected in 2024 [21,22]. However, CRC can be prevented through early detection and the removal of polyps using techniques such as colonoscopy and wireless capsule endoscopy (WCE). Recently, with the advancement of deep learning, polyp segmentation has been actively leveraged for the automatic detection of polyps in colonoscopy[30]. However, due to national and regional regulations on the public release of medical data[27,7], there are significant challenges in constructing high-quality, large-scale datasets for training polyp segmentation models. To address this issue, research has been actively conducted on medical image synthesis using generative models to quantitatively increase the amount of available medical data[9,14,3,4].

As polyp segmentation requires a paired dataset of colonoscopy images and corresponding masks, which annotate lesion areas, most generative models generate images conditioned on a given mask [5,6,28,16,17,25]. For instance, PolypDDPM[5] leveraged a diffusion probabilistic model[11] by incorporating the mask into the channel dimension as a conditioning factor, resulting in a 4% Intersection over Union (IoU) improvement in U-Net++[31] segmentation performance after data augmentation. Similarly, ArSDM[6], based on the SDM[28], used SPADE[16] to leverage masks as conditioning inputs for image generation while employing adaptive loss and refinement techniques. This approach demonstrated segmentation performance improvements in PraNet[8], SANet[29], and PVT[2].

However, existing mask-to-image generative models generate colonoscopy images based on only given masks, allowing variations in images but keeping the lesion regions constant. As shown in Fig. 1(b), this results in more diverse images than the original data Fig. 1(a) but remains constrained by the fixed masks.

To address these issues, we propose the Semantic Interpolative Diffusion Model (SIDM), which applies interpolation to both the given masks and the background of colonoscopy images to generate pairs of interpolated masks and images. Background semantic labels are newly defined based on the imaging environment, such as colonoscopy videos or snapshot images, and are applied as a condition to the diffusion model alongside masks for image generation.

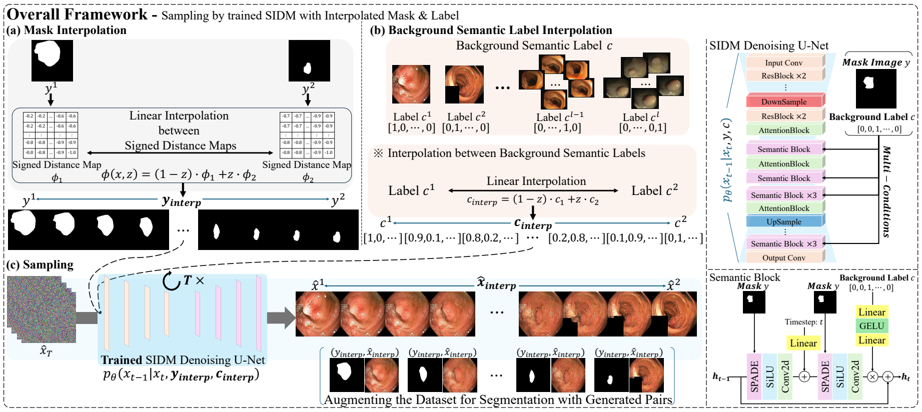
Overall Framework - Sampling by trained SIDM with Interpolated Mask & Label
(a) Mask Interpolation
y^1
y^2
Linear Interpolation between Signed Distance Maps
Signed Distance Map ϕ_1
Signed Distance Map ϕ_2
$\phi(x,z) = (1-z)\cdot\phi_1 + z\cdot\phi_2$
y_{interp}
(b) Background Semantic Label Interpolation
Background Semantic Label c
Label c^1 $[1,0,\cdots,0]$
Label c^2 $[0,1,\cdots,0]$
Label c^{l-1} $[0,\cdots,1,0]$
Label c^l $[0,\cdots,0,1]$
※ Interpolation between Background Semantic Labels
Label c^1
Linear Interpolation
$c_{interp} = (1-z)\cdot c_1 + z\cdot c_2$
Label c^2
c^1
c_{interp}
c^2
$[1,0,\cdots]$ $[0.9,0.1,\cdots]$ $[0.8,0.2,\cdots]$ $\cdots$ $[0.2,0.8,\cdots]$ $[0.1,0.9,\cdots]$ $[0,1,\cdots]$
(c) Sampling
$\hat{x}_T$
$T\times$
Trained SIDM Denoising U-Net
$p_\theta(x_{t-1}|x_t, y_{interp}, c_{interp})$
$\hat{x}^1$
$\hat{x}_{interp}$
$\hat{x}^2$
$(y_{interp}, \hat{x}_{interp})$
Augmenting the Dataset for Segmentation with Generated Pairs
SIDM Denoising U-Net
$p_\theta(x_{t-1}|x_t, y, c)$
Input Conv
ResBlock ×2
DownSample
ResBlock ×2
AttentionBlock
Semantic Block
AttentionBlock
Semantic Block
Semantic Block ×3
AttentionBlock
UpSample
Semantic Block ×3
Output Conv
Mask Image y
Background Label c
$[0,0,1,\cdots,0]$
Multi − Conditions
Semantic Block
Mask y
Mask y
Background Label c
$[0,0,1,\cdots,0]$
Timestep: t
Linear
Linear
GELU
Linear
h_{t-1}
SPADE
SiLU
Conv2d
SPADE
SiLU
Conv2d
h_t

Fig. 2. Overall framework of SIDM. (a) Mask interpolation. (b) Background semantic label interpolation. (c) The process where the trained SIDM generates images by incorporating the interpolated conditions from (a) and (b). The interpolated masks and generated images are then paired and used for generative augmentation.

After training, interpolation in masks and background semantic labels at a chosen ratio enables the generation of intermediate colonoscopy images that blend features from both conditions. Consequently, interpolated masks and generated colonoscopy images with interpolated conditions effectively expand the diversity of colonoscopic scenarios, resulting in a more diverse augmented distribution, as illustrated in Fig. 1(c), and mitigates issues related to fixed mask constraints. By augmenting data with these interpolated masks and images, segmentation models achieve improved generalization.

Our main contributions can be summarized as follows: 1) We propose the SIDM, which interpolates both lesion masks and background semantic labels to generate diverse pairs of interpolated masks and images. 2) SIDM newly defines background semantic labels and devises multi-conditional diffusion model with masks and background semantic labels for colonoscopy image synthesis. 3) Experimental results demonstrate the superiority of our model over existing generative models in segmentation performance after generative augmentation, evaluating generalization using the same test data across different segmentation models and different test datasets with the same segmentation model.

2 Method

The proposed SIDM first trains a diffusion model with masks and newly defined background semantic labels as multi-conditions, as shown in Fig. 2. Interpolation is then applied to both masks and background semantic labels, and the interpolated masks and labels are fed into the trained model to generate corresponding images. Finally, pairs of interpolated masks and generated images with interpolated conditions are used as an augmenting dataset for segmentation.

2.1 Mask Interpolation

The mask image y, representing the lesion region in a medical image, has a binary structure and is denoted as $I(\mathbf{x})$, where $\mathbf{x}$ represents pixel coordinates. If a pixel belongs to the foreground, which represents the lesion region, $I(\mathbf{x}) = 1$; otherwise, for the background, $I(\mathbf{x}) = 0$. Instead of directly interpolating binary masks, each mask is converted into a signed distance map that encodes the distance from each pixel to the mask boundary, enabling more effective interpolation by leveraging the continuous distance representation of the mask's foreground, as shown in Fig.2(a). The signed distance map is defined as:

$$\phi(\mathbf{x}) = \begin{cases} d_{\text{inside}}(\mathbf{x}), & \text{if } I(\mathbf{x}) = 1 \\ -d_{\text{outside}}(\mathbf{x}), & \text{if } I(\mathbf{x}) = 0 \end{cases}. \tag{1}$$

The distances are computed using the Euclidean distance defined as $d(\mathbf{x}) = \min_{\mathbf{y} \in \partial I} \|\mathbf{x} - \mathbf{y}\|$, where ∂I means the boundary of the mask and $d_{\text{inside}}(\mathbf{x})$ is the minimum distance from a pixel inside the lesion $I(\mathbf{x}) = 1$ to ∂I, while $d_{\text{outside}}(\mathbf{x})$ is the minimum distance from a pixel outside the lesion $I(\mathbf{x}) = 0$ to ∂I.

To interpolate between two signed distance maps ϕ_1 and ϕ_2 from masks y_1 and y_2, we introduce an auxiliary dimension $z \in [0, 1]$ where $z = 0$ corresponds to ϕ_1 and $z = 1$ corresponds to ϕ_2, and perform linear interpolation according to:

$$\phi(\mathbf{x}, z) = (1 - z) \cdot \phi_1(\mathbf{x}) + z \cdot \phi_2(\mathbf{x}). \tag{2}$$

Finally, the interpolated distance field $\phi(\mathbf{x}, z)$ is converted into a binary mask by thresholding at zero, which can be formulated as follows:

$$I(\mathbf{x}, z) = \begin{cases} 1, & \text{if } \phi(\mathbf{x}, z) \geq 0 \\ 0, & \text{if } \phi(\mathbf{x}, z) < 0 \end{cases}, \tag{3}$$

thereby generating interpolated binary mask $I(\mathbf{x}, z)$, which is y_{interp}.

2.2 Background Semantic Label Interpolation

The background semantic label is designed to control the background region. The method of defining this label varies depending on the imaging technique used in the colonoscopy dataset for training.

For datasets composed of independent 2D snapshot images, each background region represents a different environment, so each image is assigned a unique label. In contrast, for video-based sequence data, multiple frames are captured within the same colon, meaning that while different background images exist for each sequence, they globally share the same colon wall characteristics. Therefore, all frames within the same sequence are assigned the same label.

Each label is designated as an integer and applied to model training using one-hot encoding, as illustrated in Fig.2(b). During training, each image is associated with its corresponding label, whereas during interpolation, linear interpolation is applied between two labels c_1, c_2 with interpolation factor z as follows:

$$c_{interp} = (1 - z) \cdot c_1 + z \cdot c_2. \tag{4}$$

2.3 Sampling with Semantic Interpolative Diffusion Model

Training The proposed SIDM is based on a diffusion model[11,18], which generates new data from an estimated data distribution $p_\theta(x_0)$ that approximates the original data distribution $x \sim q(x_0)$. SIDM estimates this distribution conditioned on mask image y and background semantic label c. Diffusion model consists of a forward and a reverse process. The forward process is a Markov chain that involves gradually adding noise to the original data x_0 over T steps, which is defined as $q(x_t|x_{t-1}) = \mathcal{N}\left(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I}\right)$, where β_t is the predefined noise schedule. The reverse process starts with Gaussian noise x_T and aims to estimate the original data by predicting the added noise using model θ at each step t over T steps, following the reverse Markov chain, formulated as:

$$p_\theta(x_{0:T}) := p(x_T)\prod_{t=1}^{T} p_\theta(x_{t-1}|x_t, y, c). \tag{5}$$

In this process, the noise added at each step t is predicted conditioned on y and c thereby enabling the model to take masks and labels as input to generate corresponding images, formulated as:

$$p_\theta(x_{t-1}|x_t, y, c) := \mathcal{N}\left(x_{t-1}; \mu_\theta(x_t, y, c, t), \varSigma_\theta(x_t, y, c, t)\right). \tag{6}$$

Here, $\mu_\theta(x_t, y, c, t)$ is defined as $\frac{1}{\sqrt{\alpha_t}}(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, y, c, t))$ where $\alpha_t := 1 - \beta_t$, $\bar{\alpha}_t := \prod_{s=1}^{t}\alpha_s$ and ϵ_θ is predicted using a SIDM Denoising U-Net, as shown in Fig. 2. At each decoding stage of the SIDM Denoising U-Net, the mask is integrated using SPADE[16], while the background semantic label is incorporated through Linear and GELU[10] layers, as shown in Fig.2's Semantic Block. This facilitates the proper reflection of the conditions with the Classifier-Free Guidance(CFG)[12] method employed. Additionally, to prevent overfitting to the given mask, the adaptive loss with $\mathbf{W}^\lambda$ designed in ArSDM[6] is applied. $\mathbf{W}^\lambda$ takes the value $1 - r$ when $p = 1$ and r when $p = 0$, where r is defined as $r = \frac{\#(p=1)}{H\times W}$ and $\#(p = 1)$ means the number of pixel p at (h, w) belongs to the foreground. The loss function is defined as follows:

$$\mathcal{L} = \mathbb{E}_{x_0,y,c,\epsilon}\left[\mathbf{W}^\lambda \cdot \left\|\epsilon - \epsilon_\theta\left(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, y, c, t\right)\right\|^2\right]. \tag{7}$$

Sampling The sampling process applies interpolation to the two masks and background semantic labels at specific ratios, such as 1:1, 1:3, and 3:1, as described in Sections 2.1 and 2.2, to extract y_{interp} and c_{interp}. These y_{interp}, c_{interp} are then input into the SIDM, and generate corresponding image after iterative reverse process as formulaed as $p_\theta(x_{0:T}) := p(x_T)\prod_{t=1}^{T} p_\theta(x_{t-1}|x_t, y_{interp}, c_{interp})$. For details, after T times following $\hat{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}(\hat{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(\hat{x}_t, y_{interp}, c_{interp}, t)) + \sigma_t\mathbf{z}$ start from $\hat{x}_T \sim \mathcal{N}(0, \mathbf{I})$, the corresponding image $\hat{x}_{interp}$ is generated.

Finally, the $(y_{interp}, \hat{x}_{interp})$ mask-image pairs are constructed and augmented into the existing dataset for segmentation, as illustrated in Fig.2(c).

3 Experiments and Results

Data Experiments were conducted on five polyp segmentation datasets: ETIS[23], CVC-ClinicDB[1], CVC-ColonDB [24], EndoScene-CVC300[26], and Kvasir-SEG[13]. The segmentation model was initially trained on 1,450 images from CVC-ClinicDB (550) and Kvasir (900) without augmentation. The generative model was then trained on the same dataset, generating an equal number of additional samples, resulting in augmented 2,900 image-mask pairs for segmentation training. Segmentation performance on unseen test datasets was evaluated using all images from EndoScene (60 images), CVC-ColonDB (380 images), and ETIS (196 images), along with the remaining 100 images from Kvasir and 62 images from CVC-ClinicDB. For evaluation on the same test data with different segmentation models, the remaining 162 images from Kvasir and CVC-ClinicDB were used.

Evaluation The experiments involved augmenting the dataset with generated images and evaluating segmentation performance before and after augmentation. The comparative generative models included ArSDM[6], SDM[28], SPADE[16], and PolypDDPM [5]. To assess generalization, U-Net[19], U-Net++[31], and FPN[15] were evaluated on the same split test data, while PraNet[8] and FCBFormer[20] were tested on unseen datasets. This analysis examined whether different models exhibited consistent trends with generated images and whether performance improved on unseen colonoscopy datasets, demonstrating generalization.

Implementation Details Ratios of 1:1, 1:3, and 3:1 were used in SIDM experiments to represent balanced and biased interpolations. The 1:1 ratio blends both inputs equally, while 1:3 and 3:1 emphasize one input and correspond to the outer quartiles of the interpolation space. Image-mask pairs were chosen based on differences in both components. Pairs with different background semantic labels were considered distinct, regardless of mask similarity. For masks, selection was based on size comparison with a threshold at zero. SIDM was implemented

Table 1. Quantitative results of segmentation after generative augmentation, evaluated on the same test data across different models. "w/o Label" refers to the case without a proposed background semantic label, implying that no image interpolation is applied.

Method	Training Data		U-Net[19]			U-Net++[31]			FPN[15]			Average		
	Real	Gen.	Dice	IoU	F1	Dice	IoU	F1	Dice	IoU	F1	Dice	IoU	F1
No Aug	1450	0	75.1	65.9	85.8	73.8	64.8	85.2	76.4	64.8	86.4	75.1	65.1	85.8
+ArSDM[6]	1450	1450	78.4	70.6	87.8	78.7	70.3	87.9	86.3	78.6	92.1	81.1	73.2	89.3
+SDM[28]	1450	1450	78.4	69.9	87.7	78.0	70.2	87.6	87.1	80.7	92.6	81.2	73.6	89.3
+SPADE[16]	1450	1450	77.7	69.7	87.4	79.0	71.4	88.1	87.5	81.0	92.8	81.4	74.0	89.4
+PolypDDPM[5]	1450	1450	78.8	71.2	88.0	79.7	71.9	88.4	86.4	79.5	91.9	81.6	74.2	89.4
+Ours(w/o Label)(1:1)	1450	1450	79.8	72.0	88.5	77.9	70.0	87.5	84.2	75.9	90.9	80.7	72.6	89.0
+Ours(1:3)	1450	1450	79.0	71.2	88.0	77.7	69.8	87.4	**88.1**	**81.6**	93.1	81.6	74.2	89.5
+Ours(3:1)	1450	1450	80.3	72.3	88.7	**81.3**	**73.5**	**89.3**	86.5	79.0	92.2	82.7	74.9	90.1
+Ours(1:1)	1450	1450	**80.8**	**72.8**	**89.0**	79.8	72.1	88.5	87.7	80.4	**93.3**	**82.8**	**75.1**	**90.3**

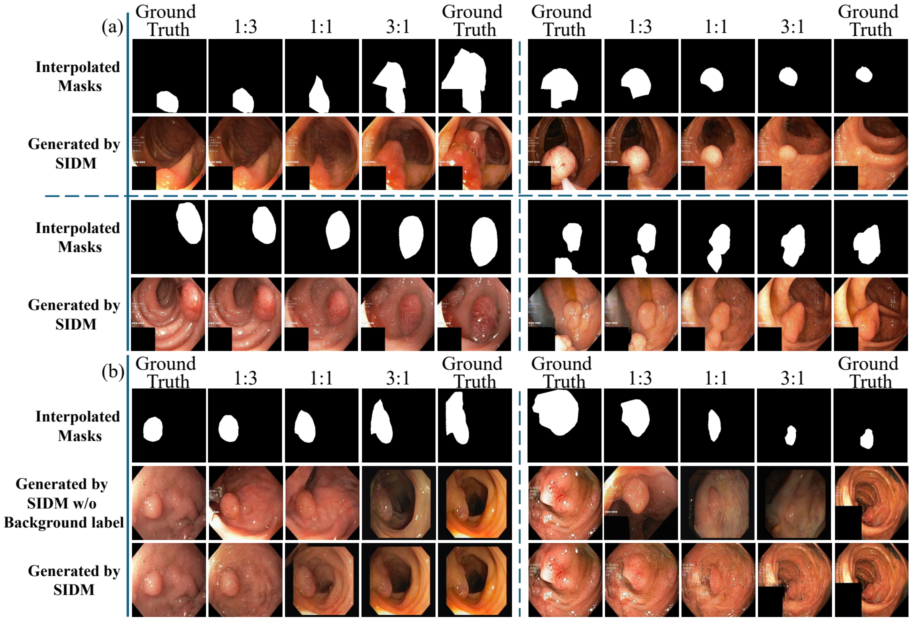

Fig. 3. Qualitative results of interpolated masks and generated images with background interpolation. (a) illustrates examples of pairs. (b) compares images with and without the background semantic label.

using PyTorch 1.11.0 and trained on an NVIDIA RTX 4080 GPU. The Adam optimizer was used with a learning rate of 0.0001. The timestep t was set to 1000, β ranged from 0.0001 to 0.02, and a CFG scale of 1.5 was used for sampling.

Results Fig. 3 illustrates how the proposed method constructs interpolated generated pairs. Interpolation was applied to two masks from the 1,450 training samples at 1:3, 1:1, and 3:1 ratios to generate interpolated masks, which were then used to generate images with interpolated background semantic labels. A qualitative evaluation confirms that interpolation occurs in both masks and background regions, following the specified ratios. Notably, in the second column of the last row in Fig. 3(a), the transition from two masks merging into one is clearly visible, along with background changes. Additionally, in all examples, background areas with and without colon folds transition smoothly, demonstrating effective background interpolation.

The interpolated mask-image pairs were then used for evaluation across different models on the same test data, as summarized in Table 1. The results show that, on average, the SIDM with 1:1 interpolated augmentation achieved the highest performance improvement, increasing from a pre-augmentation Dice score of 75.1% to 82.8%, marking a 7.7% gain and setting a new state-of-the-art (SOTA) among comparative models. Although U-Net performs best at a 1:1 ra-

Table 2. Quantitative results of segmentation after generative augmentation, evaluated on the same segmentation model across different test datasets. “w/o Label” refers to the case without a proposed background semantic label, implying that no image interpolation is applied.

Method	Training Data		EndoScene[26]		ClinicDB[1]		Kvasir[13]		ColonDB[24]		ETIS[23]		Average	
	Real	Gen.	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU
PraNet[8] (No Aug)	1450	0	88.5	81.7	89.8	85.1	89.3	84.2	68.2	61.6	60.0	54.0	79.2	73.3
+ArSDM[6]	1450	1450	85.7	77.9	92.9	87.8	87.2	81.3	70.3	61.6	**69.1**	**60.9**	81.0	73.9
+SDM[28]	1450	1450	89.1	81.7	91.6	86.7	89.7	84.2	72.6	65.9	66.1	59.3	81.8	75.6
+SPADE[16]	1450	1450	86.0	79.2	91.7	87.2	90.8	85.4	72.1	65.5	63.6	57.1	80.8	74.9
+PolypDDPM[5]	1450	1450	88.9	82.5	91.1	85.9	90.6	85.4	74.2	67.0	68.6	61.8	82.7	76.5
+Ours(w/o Label)(1:1)	1450	1450	86.1	79.1	89.8	85.4	90.6	85.2	68.9	62.0	66.2	59.6	80.3	74.3
+Ours(1:3)	1450	1450	89.2	82.5	89.0	84.0	89.9	84.7	71.3	64.3	66.6	60.0	81.2	75.1
+Ours(3:1)	1450	1450	88.5	81.7	89.9	85.5	89.5	84.2	72.7	65.3	66.8	60.0	81.5	75.3
+Ours(1:1)	1450	1450	**90.1**	**83.4**	**93.0**	**88.5**	**91.5**	**86.6**	**74.3**	**67.0**	66.6	59.3	**83.1**	**77.0**
FCBFormer[20] (No Aug)	1450	0	89.2	82.5	90.6	85.8	88.7	82.9	79.3	70.8	72.6	64.6	84.1	77.3
+ArSDM[6]	1450	1450	87.9	79.9	92.1	87.1	**91.8**	**86.6**	77.4	69.3	**74.8**	65.4	84.8	77.7
+SDM[28]	1450	1450	86.7	79.5	90.8	85.9	91.3	86.2	77.8	70.1	71.4	64.4	83.6	77.2
+SPADE[16]	1450	1450	88.2	81.0	88.5	83.7	90.9	85.1	75.5	68.0	74.6	**67.6**	83.5	77.1
+PolypDDPM[5]	1450	1450	88.7	81.4	89.0	84.1	90.5	85.0	80.2	72.4	74.1	66.4	84.5	77.9
+Ours(w/o Label)(1:1)	1450	1450	86.6	79.8	90.8	85.9	91.2	85.8	74.6	67.3	69.6	62.5	82.6	76.2
+Ours(1:3)	1450	1450	89.7	82.9	88.2	82.5	89.9	84.3	74.9	67.0	72.2	63.8	83.0	76.1
+Ours(3:1)	1450	1450	86.4	79.3	92.2	87.4	90.9	85.4	79.5	71.2	74.5	66.9	84.7	78.0
+Ours(1:1)	1450	1450	**90.6**	**84.0**	**93.6**	**88.7**	91.1	86.0	**80.9**	**73.3**	72.8	65.5	**85.8**	**79.5**

tio, U-Net++ at 3:1, and FPN at 1:3, the average performance across all models indicates that 1:1 is the most generalizable ratio.

Additionally, Table 2 presents the evaluation results of the same segmentation model across multiple test datasets before and after augmentation. For PraNet, the Dice score averaged over all test datasets improved from 79.2% to 83.1% after augmentation with the proposed model at a 1:1 interpolation ratio, achieving SOTA performance. Similarly, for FCBFormer, the score increased from 84.1% to 85.8% under the same conditions, also setting a new SOTA performance.

Although ArSDM recorded the highest performance gain on the ETIS dataset for PraNet and on the Kvasir and ETIS datasets for FCBFormer, the proposed model with a 1:1 interpolation ratio consistently demonstrated the highest improvement, averaged over all test datasets, and outperformed other methods on the remaining test datasets. This confirms that augmentation with the 1:1 ratio of the proposed model provides the most robust generalization capability for segmentation models, as it offers a balanced interpolation between different data, unlike 1:3 or 3:1, which are more biased toward one mask.

Ablation Study The qualitative results in Fig. 3(b) show that w/o the background semantic label, interpolation does not occur in the background region, meaning the model generates only from the mask-conditioned estimated distribution without incorporating background variations. Similarly, Tables 1 and 2 confirm that even with 1:1 interpolation, which gained the highest performance with SIDM, performance improvement is minimal in w/o background semantic label.

This highlights that mask interpolation alone is insufficient; both masks and images must capture intermediate perspectives for better performance. These results validate the effectiveness of the proposed background semantic label in enhancing segmentation performance.

4 Conclusion

This study proposed SIDM, which interpolates masks and background semantic labels to generate interpolated colonoscopy mask-image pairs. Background semantic labels enable effective image interpolation and lead to improved generalization of downstream models. Experimental results showed that a 1:1 interpolation ratio consistently achieved the most improvement in segmentation performance through generative augmentation across multiple models and datasets. Overall, SIDM demonstrates strong potential as a generative augmentation method to improve the robustness and generalization of segmentation models in colonoscopy analysis.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

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)
2. Bo, D., Wenhai, W., Deng-Ping, F., Jinpeng, L., Huazhu, F., Ling, S.: Polyp-pvt: Polyp segmentation with pyramidvision transformers (2023)
3. Chen, Y., Yang, X.H., Wei, Z., Heidari, A.A., Zheng, N., Li, Z., Chen, H., Hu, H., Zhou, Q., Guan, Q.: Generative adversarial networks in medical image augmentation: a review. Computers in Biology and Medicine **144**, 105382 (2022)
4. Dayarathna, S., Islam, K.T., Uribe, S., Yang, G., Hayat, M., Chen, Z.: Deep learning based synthesis of MRI, CT and PET: Review and analysis. Medical Image Analysis p. 103046 (2023)
5. Dorjsembe, Z., Pao, H.K., Xiao, F.: Polyp-ddpm: Diffusion-based semantic polyp synthesis for enhanced segmentation. In: 2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). pp. 1–7 (2024)
6. Du, Y., Jiang, Y., Tan, S., Wu, X., Dou, Q., Li, Z., Li, G., Wan, X.: ArSDM: colonoscopy images synthesis with adaptive refinement semantic diffusion models. In: International conference on medical image computing and computer-assisted intervention. pp. 339–349. Springer (2023)
7. Edemekong, P.F., Annamaraju, P., Haydel, M.J.: Health Insurance Portability and Accountability Act (2024)
8. 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)

9. Garcea, F., Serra, A., Lamberti, F., Morra, L.: Data augmentation for medical imaging: A systematic literature review. Computers in Biology and Medicine **152**, 106391 (2023)
10. Hendrycks, D., Gimpel, K.: Gaussian error linear units (gelus). arXiv preprint arXiv:1606.08415 (2016)
11. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
12. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
13. 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)
14. Kazerouni, A., Aghdam, E.K., Heidari, M., Azad, R., Fayyaz, M., Hacihaliloglu, I., Merhof, D.: Diffusion models in medical imaging: A comprehensive survey. Medical Image Analysis **88**, 102846 (2023)
15. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2117–2125 (2017)
16. 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)
17. Qadir, H.A., Balasingham, I., Shin, Y.: Simple U-net based synthetic polyp image generation: Polyp to negative and negative to polyp. Biomedical Signal Processing and Control **74**, 103491 (2022)
18. 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)
19. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. pp. 234–241. Springer (2015)
20. Sanderson, E., Matuszewski, B.J.: Fcn-transformer feature fusion for polyp segmentation. In: Annual Conference on Medical Image Understanding and Analysis. pp. 892–907. Springer (2022)
21. Siegel, R.L., Giaquinto, A.N., Jemal, A.: Cancer statistics, 2024. CA: a cancer journal for clinicians **74**(1), 12–49 (2024)
22. Siegel, R.L., Miller, K.D., Goding Sauer, A., Fedewa, S.A., Butterly, L.F., Anderson, J.C., Cercek, A., Smith, R.A., Jemal, A.: Colorectal cancer statistics, 2020. CA: a cancer journal for clinicians **70**(3), 145–164 (2020)
23. 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)
24. 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)
25. Thambawita, V., Salehi, P., Sheshkal, S.A., Hicks, S.A., Hammer, H.L., Parasa, S., Lange, T.d., Halvorsen, P., Riegler, M.A.: SinGAN-Seg: Synthetic training data generation for medical image segmentation. PloS one **17**(5), e0267976 (2022)

26. 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**(1), 4037190 (2017)
27. Voigt, P., Von dem Bussche, A.: The eu general data protection regulation (gdpr). A Practical Guide, 1st Ed., Cham: Springer International Publishing **10**(3152676), 10–5555 (2017)
28. 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)
29. Wei, J., Hu, Y., Zhang, R., Li, Z., Zhou, S.K., Cui, S.: Shallow attention network for polyp segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24. pp. 699–708. Springer (2021)
30. Wu, Z., Lv, F., Chen, C., Hao, A., Li, S.: Colorectal Polyp Segmentation in the Deep Learning Era: A Comprehensive Survey. arXiv preprint arXiv:2401.11734 (2024)
31. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: UNet++: A nested U-Net architecture for medical image segmentation. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4. pp. 3–11. Springer (2018)