Accurate Boundary Alignment and Realism Enhancement for Colonoscopic Polyp Image-Mask Pair Generation

Riyu Qiu[1], Kun Xia[2], Feng Gao[3], Shuting Yang[3], Du Cai[3], Jiacheng Wang[4], Yinran Chen[1,✉], and Liansheng Wang[1,✉]

[1] Department of Computer Science and Technology, School of Informatics, Xiamen University, Xiamen 361005, China
qiuriyu@stu.xmu.edu.cn,yinran_chen@xmu.edu.cn,lswang@xmu.edu.cn

[2] Department of Gastroenterology, The National Key Clinical Specialty, Zhongshan Hospital of Xiamen University, School of Medicine, Xiamen University, Xiamen 361004, China
xiakun@stu.xmu.edu.cn

[3] The Sixth Affiliated Hospital, Sun Yat-sen University, Guangzhou 510655, China
gaof57@mail.sysu.edu.cn, yangsht35@mail.sysu.edu.cn, caid28@mail.sysu.edu.cn

[4] Manteia Technology Co. Ltd., Xiamen 361005, China
jiachengw@stu.xmu.edu.cn

Abstract. Polyp segmentation is the foundation of colonoscopic lesion screening, diagnosis, and therapy. However, the data size of images and annotations is limited. The latent diffusion model (LDM) has emerged as a powerful tool in synthesizing high-quality medical images with low computational costs. However, the challenges of boundary-aligned image-mask pairs and image realism remain unresolved, showing that (i) the spatial relationship between the boundaries is easily distorted in the latent space; (ii) the diversity of colors, shapes, and textures, along with low boundary contrast and textures similar to surrounding tissue, makes boundary distinction of the polyps difficult. This paper proposes Polyp-LDM that encodes polyps and masks into the same latent space via a unified variational autoencoder (VAE) to align their boundaries. Furthermore, Polyp-LDM refines texture and lighting while preserving the structure by fine-tuning the VAE decoder with data augmentation and applying the style cloning module to enhance image realism. Quantitative evaluations and user preference study demonstrate that our method outperforms existing methods in image-mask pair generation. Moreover, segmentation models trained with augmented data generated by polyp-LDM achieve the best performance on three public polyp datasets. The code is available at https://github.com/16rq/Polyp-LDM.

Keywords: diffusion model · paired images generation · boundary alignment · polyp segmentation.

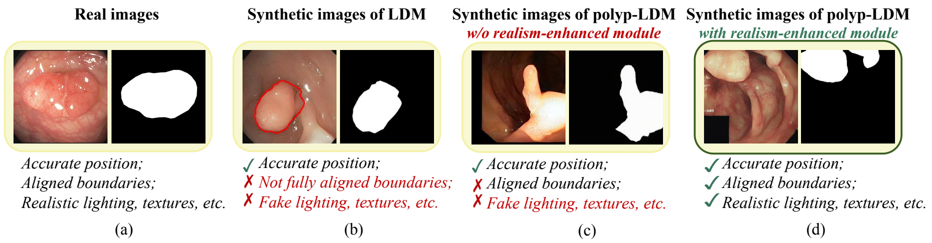
Real images
Synthetic images of LDM
Synthetic images of polyp-LDM
w/o realism-enhanced module
Synthetic images of polyp-LDM
with realism-enhanced module
Accurate position;
Aligned boundaries;
Realistic lighting, textures, etc.
✓ Accurate position;
✗ Not fully aligned boundaries;
✗ Fake lighting, textures, etc.
✓ Accurate position;
✗ Aligned boundaries;
✗ Fake lighting, textures, etc.
✓ Accurate position;
✓ Aligned boundaries;
✓ Realistic lighting, textures, etc.
(a)
(b)
(c)
(d)

Fig. 1. Comparison of synthetic results in LDM, polyp-LDM without and with realism-enhanced module and real images. Red curves sketch synthetic boundary masks.

1 Introduction

Colonic polyp segmentation is crucial in screening, diagnosis, and therapy in clinical endoscopy, as the polyps are lesions closely associated with colorectal cancer (CRC) [5]. In recent years, deep learning has been extensively applied to polyp segmentation [6, 19]. However, their effectiveness is still limited by the scarcity of data and annotations, primarily due to privacy concerns and the high cost of manual labeling [11]. Although traditional data augmentation techniques, such as rotation and flipping, could, to some extent, help expand the dataset. These methods are limited in their ability to scale datasets and increase diversity.

Recently, diffusion models (DMs) have emerged as powerful generative models for producing high-quality and diverse images [18]. DMs have been applied to augment medical image datasets like skin lesions and chest X-rays [15]. The models have also been used to generate more challenging labeled data, such as polyp images and breast MRI with masks [12, 7]. The strength of DMs lies in their stochastic process that generates multiple predictions over multiple time steps. This potential allows them to capture the ambiguous boundaries of polyps, which has been confirmed by several DM-based segmentation models [21, 20].

However, several challenges remain in generating boundary-aligned and realistic polyp-mask pairs. First, existing generative methods often fail to precisely align the boundaries between the generated masks and polyps, as illustrated in Fig. 1(b). In this aspect, current techniques, such as latent diffusion models (LDM), can generate polyps with accurate locations but misaligned boundaries. Misalignment will significantly affect the performance of the downstream tasks. Second, traditional generative models often do not take into account the fundamental variations in lighting, color, and texture of endoscopic images. Fig. 1(b) illustrates that the synthetic polyp images do not exhibit realistic lighting, texture, and other details. As a result, the generated polyp images may fail to capture the complexity of real-world scenarios and lose realism.

In this paper, we propose an LDM-based model, Polyp-LDM, to ensure boundary alignment and realism enhancement of generated polyp image-mask pairs in colonoscopy. First, we introduce a novel sampling scheme grounded in a unified latent space to strengthen the spatial relationship between polyps and masks. Specifically, we train a shared Variational Autoencoder (VAE) that maps polyp images and masks from pixel space to a unified latent space. During sampling, we generate polyp image-mask pairs from the shared initial noise to guarantee precise boundary alignment. This approach effectively preserves the structural coherence of the polyp-mask relationship and realizes more accurate generation. Second, we enhance the realism of generated images by leveraging the full potential of VAE decoder. Specifically, we fine-tune the VAE decoder by adjusting the visual attributes of the polyp and background to simulate lighting and contrast variations. This approach ensures realistic image reconstructions that are closer to real-world scenarios. In post-processing, style cloning is integrated into the pipeline to enhance image realism while preserving polyp structure.

2 Methods

In this section, we first introduce the basic image-mask pair generation model, extended by LDM. Then, we describe the novel extensions in Polyp-LDM that tackle the problem of boundary misalignment and lack of realism.

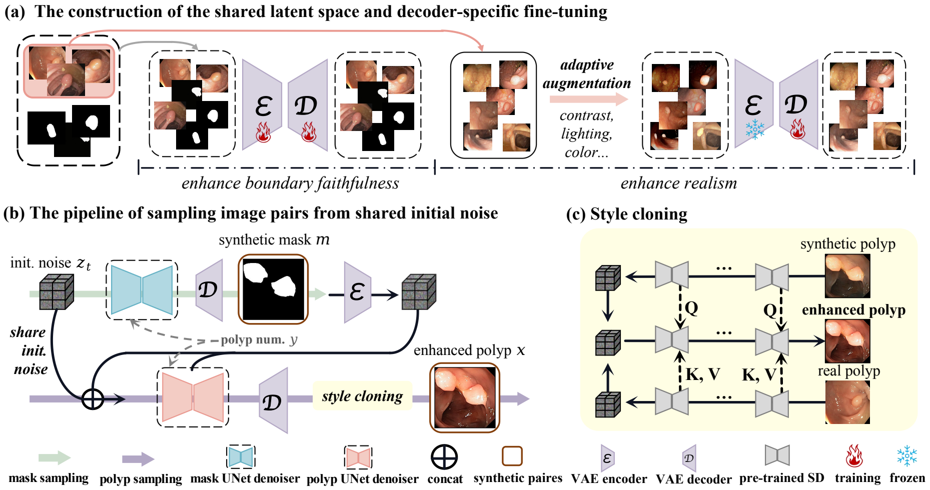

Fig. 2. The overview of polyp-LDM. **(a)** The two-stage VAE fine-tuning strategy enables precise polyp and mask reconstruction in a unified latent space. The decoder is refined to enhance realism by correcting color, brightness, and contrast. **(b)** The synthetic masks are first generated and followed by polyp generation conditioned on masks. Both generators start from shared initial noise. **(c)** Style cloning incorporates real-world features into synthetic polyps by modifying the self-attention layers.

2.1 Basic Paired Polyp Image-Mask Generation Model

The basic LDM includes a perceptual compression model VAE and a U-Net denoiser. The VAE consists of an encoder $\mathcal{E}$ that encodes an image $x \in \mathbb{R}^{3\times256\times256}$

in pixel space into a latent representation $z = \mathcal{E}(x) \in \mathbb{R}^{4\times 32\times 32}$ and a decoder $\mathcal{D}$ that reconstructs the image $\tilde{x} = \mathcal{D}(z) = \mathcal{D}(\mathcal{E}(x))$. A modified U-Net denoiser ϵ_θ generates the latent embeddings and performs diffusion process in the latent space. Two separate LDMs, each with its own VAE and U-Net denoiser, are used to generate polyp image–mask pairs. In the polyp generator, the synthetic mask m acts as an additional condition $\mathcal{E}_m(m)$ to guide polyp generation by channel concatenation. The fused features are fed into the U-Net backbone ϵ_{θ_x} to generate x. Two generators both incorporate a text prompt y via a text embedder τ_θ and cross-attention in U-Net denoiser ϵ_{θ_m}. The prompt specifies the desired number of polyps in the images.

2.2 Boundary Alignment via Unified Latent Space and Shared Initial Noise

Spatial relationships between image and mask are easily distorted during compression, decoding, and diffusion in latent space, resulting in imprecise boundary alignment. Two strategies are proposed to improve the alignment. First, the polyp and mask are mapped to the unified latent space using a shared VAE with an encoder $\mathcal{E}_u$ and a decoder $\mathcal{D}_u$. This unified VAE learns the features of the image and mask while preserving their boundary relationship during reconstruction. To build the unified latent space, we fine-tune the pre-trained Stable Diffusion VAE (SD-VAE) [18] with full parameter fine-tuning, as illustrated in Fig. 2(a). By fine-tuning, we adapt SD-VAE into a perceptual model that enables more accurate reconstruction of polyp and mask images. Based on this shared VAE, we train the UNet denoisers ϵ_{θ_m} and ϵ_{θ_x} for mask m and polyp x, respectively. As shown in Fig. 2(b), the synthetic mask and its corresponding initial noise are passed to the polyp generator as conditions and starting points during sampling. This strategy enables the generation of polyp image-mask pairs from the shared initial noise in the unified VAE, thereby improving boundary alignment. Accordingly, the objective could be formulated as:

$$\mathcal{L}_{mask} := \mathbb{E}_{\mathcal{E}_u(m),y,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_{\theta_m}(c_t, t, \tau_\theta(y))\|_2^2\right] \tag{1}$$

$$\mathcal{L}_{polyp} := \mathbb{E}_{\mathcal{E}_u(x),\mathcal{E}_u(m),y,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_{\theta_x}(z_t \oplus \mathcal{E}_u(m), t, \tau_\theta(y))\|_2^2\right] \tag{2}$$

where c and z are masks and polyps in the latent space. $\mathcal{E}_u$ is the encoder of the unified VAE. τ_θ is the text embedder. t denotes the time steps.

2.3 Realism Enhancement via Augmented Fine-tuning and Style Cloning

Endoscopic images exhibit special lighting, contrast, and texture features due to the single-light source, tubular imaging scenario, and structure of the intestinal inner surface. Image realism is particularly important for clinical endoscopic applications. We fine-tune the SD-VAE decoder and use style cloning to enhance image realism. During fine-tuning, we introduce more variations in lighting and color to further exploit the generative potential of SD-VAE. We separate the polyp from its background to independently adjust its brightness and contrast, while still requiring the model to restore the original appearance. This strategy corrects artifacts in generated images, a departure from conventional whole-image transforms. The objective applied to the augmented image $\hat{x}$ is a variant of the VAE loss used in LDM [18], and is defined as:

$$\mathcal{L}_{\mathcal{D}} := \min_{\mathcal{D}} \max_{\mathcal{D}_\psi} (\mathcal{L}_{rec}(x, \mathcal{D}(\mathcal{E}(\hat{x}))) - \mathcal{L}_{adv}(\mathcal{D}(\mathcal{E}(\hat{x}))) + \log \mathcal{D}_\psi(x) + \mathcal{L}_{reg}(\hat{x}; \mathcal{E}, \mathcal{D})) \tag{3}$$

where $\hat{x} = f(x, \alpha, \beta)$. f is the augmentation function, x is the original polyp image, and α and β control brightness and contrast (each with 0.5 probability). $\mathcal{L}_{rec}$ is the reconstruction loss, $\mathcal{L}_{adv}$ is the adversarial loss, $\mathcal{L}_{reg}$ is the regularization loss, and $\mathcal{D}_\psi$ the discriminator with parameters ψ.

During post-processing, the synthetic and real polyp noises are first obtained via DDIM inversion. Then, the texture and style of real polyps are injected by modifying self-attention layers during diffusion, as illustrated in Fig. 2(c). The self-attention mechanism is modified as follows:

$$\tilde{Q}_t = \alpha \cdot Q_t^{syn} + (1 - \alpha) \cdot Q_t, \quad K_t = K_t^{real}, \quad V_t = V_t^{real} \tag{4}$$

where $\tilde{Q}_t$ means the mixed query from queries of synthetic polyp representation Q_t^{syn} and Q_t itself at time step t. K_t and V_t are directly replaced with key K_t^{real} and value V_t^{real} of the real polyp representation. α is the preservation ratio.

3 Experiments and Results

3.1 Dataset and Preprocessing

The experiments were conducted on five public datasets containing polyp image-mask pairs, namely Kvasir-SEG (1,000 pairs) [10], PolypGen (1,537 pairs) [1], CVC-ClinicDB (601 pairs) [2], HyperKvasir (1000 pairs) [4], and CVC-ColonDB (380 pairs) [3]. The training set comprised data from Kvasir-SEG and five centers (C1, C3, C4, C5, and C6) in PolypGen. CVC-ClinicDB was utilized as the validation set. The C2 (301 pairs) of PolypGen, HyperKvasir, and CVC-ColonDB were used as test sets to evaluate the performance of polyp segmentation.

In data preprocessing, all images were resized to 256×256 through scaling and random cropping. In VAE decoder fine-tuning, data augmentation was applied by independently adjusting the brightness and contrast of the foreground (polyps) and background with a probability of 0.5.

3.2 Implementation Details and Evaluation Metrics

We implemented the proposed polyp-LDM and the compared methods in PyTorch using an NVIDIA RTX 3090 GPU. Specifically, polyp-LDM was trained for 1000 epochs using default settings. The downstream segmentation tasks were

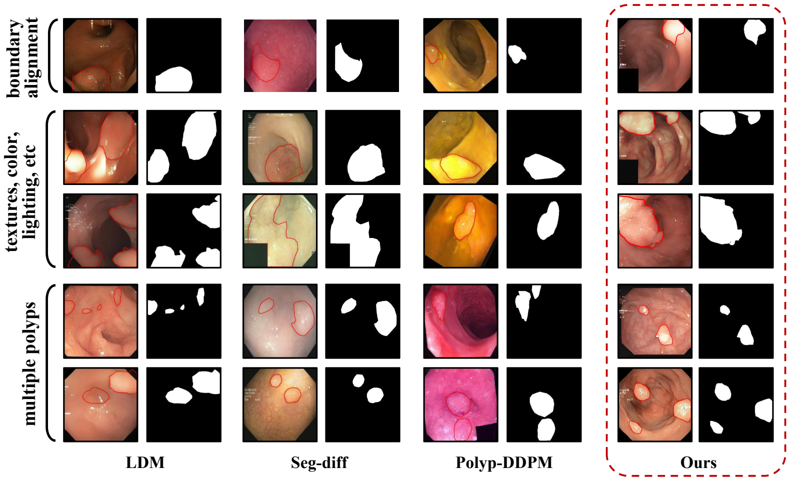

Fig. 3. Visual comparison of polyp-LDM with other mask-conditional image generation methods. The masks guide polyp generation. Red curves sketch the expected boundary. Although consistent masks across methods are unachievable as not all methods use real masks for polyp generation, we have minimized mask variability in specific conditions.

Table 1. Qualitative evaluation and average user preference of the synthetic images. The bolded values indicate the best and the underlined values the second-best.

Configuration	Qualitative evaluation					User study	
	Dice(m, m_s)↑	FID↓	CLIP-FID↓	KID↓	CMMD↓	Quality↑	Mask Fidelity↑
LDM [18]	44.20	104.07	14.27	**0.07**	1.31	1.38	1.50
Seg-Diff [13]	58.90	136.96	20.90	0.13	0.80	1.62	1.69
Polyp-DDPM [7]	42.50	163.65	36.46	0.15	2.44	1.60	1.65
w/o realism	67.66	103.85	14.64	0.09	1.40	-	-
w/o boundary	42.97	88.18	10.86	**0.07**	0.71	-	-
Ours	**79.00**	**82.92**	**6.95**	**0.07**	**0.58**	**2.22**	**2.46**

realized using Polyp-PVT [6] through 100-epoch training. The compared methods were trained using default settings.

Frechet Inception Distance (FID), CLIP-FID [14], Kernel Inception Distance (KID), and CLIP-enhanced Maximum Mean Discrepancy (CMMD) [9] are adopted to evaluate the quality of generated images. FID and KID are improved versions that address image resizing and compression [17]. CLIP-FID and CMMD use CLIP features to measure the distribution distance between real and synthetic images. Similarly to ControlNet [22], 100 synthetic images generated by each method were graded by an endoscopist according to the quality of generation and mask fidelity (1-3 score, lower is worse). To evaluate boundary alignment, a classic segmentation model Polyp-PVT [6] was trained on real data to predict m_s for synthetic polyp, and compute Dice(m, m_s) with mask guidance m. In the downstream tasks, Dice and Average Surface Distance (ASD) were used to assess segmentation performance.

3.3 Enhanced Boundary Alignment and Realism

Table 1 lists the quantitative evaluations and average user preference on the synthetic images. polyp-LDM significantly improves Dice to 79.00 when compared with Seg-Diff (58.90) [13] and Polyp-DDPM (42.50) [7]. In the user preference study, our method attains the highest generative quality score of 2.22 and mask fidelity score of 2.46, showing significant improvements when compared with Seg-Diff and Polyp-DDPM.

In the realism study, polyp-LDM demonstrates superior performance in generating realistic and diverse images, as evidenced by the lowest FID of 82.92, which outperforms LDM (104.07) [18] and Seg-Diff (136.96) [13]. In addition, our method performs well with the lowest KID of 0.07, CLIP-FID of 6.95, and CMMD of 0.58, indicating that our synthetic images contribute to more natural and convincing visualizations.

Fig. 3 visually compares polyp-LDM with existing mask-conditional generation models. One can see that the existing methods struggle to align complex boundaries beyond simple ellipses. Moreover, their lighting, colors, and textures appear artificial, which may deteriorate their realism. In contrast, our method achieves precise boundary alignment and generates colors, textures, and lighting more consistently with real-world scenarios.

3.4 Improvement in Downstream Polyp Segmentation

We explored the impact of synthetic data along with unchanged real data on downstream polyp segmentation tasks. We trained a classic polyp segmentation model, Polyp-PVT [6], with synthetic and real data in different ratios denoted as R, R2S1, R1S1, and R1S2, and R and S denote real and synthetic, respectively.

Table 2 presents the segmentation results of Polyp-PVT trained with different dataset configurations. In general, the segmentation performance improves substantially with the introduction of synthetic data. Notably, polyp-LDM achieves the best performance across all the configurations. Particularly on the HyperKvasir dataset, Dice is improved up to 99.09 from 96.90 when adding the synthetic data to training. In the CVC-ColonDB dataset with R1S1, polyp-LDM is slightly worse than Seg-Diff in Dice. However, it outperforms the compared methods with a Dice of 78.34 and an ASD of 21.07 in the R1S2 configuration.

In summary, polyp-LDM maintains stable segmentation performances without degradation, demonstrating that the data generated by our method does not mislead the learning capacity of the segmentation model.

3.5 Ablation Study

The ablation study was performed by removing the boundary alignment module (*w/o boundary*) or the realism-enhanced module (*w/o realism*). The corresponding results have been integrated into Table 1 and Table 2.

As shown in Table 1, removing the realism-enhanced module deteriorates the alignment between generated pairs, showing a decrease of Dice from 79.00

Table 2. Segmentation performance of Polyp-PVT on PolypGen (C2), HyperKvasir, and CVC-ColonDB. RxSy denotes the training dataset that mixes real-world (R) and synthetic (S) data in a ratio of $x : y$, with real data fixed. Bold denotes the best at each ratio, underline the second-best, and bold underline the best overall.

Dataset	**Model**	PolypGen (C2)		HyperKvasir		CVC-ColonDB	
		Dice↑	ASD↓	Dice↑	ASD↓	Dice↑	ASD↓
R		61.55	51.35	96.90	3.47	74.75	33.09
R2S1	LDM [18]	76.12	20.38	92.69	13.92	75.14	43.79
	Seg-Diff [13]	77.23	15.53	95.19	8.48	78.13	28.73
	Polyp-DDPM [7]	80.14	14.31	97.64	6.49	75.32	25.92
	w/o realism	77.45	17.04	95.71	10.65	77.05	29.80
	w/o boundary	77.77	16.61	97.21	6.27	77.16	27.79
	Ours	**81.58**	**10.42**	**99.09**	**1.65**	**78.34**	**21.07**
R1S1	LDM [18]	77.77	15.01	96.64	7.35	73.68	28.00
	Seg-Diff [13]	79.63	14.69	96.95	4.76	**78.09**	25.14
	Polyp-DDPM [7]	78.71	19.29	95.21	12.53	74.49	34.08
	w/o realism	78.63	13.04	**98.60**	6.06	77.23	28.70
	w/o boundary	79.63	11.15	98.16	**3.30**	73.76	**17.99**
	Ours	**81.09**	**10.64**	98.31	3.44	76.29	22.40
R1S2	LDM [18]	80.83	15.00	97.51	6.69	74.30	19.98
	Seg-Diff [13]	78.01	15.97	96.87	9.15	73.53	34.46
	Polyp-DDPM [7]	81.34	12.78	97.52	6.19	75.39	**19.70**
	w/o realism	78.78	13.27	97.32	7.97	**77.54**	30.54
	w/o boundary	80.20	13.57	97.43	5.59	74.94	21.27
	Ours	**82.43**	**10.75**	**98.47**	**2.59**	77.18	21.55

to 67.66. Furthermore, FID and KID, which evaluate the realism and diversity of generated images using Inception features, increase from 82.92 to 103.85 and from 6.95 to 14.64, respectively. Similarly, CLIP-FID and CMMD, based on CLIP features, also improve with the realism-enhanced module. Table 2 indicates that the realism-enhanced module improves segmentation performance across all datasets and training configurations.

On the other hand, removing the boundary alignment module results in poor alignment in generated image pairs (see Table 1). The Dice drops from 79.00 to 42.97, reflecting a decrease in boundary faithfulness. FID and KID also decrease, indicating a decline in realism and diversity. Table 1 indicates that *w/o boundary* has a less significant impact on realism compared to *w/o realism*. Similarly, Table 2 also illustrates the effectiveness of this module in enhancing segmentation performance across all datasets and training configurations.

4 Conclusion

This paper proposes Polyp-LDM for paired polyp image-mask generation in colonoscopy. Recently, generative models for complex and critical endoscopic scenes have drawn significant attention [16, 8]. Our method achieves precise boundary alignment by unifying the latent spaces and starting points of images and masks. The realism of generated polyps is significantly enhanced through VAE decoder fine-tuning and style cloning. Experiments demonstrate that our method can generate realistic polyp images with aligned boundaries. The model also serves as an effective data augmentation tool to improve the performance of the downstream tasks. This study focuses on the generative capabilities of diffusion models and their data augmentation potential, without extensive comparison to other generative paradigms or conventional augmentation methods. Future work will extend this method to diverse medical imaging modalities and downstream tasks to validate its broader applicability, including comprehensive comparisons with conventional augmentation methods.

Acknowledgments. This work was supported by National Natural Science Foundation of China (Grant No. 62371409) and Fujian Provincial Natural Science Foundation of China (Grant No. 2023J01005).

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

References

1. Ali, S., Jha, D., Ghatwary, N., Realdon, S., Cannizzaro, R., Salem, O.E., Lamarque, D., Daul, C., Riegler, M.A., Anonsen, K.V., et al.: A multi-centre polyp detection and segmentation dataset for generalisability assessment. Scientific Data **10**(1), 75 (2023)
2. 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)
3. Bernal, J., Sánchez, J., Vilarino, F.: Towards automatic polyp detection with a polyp appearance model. Pattern Recognition **45**(9), 3166–3182 (2012)
4. Borgli, H., Thambawita, V., Smedsrud, P.H., Hicks, S., Jha, D., Eskeland, S.L., Randel, K.R., Pogorelov, K., Lux, M., Nguyen, D.T.D., et al.: Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. Scientific Data **7**(1), 283 (2020)
5. De Leon, M.P., Di Gregorio, C.: Pathology of colorectal cancer. Digestive and Liver Disease **33**(4), 372–388 (2001)
6. Dong, B., Wang, W., Fan, D.P., Li, J., Fu, H., Shao, L.: Polyp-pvt: Polyp segmentation with pyramid vision transformers. CAAI Artificial Intelligence Research **2**, 9150015 (2023). https://doi.org/10.26599/AIR.2023.9150015
7. 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). p. 1–7. IEEE (Jul 2024). https://doi.org/10.1109/embc53108.2024.10782077, http://dx.doi.org/10.1109/EMBC53108.2024.10782077
8. Golhar, M.V., Bobrow, T.L., Ngamruengphong, S., Durr, N.J.: Gan inversion for data augmentation to improve colonoscopy lesion classification. IEEE Journal of Biomedical and Health Informatics (2024)

9. Jayasumana, S., Ramalingam, S., Veit, A., Glasner, D., Chakrabarti, A., Kumar, S.: Rethinking fid: Towards a better evaluation metric for image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9307–9315 (2024)
10. 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)
11. 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)
12. Konz, N., Chen, Y., Dong, H., Mazurowski, M.A.: Anatomically-controllable medical image generation with segmentation-guided diffusion models. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (2024)
13. Konz, N., Chen, Y., Dong, H., Mazurowski, M.A.: Anatomically-controllable medical image generation with segmentation-guided diffusion models. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 88–98. Springer Nature Switzerland (2024)
14. Kynkäänniemi, T., Karras, T., Aittala, M., Aila, T., Lehtinen, J.: The role of imagenet classes in fr\'echet inception distance. arXiv preprint arXiv:2203.06026 (2022)
15. Luo, Y., Yang, Q., Fan, Y., Qi, H., Xia, M.: Measurement guidance in diffusion models: Insight from medical image synthesis. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)
16. Mathew, S., Nadeem, S., Kaufman, A.: Clts-gan: color-lighting-texture-specular reflection augmentation for colonoscopy. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 519–529. Springer (2022)
17. Parmar, G., Zhang, R., Zhu, J.Y.: On aliased resizing and surprising subtleties in gan evaluation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (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 (June 2022)
19. Wang, H., Wang, K.N., Hua, J., Tang, Y., Chen, Y., Zhou, G.Q., Li, S.: Dynamic spectrum-driven hierarchical learning network for polyp segmentation. Medical Image Analysis **101**, 103449 (2025). https://doi.org/https://doi.org/10.1016/j.media.2024.103449
20. Wang, J., Yang, J., Zhou, Q., Wang, L.: Medical boundary diffusion model for skin lesion segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 427–436. Springer (2023)
21. Wu, J., Ji, W., Fu, H., Xu, M., Jin, Y., Xu, Y.: Medsegdiff-v2: Diffusion-based medical image segmentation with transformer. Proceedings of the AAAI Conference on Artificial Intelligence **38**(6), 6030–6038 (Mar 2024). https://doi.org/10.1609/aaai.v38i6.28418, https://ojs.aaai.org/index.php/AAAI/article/view/28418
22. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3836–3847 (2023)