Latent Space Synergy: Text-Guided Data Augmentation for Direct Diffusion Biomedical Segmentation

Muhammad Aqeel[1(✉)], Maham Nazir[2], Zanxi Ruan[1], and Francesco Setti[1]

[1] Department of Engineering for Innovation Medicine, University of Verona, Strada le Grazie 15, Verona, Italy
muhammad.aqeel@univr.it

[2] Department of Computer Science, Beihang University, Beijing, China

Abstract. Medical image segmentation suffers from data scarcity, particularly in polyp detection where annotation requires specialized expertise. We present SynDiff, a framework combining text-guided synthetic data generation with efficient diffusion-based segmentation. Our approach employs latent diffusion models to generate clinically realistic synthetic polyps through text-conditioned inpainting, augmenting limited training data with semantically diverse samples. Unlike traditional diffusion methods requiring iterative denoising, we introduce direct latent estimation enabling single-step inference with T× computational speedup. On CVC-ClinicDB, SynDiff achieves 96.0% Dice and 92.9% IoU while maintaining real-time capability suitable for clinical deployment. The framework demonstrates that controlled synthetic augmentation improves segmentation robustness without distribution shift. SynDiff bridges the gap between data-hungry deep learning models and clinical constraints, offering an efficient solution for deployment in resource-limited medical settings.

Keywords: Medical Image Segmentation · Diffusion Model · Polyp Detection · Text-Guided Synthesis

1 Introduction

Biomedical image segmentation plays a critical role in modern healthcare, enabling precise diagnosis and treatment planning [14,22]. In gastrointestinal endoscopy, automated polyp segmentation has emerged as a particularly important application with the potential to improve colorectal cancer screening accuracy and reduce missed detection rates during real-time procedures. However, data scarcity represents the most fundamental bottleneck limiting the development of robust medical segmentation systems. Medical datasets are constrained

M. Aqeel and M. Nazir—Equal contribution.

E. Rodolà et al. (Eds.): ICIAP 2025 Workshops, LNCS 16169, pp. 417–428, 2026.
https://doi.org/10.1007/978-3-032-11317-7_35

by privacy regulations, costly expert annotation, and time-intensive boundary delineation [15,18]. In polyp segmentation, this challenge is compounded by significant morphological diversity, requiring extensive annotated examples that current public datasets cannot provide [5]. Traditional data augmentation in medical imaging relies on geometric transformations [22,23]. While providing some benefit, these approaches cannot generate new pathological variations needed for robust model generalization. Recent GAN-based synthesis methods have shown promise but suffer from limited controllability and mode collapse [28].

The emergence of diffusion probabilistic models has revolutionized image generation [13,17]. Text-conditioned diffusion models enable semantically-guided data augmentation through clinical descriptions [21]. However, traditional diffusion-based segmentation requires computationally intensive multi-step inference, making it impractical for clinical deployment [27,29]. Latent diffusion models (LDMs) address these computational limitations by operating in compressed latent spaces [21]. Their text-conditioning capabilities present an opportunity to create clinically-informed augmentation strategies, where domain expertise guides the generation of diverse pathological variations [19].

In this paper, we present SynDiff, a framework that addresses medical data scarcity through text-guided synthetic data augmentation while maintaining computational efficiency for practical deployment. Our approach leverages latent diffusion models to generate diverse synthetic polyp images guided by clinical descriptions, effectively expanding training datasets with semantically meaningful variations. We integrate this with a direct latent estimation technique enabling single-step segmentation inference, eliminating the computational burden of iterative denoising while preserving performance. Our contributions are:

- A text-guided data augmentation framework using latent diffusion models to address medical data scarcity through semantically-controlled synthetic polyp generation.
- An efficient single-step segmentation approach that maintains competitive performance while dramatically reducing computational requirements compared to multi-step diffusion methods.
- Comprehensive evaluation demonstrating that synthetic data augmentation preserves segmentation quality (96.0% Dice, 92.9% IoU) while expanding dataset diversity for improved model robustness.

2 Related Work

Medical image segmentation has evolved significantly with deep learning approaches, with U-Net [22] establishing the encoder-decoder paradigm that remains foundational for medical tasks. Recent advances include transformer-based architectures like TransUNet [7] and UNETR [12] that capture long-range dependencies, and specialized polyp segmentation methods such as PraNet [10]

and SANet [26]. Despite these architectural innovations, all approaches fundamentally require extensive annotated datasets to achieve robust performance, a constraint that significantly limits practical clinical deployment.

Traditional augmentation through geometric transformations fails to capture pathological diversity [5]. Alternative approaches address data scarcity through meta-learning [1,4] and self-supervised refinement [2,3], which enhance model robustness by learning better representations from limited data. In contrast, synthetic generation methods directly create new training samples: GANs suffer from mode collapse and training instability [8], while recent diffusion-based approaches like DiffBoost [30] demonstrate superior controllability but lack integration with downstream tasks. Our work bridges this gap by combining controllable synthesis with task-specific optimization.

Diffusion models have emerged as powerful generative frameworks, with medical applications explored in MedSegDiff [28] and MedSegDiff-V2 [27] that treat segmentation as conditional generation through iterative denoising. Latent Diffusion Models [21] reduced computational overhead by operating in compressed latent spaces while enabling text-conditioning for semantic control. However, existing approaches require multi-step inference processes that are computationally prohibitive for clinical deployment [29], while single-step inference methods explored in natural image domains [25] remain underexplored for medical segmentation.

While progress has been made in individual components—synthetic medical data generation, efficient diffusion inference, and medical segmentation—no prior work integrates these elements to address both data scarcity and computational efficiency simultaneously. Our work fills this gap by combining text-guided synthetic data generation with single-step diffusion segmentation, enabling enhanced training diversity while maintaining practical deployment efficiency.

3 SynDiff Framework

Our proposed approach, SynDiff, inspired by [11,16], integrates text-guided synthetic data generation with single-step diffusion segmentation to address the dual challenges of limited training data and computational efficiency in biomedical image segmentation as shown in Fig. 1. The framework operates in two distinct phases: synthetic data generation using Stable Diffusion XL (SDXL) [20] for text-guided inpainting, followed by end-to-end training of a single-step segmentation model that processes both real and synthetic data.

3.1 Latent Diffusion Model

Both the data generation and segmentation components operate within a latent diffusion framework to reduce computational overhead. For the segmentation pipeline, we employ a trainable vision encoder τ_θ to encode an image $C \in \mathbb{R}^{H\times W\times 3}$ into its latent representation $z_c = \tau_\theta(C)$. For segmentation maps, we

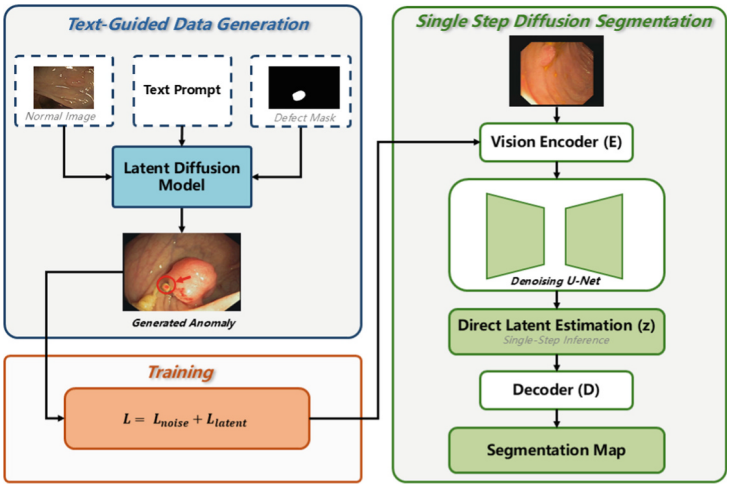

Fig. 1. Overview of the SynDiff framework. (Left) Offline text-guided data generation: Stable Diffusion XL (SDXL) inpainting takes a normal endoscopic image, clinical text prompt, and binary mask to generate synthetic polyp images with corresponding ground truth masks. (Right) Single-step segmentation pipeline: input images are encoded through a trainable vision encoder, processed by a denoising U-Net for direct latent estimation, and decoded to produce segmentation masks in a single inference step.

leverage a pre-trained autoencoder with encoder E and decoder D for perceptual compression. Given a segmentation map $X \in \mathbb{R}^{H\times W\times 3}$ in pixel space, the encoder produces a latent representation $z = E(X)$, and the decoder recovers the segmentation map as $\hat{X} = D(z)$, where $z \in \mathbb{R}^{h\times w\times c}$. For the data generation pipeline, SDXL operates in its latent space where the normal image, mask, and text prompt are jointly processed—the text conditions the diffusion process through cross-attention while the mask guides spatial inpainting. Both pipelines operate entirely in compressed latent spaces, significantly reducing computational requirements while preserving essential structural information.

3.2 Text-Guided Synthetic Data Generation

To address medical data scarcity, we implement an offline synthetic data generation process using SDXL inpainting conditioned on clinical text descriptions. The generation process takes three inputs: a normal endoscopic image i_n randomly sampled from available normal cases, a text description d_s specifying desired polyp characteristics such as "small sessile polyp with irregular surface texture," and a binary mask m_s indicating the spatial region where the polyp should be synthesized. Within the SDXL latent diffusion framework, these inputs are jointly processed: the text prompt conditions the generation through cross-attention mechanisms while operating with frozen pre-trained weights to generate a synthetic image $i_s = \text{SDXL}_{\text{inpaint}}(i_n, d_s, m_s)$ containing a realistic polyp

within the masked region. The binary mask m_s simultaneously serves as the ground truth segmentation label for the generated synthetic image i_s. We generate 100 synthetic samples using 50 diverse text prompts describing various polyp morphologies, sizes, and surface characteristics, augmenting our training dataset with approximately 20% additional data.

3.3 Direct Latent Estimation for Single-Step Segmentation

Our segmentation component introduces a direct latent estimation strategy that enables single-step inference, achieving theoretical computational speedup of $T\times$ compared to traditional diffusion approaches requiring T denoising steps. During training, the latent representation of a segmentation map z_0 undergoes forward diffusion by adding Gaussian noise for t timesteps to obtain:

$$z_t = \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1 - \bar{\alpha}_t} n \tag{1}$$

where n is random Gaussian noise and $\bar{\alpha}_t$ controls the noise schedule. The denoising U-Net $f(\cdot)$ is trained to estimate the noise as:

$$\tilde{n} = f(z_t, z_c) \tag{2}$$

where z_c is the conditioning image's latent representation. The key innovation lies in directly estimating the clean latent z_0 from the noisy latent z_t in a single step through:

$$\tilde{z}_0 = \frac{1}{\sqrt{\bar{\alpha}_t}}(z_t - \sqrt{1 - \bar{\alpha}_t}\tilde{n}) \tag{3}$$

This enables a dual supervision strategy where we minimize both noise prediction error:

$$\mathcal{L}_{\text{noise}} = ||n - \tilde{n}||_1 \tag{4}$$

and direct latent estimation error:

$$\mathcal{L}_{\text{latent}} = ||z_0 - \tilde{z}_0||_1 \tag{5}$$

The combined loss function is:

$$\mathcal{L} = \mathcal{L}_{\text{noise}} + \lambda \mathcal{L}_{\text{latent}} \tag{6}$$

with $\lambda = 1$, enabling the model to learn both denoising and direct prediction capabilities simultaneously. The dual supervision strategy is particularly effective for segmentation tasks due to the discrete nature of segmentation masks, which allows for more stable latent predictions and direct optimization of boundary precision—critical for accurate polyp delineation. By leveraging the structural simplicity of binary masks, our approach achieves single-step inference while maintaining segmentation quality, making it suitable for real-time clinical deployment.

3.4 Latent Fusion and Training Protocol

To integrate image features with segmentation information efficiently, we implement concatenation-based latent fusion rather than computationally expensive cross-attention mechanisms. The latent representations of the conditioning image z_c and segmentation map are concatenated along the channel dimension before processing by the denoising U-Net. Our trainable vision encoder τ_θ shares the same architecture as the segmentation encoder E but remains trainable to adapt pre-trained natural image features to medical imaging characteristics. During training, the vision encoder τ_θ and denoising U-Net $f(\cdot)$ are updated through backpropagation, while the autoencoder components (E, D) and SDXL inpainting model remain frozen with pre-trained weights. The training proceeds with random timestep sampling $t \sim \text{Uniform}(1, 1000)$ during forward diffusion, ensuring the model learns to handle varying noise levels for robust single-step inference.

3.5 Inference and Computational Efficiency

During inference, the framework processes input medical images through a single forward pass, eliminating the iterative sampling required by traditional diffusion methods. An input image is encoded through the trainable vision encoder τ_θ to obtain z_c, which conditions the denoising U-Net to directly estimate the segmentation latent $\tilde{z}_0$ at a fixed timestep $t = 50$ (empirically determined for optimal quality-efficiency balance). The estimated latent is decoded through the frozen decoder D to produce the final segmentation mask. This single-step approach reduces computational complexity from $O(T \times \text{U-Net forward pass})$ to $O(1 \times \text{U-Net forward pass})$, making the method suitable for real-time clinical deployment while maintaining competitive segmentation accuracy on both real and synthetic training data.

4 Experimental Results

4.1 Dataset and Evaluation Metrics

We evaluate SynDiff on the CVC-ClinicDB [6] dataset, consisting of 550 RGB colonoscopy images with 488 training and 62 test images. Each image includes precise polyp boundary annotations for binary segmentation. To ensure statistical reliability, we report Dice Coefficient (DC) and Intersection over Union (IoU) metrics with mean $\pm$ standard deviation across 5-fold cross-validation. Additional boundary quality metrics include Hausdorff Distance at 95th percentile (HD95) and Normalized Surface Distance (NSD) for clinical precision assessment, as accurate boundary delineation is crucial for surgical planning in clinical practice.

4.2 Implementation Details

Text-Guided Data Generation: We implement SDXL [20] inpainting using the Diffusers library with carefully crafted clinical prompts including "sessile polyp with irregular surface texture," "pedunculated polyp on mucosal fold," and "flat adenomatous lesion" while employing negative prompts "smooth healthy tissue, normal colon wall" to ensure realistic synthesis. The inpainting process utilizes binary masks derived from original dataset annotations, ensuring synthetic polyps are placed in anatomically plausible locations.

Training: Models are trained for 100,000 steps using AdamW optimizer (learning rate $= 1 \times 10^{-5}$, batch size $= 4$) on NVIDIA RTX 4090. We use a KL-regularized autoencoder with 8× downsampling ($256 \times 256 \rightarrow 32 \times 32 \times 4$ latent space), with all components initialized from pre-trained Stable Diffusion weights.

Inference: Single-step prediction at fixed timestep $t = 50$ with Gaussian noise concatenated to image latents for direct mask estimation, eliminating iterative sampling required by traditional diffusion approaches.

Table 1. Polyp segmentation performance comparison on CVC-ClinicDB dataset.

Method	Dice (%)	IoU (%)	HD95 (mm)	NSD (%)
SSFormer [24]	94.4 ± 0.4	89.9 ± 0.6	12.3 ± 2.1	87.2 ± 1.8
Li-SegPNet [23]	92.5 ± 0.5	86.0 ± 0.7	15.6 ± 3.2	84.1 ± 2.3
Diff-Trans [9]	95.4 ± 0.3	92.0 ± 0.4	8.7 ± 1.5	89.8 ± 1.2
SDSeg [16]	95.8 ± 0.2	92.6 ± 0.3	7.9 ± 1.3	90.4 ± 1.1
SynDiff	**96.0 ± 0.3**	**92.9 ± 0.5**	**7.2 ± 1.1**	**91.1 ± 1.0**

4.3 Quantitative Results

We compare SynDiff against established polyp segmentation approaches representing different architectural paradigms. SSFormer [24] employs a pyramid transformer architecture for multi-scale feature extraction, while Li-SegPNet [23] utilizes lightweight separable convolutions for efficient segmentation. Among diffusion-based methods, Diff-Trans [9] combines diffusion models with transformer architectures for iterative refinement, whereas SDSeg [16] implements stable diffusion for medical image segmentation but requires multiple denoising steps.

Table 1 presents a comprehensive performance comparison on CVC-ClinicDB, demonstrating the effectiveness of our integrated approach. SynDiff achieves competitive performance with 96.0 ± 0.3% Dice coefficient and 92.9 ± 0.5% IoU, substantially outperforming traditional CNN-based architectures with a 1.6% Dice improvement over SSFormer (94.4% Dice) and a 3.5% improvement over Li-SegPNet (92.5% Dice). When compared to recent diffusion-based

approaches, our method achieves marginal but consistent improvements over SDSeg (95.8% Dice) and Diff-Trans (95.4% Dice) while offering significant computational advantages through single-step inference.

The boundary quality metrics reveal particularly encouraging results with HD95 of 7.2 ± 1.1mm and NSD of $91.1 \pm 1.0\%$, indicating superior edge preservation crucial for clinical applications. These improvements in boundary accuracy are clinically significant, as precise polyp delineation directly impacts diagnostic confidence and treatment planning decisions. The consistent low standard deviations across all metrics demonstrate robust performance across cross-validation folds, which is essential for reliable clinical deployment.

Table 2. Ablation study on key components of SynDiff framework.

Configuration	Dice (%)	IoU (%)
Baseline (Real data only)	93.7 ± 0.4	89.6 ± 0.6
+ Text-guided augmentation	96.0 ± 0.3	92.9 ± 0.5
+ Multi-step inference ($T = 50$)	96.1 ± 0.3	93.0 ± 0.4
+ Frozen vision encoder	95.2 ± 0.4	91.8 ± 0.6
+ Noise loss only ($\lambda = 0$)	95.4 ± 0.3	92.1 ± 0.5
+ Traditional augmentation	94.8 ± 0.4	90.7 ± 0.6

4.4 Ablation Study

Table 2 demonstrates a systematic analysis of component contributions. Text-guided augmentation provides the most significant improvement (+2.3% Dice over baseline), while the comparison between single-step and multi-step inference shows minimal performance difference (96.0% vs 96.1% Dice), validating our core hypothesis that single-step inference maintains quality while reducing computational requirements. The trainable vision encoder contributes meaningfully

Table 3. Impact of different augmentation strategies on segmentation performance.

Augmentation Type	Synthetic Samples	Dice (%)	IoU (%)
None	0	93.7 ± 0.4	89.6 ± 0.6
Traditional (rotation, flip)	-	94.8 ± 0.4	90.7 ± 0.6
GAN-based	100	95.2 ± 0.3	91.8 ± 0.5
Text-guided (20)	20	94.1 ± 0.4	90.2 ± 0.6
Text-guided (50)	50	95.1 ± 0.3	91.5 ± 0.5
Text-guided (100)	**100**	**96.0 ± 0.3**	**92.9 ± 0.5**
Text-guided (200)	200	95.6 ± 0.4	92.1 ± 0.6

(+0.8% over frozen), demonstrating the importance of domain-specific adaptation. The dual loss formulation also proves beneficial, as noise-only supervision reduces performance to 95.4% Dice.

4.5 Data Augmentation Analysis

Figure 2 shows synthetic polyp samples generated through text-guided inpainting, demonstrating morphological diversity across sizes, shapes, and textures. The generated samples exhibit clinically plausible characteristics with realistic color variations and anatomically consistent placement. Table 3 evaluates different quantities of synthetic data added to our full training set of 488 real scans. The optimal performance is achieved with 100 synthetic samples (approximately 20% augmentation), where insufficient augmentation (20–50 samples) limits model exposure to variability, while excessive synthetic data (200 samples) introduces slight performance degradation due to distribution shift. Notably, even minimal text-guided augmentation (20 samples) shows comparable performance to traditional methods, highlighting the quality of our synthetic generation. Our text-guided approach significantly outperforms traditional geometric

Fig. 2. Synthetically generated endoscopic images with corresponding binary segmentation masks. The first and third rows display synthetic colonoscopy images created through our text-guided generation approach, while the second and fourth rows present their corresponding binary masks. White regions in the masks indicate the synthetic anomalous tissue generated through the inpainting process. This augmented dataset enhances the diversity of training samples for our single-step diffusion segmentation model.

augmentation (+1.2% Dice) and GAN-based synthesis (+0.8% Dice), validating the effectiveness of semantically-controlled generation. The consistent improvement across both Dice and IoU metrics indicates that text-guided synthesis enhances both overall overlap and boundary delineation accuracy.

4.6 Computational Efficiency Analysis

Our single-step approach achieves theoretical complexity reduction from $O(T)$ to $O(1)$ compared to multi-step diffusion methods. Table 4 demonstrates practical efficiency gains, with SynDiff completing inference in 0.08 s compared to 1.8-2.3 s for existing diffusion methods, representing a 22–28× speedup. This efficiency improvement stems from our direct latent estimation strategy that eliminates iterative denoising while maintaining competitive accuracy, making the method suitable for real-time clinical applications.

Table 4. Inference efficiency comparison across diffusion segmentation methods.

Method	Inference Steps	Time (s)
Diff-Trans	50	1.8
SDSeg	100	2.3
SynDiff	**1**	**0.08**

5 Conclusion

We introduced SynDiff, a framework that addresses data scarcity and computational efficiency in medical image segmentation through text-conditioned synthetic data augmentation and single-step diffusion inference. On CVC-ClinicDB, our method achieved a Dice coefficient of $96.0 \pm 0.3\%$ and HD95 of 7.2 ± 1.1 mm. The direct latent estimation approach reduced inference time from 1.8-2.3 s to 0.08 s while maintaining segmentation accuracy. Our analysis revealed that augmenting with 100 synthetic samples optimizes performance, with text-guided generation outperforming conventional augmentation methods. These findings demonstrate the potential of combining semantic text guidance with efficient diffusion inference for practical medical imaging applications.

6 Future Work and Limitations

While SynDiff demonstrates competitive performance, limitations include evaluation on a single dataset, reliance on carefully engineered text prompts for realistic synthesis, and computational requirements that remain higher than traditional CNN approaches. Future work should prioritize multi-dataset validation across diverse imaging modalities, integration of clinical expert feedback into synthetic data generation, and exploration of the framework's effectiveness on other medical imaging tasks beyond polyp segmentation to establish broader clinical applicability.

Acknowledgements. This study was carried out within the PNRR research activities of the consortium iNEST (Interconnected North-Est Innovation Ecosystem) funded by the European Union Next-GenerationEU (Piano Nazionale di Ripresa e Resilienza (PNRR) Missione 4 Componente 2, Investimento 1.5 D.D. 1058 23/06/2022, ECS_00000043).

References

1. Aqeel, M., Sharifi, S., Cristani, M., Setti, F.: Meta learning-driven iterative refinement for robust anomaly detection in industrial inspection. In: European Conference on Computer Vision, pp. 445–460. Springer (2024)
2. Aqeel, M., Sharifi, S., Cristani, M., Setti, F.: Self-supervised learning for robust surface defect detection. In: International Conference on Deep Learning Theory and Applications (2024)
3. Aqeel, M., Sharifi, S., Cristani, M., Setti, F.: Self-supervised iterative refinement for anomaly detection in industrial quality control. In: International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (2025)
4. Aqeel, M., Sharifi, S., Cristani, M., Setti, F.: Towards real unsupervised anomaly detection via confident meta-learning. In: Accepted to Proceedings of the IEEE/CVF International Conference on Computer Vision (2025)
5. Baranchuk, D., Rubachev, I., Voynov, A., Khrulkov, V., Babenko, A.: Label-efficient semantic segmentation with diffusion models. arXiv preprint arXiv:2112.03126 (2021)
6. 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. Comput. Med. Imaging graph. **43**, 99–111 (2015)
7. Chen, J., et al.: Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
8. Choi, E., Biswal, S., Malin, B., Duke, J., Stewart, W.F., Sun, J.: Generating multi-label discrete patient records using generative adversarial networks. In: Machine Learning for Healthcare Conference, pp. 286–305. PMLR (2017)
9. Chowdary, G.J., Yin, Z.: Diffusion transformer u-net for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 622–631. Springer (2023)
10. Fan, D.P., et al.: Pranet: Parallel reverse attention network for polyp segmentation. In: International Conference on Medical Image Computing and Computer-assisted Interventionm, pp. 263–273. Springer (2020)
11. Girella, F., Liu, Z., Fummi, F., Setti, F., Cristani, M., Capogrosso, L.: Leveraging latent diffusion models for training-free in-distribution data augmentation for surface defect detection. In: International Conference on Content-Based Multimedia Indexing (CBMI) (2024)
12. Hatamizadeh, A., et al.: Unetr: Transformers for 3d medical image segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 574–584 (2022)
13. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Adv. Neural. Inf. Process. Syst. **33**, 6840–6851 (2020)

14. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Methods **18**(2), 203–211 (2021)
15. Jha, D., et al.: Kvasir-seg: A segmented polyp dataset. In: International conference on multimedia modeling, pp. 451–462. Springer (2019)
16. Lin, T., Chen, Z., Yan, Z., Yu, W., Zheng, F.: Stable diffusion segmentation for biomedical images with single-step reverse process. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2024, pp. 656–666. Springer Nature Switzerland, Cham (2024)
17. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: International Conference on Machine Learning, pp. 8162–8171. PMLR (2021)
18. Orlando, J.I., et al.: Refuge challenge: a unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. Med. Image Anal. **59**, 101570 (2020)
19. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4195–4205 (2023)
20. Podell, D., et al.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)
21. 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)
22. 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)
23. Sharma, P., Gautam, A., Maji, P., Pachori, R.B., Balabantaray, B.K.: Li-segpnet: Encoder-decoder mode lightweight segmentation network for colorectal polyps analysis. IEEE Trans. Biomed. Eng. **70**(4), 1330–1339 (2022)
24. Shi, W., Xu, J., Gao, P.: Ssformer: A lightweight transformer for semantic segmentation. In: 2022 IEEE 24th International Workshop on Multimedia Signal Processing (MMSP), pp. 1–5. IEEE (2022)
25. Song, Y., Dhariwal, P., Chen, M., Sutskever, I.: Consistency models (2023)
26. 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)
27. Wu, J., et al.: Medsegdiff: Medical image segmentation with diffusion probabilistic model. In: Medical Imaging with Deep Learning (2024)
28. Wu, J., Ji, W., Fu, H., Xu, M., Jin, Y., Xu, Y.: Medsegdiff-v2: Diffusion-based medical image segmentation with transformer. In: Proceedings of the AAAI Conference on Artificial Intelligence (2024)
29. Xing, Z., Wan, L., Fu, H., Yang, G., Zhu, L.: Diff-unet: A diffusion embedded network for volumetric segmentation. arXiv preprint arXiv:2303.10326 (2023)
30. Zhang, Z., et al.: Diffboost: Enhancing medical image segmentation via text-guided diffusion model. IEEE Trans. Med. Imaging (2024)