# Dual-Stream Diffusion with Structure-Aware Fusion for Joint Synthesis of Coherent Medical Image-Mask Pairs

Yu Li[1], Jiaqing Liu[1], Rahul Kumar Jain[1], Chujie Zhang[1], and Yen-Wei Chen[1*]

Graduate School of Information Science and Engineering, Ritsumeikan University, Osaka, Japan
[*]Corresponding author: chen@is.ritsumei.ac.jp

**Abstract.** High-quality paired medical image-mask data are scarce due to privacy constraints and the cost of expert annotation. While joint generative modeling offers a scalable solution for paired image-mask augmentation, existing methods often couple structure and appearance, leading to degraded region-of-interest (ROI) geometry, boundary inconsistency, and synthetic-to-real distribution shift. We propose Dual-Stream Diffusion (DSD), a joint generative framework that models paired data with two decoupled denoising streams: a structure stream for mask prediction and an appearance stream for image synthesis, thereby mitigating optimization conflict. To further preserve image-mask coherence, DSD introduces two structure-aware modules: Time-Gated Fusion provides progressive structural guidance during denoising, while Structure-Aware Heads enforce boundary-aligned prediction with improved image fidelity. We evaluate DSD on three public datasets (Kvasir-SEG, ISIC 2016, and 3D-IRCADb-01) and compare against competitive generation baselines. Across all datasets, DSD consistently improves both generation quality and image-mask consistency, demonstrating robust paired synthesis for data augmentation in medical image segmentation.

**Keywords:** Medical Image Synthesis · Diffusion Models · Dual-Stream Diffusion.

## 1 Introduction

High-quality medical image-mask pairs are essential for training robust segmentation models and supporting downstream medical image analysis. However, acquiring such paired data at scale remains challenging due to factors such as the cost of expert labeling [1] and privacy regulations [2]. Meanwhile, advances in generative modeling, from earlier GAN-based [3] synthesis to more recent diffusion-based [4–6] methods, have made synthetic data generation increasingly practical for medical imaging. Some recent work has explored generative augmentation to synthesize images [7–10], masks [11, 12], and image-mask pairs [13–15] for downstream segmentation.

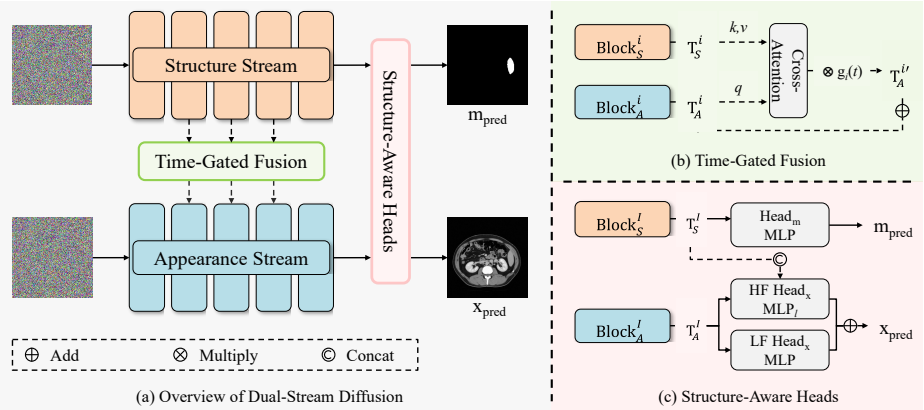



Fig. 1: **Overview of the proposed Dual-Stream Diffusion (DSD).** (a) DSD models the joint image-mask distribution with two decoupled streams for mask prediction and image synthesis. Each stream adopts a JiT-style transformer backbone [6]. To improve image-mask coherence, we design the following two modules: (b) **Time-Gated Fusion** injects structure tokens into the appearance stream via cross-attention with a time-dependent gate $g(t)$. (c) **Structure-Aware Heads** perform mask prediction and coarse-to-fine image prediction.

In practice, synthesizing clinically valid image-mask pairs remains challenging. First, a synthetic-to-real joint distribution shift can arise when generative priors fail to capture dataset-specific anatomy and acquisition characteristics. This can produce unrealistic image-mask co-occurrence patterns (e.g., object-count priors or boundary correspondence) [16, 17]. Second, in some methods, jointly modeling structure and appearance within a single-stream formulation can introduce optimization conflict, which may cause fragmented masks and boundary misalignment [15, 18]. Third, many existing approaches rely on strong conditional inputs (e.g., an image, a mask), which can reduce scalability and applicability [7–12, 15]. Therefore, addressing these challenges is crucial for translating synthetic pair quality into effective downstream performance.

To address these challenges, we propose Dual-Stream Diffusion (DSD), a joint diffusion framework for medical image-mask generation. DSD employs two JiT-style [6] Transformer [19] denoisers for direct image prediction rather than noise prediction, supporting structure-guided synthesis and coarse-to-fine refinement. Moreover, DSD is centered on two modules: Time-Gated Fusion, which provides progressive structural guidance during denoising, and Structure-Aware Heads, which improve boundary-aligned image-mask generation and fine-detail fidelity. As a result, DSD achieves superior generative quality and improved image-mask coherence. Our contributions are summarized as follows:

1. We propose **Dual-Stream Diffusion (DSD)**, a joint generative framework that models coherent medical image-mask pairs with decoupled structure

and appearance streams, reducing optimization conflict between geometric and appearance modeling.

2. We propose a **Time-Gated Fusion** module that injects structure features into the appearance stream via cross-attention and a time-dependent gate, enabling strong early structural control with progressively relaxed guidance during denoising.

3. We propose **Structure-Aware Heads** for joint mask and image prediction, including a mask head, a low-frequency image head, and a structure-conditioned high-frequency head for improved boundary fidelity and image-mask alignment.

4. Experiments on three public datasets show that DSD improves generative quality, image-mask consistency, and downstream segmentation performance over strong baselines.

## 2 Methodology

*Problem Setting.* Let $(x, m) \sim p_{\text{data}}$ denote a paired medical image and its ROI mask, where $x \in \mathbb{R}^{3\times H\times W}$ and $m \in \{0,1\}^{1\times H\times W}$. Our goal is to learn the joint distribution $p(x, m)$ and generate pixel-aligned image-mask pairs. We concatenate the clean mask and image as a 4-channel target

$$x_0 = [m,\, x] \in \mathbb{R}^{4\times H\times W}. \tag{1}$$

### 2.1 Preliminaries

*Continuous-Time Diffusion with Direct Target Prediction.* Let $x_0$ denote the paired mask–image target (concatenated along channels). Following a JiT-style [6] continuous-time formulation, we sample a continuous noise level $t \in (0, 1)$ and construct a corrupted state by interpolating between Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$ and the clean target:

$$z_t = t\, x_0 + (1 - t)\, \epsilon. \tag{2}$$

Thus, $t \approx 0$ corresponds to highly noisy states, while $t \approx 1$ corresponds to near-clean states. The denoiser directly predicts the clean target rather than noise:

$$\hat{x}_0 = f_\theta(z_t, t). \tag{3}$$

*Derived Velocity-Space Supervision.* Although the denoiser directly predicts $\hat{x}_0$, training follows the JiT-style derived velocity parameterization:

$$v_\theta(t) = \frac{\hat{x}_0 - z_t}{1 - t}, \qquad v^*(t) = \frac{x_0 - z_t}{1 - t}. \tag{4}$$

We optimize a channel-weighted MSE in this velocity space:

$$\mathcal{L} = \lambda_m \|v_{\theta,m}(t) - v^*_m(t)\|_2^2 + \lambda_x \|v_{\theta,x}(t) - v^*_x(t)\|_2^2. \tag{5}$$

This preserves direct target prediction while using a reparameterized training objective; implementation follows JiT with standard numerical stabilization.

### 2.2 Dual-Stream Diffusion

In paired medical image-mask synthesis, structural spatial features (e.g., ROI shapes) and image appearance (e.g., intensity patterns and textures) follow different learning dynamics. Modeling them within a single denoising stream can cause optimization conflict, potentially degrading mask integrity and image-mask alignment. To address this, we design a dual-stream diffusion that explicitly separates structure and appearance, and introduce (1) time-gated fusion and (2) structure-aware heads for structure-aware fusion, as shown in Fig. 1.

Each stream of DSD adopts the same denoiser as JiT (i.e., direct image prediction) [6]. Let $T_S^i, T_A^i \in \mathbb{R}^{B\times N\times D}$ denote the structure and appearance token sequences at transformer block $i$, $(i = 1, \ldots, I)$. Where $N = (H/P) \times (W/P)$ is the number of patches for patch size $P$, and $D$ is the hidden dimension. Each block updates both streams with standard token mixing (self-attention and MLP), followed by a structure-to-appearance cross-stream interaction.

### 2.3 Time-Gated Fusion

Decoupling the two streams reduces optimization conflict, but fully independent updates can weaken image-mask coherence, particularly around ROI boundaries. In medical images, strong structural guidance is crucial in early denoising stages to suppress spurious regions, while later stages require greater appearance flexibility to recover realistic details. To achieve this, we inject structure information into the appearance stream using a cross-attention (CA) [20] mechanism weighted by a time-gate.

*Structure-to-appearance cross-attention.* As shown in Fig. 1(b), for each transformer block $(i = 1, \ldots, I)$ in each stream, we inject structural information into the appearance stream after intra-stream updates:

$$T_A^{i'} \leftarrow T_A^i + g_i(t)\,\mathrm{CA}\big(Q = T_A^i, K = T_S^i, V = T_S^i\big)\,, \tag{6}$$

$T_A^{i'}$ replaces $T_A^i$ as the output token sequence of appearance stream block $i$.

*Dual gating.* The fusion strength is modulated by the product of a fixed time-dependent gate and a learnable per-block gate:

$$g_i(t) = (1-t)^p \cdot g_{\mathrm{ca}}^i, \qquad p = 1.5, \tag{7}$$

where $(1-t)^p$ enforces stronger structural coupling at earlier, noisier steps ($t \approx 0$) and gradually weakens it toward later, cleaner steps, while $g_{\mathrm{ca}}^i$ adapts the fusion magnitude across network depth.

### 2.4 Structure-Aware Heads

Medical image synthesis requires both globally coherent appearance (e.g., overall intensity distribution and tissue context) and locally accurate high-frequency details (e.g., lesion boundaries and fine textures). A single image head may conflate

these objectives, potentially weakening boundary coherence with the generated mask. To better balance the structure–appearance trade-off, we employ a decomposed image prediction design alongside a dedicated mask head.

From the final tokens, DSD predicts separate mask and image channels (Fig. 1(c)). The structure stream outputs a single-channel mask prediction (real-valued scores), while the appearance stream generates the image using a low-/high-frequency (LF/HF)decomposition:

$$\hat{m} = h_{\text{mask}}(T_S^I, t), \tag{8}$$

$$\hat{x}_{\text{LF}} = h_{\text{low}}(T_A^I, t), \quad \hat{x}_{\text{HF}} = h_{\text{high}}(T_A^I, T_S^I, t), \quad \hat{x} = \hat{x}_{\text{LF}} + \hat{x}_{\text{HF}}. \tag{9}$$

The $h$ represents MLP, and $h_{\text{low}}$ is denoted as $\text{MLP}_l$ in Fig. 1(c), where $l$ denotes the low-frequency branch. The joint clean prediction is then

$$\hat{x}_0 = [\hat{m},\, \hat{x}] \in \mathbb{R}^{4\times H\times W}. \tag{10}$$

### 2.5 Training Objective

DSD jointly generates a binary-like mask and continuous-valued image. Using separate objectives can introduce additional tuning complexity and may bias optimization toward one modality. To keep supervision simple and consistent across modalities, we train both streams with the same channel-weighted MSE in the derived velocity space (Sec. 2.1, Eqs. (4)–(5)). Although DSD predicts $\hat{x}_0$ (Eq. (3)), supervision is applied to the corresponding velocity-space quantities for both mask and image channels. Unless otherwise stated, we use $\lambda_m = \lambda_x = 1.0$.

### 2.6 Sampling

Inference uses the same dual-stream denoiser and time-gated cross-stream fusion as in training throughout the denoising trajectory. Starting from Gaussian noise $z_{t=0} \sim \mathcal{N}(0, I)$, we integrate from $t = 0$ to $t = 1$ with the Heun solver [21]. The final output is $\hat{x}_0 = [\hat{m}, \hat{x}]$. For binary mask evaluation, we obtain

$$\tilde{m} = \mathbf{1}[\sigma(\hat{m}) > 0.5]\,. \tag{11}$$

## 3 Experiments

### 3.1 Setup

*Datasets.* We conduct experiments on three public datasets: Kvasir-SEG [22] (endoscopic polyps), ISIC 2016 [23] (dermoscopic skin lesions), and 3D-IRCADb-01 [24] (liver CT). For 3D-IRCADb-01, we extract 2D axial slices from CT volumes. All inputs are represented as three-channel tensors; CT slices are replicated from one channel during preprocessing. No new patient data were collected; ethics approval was not required for these public datasets. The train/val/test splits are 700/100/200 (Kvasir-SEG), 700/200/379 (ISIC 2016), and 1451/207/416 (3D-IRCADb-01); the CT split is performed at the patient level. All images are resized to $256 \times 256$. Masks are binarized and stored as single-channel annotations.

Table 1: Generative fidelity and image-mask pair consistency comparison. KID (↓) measures image generation fidelity, while Dice/IoU (↑) quantify consistency using a frozen ResNet34 [28] trained on real data. "Real Data" refers to evaluation on real test samples, whereas all other methods are evaluated on generated samples. The best result excluding "Real Data" is highlighted in bold. "–" indicates not applicable.

| Method | Kvasir-SEG [22] | | | ISIC 2016 [23] | | | 3D-IRCADb-01 [24] | | |
|---|---|---|---|---|---|---|---|---|---|
| | Dice↑ | IoU↑ | KID↓ | Dice↑ | IoU↑ | KID↓ | Dice↑ | IoU↑ | KID↓ |
| Real Data | 0.7020 | 0.5999 | – | 0.9012 | 0.8359 | – | 0.9676 | 0.9454 | – |
| SatSynth [30] | 0.4114 | 0.3060 | 0.1149 | 0.4530 | 0.3404 | 0.7857 | 0.5938 | 0.5010 | 0.3933 |
| DiffGen [14] | 0.1992 | 0.1372 | 0.1604 | 0.4057 | 0.3265 | 0.2891 | 0.1860 | 0.1272 | 0.2146 |
| DiffAtlas [12] | 0.5903 | 0.5224 | 0.0830 | 0.6935 | 0.6126 | 0.0449 | 0.7300 | 0.6667 | **0.0370** |
| MSF [15] | 0.5968 | 0.4579 | 0.0788 | 0.7559 | 0.6304 | 0.0832 | 0.8365 | 0.7957 | 0.0487 |
| JiT [6] | 0.6598 | 0.5607 | 0.1373 | 0.8934 | 0.8317 | 0.1117 | 0.8621 | 0.8140 | 0.0733 |
| Ours | **0.7200** | **0.6309** | **0.0358** | **0.9150** | **0.8556** | **0.0407** | **0.9414** | **0.8750** | 0.0466 |

*Settings.* Our DSD uses a dual-stream ViT-M/16 denoiser [19] (depth 10, embedding dimension 640, and 10 attention heads). We train the model with AdamW. Timesteps are sampled using a logit-normal schedule, $t = \sigma(\mathcal{N}(-0.8, 0.8^2))$. For sampling, we use Heun's method with 50 steps. We jointly train DSD on the three datasets and use classifier-free guidance (CFG) [25], with 0.1 label dropout during training and a guidance scale of 1.0 at inference.

*Evaluation.* We evaluate fidelity, image–mask consistency, and usability for segmentation. **Fidelity:** We measure fidelity using Kernel Inception Distance (KID) between generated and real image samples. **Pair consistency:** Following [26, 27], for each dataset we train a ResNet34 [28] segmentation evaluator $\Phi_{\text{seg}}$ *only* on the *real* training split. We then fix $\Phi_{\text{seg}}$ and predict $m' = \Phi_{\text{seg}}(\hat{x})$ from a generated image $\hat{x}$, and compute Dice and IoU between $m'$ and the generated mask $\hat{m}$. **Replacement experiment:** Following [29], we train the same ResNet34 from scratch using only synthetic image–mask pairs and evaluate it on the real test set to probe synthetic-to-real distribution shift.

### 3.2 Results

We compare our method with state-of-the-art baselines, including SatSynth [30], DiffGen [14], DiffAtlas [12], MedSegFactory (MSF) [15], and JiT [6]. All methods are evaluated under the same pre-processing.

*Quantitative results.* Table 1 shows a consistent trend: DSD improves both image-mask alignment (Dice/IoU) and image fidelity (KID), without an evident fidelity-consistency trade-off in our settings.

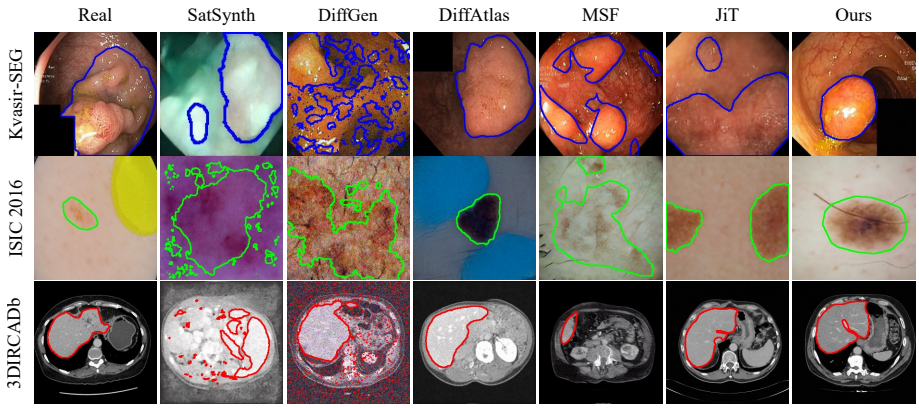
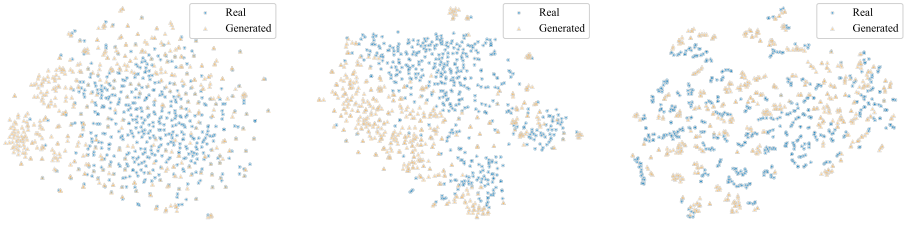



Fig. 2: Qualitative comparison of generated results on Kvasir-SEG, ISIC 2016, and 3D-IRCADb-01 (labeled as 3DIRCADb).




(a) t-SNE on Kvasir-SEG (b) t-SNE on ISIC 2016 (c) t-SNE on 3DIRCADb

Fig. 3: t-SNE visualization comparing real and generated images by our method.

*Qualitative results.* Fig. 2 shows example image-mask pairs. Compared with prior baselines, DSD produces closer boundary alignment between $\hat{x}$ and $\hat{m}$, reducing mask leakage into the background and shape distortion for endoscopic polyps and dermoscopic lesions.

*t-SNE visualization.* As shown in Fig. 3, our generated samples overlap with real data in the learned t-SNE feature space. Minor shifts remain, but the overall embedding structure is well preserved.

*Full replacement evaluation.* Table 2 directly measures the usability of generated labels, with training on real data serving as an upper bound. DSD performs best among synthetic-data methods on ISIC 2016 and 3D-IRCADb-01, and remains competitive on Kvasir-SEG. This trend suggests that the synthetic pairs generated by DSD transfer effectively to downstream segmentation.

Table 2: Full replacement evaluation using ResNet34 [28] across datasets. "Real Data" denotes training and testing on real data, while all other methods are trained on synthetic data and evaluated on the real test set. The best result excluding "Real Data" is highlighted in bold.

| | Kvasir-SEG [22] | | ISIC 2016 [23] | | 3D-IRCADb-01 [24] | |
|---|---|---|---|---|---|---|
| Method | Dice↑ | IoU↑ | Dice↑ | IoU↑ | Dice↑ | IoU↑ |
| Real Data | 0.7183 | 0.5514 | 0.9043 | 0.8268 | 0.9834 | 0.9673 |
| SatSynth [30] | 0.5820 | 0.3370 | 0.7859 | 0.5499 | 0.9315 | 0.8716 |
| DiffGen [14] | 0.6583 | 0.4806 | 0.8293 | 0.6216 | 0.9339 | 0.7084 |
| DiffAtlas [12] | **0.6950** | **0.5132** | 0.8851 | 0.7764 | 0.9401 | 0.8813 |
| MSF [15] | 0.4654 | 0.3782 | 0.8081 | 0.6820 | 0.9304 | 0.8671 |
| JiT [6] | 0.5976 | 0.4158 | 0.8468 | 0.7345 | 0.9654 | 0.9330 |
| Ours | 0.6441 | 0.4623 | **0.8867** | **0.7879** | **0.9757** | **0.9517** |

Table 3: Ablation study of key components. Components include Dual (dual-stream), Gate (time-gate), CA (cross-attention), SH (structure-aware heads), and HF (only high-frequency head).

| Components | | | | | KID↓ | | |
|---|---|---|---|---|---|---|---|
| Dual | Gate | CA | SH | HF | Kvasir-SEG [22] | ISIC 2016 [23] | 3D-IRCADb-01 [24] |
| ✓ | ✗ | ✗ | ✗ | – | 0.2808 | 0.3008 | 0.1059 |
| ✓ | ✓ | ✓ | ✗ | – | 0.2313 | 0.2962 | 0.2141 |
| ✓ | ✗ | ✓ | ✓ | ✓ | 0.0868 | 0.1028 | 0.0686 |
| ✓ | ✓ | ✗ | ✓ | ✓ | 0.0889 | 0.1089 | 0.0681 |
| ✓ | ✓ | ✓ | ✓ | ✗ | 0.0924 | 0.1076 | 0.0659 |
| ✓ | ✓ | ✓ | ✓ | ✓ | **0.0358** | **0.0407** | **0.0466** |

*Ablation and Analysis.* Table 3 reports an ablation of the key design components. Starting from the basic dual-stream configuration, progressively adding cross-stream fusion and structure-aware heads consistently improves performance across the evaluated metrics. The full configuration achieves the best overall results on all three datasets.

## 4 Conclusion

We proposed Dual-Stream Diffusion (DSD), a pixel-space joint diffusion framework for generating aligned medical image-mask pairs. By decoupling structure and appearance into two denoising streams and introducing time-gated fusion and structure-aware heads modules, DSD improves image-mask coherence and the practical utility of synthetic pairs for segmentation. Experiments demonstrate generative quality, pair-consistency, and full-replacement evaluations show consistent gains over strong baselines.

***Acknowledgements.*** This work was supported in part by the Grant in Aid for Scientific Research from the Japanese Ministry of Education, Culture, Sports, Science and Technology (MEXT) under the Grant No. 20KK0234 and in part by JST CREST, Japan, under Grant No. JPMJCR25T4.

***Disclosure of Interests.*** The manuscript is approved for publication by all authors. All authors declare no conflicts of interest regarding the submission and publication.

## References

1. Zhang, L., Tanno, R., Xu, M., et al.: Learning from multiple annotators for medical image segmentation. Pattern Recognition **138**, 109400 (2023)
2. Adnan, M., Kalra, S., Cresswell, J.C., et al.: Federated learning and differential privacy for medical image analysis. Scientific Reports **12**(1), 1953 (2022)
3. Goodfellow, I., Pouget-Abadie, J., Mirza, M., et al.: Generative adversarial networks. Communications of the ACM **63**(11), 139–144 (2020)
4. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Advances in Neural Information Processing Systems (NeurIPS), vol. 33, pp. 6840–6851 (2020)
5. Rombach, R., Blattmann, A., Lorenz, D., et al.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10684–10695 (2022)
6. Li, T., He, K.: Back to basics: Let denoising generative models denoise. arXiv preprint arXiv:2511.13720 (2025)
7. Sharma, V., Kumar, A., Jha, D., et al.: ControlPolypNet: Towards controlled colon polyp synthesis for improved polyp segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2325–2334 (2024)
8. 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. IEEE (2024)
9. Li, Y., Liu, J., Jain, R.K., et al.: Multimodal Brownian bridge diffusion model for controllable synthetic medical image generation. Biomedical Signal Processing and Control **117**, 109584 (2026)
10. Qiu, K., Gao, Z., Zhou, Z., et al.: Noise-consistent siamese-diffusion for medical image synthesis and segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 15672–15681 (2025)
11. Wu, J., Ji, W., Fu, H., et al.: MedSegDiff-V2: Diffusion-based medical image segmentation with transformer. In: Proceedings of the AAAI Conference on Artificial Intelligence, **38**(6), 6030–6038 (2024)
12. Zhang, H., Liu, Y., Yang, J., et al.: DiffAtlas: GenAI-fying atlas segmentation via image-mask diffusion. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 161–172. Springer, Cham (2025)
13. Li, W., Zhang, J., Heng, P.-A., et al.: Comprehensive generative replay for task-incremental segmentation with concurrent appearance and semantic forgetting. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 80–90. Springer, Cham (2024)

14. Saragih, D.G., Hibi, A., Tyrrell, P.N.: Using diffusion models to generate synthetic labeled data for medical image segmentation. International Journal of Computer Assisted Radiology and Surgery **19**(8), 1615–1625 (2024)
15. Mao, J., Wang, Y., Tang, Y., et al.: MedSegFactory: Text-guided generation of medical image-mask pairs. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 21525–21535 (2025)
16. Usman Akbar, M., Larsson, M., Blystad, I., et al.: Brain tumor segmentation using synthetic MR images: A comparison of GANs and diffusion models. Scientific Data **11**(1), 259 (2024)
17. Prochazka, A., Zeman, J.: Domain adaptation of stable diffusion for ultrasound inpainting: A synthetic data approach for enhanced thyroid nodule segmentation. Journal of Biomedical Informatics, 104963 (2025)
18. Pascual-González, M., Jiménez-Partinen, A., Luque-Baena, R.M., et al.: SLIM-Diff: Shared latent image-mask diffusion with Lp loss for data-scarce epilepsy FLAIR MRI. arXiv preprint arXiv:2602.03372 (2026)
19. Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (ICLR) (2021)
20. Vaswani, A., Shazeer, N., Parmar, N., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems (NeurIPS), vol. 30 (2017)
21. Karras, T., Aittala, M., Aila, T., et al.: Elucidating the design space of diffusion-based generative models. In: Advances in Neural Information Processing Systems (NeurIPS), vol. 35, pp. 26565–26577 (2022)
22. Jha, D., Smedsrud, P.H., Riegler, M.A., et al.: Kvasir-SEG: A segmented polyp dataset. In: MultiMedia Modeling (MMM 2020), LNCS 11962, pp. 451–462. Springer, Cham (2020)
23. Gutman, D., Codella, N.C.F., Celebi, E., Helba, B., Marchetti, M., Mishra, N., Halpern, A.: Skin lesion analysis toward melanoma detection: A challenge at the International Symposium on Biomedical Imaging (ISBI) 2016, hosted by the International Skin Imaging Collaboration (ISIC). arXiv:1605.01397 (2016)
24. Soler, L., Hostettler, A., Agnus, V., et al.: 3D-IRCADb-01 liver CT dataset. Available at: https://www.ircad.fr/research/data-sets/liver-segmentation-3d-ircadb-01/
25. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
26. Wang, T.-C., Liu, M.-Y., Zhu, J.-Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional GANs. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8798–8807 (2018)
27. 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 (CVPR), pp. 2337–2346 (2019)
28. He, K., Zhang, X., Ren, S., et al.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016)
29. Sushko, V., Zhang, D., Gall, J., et al.: One-shot synthesis of images and segmentation masks. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 6285–6294 (2023)
30. Toker, A., Eisenberger, M., Cremers, D., et al.: SatSynth: Augmenting image-mask pairs through diffusion models for aerial semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 27695–27705 (2024)