Computers and Artificial Intelligence；Vol.2, No.3; 2025
E-ISSN: 3006 -3728 P-ISSN：3006-371X
DOI: https://doi.org/10.70267/cai.25v2n3.5562
Published by: Zeus Press

Mamba Convolutional Hybrid Spatio-Temporal Medical Image Generation Based on Diffusion Probabilistic Models

Mei Zhang* and Zhengjie Liang

Guizhou University of Finance and Economics, Guiyang, Guizhou, China, 550025

**Corresponding author: Mei Zhang.*

Abstract

We introduce a novel hybrid deep learning module, termed the Mamba-Spatial-Temporal Generator (MSTG), which integrates the strengths of Convolutional Neural Networks (CNNs) with the advanced Mamba architecture. While conventional CNNs are effective in extracting local features within diffusion models, their limited receptive field restricts their capacity to capture long-range dependencies. To overcome this limitation, MSTG first employs CNN-based convolutional and pooling layers to extract multi-level local features, and subsequently incorporates Mamba blocks founded on State Space Models (SSMs). Owing to its linear computational complexity and powerful long-sequence modeling capability, Mamba adaptively selects and fuses global contextual information. Through this synergistic design, MSTG retains the local perceptual advantages of CNNs while simultaneously leveraging the global dynamic modeling capacity of Mamba. As a result, it significantly improves the understanding of complex spatial and sequential dependencies without compromising computational efficiency. This module has a clear structure and good scalability, providing a new and effective way to improve the performance of cardiac medical image generation tasks for 4D data.

Keywords

medical image generation, Mamba, 4-dimensional data, long-range dependencies, diffusion model

1. Introduction

Denoising Diffusion Probabilistic Models (DDPMs), as an emerging class of generative models, have demonstrated remarkable potential in medical image synthesis. In recent years, with the rapid advancement of deep learning, image generation has become a central research focus in computer vision. Traditional approaches, such as Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs), have long served as mainstream frameworks (Müller-Franzes et al., 2022). However, in the past few years (Huang, 2024; Liu et al., 2023), the superior ability of DDPMs to generate high-quality synthetic images has increasingly surpassed that of GANs in natural image generation. Their importance is particularly evident in medical imaging, where issues of confidentiality and privacy significantly constrain the acquisition, annotation, and sharing of medical data (Khadra & Türkbey, 2024; Khazrak et al., 2024; Khosravi et al., 2023). Challenges such as privacy concerns and the scarcity of disease-specific data often result in datasets that are small and imbalanced, thereby hindering the development of accurate medical image classification models (Khazrak et al., 2024). Moreover, producing high-quality medical image annotations requires not only precision but also sufficiently diverse datasets to cover the full spectrum of anatomical structures, pathological features, and

imaging modalities (Krishna et al., 2024). By generating synthetic images with corresponding annotations, DDPMs provide a controllable generative framework that facilitates this objective and enables broader applications of deep learning in medical imaging.

2. Theoretical Basis and Development Status of Generative Models

The core principle of Denoising Diffusion Probabilistic Models (DDPMs) originates from non-equilibrium thermodynamics. By defining a forward Markov chain that gradually transforms data into Gaussian noise, and learning a corresponding reverse process to reconstruct the original data from noise, DDPMs achieve their generative capability, as illustrated in Figures 1 and 2.

Figure 1: Forward Diffusion

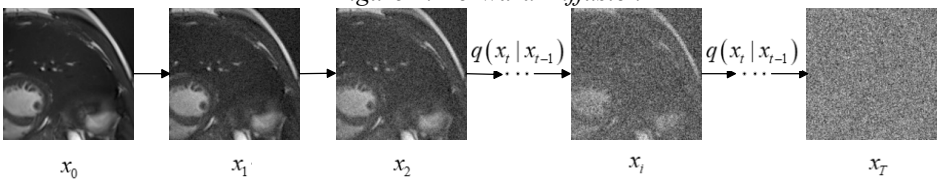
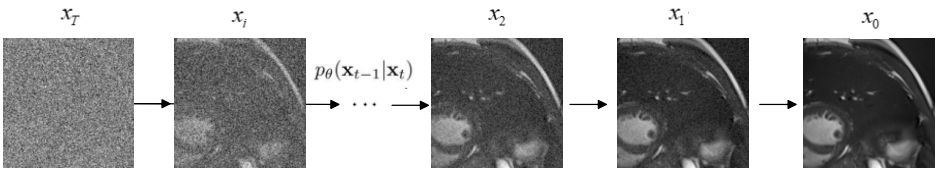


Figure 2: Reverse Diffusion



The forward diffusion process is computed as follows:

$$q\left(x_{1\urcorner T} \mid x_0\right) \coloneqq \prod_{t=1}^{T} q\left(x_t \mid x_{t-1}\right) \tag{1}$$

$$q\left(x_t \mid x_{t-1}\right) \coloneqq N\left(x_t \sqrt{1-\beta_t} x_{t-1}, \beta_t I\right) \tag{2}$$

The reverse diffusion process is computed as follows:

$$p_\theta\left(x_{T:0}\right) \coloneqq p\left(x_T\right) \prod_{t=1}^{T} N\left(x_{t-1}; \mu_\theta\left(x_t, t\right), \sum_\theta\left(x_t, t\right)\right) \tag{3}$$

Ho et al. first proposed Denoising Diffusion Probabilistic Models (DDPMs) (Ho et al., 2020) and demonstrated their ability to generate high-quality images. They achieved remarkable generative performance by training a weighted variational lower bound designed based on a novel connection between denoising score matching and Langevin dynamics. Subsequently, DDPMs and their variants have achieved considerable success in the field of computer vision (Jiang et al., 2025). The development of diffusion models has proceeded through multiple stages, aiming to optimize generation quality, training efficiency, and inference speed. For instance, improved DDPMs learn the reverse process via simple reparameterization and hybrid learning objectives (Liu et al., 2023), while Variational Diffusion Models (VDMs) introduce learnable diffusion variances, Fourier features, and architectural innovations to capture finer details. Denoising Diffusion Implicit Models (DDIMs) (Liu et al., 2023) reduce the number of autoregressive steps to generate higher-quality samples, significantly improving sampling efficiency. Analytic-DPM (Bao et al., 2022) provides a training-free inference framework that achieves substantial speedup while maintaining high-quality samples. Diffusion models continue to evolve in terms of maximum likelihood optimization, data generalization, and slice-based sampling, laying a solid foundation for their application in specialized domains such as medical image synthesis.

DDPMs can generate high-quality synthetic medical images, effectively augment training datasets and enhance model performance in tasks such as classification and segmentation (Khazrak et al., 2024). In studies on pulmonary nodule segmentation, a memory-efficient block-wise DDPM (Khadra & Türkbey, 2024) was

proposed to generate CT scans containing pulmonary nodules, addressing memory constraints while improving the practicality of synthetic images. DDPMs can also generate medical images conditioned on specific attributes, such as pathological features, anatomical masks, or imaging modalities. For example, the seg2med framework (Yang et al., 2025) uses DDPMs to synthesize CT and MR images conditioned on anatomical masks, achieving high Structural Similarity Index (SSIM). Multi-conditional DDPMs (mDDPMs) (Krishna et al., 2024) provide a controllable generative framework for medical image synthesis, capable of generating annotated synthetic images to meet the demand for highly accurate, diverse, and sufficiently large annotated datasets in medical imaging applications.

Li et al. (2025) proposed augmenting the backbone U-Net of diffusion models with Kolmogorov–Arnold Networks (KANs), enhancing the nonlinear modeling capacity of the network. Compared with the conventional U-Net used in diffusion models, Diffusion U-KAN more effectively captures and represents complex nonlinear features in medical images, exhibiting superior generative and generalization capabilities. On the other hand, Joshi A. et al. (2022, 2023) proposed the R2Net framework, introducing Lipschitz continuity constraints and multi-scale extensions to achieve efficient and flexible registration of multiple medical images, improving computational efficiency while preserving deformation properties. Although R2Net demonstrates fast inference, it shows limited preservation of fine-grained details in medical images, with partial loss of subtle information.

To mitigate mode collapse, Conditional Generative Adversarial Networks (CGANs) (Mirza & Osindero, 2014) were introduced, which incorporate conditional variables into the latent space of the generator to constrain the sample generation process, thereby partially alleviating mode collapse. Kim and Ye (2022) proposed a novel Diffusion Deformation Model (DDM), which combines DDPMs with deformation registration models to successfully generate four-dimensional (4D) temporal medical images.

3. Mamba-Spatial-Temporal Generator Model

The Mamba module is an efficient deep learning architecture specifically designed for spatiotemporal feature extraction, with the aim of enhancing a model's representational capacity for complex, high-dimensional data. It employs a multi-branch design to simultaneously process spatial and temporal information. Within each branch, attention mechanisms are seamlessly integrated with convolutional operations, enabling joint modeling of both local and global features. The module further decomposes input feature maps into multiple channel subsets, each subjected to learnable linear transformations and nonlinear activations, followed by a gating mechanism that dynamically fuses information across channels. This design not only strengthens the capture of spatiotemporal dependencies but also effectively alleviates feature redundancy and information loss in multi-frame sequence generation tasks.

Denoising diffusion probabilistic models (DDPMs) have demonstrated remarkable performance in image generation, speech synthesis, and cross-modal generation tasks. Their core architecture is typically based on U-Net, leveraging convolutional neural networks (CNNs) for local feature extraction and multi-scale representation. However, CNNs are inherently limited in modeling long-range dependencies and capturing global contextual information. While Transformers can address these limitations, their quadratic computational complexity and training instability have constrained their broad adoption in diffusion-based generative models. To overcome these challenges, we propose a novel hybrid feature extraction module, the Mamba-Spatial-Temporal Generator (MSTG), as illustrated in Figure 3. MSTG synergistically combines the local perceptual strengths of CNNs with the global modeling capabilities of state-space models, further enhanced by linear projections and convolutional operations. This hybrid design substantially improves the representational power and generative fidelity of DDPMs when modeling complex data distributions.

Owing to its linear computational complexity, Mamba offers a computational efficiency advantage in long-sequence processing. When combined with the lightweight convolutional design of CNNs, it leverages the CNN-driven selective state-space mechanism to enhance the understanding of global contextual information. The discrete formulation of the Mamba State Space Model is computed as follows:

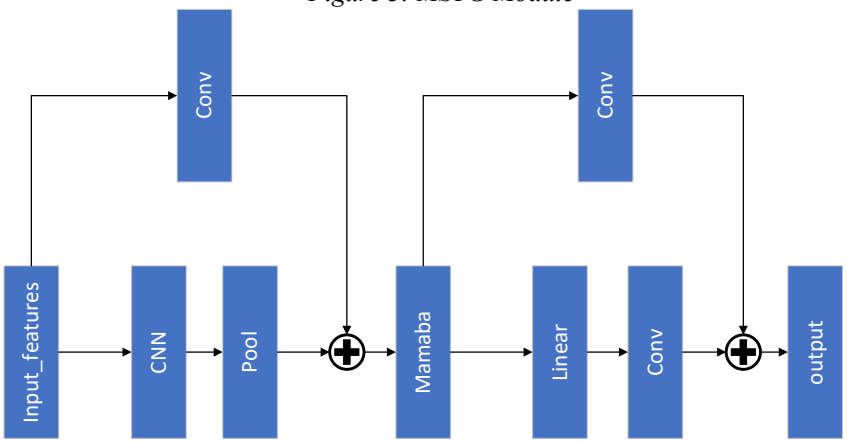
$$h_{t+1} = Ah_t + Bx_t \tag{4}$$

$$y_t = Ch_t + Dx_t \tag{5}$$

By incorporating a selective mechanism (selective scan), Mamba combines convolutional kernels with dynamic parameterization to achieve linear computational complexity, computed as follows:

$$y_t = SS(\sum_{\tau}^{t} K(t-\tau)\cdot x_\tau) \tag{6}$$

Figure 3: MSTG Module



Building upon convolutional feature extraction, this study introduces the Mamba module—a selective state-space model—as a core innovation to further enhance the capacity of generative models in capturing spatiotemporal features of high-dimensional image data. The Mamba module systematically models sequential data via state-space equations, enabling global receptive field coverage while maintaining linear computational complexity, thereby effectively capturing long-range dependencies within images. Compared with conventional Transformer models, this design exhibits notable advantages in high-resolution image generation tasks, demonstrating improved training stability, memory efficiency, and computational scalability. Its linear-complexity property is particularly suitable for processing large-scale image sequences, circumventing the exponential computational and memory overhead associated with Transformers as sequence length increases, and providing a viable solution for applications such as high-resolution medical and natural image synthesis.

Within the Mamba-Spatial-Temporal Generator (MSTG), the Mamba module first serializes convolutional feature maps, converting two- or three-dimensional spatial representations into channel-wise temporal sequences. This formulation enables Mamba to model global dependencies between pixels, thereby compensating for the inherent limitations of convolutional neural networks (CNNs) in capturing long-range structural relationships. While convolutions excel at local feature extraction and texture representation, their receptive field is constrained by kernel size and network depth, limiting their capacity to capture semantic correlations across spatially distant regions. The state-space modeling strategy of Mamba integrates information across the entire feature sequence, achieving precise representation of global semantic structures and enhancing both the overall coherence and local detail consistency of generated images.

To further optimize feature representation, a linear layer is applied to the output of the Mamba module for dimensional adjustment and information compression. This linear transformation reduces feature redundancy and strengthens the model's ability to represent nonlinear relationships, ensuring that subsequent processing can more fully exploit global semantic information. The transformed features are then processed through convolutional layers to reconstruct and refine spatial structures. These convolutional operations operate over local receptive fields to ensure spatial consistency with the target distribution while preserving semantic integrity. Additionally, residual connections are introduced to fuse the global information from Mamba with the locally reconstructed details. This residual mechanism facilitates the construction of rich hierarchical features and mitigates potential detail loss during feature extraction, enabling a smooth transition from global semantics to local structures and improving the visual coherence and realism of the generated images.

To enhance deep representation and multi-scale generation, the entire module is designed as a cascaded architecture composed of multiple stacked sub-modules. Each stage consists of three key components: CNN-based local feature extraction, Mamba-driven global sequence modeling, and linear-convolution reconstruction with feature fusion. Through this multi-stage design, the model progressively accumulates and refines feature

representations, capturing low-level textures in shallow layers as well as high-level semantic information in deeper layers. Moreover, a cross-stage feature fusion mechanism, implemented via skip connections, integrates shallow detailed features with deep semantic features, further enhancing the transmission and utilization of multi-scale information.

4. Experiments

4.1 Datasets

In this study, we utilized the multi-frame cardiac MRI dataset provided by the Automated Cardiac Diagnosis Challenge (ACDC). The dataset primarily comprises scans capturing the heart from end-diastole (ED) to end-systole (ES) and is divided into five categories: normal (NOR), dilated cardiomyopathy (DCM), right ventricular abnormality (ARV), myocardial infarction with systolic heart failure (MINF), and hypertrophic cardiomyopathy (HCM), encompassing a total of 100 subjects. For the experiments conducted in this work, the original cardiac images were resampled to a voxel spacing of $1.5 \times 1.5 \times 3.15$ mm.

4.2 Experimental Details

In this experiment, the model input channel sizes were set to 8, 16, 32, and 32. The initial learning rate was set to 0.0001, and the training batch size was 1. Network weights were updated using stochastic gradient descent with the Adam optimizer.

4.3 Experimental Analysis

4.3.1 Quantitative Analysis

Table 1: Comparison of quantitative results of various models in image reconstruction tasks

Model	NMSE↓	PSNR↑	SSIM↑
CGAN	0.316	20.047	0.565
U-KAN	0.557	18.696	0.600
R2	0.038	23.201	0.678
Ours	0.078	29.23	0.8978

In our experiments, R2 (Joshi & Hong, 2022, 2023), U-KAN (Li et al., 2025), and CGAN were employed as comparative models. The evaluation metrics included Normalized Mean Squared Error (NMSE), Peak Signal-to-Noise Ratio (PSNR), and Structural Similarity Index (SSIM), assessing the generation quality from three perspectives: error level, signal-to-noise fidelity, and perceptual structural consistency. As shown in Table 1, our method significantly outperforms the comparative models across multiple metrics, particularly excelling in metrics sensitive to visual quality. Specifically, for NMSE, our method achieved an outstanding value of 0.078, markedly lower than those of CGAN and U-KAN, indicating minimal discrepancy between the generated outputs and the ground truth images. Regarding PSNR, our approach reached 29.23 dB, substantially higher than CGAN, U-KAN, and R2. This improvement in PSNR demonstrates the method's superior ability to suppress generation noise while preserving high-frequency image details, thereby maintaining overall image signal-to-noise fidelity and producing outputs visually closer to real images. In terms of SSIM, our method attained a high score of 0.8978, significantly surpassing all comparative models. As an important metric that evaluates image structural, luminance, and contrast similarity, the high SSIM score indicates that our approach effectively preserves structural integrity, edge clarity, and local texture details, thereby better reconstructing the perceptual content of the images and meeting the practical requirements for high-quality image synthesis.

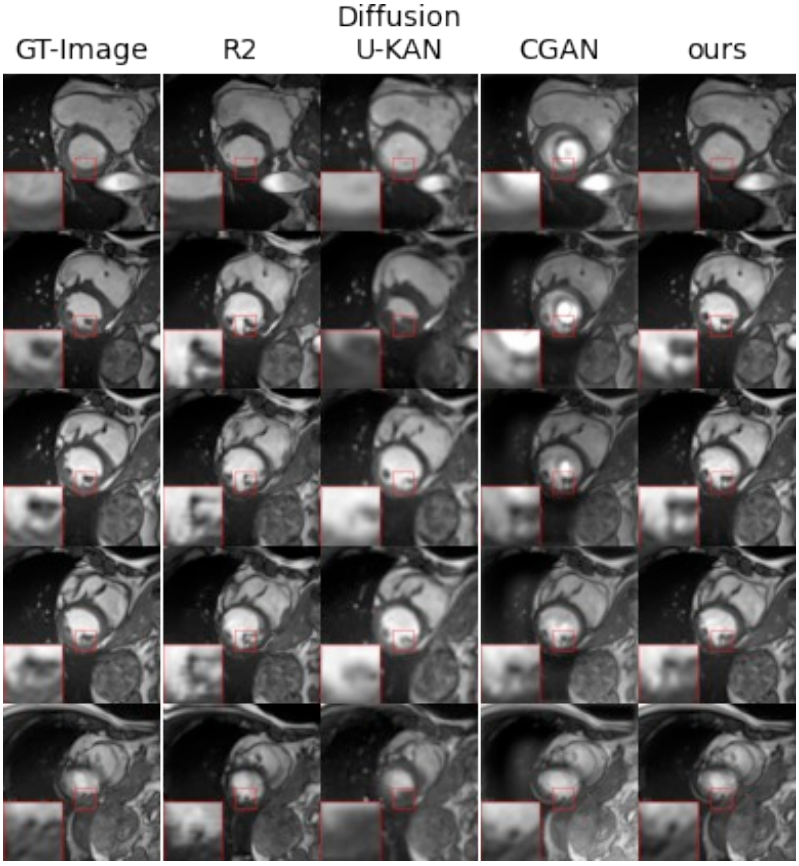
A comprehensive analysis reveals that although R2 shows slightly better performance in NMSE, our method demonstrates overall superiority in image quality, particularly in perceptual quality. This indicates that the proposed generative model not only achieves high reconstruction accuracy at the pixel level but also maintains higher-level semantic and structural information, effectively balancing the generation quality of both low-frequency content and high-frequency details.

4.3.2 Qualitative Analysis

As shown in Figure 4, in the task of continuous image generation, both the R2 and CGAN models exhibit varying degrees of deformation, compromising the structural integrity of the generated images. Additionally, these models display noticeable discontinuities in brightness distribution, resulting in overall lower image quality. This indicates that, although R2 and CGAN are capable of capturing certain local features, they are insufficient in maintaining global structural continuity and fine-grained detail consistency. The U-KAN model demonstrates relatively better performance in brightness continuity, preserving the gradient information of the original data with greater stability. However, the images it generates still deviate from the original data, particularly along edges, where blurring occurs, suggesting limitations in local structural refinement. While U-KAN partially alleviates the issue of brightness discontinuity, its generated images remain inadequate in terms of structural and textural fidelity for high-precision generation tasks.

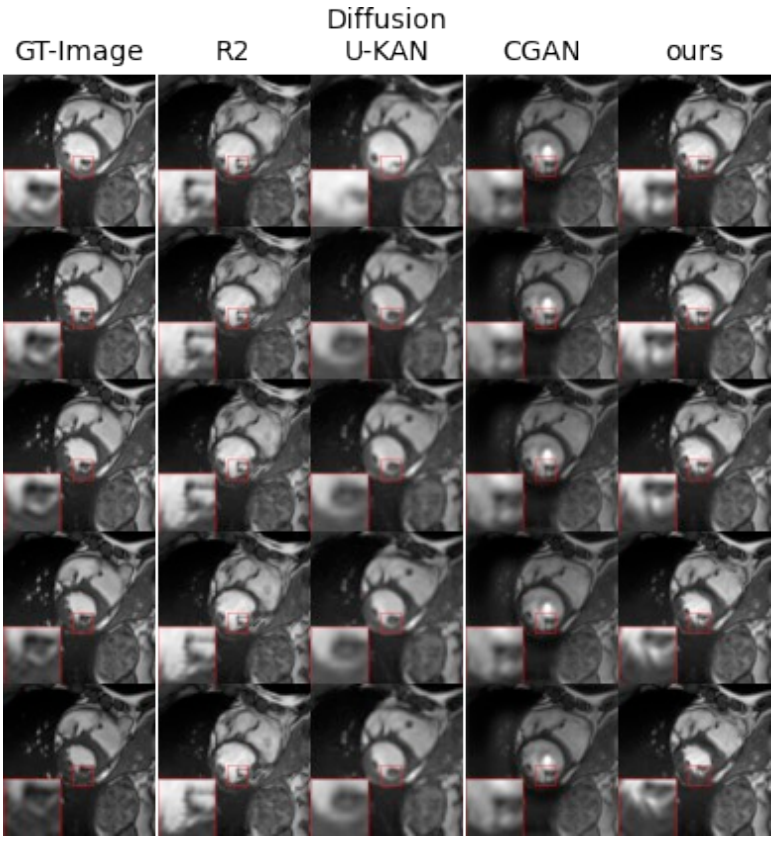
In contrast, the method proposed in this study significantly outperforms the aforementioned models in both vertical brightness distribution and edge detail fidelity. Our approach more accurately reproduces the brightness variations of the original data, ensuring continuity without abrupt changes. Moreover, in edge and detail reconstruction, the generated images closely match the originals, with well-defined structural contours and richly preserved textural details.

Figure 4: Different Slice Positions



As illustrated in Fig. 5, all models exhibit edge diffusion during the horizontal sequential generation process, with noticeable deviations in brightness compared to the original images, resulting in substantial differences between the generated samples and the ground truth. While U-KAN produces images with trends similar to the original data, the differences in fine details remain pronounced. In contrast, the images generated by our method demonstrate superior horizontal brightness consistency and edge fidelity relative to all other models, closely approximating the distribution characteristics of the original data.

Figure 5: Different Slice Positions Across Frames



5. Conclusion

In this work, we designed a hybrid architecture that integrates Convolutional Neural Networks (CNNs) with State Space Models, aiming to effectively combine local feature extraction with global sequential modeling capabilities. Experimental results demonstrate that the proposed generative module not only achieves high generation accuracy at the pixel level but also exhibits strong competitiveness in preserving higher-level semantic and structural information, effectively balancing the generation quality of low-frequency content and high-frequency details. These characteristics render it particularly favorable in visual evaluations, enabling the generation of new samples with superior visual fidelity that closely resemble real images. Future work will further investigate the model's generalization capability in complex scenarios and explore avenues for optimizing computational efficiency, with the goal of maximizing its potential in practical applications.

References

Bao, F., Li, C., Zhu, J., & Zhang, B. (2022). *ANALYTIC-DPM: An analytic estimate of the optimal reverse variance in diffusion probabilistic models* [Paper presentation]. ICLR 2022 - 10th International Conference on Learning Representations, Virtual.

Ho, J., Jain, A., & Abbeel, P. (2020). *Denoising diffusion probabilistic models* [Paper presentation]. 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

Huang, Y. (2024). Research advanced in image generation based on diffusion probability model. *Highlights in Science, Engineering and Technology, 85*, 452–456. https://doi.org/10.54097/WAYBGZ41

Jiang, H., Imran, M., Zhang, T., Zhou, Y., Liang, M., Gong, K., & Shao, W. (2025). Fast-DDPM: Fast denoising diffusion probabilistic models for medical image-to-image generation. *IEEE Journal of Biomedical and Health Informatics, 29*(10), 7326–7335. https://doi.org/10.1109/JBHI.2025.3565183

Joshi, A., & Hong, Y. (2022). Diffeomorphic image registration using lipschitz continuous residual networks. *Proceedings of Machine Learning Research, 172*, 1–13.

Joshi, A., & Hong, Y. (2023). R2Net: Efficient and flexible diffeomorphic image registration using lipschitz continuous residual networks. *Medical Image Analysis, 89*, Article 102917. https://doi.org/10.1016/J.MEDIA.2023.102917

Khadra, K., & Türkbey, U. (2024). Evaluating utility of memory efficient medical image generation: A study on lung nodule segmentation. *arXiv preprint*, arXiv:2410.12542. https://doi.org/10.48550/arXiv.2410.12542

Khazrak, I., Takhirova, S., Rezaee, M. M., Yadollahi, M., Green II, R. C., & Niu, S. (2024). Addressing small and imbalanced medical image datasets using generative models: A comparative study of DDPM and PGGANs with random and greedy K sampling. *arXiv preprint*, arXiv:2412.12532. https://doi.org/10.48550/arXiv.2412.12532

Khosravi, B., Rouzrokh, P., Mickley, J. P., Faghani, S., Mulford, K., Yang, L., Larson, A. N., Howe, B. M., Erickson, B. J., Taunton, M. J., & Wyles, C. C. (2023). Few-shot biomedical image segmentation using diffusion models: Beyond image generation. *Computer Methods and Programs in Biomedicine, 242*, Article 107832. https://doi.org/10.1016/J.CMPB.2023.107832

Kim, B., & Ye, J. C. (2022). *Diffusion deformable model for 4d temporal medical image generation* [Paper presentation]. Medical Image Computing and Computer Assisted Intervention – MICCAI 2022, Singapore.

Krishna, A., Wang, G., & Mueller, K. (2024). Multi-conditioned denoising diffusion probabilistic model (mDDPM) for medical image synthesis. *arXiv preprint*, arXiv:2409.04670. https://doi.org/10.48550/arXiv.2409.04670

Li, C., Liu, X., Li, W., Wang, C., Liu, H., Liu, Y., Chen, Z., & Yuan, Y. (2025). U-kan makes strong backbone for medical image segmentation and generation. *Proceedings of the AAAI Conference on Artificial Intelligence, 39*(5), 4652–4660. https://doi.org/10.1609/aaai.v39i5.32491

Liu, Y., Yang, Z., Yu, Z., Liu, Z., Liu, D., Lin, H., Li, M., Ma, S., Avdeev, M., & Shi, S. (2023). Generative artificial intelligence and its applications in materials science: Current situation and future perspectives. *Journal of Materiomics, 9*(4), 798–816. https://doi.org/10.1016/J.JMAT.2023.05.001

Mirza, M., & Osindero, S. (2014). Conditional generative adversarial nets. *arXiv preprint*, arXiv:1411.1784. https://doi.org/10.48550/arXiv.1411.1784

Müller-Franzes, G., Niehues, J. M., Khader, F., Arasteh, S. T., Haarburger, C., Kuhl, C., Wang, T., Han, T., Nebelung, S., Kather, J. N., & Truhn, D. (2022). Diffusion probabilistic models beat GANs on medical images. *arXiv preprint*, arXiv:2212.07501. https://doi.org/10.1038/s41598-023-39278-0

Yang, Z., Chen, Z., Sun, Y., Strittmatter, A., Raj, A., Allababidi, A., Rink, J. S., & Zöllner, F. G. (2025). seg2med: A bridge from artificial anatomy to multimodal medical images. *arXiv preprint*, arXiv:2504.09182. https://doi.org/10.48550/arXiv.2504.09182

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

This paper is an output of the science project.

Copyrights