This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

High-Fidelity Unified One-to-Many Medical Image Synthesis via Text-Conditioned Latent Diffusion

Youjian Zhang[1*], Jian Huang[2,3*], Jie Wang[1], Zezhou Li[1], Zhongya Wang[1], Guanqun Zhou[1], Zhicheng Zhang[1,3](✉), and Gang Yu[2,3](✉)

[1] JancsiLab, JancsiTech, Hongkong, 999077, China
[2] National Clinical Research Center for Child Health, National Children's Regional Medical Center, Children's Hospital, Zhejiang University School of Medicine, Hangzhou 310052, China
[3] Sino-Finland Joint AI Laboratory for Child Health of Zhejiang Province, Hangzhou 310052, China
zhangzhicheng13@mails.ucas.edu.cn; yugbme@zju.edu.cn

Abstract. Current deep learning approaches for medical image synthesis require training multiple specialized models for different modality conversions, leading to inefficient parameter utilization. In this work, we propose a unified text-conditioned latent diffusion framework that achieves one-to-many medical image synthesis through two key innovations: (1) With text-guided dynamic gating, a shared latent space construction using pre-trained modality-specific encoders is proposed, reducing model parameters compared to training several separate models. (2) An adaptive hybrid frequency processor combining wavelet decomposition and Fourier analysis is designed to preserve both local textures and global anatomical structures. Our comprehensive experimental evaluation in various datasets validates that this framework is capable of transforming a single medical imaging modality into multiple target modalities using only one model, surpassing existing methods based on Generative Adversarial Networks and diffusion models in terms of generation quality. The success of this work establishes a new paradigm for efficient multi-modal medical image synthesis through latent space unification and frequency-aware diffusion, significantly advancing the practicality of virtual medical image generation systems.

Keywords: Medical image Synthesis · One-to-many · Diffusion model · Model Pre-training

1 Introduction

Multimodal medical imaging seamlessly integrates various imaging techniques, such as magnetic resonance imaging (MRI) [8, 18], computed tomography (CT)

[*] These authors contributed equally to this work.

[14, 15] and positron emission tomography (PET) [3], dramatically improving the precision of disease diagnosis and biomarker identification [13]. However, its clinical applications are often hampered by high costs, limited equipment availability, and patient safety concerns [19].

With the advancements in deep learning, image-to-image translation has witnessed significant progress [21], which enables data-driven establishment of mapping relationships among different modalities, eliminating the reliance on specialized imaging devices. In the context of medical image synthesis, Generative Adversarial Networks (GANs) [11, 28] and Variational Autoencoders [17] are the two most prevalent deep learning-based models. However, these models face several limitations, including training challenges, mode collapse [1], and shortcomings in the visual quality of poor fidelity and lower resolution [9]. Furthermore, current deep learning approaches predominantly concentrate on one-to-one medical image synthesis [7], which typically necessitates independent setup and training for each modality, leading to increased training complexity and computational expenses.

To address the challenge of high-quality one-to-many unified image synthesis in multimodal medical imaging, this study proposes a generative framework using diffusion models [4]. Through shared latent space learning and text-conditioned encoding (*i.e.*, text prompts), it has only $\mathcal{O}(1)$ parameter growth, more efficient than training N one-to-one models with $\mathcal{O}(\mathrm{N})$ parameters. Facing the complexity of multimodal medical images and high computational needs of diffusion models, we pre-train modality-specific encoder-decoder networks on multimodal data. Using VQ-GAN [6], we compress the diffusion process into a lower-dimensional latent space. Then, a conditional diffusion model is built, taking the semantic info from a pre-trained BERT model as a guiding signal for generating multimodal medical images according to various text prompts. Besides, an adaptive frequency feature fusion module is introduced to enhance modality-specific representations while keeping anatomical consistency. This method is anticipated to boost precision medicine development via high-quality multimodal synthesis. The contribution of this paper can be summarized as: this study introduces a pioneering unified one-to-many framework for multimodal medical image synthesis, leveraging text-conditioned latent diffusion with adaptive hybrid frequency processing. Its superior performance is validated through testing on two distinct datasets.

2 Releated Work

In medical image synthesis, the key is to map images from a source modality to a target one, and deep learning methods with GANs have greatly advanced it, used in various image-modality conversions like MRI to CT. But GANs suffer from mode collapse, training issues, and have limitations in handling long-tail distributions and generating high-resolution images. Diffusion Denoising Probabilistic Models (DDPM) [12] are a powerful alternative. By simulating diffusion and reverse diffusion, they outperform GAN-based methods [5], and latent diffusion

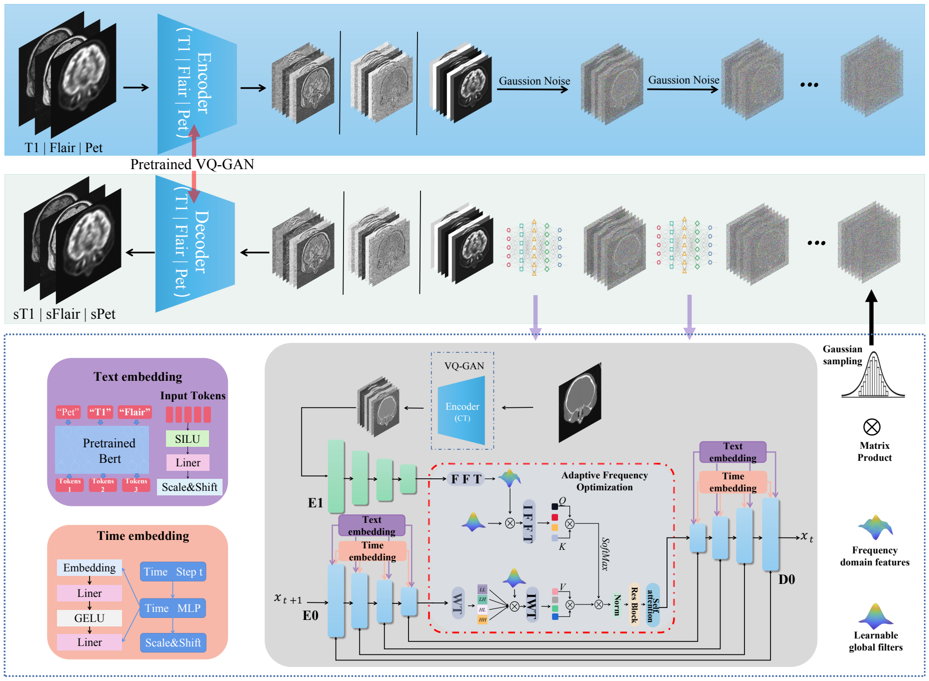

Fig. 1. The overall framework involves using three pre-trained VQ-GAN models to compress T1, FLAIR, and PET images into corresponding representations in the latent space, followed by diffusion and reverse diffusion processes in the latent space. And then, we can reconstruct the final image via the pre-trained decoder on the synthetic latent represetations.

models by Rombach *et al.* [23] are a significant improvement. For medical image synthesis, adding conditional information to diffusion models helps with accurate modality transformations. Graf *et al.*'s use of Denoising Diffusion Implicit Models (DDIM) [24] to convert MRI to CT [10] and Peng *et al.*'s conditional DDPM framework for transforming noise to CT distribution based on CBCT [22] both demonstrate the potential and effectiveness of conditional diffusion models in dealing with complex medical imaging tasks. Beyond generative architectures, frequency domain decomposition also enhances cross-modal synthesis. For instance, Wu W et al.[25] significantly improved medical image synthesis by integrating wavelet denoising into the Score-based Generative Model (SGM) framework. Cao J et al.[2] achieved accurate lumbar spine image synthesis by proposing multi-scale frequency channel attention and a dual-resolution frequency domain FFN module.

3 Methods

Fig. 1 shows the overview of our method. To facilitate effective image-to-image synthesis, this study initially involves the pre-training of modality-specific VQ-GAN with extensive datasets encompassing a range of modalities. This preparatory step is crucial to develop a competently trained encoder and decoder capable of handling various modal data. In the image-to-image synthesis process, the encoder from the pre-trained VQ-GAN is utilized to transform input images into latent space representations, which are subsequently fed into a DDPM to generate corresponding representations of target modalities. The process culminates in the reconstruction of these representations into final synthetic medical images.

As illustrated in Fig. 1, our proposed model comprises two encoders (**E0** for processing noisy input images, and **E1** for handling feature representations of conditional CT images) and a decoder **D0**. To effectively mitigate the inherent domain discrepancies between different modality-related input information, we incorporate learnable wavelet transform (WT) into **E0**. In contrast, following **E1**, we employ a learnable Fourier transform (FFT) which derived from the distinct capabilities of WT in capturing intricate local frequency features and the exceptional efficiency of FFT in conducting comprehensive global frequency analysis [26][27]. Specifically, we first utilize Fast FFT to extract the global spectral information of the input features from conditional brach and perform matrix multiplication with a learnable parameter matrix. Subsequently, the processed spectral features are reconstructed back to the spatial domain through inverse FFT, yielding feature $f_{fft} \in \mathbb{R}^{C\times H\times W}$. Concurrently, for the noisy image in **E0**, we employ learnable WT to decompose it into multi-scale features and similarly perform matrix multiplication with a learnable parameter matrix. Afterward, the multi-scale features are reconstructed back to the spatial domain through inverse WT, resulting in feature $f_{wt} \in \mathbb{R}^{C\times H\times W}$.

In the feature fusion stage, we use the conditional modal feature f_{fft} as the query matrix Q and key matrix K, and the noisy feature f_{wt} to generate the value matrix V, and compute the correlation of cross-modal features through an attention mechanism: $\text{Sim}(Q, K) = \frac{QK^T}{\sqrt{d}}$. The attention weights after $SoftMax$ normalization are used to dynamically modulate the value matrix V, thereby achieving effective fusion of structural information from other modalities and anatomical details from CT.

The text embedding section utilizes a pre-trained BERT model to encode modal labels (such as "PET", "T1") into 128-dimensional vectors, which are then processed by an MLP to generate scaling/shifting parameters for convolutional kernels. This design enables the model to adjust feature mappings based on text instructions, achieving controlled generation of multi-modal images. Simultaneously,to capture the noise characteristics at different time steps, our model, like most diffusion models, incorporates a time embedding.

DataSet: In the one-to-many generation task, the experiment uses two datasets: CERMEP-iDB-MRXFDG (CiM)[4] and BraTS 2019[5]. CiM contains multimodal brain imaging (FDG PET/CT and MRI sequences T1, T2 FLAIR) from 37

[4] https://doi.org/10.1186/s13550-021-00830-6

[5] https://www.med.upenn.edu/cbica/brats-2019/

healthy adults, collected on the same day using a Siemens Sonata 1.5T MRI and Biograph mCT64 PET/CT devices to ensure temporal consistency. BraTS 2019 provides a complete imaging set for 335 brain cases (non-enhanced T1, contrast-enhanced T1CE, T2, and FLAIR). In addition, for VQ-GAN pre-training, a modality-specific pretraining strategy is designed, aiming to enhance the feature representation capabilities of multimodal medical imaging. The training data for each modality are constructed using a multi-source fusion strategy: T1-weighted images are sourced from the T1 sequences of both the CiM and BraTS 2019 datasets; FLAIR images use the FLAIR sequences from both CiM and BraTS 2019; PET images are selected from the CiM PET images and the iFlytek Medical Image Analysis Challenge[6]; T1CE sequences use T1CE images from the BraTS 2019 dataset.

Implementation Details: For data pre-processing, all images from the CiM and BraTS 2019 datasets were resized to 256 × 256 and subjected to Z-score standardization using pre-calculated modality-specific means and standard deviations. For paired data (CiM), rigid registration was applied to align CT with MRI/PET. During VQ-GAN pre-training, modality-specific models used external data (except CT), while the CT model was trained solely on internal CiM data for encoding-only purposes. This study was conducted on an Ubuntu 18.04 system equipped with four NVIDIA RTX 3090 GPUs, utilizing Python 3.8 and PyTorch 1.10. For diffusion model training, the AdamW optimizer was employed, with diffusion temporal steps set to T=1000 and β values linearly increasing from 10^{-4} to 0.02. Inference efficiency was enhanced using DDIM with 100 sampling steps. A cosine annealing strategy was applied to the learning rate, and automatic mixed precision was adopted to optimize computational efficiency and reduce memory usage. The pre-trained VQ-GAN encoders/decoders and BERT text encoder remained frozen during training, while the DDPM and adaptive frequency modules (FFT/WT) were trained in an end-to-end manner. To prevent data leakage, subjects were partitioned at the patient level into training, validation, and test sets in an 8:1:1 ratio. The source code is available at `https://github.com/zyj15416/One-to-Many-Medical-Image-Synthesis`.

Evaluation Method: To evaluate the performance of the proposed method quantitatively, three commonly-used metrics were used: mean absolute error (MAE), structural similarity index (SSIM) and peak signal-to-noise ratio (PSNR). For metric calculation, synthesized images and ground truth were denormalized to original intensity ranges. We utilized three representative methods that have demonstrated excellent performance in the field of medical image generation as our comparative experiments: Pix2Pix [7], ALDM [16], SynDiff [20], CDDPM [22]. Among these, Pix2Pix is a GAN-based image generation model, while SynDiff, CDDPM, ALDM, and our proposed method are diffusion-based generative models. Notably, our model and ALDM operate as many-to-one image synthesis

[6] https://challenge.xfyun.cn/topic/info?type=pet-2023

frameworks, whereas Pix2Pix, SynDiff, and CDDPM function as one-to-one mapping architectures, utilizing individual models for single-modal translation tasks.

4 Experimental Results

Performance on CiM Dataset: Table 1 presents the quantitative evaluation of various models in the CiM tesing dataset, reflecting their performance in CT-to-T1, CT-to-FLAIR, and CT-to-PET synthesis. It is evident that our proposed method outperforms the other approaches across all three evaluation metrics. To be specific, for the CT→T1 synthesis task, our method outperforms all comparative methods with results of $\text{MAE} = 9.8 \pm 1.67$, $\text{SSIM} = 0.88 \pm 0.03$, and $\text{PSNR} = 39.08 \pm 3.92$. Compared with the second-best method ALDM, it reduces the MAE by 21.10%, increases the SSIM by 2.30%, and raises the PSNR by 5.04%. In the more challenging CT→FLAIR conversion, our method achieves the current best level with an MAE of 6.97 ± 1.99, which is 21.69% lower than that of the best comparative method ALDM (8.90 ± 2.41). It is particularly noteworthy that in the CT→PET synthesis task, our method, while maintaining $\text{SSIM} = 0.96 \pm 0.16$ (higher than other methods), improves the PSNR to 42.52 ± 6.83. Compared with the DDPM-based baselines CDDPM (38.62 ± 7.19), SynDiff (36.52 ± 7.24), and ALDM (38.55 ± 7.01), it increases by 10.10%, 12.90%, and 9.34% respectively. The comprehensive evaluation of the three tasks shows that the GAN-based method (Pix2Pix) performs worse than all DDPM-based methods. However, our framework obtains the lowest MAE, the highest SSIM and PSNR among all methods, and the standard deviation is generally smaller than that of the comparative methods, which proves its robustness.

Table 1. Experimental Results on The CiM Testing Dataset.

	CT-to-T1			CT-to-FLAIR			CT-to-PET		
	MAE	SSIM	PSNR	MAE	SSIM	PSNR	MAE	SSIM	PSNR
Pix2Pix	19.80 ±1.99	0.79 ±0.08	31.44 ±5.99	16.27 ±2.91	0.74 ±0.07	28.71 ±2.91	10.14 ±2.51	0.92 ±0.18	34.62 ±6.90
SynDiff	15.17 ±2.16	0.84 ±0.09	36.41 ±5.40	12.55 ±2.68	0.82 ±0.11	31.90 ±6.33	8.96 ±1.88	0.93 ±0.15	36.52 ±7.24
CDDPM	13.08 ±2.66	0.83 ±0.07	36.24 ±5.02	9.58 ±2.32	0.81 ±0.06	33.62 ±2.57	5.92 ±2.08	0.94 ±0.11	38.62 ±7.19
ALDM	12.42 ±1.88	0.86 ±0.07	37.11 ±4.83	8.90 ±2.41	0.84 ±0.08	34.02 ±2.70	6.04 ±1.31	0.94 ±0.18	38.55 ±7.01
Ours	9.80 ±1.67	0.88 ±0.03	39.08 ±3.92	6.97 ±1.99	0.86 ±0.04	37.94 ±2.49	5.40 ±1.95	0.96 ±0.16	42.52 ±6.83

Fig. 2 shows qualitative results from selected sample from the testing dataset. From the generated results across various modalities, ALDM produces images that, apart from our method, are the closest to the ground truth in terms of detail; however, its images are relatively blurry. Pix2Pix not only exhibits blurry images but also suffers from varying degrees of detail loss across all modalities.

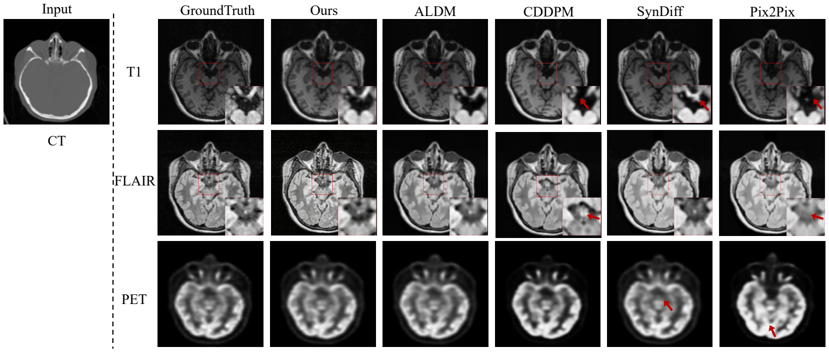

Fig. 2. Visual inspection for experimental results on the CiM testing dataset. On the left are the input CT images and on the right are synthetic images of three modalities generated by all the related methods.

Regarding CDDPM and SynDiff, their generated results show detail deficiencies in certain modalities. In the visual representations, red arrows indicate areas with noticeable detail loss.

Performance on BraTS 2019 Dataset: To further evaluate our method, we re-trained our model using the BraTS 2019 dataset for one-to-many synthesis from T1 to T1CE, FLAIR, and T2. Fig. 3 shows the test results of all related models in the BraTS 2019 dataset. From Fig. 3, we can see that our model outperforms the other four models in terms of quantitative metrics, which is consistent with the results in Table 1. Furthermore, based on the visual results, our model maintains the best integrity of lesion generation in the generated images.

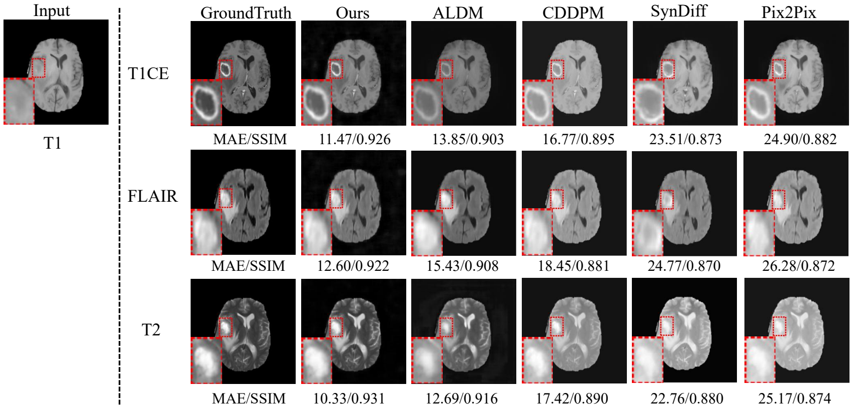

Fig. 3. Visual inspection for experimental results on the BraTS 2019 dataset .The red rectangular boxes highlight magnified views of pathological regions. Quantitative metrics (MAE/SSIM) corresponding to each imaging modality are displayed below the respective images.

Ablation study on key components in the proposed method: To investigate the impact of key components in the proposed method, we conducted multiple ablation experiments (see Table 2). By introducing the Adaptive Frequency Feature Processing mechanism (AF), the method reduced MAE by 15.95%, increased SSIM by 2.78%, and improved PSNR by 10.2% in the CT-to-T1 synthesis task. The experimental results also show that, after incorporating the pre-training process of the autoencoder (PR), MAE decreased by 14.86%, SSIM increased by 1.37%, and PSNR improved by 8.34% in the CT-to-T1 synthesis task. These performance improvements were also observed in the CT-to-FLAIR and CT-to-PET synthesis tasks, further validating the crucial role of the adaptive feature processing mechanism and model structure optimization in improving the quality of medical image synthesis.

Ablation study on BERT tokenizer: Specifically, to verify the effectiveness of the BERT tokenizer (BT) in text feature embedding, we removed BERT and conducted experiments using one-hot vectors as a control. The results demonstrated that the model with BT reduced MAE by 13.27%, increased SSIM by 1.02%, and improved PSNR by 6.08% compared to the model without BT in the CT-to-T1 synthesis task.

Table 2. Experimental Results of Ablation Study

			CT-to-T1			CT-to-FLAIR			CT-to-PET		
AF	PR	BT	MAE	SSIM	PSNR	MAE	SSIM	PSNR	MAE	SSIM	PSNR
✗	✓	✓	11.66	0.862	35.44	8.71	0.852	35.44	6.70	0.945	39.02
			±1.73	±0.05	±3.77	±2.70	±0.07	±2.92	±2.51	±0.19	±5.11
✓	✗	✓	11.51	0.874	36.07	9.34	0.824	32.28	5.52	0.936	37.06
			±2.20	±0.07	±4.14	±3.01	±0.09	±3.52	±1.77	±0.16	±6.65
✓	✓	✗	11.30	0.877	36.84	8.82	0.859	35.91	6.70	0.949	39.42
			±1.94	±0.06	±4.02	±2.81	±0.05	±3.62	±1.98	±0.10	±6.27
✓	✓	✓	9.80	0.886	39.08	6.97	0.865	37.94	5.40	0.962	42.52
(Ours)			±1.67	±0.03	±3.92	±1.99	±0.04	±2.49	±1.95	±0.16	±6.83

5 Conclusion

This paper introduces a novel one-to-many multimodal medical image synthesis method employing text-guided diffusion models. By integrating frequency-domain feature processing and leveraging modality-specific pre-trained VQ-GANs, the diffusion model's performance in medical image synthesis is significantly enhanced, demonstrating notable advantages over existing GAN-based and diffusion-model-based approaches. However, while the current method excels in standard image quality metrics, these metrics primarily emphasize global structural fidelity and inadequately capture critical diagnostic details. Additionally, the model presents certain limitations: it relies on modality-specific pre-trained VQ-GANs and text prompts, assuming the target modality name is a known condition. Future research aims to incorporate physical information constraints into the diffusion process, extend the approach to 3D volume synthesis, and validate its efficacy in clinical trials for lesion detection tasks. Furthermore, evaluations specific to particular tasks, the integration of domain-specific constraints, and the exploration of few-shot adaptation or unified latent spaces for novel modalities will be pursued.

Acknowledgments. This work was partially supported by the China National Key R&D Program (2023YFC2706400), the Beijing Natural Science Foundation (L245015), the Medical Health Science and Technology Project of Zhejiang Provincial Health Commission (2024KY1172), the Zhejiang Provincial Natural Science Foundation of China (LTGY23F020003), and the Chongqing Natural Science Foundation (CSTB2024NSCQ-MSX0451).

Disclosure of Interests. All authors declare that they have no conflicts of interest in terms of competing financial interests or personal relationships that could influence or are relevant to the work reported in this document.

References

1. David Bau, Jun-Yan Zhu, Jonas Wulff, William Peebles, Hendrik Strobelt, Bolei Zhou, and Antonio Torralba. Seeing what a gan cannot generate. In *International Conference on Computer Vision*, pages 4502–4511, 2019.
2. Jie Cao, Qingxuan Jiang, and Boshuo Li. Diff-fft: Synthetic lumbar magnetic images from frequency domain-based. In *2024 IEEE 16th International Conference on Computational Intelligence and Communication Networks (CICN)*, pages 839–844. IEEE, 2024.
3. Gaël Chételat, Javier Arbizu, Henryk Barthel, Valentina Garibotto, Ian Law, Silvia Morbelli, et al. Amyloid-pet and 18f-fdg-pet in the diagnostic investigation of alzheimer's disease and other dementias. *The Lancet Neurology*, 19(11):951–962, 2020.
4. Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.
5. Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.
6. Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Computer Vision and Pattern Recognition*, pages 12873–12883, 2021.
7. Eryan Feng, Pinle Qin, Rui Chai, Jianchao Zeng, Qi Wang, Yanfeng Meng, and Peng Wang. Mri generated from ct for acute ischemic stroke combining radiomics and generative adversarial networks. *IEEE Journal of Biomedical and Health Informatics*, 26(12):6047–6057, 2022.
8. Massimo Filippi, Maria A Rocca, Olga Ciccarelli, Nicola De Stefano, Nikos Evangelou, Ludwig Kappos, et al. Mri criteria for the diagnosis of multiple sclerosis: Magnims consensus guidelines. *The Lancet Neurology*, 15(3):292–303, 2016.

9. Meiqin Gong, Siyu Chen, Qingyuan Chen, Yuanqi Zeng, and Yongqing Zhang. Generative adversarial networks in medical image processing. *Current Pharmaceutical Design*, 27(15):1856–1868, 2021.
10. Robert Graf, Joachim Schmitt, Sarah Schlaeger, Hendrik Kristian Möller, Vasiliki Sideri-Lampretsa, Anjany Sekuboyina, et al. Denoising diffusion-based mri to ct image translation enables automated spinal segmentation. *European Radiology Experimental*, 7(1):70, 2023.
11. Jie Gui, Zhenan Sun, Yonggang Wen, Dacheng Tao, and Jieping Ye. A review on generative adversarial networks: Algorithms, theory, and applications. *IEEE Transactions on Knowledge and Data Engineering*, 35(4):3313–3332, 2021.
12. Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.
13. Bing Huang, Feng Yang, Mengxiao Yin, Xiaoying Mo, Cheng Zhong, et al. A review of multimodal medical image fusion techniques. *Computational and Mathematical Methods in Medicine*, 2020, 2020.
14. Yuming Jiang, Zhicheng Zhang, Wei Wang, Weicai Huang, Chuanli Chen, Sujuan Xi, et al. Biology-guided deep learning predicts prognosis and cancer immunotherapy response. *Nature Communications*, 14(1):5135, 2023.
15. Yuming Jiang, Zhicheng Zhang, Qingyu Yuan, Wei Wang, Hongyu Wang, Tuanjie Li, Weicai Huang, Jingjing Xie, et al. Predicting peritoneal recurrence and disease-free survival from ct images in gastric cancer with multitask deep learning: a retrospective study. *The Lancet Digital Health*, 4(5):e340–e350, 2022.
16. Jonghun Kim and Hyunjin Park. Adaptive latent diffusion model for 3d medical image to image translation: Multi-modal magnetic resonance imaging study. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 7604–7613, 2024.
17. Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
18. Monica Morrow, Janet Waters, and Elizabeth Morris. Mri for breast cancer screening, diagnosis, and treatment. *The Lancet*, 378(9805):1804–1811, 2011.
19. Reabal Najjar. Redefining radiology: a review of artificial intelligence integration in medical imaging. *Diagnostics*, 13(17):2760, 2023.
20. Muzaffer Özbey, Onat Dalmaz, Salman UH Dar, Hasan A Bedel, Şaban Özturk, Alper Güngör, and Tolga Çukur. Unsupervised medical image translation with adversarial diffusion models. *IEEE Transactions on Medical Imaging*, 2023.
21. Yingxue Pang, Jianxin Lin, Tao Qin, and Zhibo Chen. Image-to-image translation: Methods and applications. *IEEE Transactions on Multimedia*, 24:3859–3881, 2021.
22. Junbo Peng, Richard LJ Qiu, Jacob F Wynne, Chih-Wei Chang, Shaoyan Pan, Tonghe Wang, et al. Cbct-based synthetic ct image generation using conditional denoising diffusion probabilistic model. *arXiv preprint arXiv:2303.02649*, 2023.
23. Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.
24. Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.
25. Weiwen Wu, Yanyang Wang, Qiegen Liu, Ge Wang, and Jianjia Zhang. Wavelet-improved score-based generative model for medical imaging. *IEEE transactions on medical imaging*, 43(3):966–979, 2023.
26. Youjian Zhang, Li Li, Jie Wang, Xinquan Yang, Haotian Zhou, Jiahui He, Yaoqin Xie, Yuming Jiang, Wei Sun, Xinyuan Zhang, et al. Texture-preserving diffusion model for cbct-to-ct synthesis. *Medical Image Analysis*, 99:103362, 2025.

27. Man Zhou, Jie Huang, Chun-Le Guo, and Chongyi Li. Fourmer: An efficient global modeling paradigm for image restoration. In *International Conference on Machine Learning*, pages 42589–42601. PMLR, 2023.
28. Tao Zhou, Qi Li, Huiling Lu, Qianru Cheng, and Xiangxiang Zhang. Gan review: Models and medical image fusion applications. *Information Fusion*, 91:134–148, 2023.