Make-A-Volume: Leveraging Latent Diffusion Models for Cross-Modality 3D Brain MRI Synthesis

Lingting Zhu[1], Zeyue Xue[1], Zhenchao Jin[1], Xian Liu[2], Jingzhen He[3(✉)], Ziwei Liu[4], and Lequan Yu[1(✉)]

[1] The University of Hong Kong, Hong Kong SAR, China
ltzhu99@connect.hku.hk, lqyu@hku.hk
[2] The Chinese University of Hong Kong, Hong Kong SAR, China
[3] Qilu Hospital of Shandong University, Jinan, China
hjzhhjzh@163.com
[4] S-Lab, Nanyang Technological University, Singapore, Singapore

Abstract. Cross-modality medical image synthesis is a critical topic and has the potential to facilitate numerous applications in the medical imaging field. Despite recent successes in deep-learning-based generative models, most current medical image synthesis methods rely on generative adversarial networks and suffer from notorious mode collapse and unstable training. Moreover, the 2D backbone-driven approaches would easily result in volumetric inconsistency, while 3D backbones are challenging and impractical due to the tremendous memory cost and training difficulty. In this paper, we introduce a new paradigm for volumetric medical data synthesis by leveraging 2D backbones and present a diffusion-based framework, **Make-A-Volume**, for cross-modality 3D medical image synthesis. To learn the cross-modality slice-wise mapping, we employ a latent diffusion model and learn a low-dimensional latent space, resulting in high computational efficiency. To enable the 3D image synthesis and mitigate volumetric inconsistency, we further insert a series of volumetric layers in the 2D slice-mapping model and fine-tune them with paired 3D data. This paradigm extends the 2D image diffusion model to a volumetric version with a slightly increasing number of parameters and computation, offering a principled solution for generic cross-modality 3D medical image synthesis. We showcase the effectiveness of our Make-A-Volume framework on an in-house SWI-MRA brain MRI dataset and a public T1-T2 brain MRI dataset. Experimental results demonstrate that our framework achieves superior synthesis results with volumetric consistency.

Keywords: Cross-modality medical image synthesis · Volumetric data · Latent diffusion model · Brain MRI

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14229, pp. 592–601, 2023.
https://doi.org/10.1007/978-3-031-43999-5_56

1 Introduction

Medical images are essential in diagnosing and monitoring various diseases and patient conditions. Different imaging modalities, such as computed tomography (CT) and magnetic resonance imaging (MRI), and different parametric images, such as T1 and T2 MRI, have been developed to provide clinicians with a comprehensive understanding of the patients from multiple perspectives [7]. However, in clinical practice, it is commonly difficult to obtain a complete set of multiple modality images for diagnosis and treatment due to various reasons, such as modality corruption, incorrect machine settings, allergies to specific contrast agents, and limited available time [5,10]. Therefore, cross-modality medical image synthesis is useful by allowing clinicians to acquire different characteristics across modalities and facilitating real-world applications in radiology and radiation oncology [28,32].

With the rise of deep learning, numerous studies have emerged and are dedicated to medical image synthesis [4,7,18]. Notably, generative adversarial networks (GANs) [8] based approaches have garnered significant attention in this area due to their success in image generation and image-to-image translation [11,33]. Moreover, GANs are also closely related to cross-modality medical image synthesis [2,10,32]. However, despite their efficacy, GANs are susceptible to mode collapse and unstable training, which can negatively impact the performance of the model and decrease the reliability in practice [1,17]. Recently, the advent of denoising diffusion probabilistic models (DDPMs) [9,24] has introduced a new scheme for high-quality generation, offering desirable features such as better distribution coverage and more stable training when compared to GAN-based counterparts. Benefiting from the better performance [6], diffusion-based models may be deemed much more reliable and dominant and recently researchers have made the first attempts to employ diffusion models for medical image synthesis [12–14,19].

Different from natural images, most medical images are volumetric. Previous studies employ 2D networks as backbones to synthesize slices of medical volumetric data due to their ease of training [18,32] and then stack 2D results for 3D synthesis. However, this fashion induces volumetric inconsistency, particularly along the z-axis when following the standard way of placing the coordinate system. Although training 3D models may avoid this issue, it is challenging and impractical due to the massive amount of volumetric data required, and the higher dimension of the data would result in costly memory requirements [3,16,26]. To sum up, balancing the trade-off between training and volumetric consistency remains an open question that requires further investigation.

In this paper, we propose **Make-A-Volume**, a diffusion-based pipeline for cross-modality 3D brain MRI synthesis. Inspired by recent works that factorize video generation into multiple stages [23,31], we introduce a new paradigm for volumetric medical data synthesis by leveraging 2D backbones to simultaneously facilitate high-fidelity cross-modality synthesis and mitigate volumetric inconsistency for medical data. Specifically, we employ a latent diffusion model (LDM) [20] to function as a slice-wise mapping that learns cross-modality translation in an image-to-image manner. Benefiting from the low-dimensional latent space of LDMs, the high memory requirements for training are mitigated. To enable the 3D image synthesis and enhance volumetric smoothness among medical slices, we further insert and fine-tune a series of volumetric layers to upgrade the slice-wise model to a volume-wise model. In summary, our contributions are three-fold: (1) We introduce a generic paradigm for 3D image synthesis with 2D backbones, which can mitigate volumetric inconsistency and training difficulty related to 3D backbones. (2) We propose an efficient latent diffusion-based framework for high-fidelity cross-modality 3D medical image synthesis. (3) We collected a large-scale high-quality dataset of paired susceptibility weighted imaging (SWI) and magnetic resonance angiography (MRA) brain images. Experiments on these in-house and public T1-T2 brain MRI datasets show the volumetric consistency and superior quantitative result of our framework.

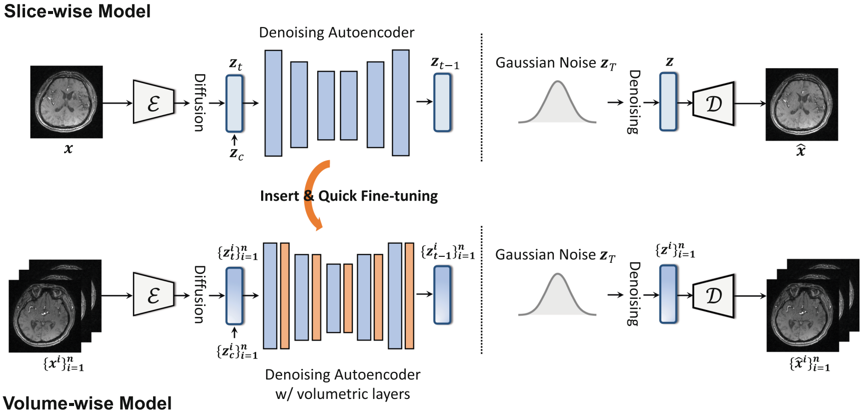

Fig. 1. Overview of our proposed two-stage Make-A-Volume framework. A latent diffusion model is used to predict the noises added to the image and synthesize independent slices from Gaussian noises. We insert volumetric layers and quickly fine-tune the model, which extends the slice-wise model to be a volume-wise model and enables synthesizing volumetric data from Gaussian noises.

2 Method

2.1 Preliminaries of DDPMs

In the diffusion process, DDPMs produce a series of noisy inputs $x_0, x_1, ..., x_T$, via sequentially adding Gaussian noises to the sample over a predefined number of timesteps T. Formally, given clean data samples which follow the real distribution $x_0 \sim q(x)$, the diffusion process can be written down with variances $\beta_1, ..., \beta_T$ as

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I}). \tag{1}$$

Employing the property of DDPMs, the corrupted data x_t can be sampled easily from x_0 in a closed form:

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\mathbf{I}); \quad x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \tag{2}$$

where $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$, and $\epsilon \sim \mathcal{N}(0, 1)$ is the added noise.

In the reverse process, the model learns a Markov chain process to convert the Gaussian distribution into the real data distribution by predicting the parameterized Gaussian transition $p(x_{t-1}|x_t)$ with the learned model θ:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \sigma_t^2\mathbf{I}). \tag{3}$$

In the model training, the model tries to predict the added noise ϵ with the simple mean squared error (MSE) loss:

$$L(\theta) = \mathbb{E}_{x_0 \sim q(x), \epsilon \sim \mathcal{N}(0,1), t}\left[\left\|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t)\right\|^2\right]. \tag{4}$$

2.2 Slice-Wise Latent Diffusion Model

To improve the computational efficiency of DDPMs that learn data in pixel space, Rombach *et al.* [20] proposes training an autoencoder with a KL penalty or a vector quantization layer [15,27], and introduces the diffusion model to learn the latent distribution. Given calibrated source modality image x_c and target modality image x, we leverage a slice-wise latent diffusion model to learn the cross-modality translation. With the pretrained encoder $\mathcal{E}$, x_c and x are compressed into a spatially lower-dimensional latent space of reduced complexity, generating z_c and z. The diffusion and denoising processes are then implemented in the latent space and a U-Net [21] is trained to predict the noise in the latent space. The input consists of the concatenated z_c and z and the network learns the parameterized Gaussian transition $p_\theta(z_{t-1}|z_t, z_c) = \mathcal{N}(z_{t-1}; \mu_\theta(z_t, t, z_c), \sigma_t^2\mathbf{I})$. After learning the latent distribution, the slice-wise model can synthesize target latent $\hat{z}$ from Gaussian noise, given the source latent z_c. Finally, the decoder $\mathcal{D}$ restores the slice to the image space via $\hat{x} = \mathcal{D}(\hat{z})$.

2.3 From Slice-Wise Model to Volume-Wise Model

Figure 1 illustrates an overview of the Make-A-Volume framework. The first stage involves a latent diffusion model that learns the cross-modality translation in an image-to-image manner to synthesize independent slices from Gaussian noises. Then, to extend the slice-wise model to be a volume-wise model, we insert volumetric layers and quickly fine-tune the U-Net. As a result, the volume-wise model synthesizes volumetric data without inconsistency from Gaussian noises.

In the slice-wise model, distribution of the latent $z \in \mathbb{R}^{b_s \times c \times h \times w}$ is learned by the U-Net, where b_s, c, h, w are the batch size of slice, channels, height, and width dimensions respectively, and there is where little volume-awareness is introduced to the network. Since we target in synthesizing volumetric data and assume each volume consists of N slices, we can factorize the batch size of slices as $b_s = b_v n$, where B_v represents the batch size of volumes. Now, volumetric layers are injected and help the U-Net learn to latent feature $f \in \mathbb{R}^{(b_v \times n) \times c \times h \times w}$ with volumetric consistency. The volumetric layers are basic 1D convolutional layers and the $i-$th volumetric layer l_v^i takes in feature f and outputs f' as:

$$f' \leftarrow \text{Rearrange}(f, (b_v \times n)\ c\ h\ w \rightarrow (b_v \times h \times w)\ c\ n), \tag{5}$$

$$f' \leftarrow l_v^i(f'), \tag{6}$$

$$f' \leftarrow \text{rearrange}(f, (b_v \times h \times w)\ c\ n \rightarrow (b_v \times n)\ c\ h\ w). \tag{7}$$

Here, the 1D conv layers combined with the pretrained 2D conv layers, serve as pseudo 3D conv layers with little extra memory cost. We initialize the volumetric 1D convolution layers as Identity Functions for more stable training and we empirically find tuning is efficient. With the volume-aware network, the model learns volume data $\{x^i\}_{i=1}^n$, predicts $\{z^i\}_{i=1}^n$, and reconstruct $\{\hat{x}^i\}_{i=1}^n$. For diffusion model training, in the first stage, we randomly sample timestep t for each slice. However, when tuning the second stage, the U-Net with volumetric layers learns the relationship between different slices in one volume. As a result, fixing t for each volume data is necessary and we encourage the small t values to be sampled more frequently for easy training. In detail, we sample the timestep t with replacement from multinomial distribution, and the pre-normalized weight (used for computing probabilities after normalization) for timestep t equals $2T - t$, where T is the total number of timesteps. Therefore, we enable a seamless translation from the slice-wise model which processes slices individually, to a volume-wise model with better volumetric consistency.

3 Experiments

Datasets. The experiments were conducted on two brain MRI datasets: SWI-to-MRA (S2M) dataset and RIRE [30][1] T1-to-T2 dataset. To facilitate SWI-to-MRA brain MRI synthesis applications, we collected a high-quality SWI-to-MRA dataset. This dataset comprises paired SWI and MRA volume data of 111 patients that were acquired at Qilu Hospital of Shandong University using one 3.0T MRI scanner (*i.e.,* Verio from Siemens). The SWI scans have a voxel spacing of $0.3438 \times 0.3438 \times 0.8$ mm and the MRA scans have a voxel spacing of $0.8984 \times 0.8984 \times 2.0$ mm. While most public brain MRI datasets lack high-quality details along z-axis and therefore are weak to indicate volumetric inconsistency, this volume data provides a good way to illustrate the performances for volumetric synthesis due to the clear blood vessels. We also evaluate our method on the public RIRE dataset [30]. The RIRE dataset includes T1 and T2-weighted MRI volumes, and 17 volumes were used in the experiments.

Implementation Details. To summarize, for the S2M dataset, we randomly select 91 paired volumes for training and 20 paired volumes for inference; for the

[1] https://rire.insight-journal.org/index.html.

RIRE T1-to-T2 dataset, 14 volumes are randomly selected for training and 3 volumes are used for inference. All the volumes are resized to 256 × 256 × 100 for S2M and 256 × 256 × 35 for RIRE, where the last dimension represents the z-axis dimension, *i.e.,* the number of slices in one volume for 2D image-to-image setting. Our proposed method is built upon U-Net backbones. We use a pretrained KL autoencoder with a downsampling factor of $f = 4$. We train our model on an NVIDIA A100 80 GB GPU.

Table 1. Quantitative comparison on S2M and RIRE datasets.

Methods	S2M			RIRE [30]		
	MAE ↓	SSIM ↑	PSNR ↑	MAE ↓	SSIM ↑	PSNR ↑
Pix2pix [11]	8.175	0.739	25.663	16.812	0.538	20.106
Palette [22]	26.806	0.141	15.643	36.131	0.251	14.269
Pix2pix 3D [11]	6.234	0.765	28.395	11.369	0.650	22.854
CycleGAN 3D [33]	7.621	0.755	26.908	13.794	0.542	20.627
Ours 200 steps	**5.243**	**0.788**	**29.446**	**10.794**	**0.676**	**24.332**
Ours 1000 steps	**4.801**	**0.801**	**30.143**	**10.619**	**0.684**	**25.458**

Quantitative Results. We compare our pipeline to several baseline methods, including 2D-based methods: (1) Pix2pix [11], a solid baseline for image-to-image translation; (2) Palette [22], a diffusion-based method for 2D image translation; 3D-based methods: (3) a 3D version of Pix2pix, created by modifying the 2D backbone as a 3D backbone in the naive Pix2pix approach; and (4) a 3D version of CycleGAN [33]. Naive 3D diffusion-based models are not included due to the lack of efficient backbones and the matter of timesteps' sampling efficiency. We report the results in terms of mean absolute error (MAE), Structural Similarity Index (SSIM) [29], and peak signal-to-noise ratio (PSNR).

Table 1 presents a quantitative comparison of our method and baseline approaches on the S2M and RIRE datasets. Our method achieves better performance than the baselines in terms of various evaluation metrics. To accelerate the sampling of diffusion models, we implement DDIM [25] with 200 steps and report the results accordingly. It is worth noting that for the baseline approaches, the 3D version method (Pix2pix 3D) outperforms the corresponding 2D version (Pix2pix) at the cost of additional memory usage. For the Palette method, we implemented the 2D version but were unable to produce high-quality slices stably and failure cases dramatically affected the metrics results. Nonetheless, we included this method due to its great illustration of volumetric inconsistency.

Qualitative Results. Figure 2 presents a qualitative comparison of different methods, showcasing two axial slices of clear vessels. Our method synthesizes better images with more details, as shown in the qualitative results. The areas requiring special attention are highlighted with red arrows and red rectangles. It is worth noting that the synthesized axial slices not only depend on the source slice but also on the volume knowledge. For instance, for S2M case 1, the target slice shows a clear vessel cross-section that is based on the shape of the vessels in the volume. In Fig. 3, we provide coronal and sagittal views. For methods that rely on 2D generation, we synthesize individual slices and concatenate them to create volumes. It is clear to observe the volumetric inconsistency examining the coronal and sagittal views of these volumes. For instance, Palette synthesizes 2D slices unstably, where some good slices are synthesized but others are of poor quality. As a result, volumetric inconsistency severely impacts the performance of volumes. While 2D baselines inherently introduce inconsistency in the coronal and sagittal views, 3D baselines also generate poor results than ours, particularly in regard to blood vessels and ventricles.

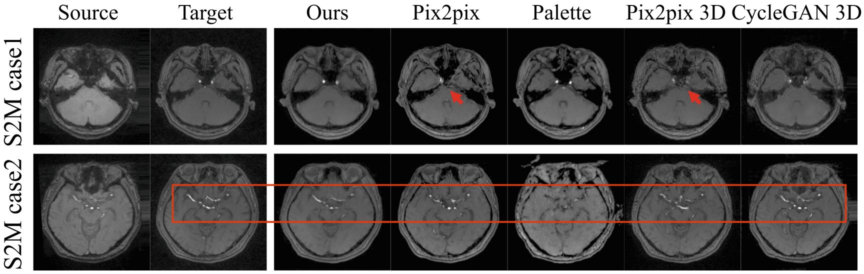

Fig. 2. Qualitative comparison. We compare our methods with baselines on two cases.

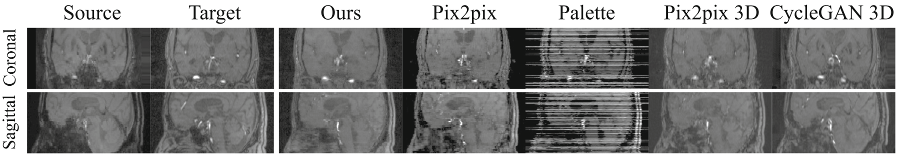

Fig. 3. Coronal view and sagittal view. To clearly indicate the volumetric consistency, we show a coronal view and a sagittal view of the volumes synthesized and the ground truth volumes.

Table 2. Ablation Quantitative Results.

Methods	S2M			RIRE [30]		
	MAE ↓	SSIM ↑	PSNR ↑	MAE ↓	SSIM ↑	PSNR ↑
w/o volumetric layers	5.128	0.792	29.894	10.925	0.667	24.623
w/ volumetric layers	4.801	0.801	30.143	10.619	0.684	25.458

Ablation Analysis. We conduct an ablation study to show the effectiveness of volumetric fine-tuning. Table 2 presents the quantitative results, demonstrating that our approach is able to increase the model's performance beyond that of the slice-wise model, without incurring significant extra training expenses. Figure 4 illustrates that fine-tuning volumetric layers helps to mitigate volumetric artifacts and produce clearer vessels, which is crucial for medical image synthesis.

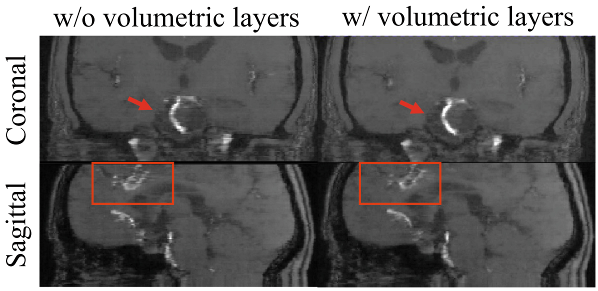

Fig. 4. Ablation qualitative results with coronal view and sagittal view.

4 Conclusion

In this paper, we propose Make-A-Volume, a diffusion-based framework for cross-modality 3D medical image synthesis. Leveraging latent diffusion models, our method achieves high performance and can serve as a strong baseline for multiple cross-modality medical image synthesis tasks. More importantly, we introduce a generic paradigm for volumetric data synthesis by utilizing 2D backbones and demonstrate that fine-tuning volumetric layers helps the two-stage model capture 3D information and synthesize better images with volumetric consistency. We collected an in-house SWI-to-MRA dataset with clear blood vessels to evaluate volumetric data quality. Experimental results on two brain MRI datasets demonstrate that our model achieves superior performance over existing baselines. Generating coherent 3D and 4D data is at an early stage in the diffusion models literature, we believe that by leveraging slice-wise models and extending them to 3D/4D models, more work can help achieve better volume synthesis with reasonable memory requirements. In the future, we will investigate more efficient approaches for more high-resolution volumetric data synthesis.

Acknowledgement. The work described in this paper was partially supported by grants from the Research Grants Council of the Hong Kong Special Administrative Region, China (T45-401/22-N), the National Natural Science Fund (62201483), HKU Seed Fund for Basic Research (202009185079 and 202111159073), RIE2020 Industry Alignment Fund - Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s).

References

1. Bau, D., et al.: Seeing what a GAN cannot generate. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4502–4511 (2019)
2. Ben-Cohen, A., et al.: Cross-modality synthesis from CT to pet using FCN and GAN networks for improved automated lesion detection. Eng. Appl. Artif. Intell. **78**, 186–194 (2019)
3. Chung, H., Ryu, D., McCann, M.T., Klasky, M.L., Ye, J.C.: Solving 3D inverse problems using pre-trained 2D diffusion models. arXiv preprint arXiv:2211.10655 (2022)
4. Dalmaz, O., Yurt, M., Çukur, T.: ResViT: residual vision transformers for multimodal medical image synthesis. IEEE TMI **41**(10), 2598–2614 (2022)
5. Dar, S.U., Yurt, M., Karacan, L., Erdem, A., Erdem, E., Cukur, T.: Image synthesis in multi-contrast MRI with conditional generative adversarial networks. IEEE TMI **38**(10), 2375–2388 (2019)
6. Dhariwal, P., Nichol, A.: Diffusion models beat GANs on image synthesis. Adv. Neural. Inf. Process. Syst. **34**, 8780–8794 (2021)
7. Filippou, V., Tsoumpas, C.: Recent advances on the development of phantoms using 3d printing for imaging with CT, MRI, PET, SPECT, and ultrasound. Med. Phys. **45**(9), e740–e760 (2018)
8. Goodfellow, I., et al.: Generative adversarial networks. Commun. ACM **63**(11), 139–144 (2020)
9. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Adv. Neural. Inf. Process. Syst. **33**, 6840–6851 (2020)
10. Hu, X., Shen, R., Luo, D., Tai, Y., Wang, C., Menze, B.H.: AutoGAN-synthesizer: neural architecture search for cross-modality MRI synthesis. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022, Part VI. LNCS, vol. 13436, pp. 397–409. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16446-0_38
11. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1125–1134 (2017)
12. Kazerouni, A., et al.: Diffusion models for medical image analysis: a comprehensive survey. arXiv preprint arXiv:2211.07804 (2022)
13. Khader, F., et al.: Medical diffusion-denoising diffusion probabilistic models for 3D medical image generation. arXiv preprint arXiv:2211.03364 (2022)
14. Kim, B., Ye, J.C.: Diffusion deformable model for 4D temporal medical image generation. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022, Part I. LNCS, vol. 13431, pp. 539–548. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16431-6_51
15. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
16. Lee, S., Chung, H., Park, M., Park, J., Ryu, W.S., Ye, J.C.: Improving 3D imaging with pre-trained perpendicular 2D diffusion models. arXiv preprint arXiv:2303.08440 (2023)
17. Li, K., Malik, J.: On the implicit assumptions of GANs. arXiv preprint arXiv:1811.12402 (2018)
18. Nie, D., et al.: Medical image synthesis with deep convolutional adversarial networks. IEEE Trans. Biomed. Eng. **65**(12), 2720–2730 (2018)

19. Pinaya, W.H., et al.: Brain imaging generation with latent diffusion models. In: Mukhopadhyay, A., Oksuz, I., Engelhardt, S., Zhu, D., Yuan, Y. (eds.) DGM4MICCAI 2022. LNCS, vol. 13609, pp. 117–126. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-18576-2_12
20. 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)
21. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015, Part III. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
22. Saharia, C., et al.: Palette: image-to-image diffusion models. In: ACM SIGGRAPH 2022 Conference Proceedings, pp. 1–10 (2022)
23. Singer, U., et al.: Make-a-video: text-to-video generation without text-video data. arXiv preprint arXiv:2209.14792 (2022)
24. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International Conference on Machine Learning, pp. 2256–2265. PMLR (2015)
25. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
26. Uzunova, H., Ehrhardt, J., Handels, H.: Memory-efficient GAN-based domain translation of high resolution 3d medical images. Comput. Med. Imaging Graph. **86**, 101801 (2020)
27. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. Adv. Neural Inf. Process. Syst. **30** (2017)
28. Wang, T., et al.: A review on medical imaging synthesis using deep learning and its clinical applications. J. Appl. Clin. Med. Phys. **22**(1), 11–36 (2021)
29. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)
30. West, J., et al.: Comparison and evaluation of retrospective intermodality brain image registration techniques. J. Comput. Assist. Tomogr. **21**(4), 554–568 (1997)
31. Wu, J.Z., et al.: Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. arXiv preprint arXiv:2212.11565 (2022)
32. Yu, B., Zhou, L., Wang, L., Shi, Y., Fripp, J., Bourgeat, P.: EA-GANs: edge-aware generative adversarial networks for cross-modality MR image synthesis. IEEE TMI **38**(7), 1750–1762 (2019)
33. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2223–2232 (2017)