Flip Distribution Alignment VAE for Multi-Phase MRI Synthesis

Xiaoyan Kui[1], Qianmu Xiao[1⋆], Qinsong Li[2], Zexin Ji[1], Jielin Zhang[3], and Beiji Zou[1]

[1] School of Computer Science and Engineering, Central South University, ChangSha, China
{xykui, qianmu, zexin.ji, bjzou}@csu.edu.cn
[2] Big Data Institute, Central South University, ChangSha, China
qinsli.cg@foxmail.com
[3] Department of Radiology, The Second Xiangya Hospital, Central South University, ChangSha, China
238212362@csu.edu.cn

Abstract. Separating shared and independent features is crucial for multi-phase contrast-enhanced (CE) MRI synthesis. However, existing methods use deep autoencoder generators with low parameter efficiency and lack interpretable training strategies. In this paper, we propose Flip Distribution Alignment Variational Autoencoder (FDA-VAE), a lightweight feature-decoupled VAE model for multi-phase CE MRI synthesis. Our method encodes input and target images into two latent distributions that are symmetric concerning a standard normal distribution, effectively separating shared and independent features. The Y-shaped bidirectional training strategy further enhances the interpretability of feature separation. Experimental results show that compared to existing deep autoencoder-based end-to-end synthesis methods, FDA-VAE significantly reduces model parameters and inference time while effectively improving synthesis quality. The source code is publicly available at https://github.com/QianMuXiao/FDA-VAE.

Keywords: Multi-Phase MRI Synthesis · Variational Autoencoder · Feature Alignment · Medical Image Synthesis.

1 Introduction

Multi-phase contrast-enhanced (CE) MRI provides essential diagnostic information for assessing organ lesions, tumors, and vascular abnormalities. However, this imaging technique is still limited by the long scanning time and nephrotoxicity risk caused by gadolinium-based contrast agents. Medical image super-resolution [13,3] and synthesis [8,21] techniques address these challenges by generating high-quality or missing images from low-quality or existing scans. In CE MRI, synthesizing different enhanced-phase images from unenhanced-phase can

⋆ Corresponding author.

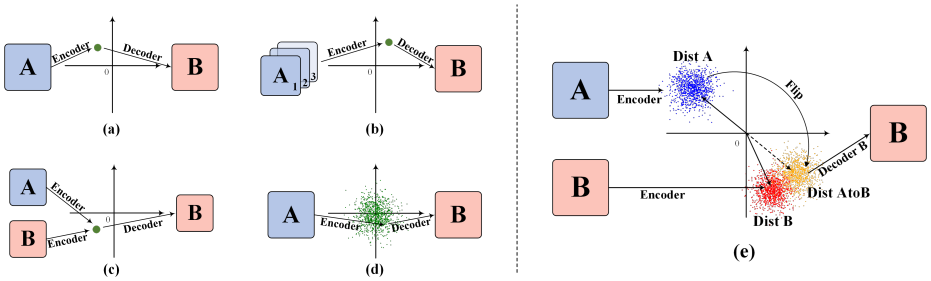

Fig. 1. Overview of existing AE-based medical image synthesis strategies. **(a)** Basic one-to-one autoencoder. **(b)** Multi-phase many-to-one autoencoder. **(c)** Autoencoder with latent space comparative learning. **(d)** Variational Autoencoder (VAE). **(e)** Our proposed method: Flip Distribution Alignment Variational Autoencoder (FDA-VAE).

effectively reduce scanning time and mitigate the health risks associated with contrast agents.

Current medical image synthesis methods can be categorized into Diffusion-based [23,18,28] and Autoencoder-based (AE-based) approaches [12,9,31,32,2]. Diffusion-based methods generate high-quality synthetic images but require substantial computational resources and are constrained by inflexible training strategies. In contrast, AE-based methods provide greater flexibility in training by allowing direct manipulation of latent space structures and optimization objectives.

Existing AE-based models employ different feature extractors to improve representation learning. Traditional CNN-based AEs [12] (Fig.1 (a)) suffer from limited receptive fields, making it difficult to capture long-range dependencies. Vision Transformers (ViTs) [10,31] address this issue but introduce high computational costs. Recently, state-space models (SSMs) such as Mamba [11,2] have emerged as efficient alternatives. Some studies further integrate hybrid architectures [9,32,16,29,6] that combine CNNs and Transformers to balance efficiency and performance. However, optimizing the encoding-decoding structure remains crucial for high-quality synthesis, as shown in Fig.1 (b)-(d). Some approaches integrate multi-phase input features to generate the target modality Fig.1 (b). Others use latent feature contrastive learning [19] Fig.1 (c) or structure-supervised loss [7,20] to reinforce shared features between the input and target modalities. Furthermore, some studies encode latent features as probabilistic distributions [14,4,15,5] Fig.1 (d). This improves the smoothness of the latent space. It also enhances uncertainty modeling and generation diversity.

However, there are still some problems with these strategies: (1) Many-to-one methods require paired multi-phase data for training. (2) One-to-one methods focus only on cross-modality mappings or shared features but neglect their independent features. (3) Existing methods still rely on deep autoencoders for cross-modality mapping even under limited paired training data, which may lead to suboptimal parameter utilization.

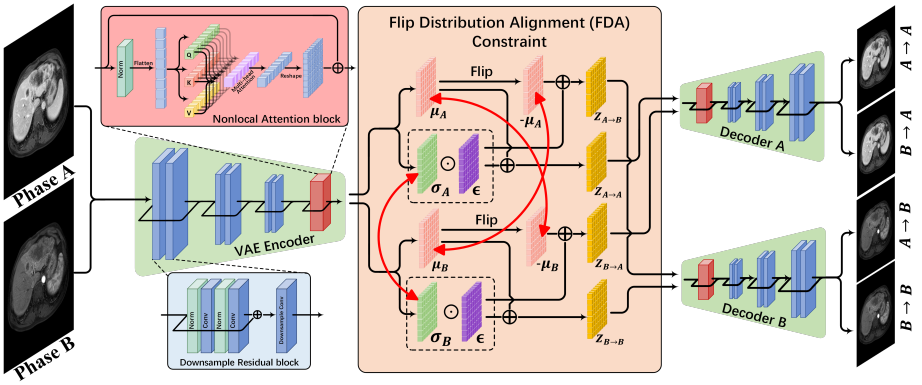

Fig. 2. Overview of the proposed Flip Distribution Alignment Variational Autoencoder (FDA-VAE). The model consists of a shared encoder, two independent decoders, and a flip distribution alignment (FDA) constraint layer. During training, a pair of different phase MRI images is input to obtain self-reconstructed and cross-phase transformed outputs. In the inference stage, only the target decoder is retained. The image is encoded to obtain the latent distribution, and the mean vector is flipped before decoding, generating the target-phase image from the flipped distribution.

To address these problems, we propose Flip Distribution Alignment VAE (FDA-VAE). It is a lightweight feature-decoupled model for multi-phase CE MRI synthesis. Our method uses a compact hybrid-architecture VAE as the generator to reduce model parameters and improve efficiency. We introduce Flip Distribution Alignment (FDA) as a structured constraint on the latent space. Specifically, our method encodes input and target images as two latent distributions, enforcing symmetry by setting opposite means and equal variances. This ensures that shared features are preserved while maximizing independent components, with transformation achieved via simple mean flipping. Additionally, we design a Y-shaped bidirectional training strategy, enabling both self-reconstruction and cross-phase synthesis through mean flipping. This enhances the interpretability and stability of latent space modeling. Compared to existing methods, FDA-VAE provides a structured and interpretable latent space representation, significantly improving synthesis quality and parameter efficiency.

2 Method

Lightweight VAE vs Deep AE. Pre-trained VAE [14] demonstrates excellent data compression and decoding capabilities in high-resolution image synthesis tasks. Recent approaches, such as Latent Diffusion (LDM) [24] and Visual Autoregressive [25] (VAR), utilize pre-trained VAE or Vector Quantised-VAE (VQ-VAE) [26] models. These models typically contain around 100M parameters and are trained on 1.2 million natural images. They are first trained for image self-reconstruction, providing a latent representation that facilitates sub-

sequent feature generation. In medical image synthesis, MONAI's LDM-based approach [18] uses about 38,000 brain MRI slices to train a VAE with about 12M parameters. In contrast, methods like ResVit [9] and I2I-Mamba [2] train cross-modality mappings using only about 2,500 paired slices, yet rely on generators exceeding 100M parameters. Although the self-reconstruction task is relatively simple, cross-modality medical images typically exhibit strong structural correlations. In terms of efficiency, existing deep AE generators tend to have excessive parameters, resulting in inefficient utilization, particularly given the limited size of medical datasets.

In this paper, we propose utilizing a shallow, lightweight VAE backbone directly as the generator, aiming to enhance image synthesis quality and model interpretability through structured latent space modeling. As shown in Fig. 2, we construct a hybrid-architecture VAE backbone, where both the encoder and decoder consist of three residual convolutional blocks and one non-local attention block to capture local and global features. Compared to existing deep AE generators, our backbone has fewer layers and a narrower model width. The formula for the model is as follows:

$$\mu, \sigma = \text{Encoder}(x), \quad z = \mu + \sigma * \epsilon, \quad \epsilon \sim \mathcal{N}(0,1), \quad \hat{x} = \text{Decoder}(z) \tag{1}$$

$$\mathcal{L}_{\text{Kullback-Leibler}}\big(\mathcal{N}(\mu,\sigma^2) \parallel \mathcal{N}(0,1)\big) = \frac{1}{2}\Big(\mu^2 + \sigma^2 - \log(\sigma^2) - 1\Big) \tag{2}$$

Given an input image x, the encoder outputs a mean vector μ and a variance vector σ. Latent features z are then sampled from this latent distribution $\mathcal{N}(\mu, \sigma^2)$, and subsequently decoded into the target image $\hat{x}$ (Eq.1). Additionally, we employ the Kullback–Leibler (KL) divergence constraint (Eq.2) to regularize the encoded distributions towards a standard normal distribution.

Flip Distribution Alignment (FDA). The core idea of FDA-VAE is to build a structured and efficient latent space representation. With a shared encoder-decoder, we map the input and target images to separate latent distributions and sample from them to synthesize the target image. However, solely relying on KL divergence for regularization leads to two issues, as shown in Fig.3 (a).

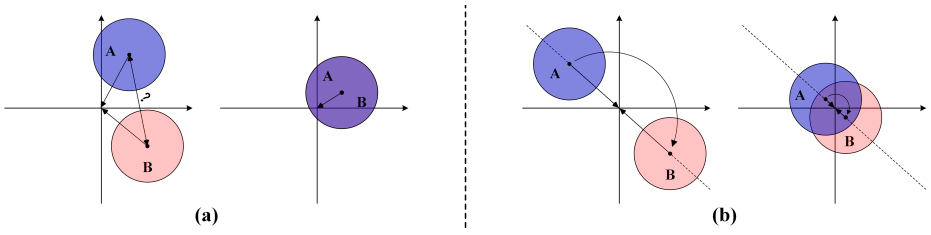

Fig. 3. Convergence process of input and target distributions: (a) KL divergence only, (b) KL divergence + FDA.

In the early stages, the input (A) and target (B) distributions move toward the standard normal distribution independently. Without explicit constraints, they approach from random locations, leading to unpredictable relative positioning. This misalignment disrupts feature correspondence and increases divergence, making feature transformation difficult and degrading synthesis quality. As training progresses, the lack of alignment causes the distributions to collapse onto each other, overemphasizing shared features while suppressing modality-specific information, reducing independent feature distinctiveness. To address these problems, we introduce an additional Flip Distribution Alignment (FDA) constraint shown in Fig.3 (b). FDA constrains the input and target distributions to remain symmetric concerning the standard normal distribution throughout training. Specifically, we enforce equal variances ($\sigma_A^2 = \sigma_B^2$) and opposite means ($\boldsymbol{\mu}_A = -\boldsymbol{\mu}_B$) as show in Eq.3.

$$\mathcal{L}_{FDA} = \left\|\boldsymbol{\mu}_A + \boldsymbol{\mu}_B\right\|_1 + \left\|\sigma_A^2 - \sigma_B^2\right\|_1 \tag{3}$$

Combining KL divergence and FDA constraints ensures that both distributions converge toward the standard normal distribution while maintaining structural symmetry. This design ensures that the input and target features remain maximally separated during convergence while preserving alignment with the standard normal distribution. Additionally, the symmetric relative positioning allows feature transformation to be efficiently performed via a simple mean-flipping operation.

Y-shaped Bidirectional Training. To further enhance feature disentanglement, we design a Y-shaped bidirectional training strategy, consisting of a shared encoder that maps both modalities to symmetric latent distributions and two phase-specific decoders to synthesize images. Given input phase x_A, the process involves encoding, flipping, and decoding to obtain $x_{A\to A}$ and $x_{A\to B}$, while input x_B follows the same process for $x_{B\to B}$ and $x_{B\to A}$.

$$\mu_A, \sigma_A = \text{Encoder}(x_A), \quad z_{A\to A} \sim \mathcal{N}(\mu_A, \sigma_A^2), \quad z_{A\to B} \sim \mathcal{N}(-\mu_A, \sigma_A^2) \tag{4}$$

$$\hat{x}_{A\to A} = \text{Decoder}_A(z_{A\to A}), \quad \hat{x}_{A\to B} = \text{Decoder}_B(z_{A\to B}) \tag{5}$$

For the self-reconstruction task, we use L1 loss for supervision $\mathcal{L}_{Rec}$, for the cross-phase synthesis task, we incorporate L1 loss $\mathcal{L}_{Trans}$, GAN loss $\mathcal{L}_{GAN}$, and perceptual loss $\mathcal{L}_{Perce}$ for co-supervision. The entire loss function FDA-VAE is summarized as $\mathcal{L}_{FDA-VAE}$:

$$\begin{aligned}\mathcal{L}_{FDA-VAE} =& \lambda_{\text{rec}}\mathcal{L}_{\text{Rec}} + \mathcal{L}_{\text{Tran}} + \lambda_{\text{gan}}\mathcal{L}_{\text{GAN}} \\ &+ \lambda_{\text{perce}}\mathcal{L}_{\text{Perce}} + \lambda_{\text{kl}}\mathcal{L}_{\text{KL}} + \lambda_{\text{fda}}\mathcal{L}_{\text{FDA}}\end{aligned} \tag{6}$$

where λ_{rec}, λ_{gan}, λ_{perce} and λ_{fda}, take the value of 1×10^{-2}, λ_{kl} takes the value of 1×10^{-7}.

3 Experiments

3.1 Experiment Setups

Dataset & Pre-process. FDA-VAE was trained on the LLD-MMRI 2023 dataset [17], containing 498 patients across seven liver lesion types (four benign, three malignant). We selected four T1 contrast-enhanced phases: Pre-contrast (Pre), arterial (CA), venous (CV), and delayed (Delay), designing six early-to-late phase synthesis tasks. Non-rigid registration was performed using ANTsPy [1] with the C+V phase as the reference. To ensure lesion-type consistency, images were grouped by disease category and split 4:1 for training and validation. Preprocessing included top 0.1% intensity clipping, normalization, and resizing to 256×256.
Evaluation Metrics. We evaluated our model using PSNR, SSIM [27] and LPIPS [30] to assess image quality. Additionally, we analyzed model efficiency in terms of parameter count and inference time per slice.
Training Details. All experiments were implemented in PyTorch v2.5.1 in conjunction with the MONAI [22] framework. We employed the Adam optimizer with an initial learning rate of 1e-4 and trained each model for 40 epochs on a Linux workstation with 4 × NVIDIA RTX 4090 24G GPUs. It took about six and a half hours to train FDA-VAE.

3.2 Ablation Study

We conduct an ablation study to assess the impact of each component. First, we establish the lightweight VAE backbone as a baseline (Tab.1, VAE (backbone)). While it achieves reasonable performance, its synthesis quality is constrained by the reduced model capacity. Next, we introduce the FDA constraint to enforce structured latent space alignment (Tab.1 VAE (KL+FDA)). This variant applies to mean flipping to transform the input distribution while using the target distribution for self-reconstruction. As shown in the result, the FDA constrained improves PSNR and SSIM while reducing LPIPS, demonstrating its effectiveness in feature separation and alignment. Finally, our complete model, FDA-VAE, integrates a bidirectional synthesis training strategy with two independent decoders. This further enhances synthesis quality, achieving the highest PSNR and SSIM while maintaining the lowest LPIPS scores (Tab.1). These results confirm the role of bidirectional training in stabilizing latent space modeling and improving synthesis performance.

3.3 Comparison with state-of-the-art models.

Quantitative Analysis. We evaluate the synthesis performance of FDA-VAE against Pix2Pix [12], ResVit [9], TransUnet [6], PTNet [31], and I2I-Mamba [2] across six tasks. Tab.1 presents the PSNR, SSIM, and LPIPS results, while Tab.2 compares parameter size and inference time. Among existing methods, ResVit [9] and TransUnet [6] achieve the best synthesis quality but require over

Table 1. Overview of Evaluation Results (**Bold** indicates optimal, Underline indicates sub-optimal, Box indicates optimal among the compared models, same as Table 2.)

Method/Task	Pre→CA	Pre→CV	Pre→Delay	CA→CV	CA→Delay	CV→Delay
			PSNR(dB)↑			
Pix2Pix [12]	24.79	23.74	23.60	24.73	24.56	27.27
ResVit [9]	25.34	24.06	24.33	25.88	25.41	26.63
TransUnet [6]	25.01	24.74	24.74	26.18	25.86	26.35
PTNet [31]	24.85	23.68	24.03	25.38	24.68	27.89
I2I-Mamba [2]	24.95	24.39	24.12	25.46	25.24	25.61
VAE(backbone)	25.23	24.96	23.63	26.48	24.65	27.08
VAE (KL+FDA)	25.71	24.95	**25.07**	26.54	26.10	27.99
FDA-VAE(Ours)	**25.89**	**24.98**	24.89	**26.72**	**26.33**	**28.59**
			SSIM(%)↑			
Pix2Pix [12]	80.41	68.23	77.14	78.50	78.95	84.56
ResVit [9]	81.79	76.32	78.99	79.20	81.90	84.98
TransUnet [6]	81.22	79.79	78.84	82.38	83.07	84.64
PTNet [31]	81.35	77.56	78.69	81.75	80.08	86.19
I2I-Mamba [2]	81.16	76.91	78.51	79.44	81.47	82.85
VAE(backbone)	72.72	79.60	76.29	82.57	74.23	80.42
VAE (KL+FDA)	82.99	80.10	**81.32**	83.07	82.99	86.41
FDA-VAE(Ours)	**83.70**	**80.68**	81.17	**84.01**	**83.84**	**87.48**
			LPIPS↓			
Pix2Pix [12]	0.0854	0.0903	0.0914	0.0837	0.0851	0.0568
ResVit [9]	**0.0713**	0.0776	0.0774	0.0626	0.0682	0.0534
TransUnet [6]	0.0768	0.0761	0.0791	0.0637	0.0677	0.0632
PTNet [31]	0.0771	0.0787	0.0824	0.0635	0.0764	**0.0463**
I2I-Mamba [2]	0.0739	0.0750	0.0783	0.0646	0.0698	0.0632
VAE(backbone)	0.0799	0.0779	0.1064	0.0661	0.0818	0.0576
VAE (KL+FDA)	0.0720	0.0735	**0.0716**	0.0656	0.0667	0.0490
FDA-VAE(Ours)	**0.0713**	**0.0729**	0.0723	**0.0611**	**0.0620**	0.0465

Table 2. Generator Params & Inference Times.

	Pix2Pix	ResVit	TransUnet	PTNet	I2I-Mamba	**Ours**
Params (m)	51.89	117.72	100.45	26.83	100.64	**11.78**
Inference (secs/slice)	**0.0019**	0.0121	0.0109	0.0141	0.0071	0.0050

100M parameters and have inference times exceeding 0.01s per slice. In contrast, our lightweight VAE backbone, with only 11.78M parameters, achieves comparable synthesis quality. Further improvements are observed with FDA and Y-shaped bidirectional training, significantly enhancing evaluation metrics. FDA-VAE achieves the best overall performance across most tasks, demonstrating its effectiveness in balancing synthesis quality and computational efficiency.

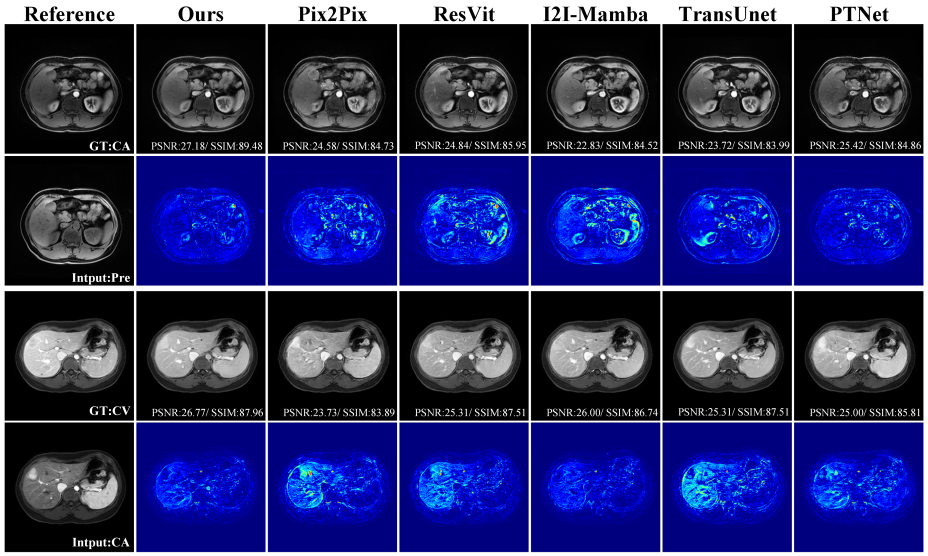

Fig. 4. Visualization of synthesis result and errors heat maps.

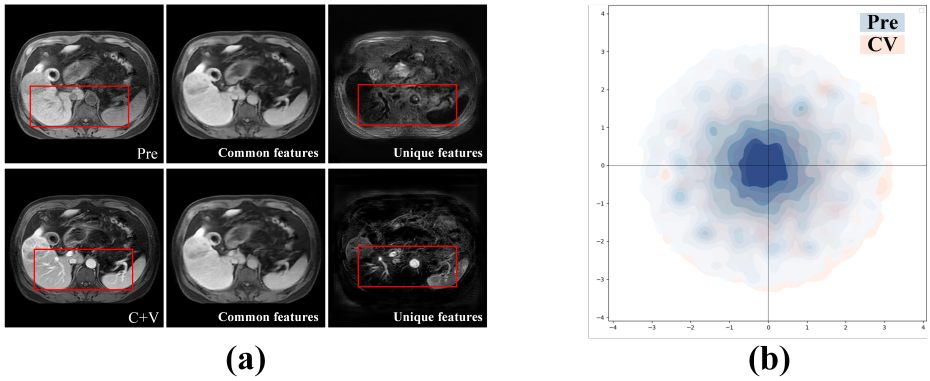

Fig. 5. FDA feature decoupling in pixel level (a) & latent space level (b).

Qualitative Analysis. Fig.4 presents the visualization results and error heat maps for six early-to-late phase synthesis tasks across all methods. Our method achieves the lowest pixel-level error, as indicated by the error heat maps. Fig.5 illustrates feature decoupling at both the pixel level (**a**) and latent space level (**b**). At the pixel level, our method effectively captures common structural and contrast features (second column in **a**) while preserving phase-specific contrast details (third column in **a**). In latent space, dimensionality reduction visualization shows overlapping regions between input and target distributions, with preserved non-overlapping areas, aligning with our FDA design objective.

4 Conclusion

In this paper, we propose a lightweight feature-decoupled VAE framework called FDA-VAE for multi-phase MRI synthesis. By incorporating the FDA constraint and a Y-shaped bidirectional training strategy, FDA-VAE simultaneously retains both common and independent features of the input and target images at the latent feature level. Compared with state-of-the-art methods, FDA-VAE achieves better synthesis quality while significantly reducing model parameters and inference time. In future work, we aim to extend this framework to unpaired data by leveraging unsupervised learning techniques for cross-phase synthesis.

Acknowledgments. This work is supported by the National Natural Science Foundation of China (No. U22A2034, 62177047, 62302530), High Caliber Foreign Experts Introduction Plan funded by MOST, Key Research and Development Programs of Department of Science and Technology of Hunan Province (No. 2024JK2135), Major Program from Xiangjiang Laboratory (No. 23XJ02005), the Scientific Research Fund of Hunan Provincial Education Department (No. 24A0018), Hunan Provincial Natural Science Foundation (No. 2023JJ40769), and Central South University Research Programme of Advanced Interdisciplinary Studies (No. 2023QYJC020).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. ANTsX Team: ANTsPy: Advanced Normalization Tools in Python (2023), https://github.com/ANTsX/ANTsPy, accessed: 2023-10-25
2. Atli, O.F., Kabas, B., Arslan, F., Yurt, M., Dalmaz, O., Çukur, T.: I2i-mamba: Multi-modal medical image synthesis via selective state space modeling. arXiv preprint arXiv:2405.14022 (2024)
3. Borges, P., Fernandez, V., Tudosiu, P.D., Nachev, P., Ourselin, S., Cardoso, M.J.: Using mr physics for domain generalisation and super-resolution. In: International Workshop on Simulation and Synthesis in Medical Imaging. pp. 177–186. Springer (2024)
4. Cackowski, S., Barbier, E.L., Dojat, M., Christen, T.: Imunity: a generalizable vae-gan solution for multicenter mr image harmonization. Medical Image Analysis **88**, 102799 (2023)
5. Cetin, I., Stephens, M., Camara, O., Ballester, M.A.G.: Attri-vae: Attribute-based interpretable representations of medical images with variational autoencoders. Computerized Medical Imaging and Graphics **104**, 102158 (2023)
6. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
7. Chen, W., Zhao, W., Chen, Z., Liu, T., Liu, L., Liu, J., Yuan, Y.: Mask-aware transformer with structure invariant loss for ct translation. Medical Image Analysis **96**, 103205 (2024)

8. Chu, Y., Yang, C., Luo, G., Qiu, Z., Gao, X.: Anatomic-constrained medical image synthesis via physiological density sampling. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 69–79. Springer (2024)
9. Dalmaz, O., Yurt, M., Çukur, T.: Resvit: residual vision transformers for multimodal medical image synthesis. IEEE Transactions on Medical Imaging **41**(10), 2598–2614 (2022)
10. Dosovitskiy, A.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
11. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752 (2023)
12. 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)
13. Ji, Z., Zou, B., Kui, X., Vera, P., Ruan, S.: Deform-mamba network for mri super-resolution. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 242–252. Springer (2024)
14. Kingma, D.P.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
15. Laptev, V.V., Gerget, O.M., Markova, N.A.: Generative models based on vae and gan for new medical data synthesis. Society 5.0: Cyberspace for advanced human-centered society pp. 217–226 (2021)
16. Li, Y., Zhou, T., He, K., Zhou, Y., Shen, D.: Multi-scale transformer network with edge-aware pre-training for cross-modality mr image synthesis. IEEE Transactions on Medical Imaging **42**(11), 3395–3407 (2023)
17. Lou, M., Ying, H., Liu, X., Zhou, H.Y., Zhang, Y., Yu, Y.: Sdr-former: A siamese dual-resolution transformer for liver lesion classification using 3d multi-phase imaging. Neural Networks p. 107228 (2025)
18. MONAI, P.: Brats mri axial slices generative diffusion model. https://github.com/Project-MONAI/model-zoo/tree/dev/models/brats_mri_axial_slices_generative_diffusion (2024), accessed: 2024-05-22
19. Park, T., Efros, A.A., Zhang, R., Zhu, J.Y.: Contrastive learning for unpaired image-to-image translation. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16. pp. 319–345. Springer (2020)
20. Phan, V.M.H., Liao, Z., Verjans, J.W., To, M.S.: Structure-preserving synthesis: Maskgan for unpaired mr-ct translation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 56–65. Springer (2023)
21. Phan, V.M.H., Xie, Y., Zhang, B., Qi, Y., Liao, Z., Perperidis, A., Phung, S.L., Verjans, J.W., To, M.S.: Structural attention: Rethinking transformer for unpaired medical image synthesis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 690–700. Springer (2024)
22. Pinaya, W.H., Graham, M.S., Kerfoot, E., Tudosiu, P.D., Dafflon, J., Fernandez, V., Sanchez, P., Wolleb, J., Da Costa, P.F., Patel, A., et al.: Generative ai for medical imaging: extending the monai framework. arXiv preprint arXiv:2307.15208 (2023)
23. Pinaya, W.H., Tudosiu, P.D., Dafflon, J., Da Costa, P.F., Fernandez, V., Nachev, P., Ourselin, S., Cardoso, M.J.: Brain imaging generation with latent diffusion models. In: MICCAI Workshop on Deep Generative Models. pp. 117–126. Springer (2022)

24. 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)
25. Tian, K., Jiang, Y., Yuan, Z., Peng, B., Wang, L.: Visual autoregressive modeling: Scalable image generation via next-scale prediction. Advances in neural information processing systems **37**, 84839–84865 (2025)
26. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. Advances in neural information processing systems **30** (2017)
27. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
28. Xu, C., Tian, S., Wang, B., Zhang, J., Polat, K., Alhudhaif, A., Li, S.: Common-unique decomposition driven diffusion model for contrast-enhanced liver mr images multi-phase interconversion. IEEE Journal of Biomedical and Health Informatics (2024)
29. Yurt, M., Dar, S.U., Erdem, A., Erdem, E., Oguz, K.K., Çukur, T.: mustgan: multi-stream generative adversarial networks for mr image synthesis. Medical image analysis **70**, 101944 (2021)
30. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
31. Zhang, X., He, X., Guo, J., Ettehadi, N., Aw, N., Semanek, D., Posner, J., Laine, A., Wang, Y.: Ptnet: A high-resolution infant mri synthesizer based on transformer. arXiv preprint arXiv:2105.13993 (2021)
32. Zhou, T., Fu, H., Chen, G., Shen, J., Shao, L.: Hi-net: hybrid-fusion network for multi-modal mr image synthesis. IEEE transactions on medical imaging **39**(9), 2772–2781 (2020)