DisDiff: Disentanglement Diffusion Network for MR Imaging Translation

Yipin Zhang[1], Ziqi Yu[2,3,4], Xiange Zhang[5], Shengjie Zhang[2,3,4], Xiang Chen[1], Haibo Yang[1], and Xiao-Yong Zhang[2,3,4]*

[1] Institute of Science and Technology for Brain-inspired Intelligence, Fudan University, Shanghai, China.
[2] Faculty of Medical Imaging Technology, College of Health Science and Technology, Shanghai Jiao Tong University School of Medicine, Shanghai, China
[3] Department of Radiology, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China
[4] Shanghai Key Laboratory of Child Brain and Development, Shanghai, China
zhangxiaoyong@sjtu.edu.cn
[5] Global Innovation Exchange, University of Washington, Bellevue, United States

Abstract. Multi-modal MR imaging plays a crucial role in clinical diagnosis and medical research. However, its widespread adoption is hindered by significant time and hardware costs. Medical image translation, which aims to synthesize missing modalities from available data, presents a promising solution. Nevertheless, existing models often struggle to maintain the structural consistency required for clinical applications. We introduced a disentanglement diffusion network –*DisDiff*, a novel disentangled adversarial diffusion framework designed to address these challenges. DisDiff incorporates a Disentangled module that decouples content and style factors within image features, thereby enabling the generation of anatomically precise images. Conditioned on disentangled representations, compared to traditional diffusion-based models, DisDiff not only accelerates the learning process, but also improves image quality and enhances training efficiency. In addition, we proposed a content discriminator module to further enforce anatomical consistency, effectively addressing the lack of explicit structural guidance in conventional diffusion models. Experimental evaluations on multi-contrast MRI translation demonstrate that DisDiff substantially outperforms existing methods in both image quality and structural preservation, positioning it as a promising solution for real-world clinical applications.

Keywords: medical image translation · diffusion model · disentangled representations · structure preservation.

1 Introduction

Multi-modal MR imaging is essential for evaluating both anatomical and functional processes in the human body, improving diagnostic accuracy and enabling advanced applications [1, 2]. However, the widespread adoption of multi-modal

protocols is hindered by significant financial and logistical challenges [3]. Medical image translation offers a promising solution by synthesizing missing modalities from available data. Besides its potential, this task remains challenging due to the complex, nonlinear variations in tissue signals across modalities [4, 5]. Recently, deep learning-based approaches have demonstrated significant potential in the field of image translation. These methods leverage data-driven priors to address challenges and enhance translation performance.

Deep learning-based image translation typically involves training models to approximate the conditional distribution of target images given source images [6, 7, 8]. Generative adversarial network (GAN) was once the dominant framework due to their remarkable ability to generate realistic images [9, 10, 11], achieving state-of-the-art results in tasks such as cross-scanner MRI synthesis [10], multi-contrast MRI synthesis [9, 12], and cross-modal synthesis [13]. However, GAN-based networks often struggle with training instability and convergence issues. Recently, diffusion models [14] have emerged as a promising alternative, offering improved stability and image quality by progressively denoising random noise. Their strong mathematical foundation provides a more interpretable framework, which has shown superior performance in medical image translation when compared to GANs [15].

Despite these advantages, diffusion models present several limitations in medical image translation. First, their lack of clinical interpretability, resulting from the process of denoising random noise, limits their applicability in clinical settings. Second, diffusion models typically model pixel values directly, which can be inefficient for medical images that require fine-grained details. Finally, the absence of direct structural constraints hinders the enforcement of anatomical consistency during generation. These challenges highlight the need for further clinical adaptations.

To address these limitations, we proposed the disentanglement diffusion network – *DisDiff* model, a novel disentangled adversarial diffusion framework designed for efficient and high-fidelity medical image synthesis in the modality translation task. To enhance clinical interpretation, we integrated the disentanglement module, which enables the generation of precise images. For improvement of the learning efficiency, we use disentangled representations as conditional inputs to guide the diffusion process, accelerating the learning process compared to pixel-level modeling. Finally, to address the lack of structural constraints, we introduced a content discriminator module to enforce anatomical consistency, compensating for the absence of explicit structural guidance.

Our contributions are fourfold: (1) improving clinical interpretation through a disentangled network that facilitates anatomically relevant image generation; (2) enhancing learning efficiency by guiding the diffusion process with disentangled representations instead of directly modeling pixel values; (3) introducing a content discriminator to ensure structural consistency and enforce anatomical constraints; and (4) Extensive experiments on multi-contrast MRI translation tasks demonstrate that DisDiff outperforms state-of-the-art(SOTA) GAN-,

diffusion-, and disentanglement-based methods in both image quality and efficiency.

2 Methods

2.1 Adversarial Diffusion Models

Diffusion models generate realistic images by progressively transforming noise into samples. They consist of two processes: the forward process, which adds noise to the input image $\mathbf{x}_0 \sim q(\mathbf{x}_0)$ and the reverse process, which denoises the noisy samples to generate a clear image:

$$\mathbf{x}_t = (1-\beta_t)\mathbf{x}_{t-1} + \beta_t \boldsymbol{\epsilon}, \quad q(\mathbf{x}_{t-1}|\mathbf{x}_t) \sim \mathcal{N}(\mathbf{x}_{t-1}; \mu(\mathbf{x}_t, t), \Sigma(\mathbf{x}_t, t)), \tag{1}$$

where β_t is the noise variance, $\mu(\mathbf{x}_t, t)$ and $\Sigma(\mathbf{x}_t, t)$ denote the mean and covariance of the reverse process, respectively.

In contrast, Muzaffer ¨Ozbey et al. [15] introduces a novel adversarial diffusion model, which is designed to improve the efficiency of the diffusion process with a larger step size $k \gg 1$. In this variant, the forward diffusion process is described by:

$$\mathbf{x}_t = (1-\gamma_t)\mathbf{x}_{t-k} + \sqrt{\gamma_t}\boldsymbol{\epsilon}, \quad q(\mathbf{x}_t|\mathbf{x}_{t-k}) = \mathcal{N}(\mathbf{x}_t; (1-\gamma_t)\mathbf{x}_{t-k}, \gamma_t I), \tag{2}$$

where γ_t is the noise variance at timestep t controlled by the following schedule:

$$\gamma_t = 1 - e^{\beta_{\min}\left(\frac{t_k}{T} - \frac{\beta_{\max}-\beta_{\min}}{2T^2}\right)}. \tag{3}$$

Given the breakdown of the normality assumption for a large k, an adversarial framework is introduced. A generator $G_\theta(\mathbf{x}_t, y, t)$ estimates the denoised image $\hat{\mathbf{x}}_{t-k}$ while a discriminator $D_\theta(\hat{\mathbf{x}}_{t-k}, \mathbf{x}_t, t)$ distinguishes true and generated samples. The loss functions for G_{theta} and D_θ are presented by:

$$L_{G_\theta} = \mathbb{E}_{q(\mathbf{x}_t|\mathbf{x}_0,y),p_\theta(\mathbf{x}_{t-k}|\mathbf{x}_t,y)}\left[-\log(D_\theta(\hat{\mathbf{x}}_{t-k}))\right], \tag{4}$$

$$\begin{aligned} L_{D_\theta} = {} & \mathbb{E}_{q(\mathbf{x}_t|\mathbf{x}_0,y),q(\mathbf{x}_{t-k}|\mathbf{x}_t,y)}\left[-\log(D_\theta(\mathbf{x}_{t-k}))\right] + \mathbb{E}_{p_\theta(\mathbf{x}_{t-k}|\mathbf{x}_t,y)}\left[-\log(1-D_\theta(\hat{\mathbf{x}}_{t-k}))\right] \\ & + \eta\mathbb{E}_{q(\mathbf{x}_{t-k}|\mathbf{x}_t,y)}\left[\|\nabla_{\mathbf{x}_{t-k}} D_\theta(\mathbf{x}_{t-k})\|^2\right], \end{aligned} \tag{5}$$

where η is a regularization term controlling the gradient penalty. This adversarial setup enables the effective learning of the reverse diffusion process, even with a large step size k.

2.2 Disentanglement Networks

Achieving disentanglement in multi-modal medical image tasks requires appropriate inductive biases, as without these, disentanglement is infeasible [6]. In translation tasks, we hypothesize that anatomical structures, consistent across modalities, should be captured in modality-independent content factors, while modality-specific variations should be encoded as attributes. This assumption enables more accurate synthesis by separating structural and appearance-related features, aligning better with the nature of medical data [16].

The disentanglement framework consists of content encoders $\{E_c^{m_i}, E_c^{m_j}\}$, attribute encoders $\{E_a^{m_i}, E_a^{m_j}\}$, generators $\{G^{m_i}, G^{m_j}\}$ for reconstructions, discriminators D_c for content consistency and $\{D^{m_i}, D^{m_j}\}$ for modality-specific characteristics.

For an input image m_i, the content encoder and attribute encoder map the image to separate content and attribute features: $z_c^{m_i} = E_c^{m_i}(m_i)$, $z_a^{m_i} = E_a^{m_i}(m_i)$. Cross-modal synthesis $m_{i\to j}$ is achieved by swapping the attribute features between modalities while preserving the original content feature. After generating $m_{i\to j}$, the content feature is re-extracted and combined with the original attribute representation to reconstruct the input image, thereby enforcing cycle consistency:

$$m_{i\to j} = G^{m_j}(z_c^{m_i}, z_a^{m_j}), \quad m_{i\to j\to i} = G^{m_i}(E_c^{m_i}(m_{i\to j}), z_a^{m_i}). \tag{6}$$

Especially, for inputs pairs m_i and m_j, the discriminator D_c encourages that the content feature z_c remains consistent. The objective function for D_c is:

$$L_c = \mathbb{E}_{m_i\sim p(m_i)}\left[\log D_c(z_c(m_i))\right] + \mathbb{E}_{m_j\sim p(m_j)}\left[\log\left(1 - D_c(z_c(m_j))\right)\right]. \tag{7}$$

This consistency promotes the alignment of anatomical structures across different modalities, enhancing the reconstruction performance.

2.3 DisDiff

The proposed disDiff model comprises a disentanglement learning module and an adversarial diffusion module, enabling unsupervised image translation while preserving anatomical structures and modality-specific features.

The disentanglement learning module separates anatomical structures from modality-specific features, enabling cross-modal synthesis. Given image pairs (m_i and m_j) from modalities i and j, respectively, the module extracts the anatomical/content features and modality-specific/attribute features for both modalities:

$$z_c^{m_i}, z_a^{m_i} = E^{m_i}(m_i), \quad z_c^{m_j}, z_a^{m_j} = E^{m_j}(m_j). \tag{8}$$

where E^{m_i}/G^{m_i} and E^{m_j}/G^{m_j} denote encoders/generators for domains i and j respectively. Then, cross-modal synthesis is achieved by applying the disentanglement decoder with content features exchanged:

$$m_{i\to j} = G^{m_j}(z_c^{m_i}, z_a^{m_j}), \quad m_{j\to i} = G^{m_i}(z_c^{m_j}, z_a^{m_i}). \tag{9}$$

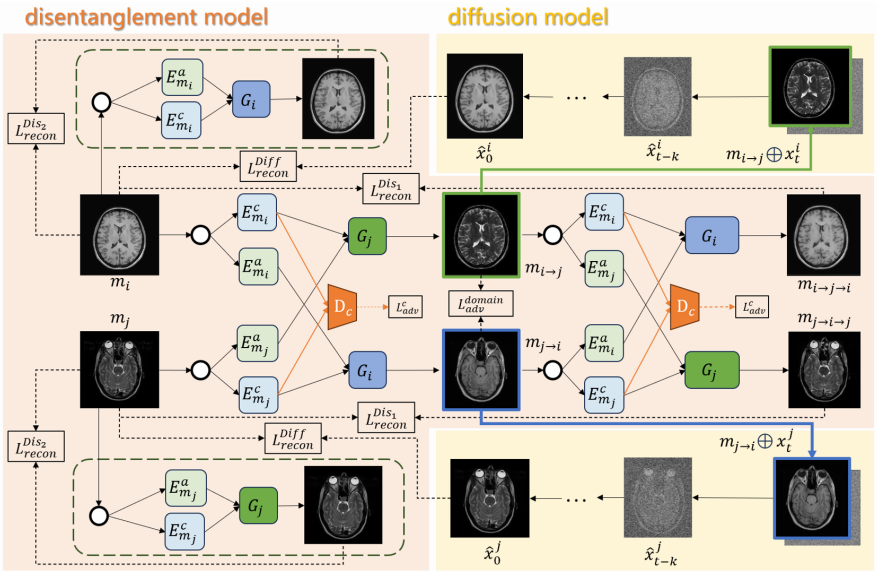

Fig. 1. The overview of DisDiff framework. Given input image pairs (m_i, m_j), the disentanglement modules first generate cross-domain images $m_{i\to j}$ and $m_{j\to i}$. These intermediate representations are then fed into the diffusion modules, in conjunction with noise images sampled from $\mathcal{N}(0,1)$, progressively denoise and reconstruct the original image $\hat{m}_i$ and $\hat{m}_j$. Disentanglement modules also generate reconstructions $m_{i\to j\to i}$ and $m_{j\to i\to j}$ for the cycle-consistency mechanism. The solid line represents the process flow of the image processing, while the dashed line represents the objects that constitute the loss.

The diffusion module reconstructs target modality images from content features provided by the disentanglement module. Starting with noise image pairs $(x_i^T, x_j^T) \sim N(0, I)$, the reverse diffusion process iteratively refines the noisy inputs. At each time step t, the generator $G_{\theta_i}, G_{\theta_j}$ produces denoised estimates $\tilde{x}_i^0, \tilde{x}_j^0$ of the target modality images:

$$\tilde{x}_i^0 = G_{\theta_i}(x_i^t, z_c^{m_i}, t), \quad \tilde{x}_j^0 = G_{\theta_j}(x_j^t, z_c^{m_j}, t). \tag{10}$$

Refinements are then applied using the denoising distributions:

$$\hat{x}_i^{t-k} \sim q(x_i^{t-k}|x_i^t, \tilde{x}_i^0), \quad \hat{x}_j^{t-k} \sim q(x_j^{t-k}|x_j^t, \tilde{x}_j^0). \tag{11}$$

This iterative process continues until the target images $\hat{x}_i^0, \hat{x}_j^0$ are synthesized at time step 0.

Consistency between the true target images and their reconstructed images is enforced through a cycle-consistency loss:

$$L_{\text{cyc}} = \mathbb{E}_{m_i, m_j}\left[\|m_i - \hat{x}_i^0\|_1 + \|m_j - \hat{x}_j^0\|_1\right], \tag{12}$$

where $\hat{x}_i^0$ and $\hat{x}_j^0$ are generated target images from the diffusion module.

During the inference stage, the diffusion module generates the images in the target modality by leveraging the anatomical encoding $z_c^{m_i}$ obtained from the disentanglement module, and modality-specific features $z_a^{m_j}$ from the target modality.

3 Experiments

3.1 Datasets

IXI Dataset In this study, we use T1 and T2 brain MRI scans from 80 healthy subjects in the IXI dataset, with 60 subjects for training (90%/10% for training and validation) and 20 for testing. We select 100 axial cross-sections containing brain tissues. Since the IXI MR images are unregistered, we apply deep learning-based tools [17, 18] for subject-level registration after skull stripping. Then, we could obtain reference images to evaluate I2I translation performance.

BraTS Dataset The 2018 Multimodal Brain Tumor Segmentation Challenge (BraTS) dataset [19] contains 285 annotated MRI scans from glioma patients with all modalities aligned. The MRI volumes have been skull-stripped and re-sampled to a resolution of $1 \times 1 \times 1mm^3$. For I2I translation, training/testing ratio is set as 7 (also 90%/10% for training and validation) /3 as 100 axial cross-sections containing brain tissue from each subject are used in our study.

3.2 Implementation Details

All models were implemented in PyTorch and trained with the Adam optimizer ($\beta_1 = 0.5$, $\beta_2 = 0.9$) and an initial learning rate of 10^{-4}, reduced by 0.1 every 10 epochs if validation loss did not improve for two consecutive epochs. Training was performed on a workstation with two Nvidia RTX A6000 GPUs. Performance was evaluated every 5 epochs on a validation set. Test set performance was assessed using PSNR and SSIM metrics, which measure pixel-level accuracy and perceptual similarity, respectively.

4 Results and Discussion

4.1 Comparison on SOTA methods

We compared the performance of our proposed model, DisDiff, with several SOTA image translation methods, including DRIT++ [16], CycleGAN [20], AttentionGAN [21], StarGAN-v2 [22], MGUIT [23], InstaFormer [24], Hifi-Syn [25], UNIT-DDPM [26] and SynDiff [15]. Specifically, we focused on one-to-one modality translation task (e.g., T1w-T2w translation task) to ensure a fair comparison. The quantitative results for the T1w-T2w modality translation are summarized in Table 1, with exemplary synthetic images shown in Fig. 1.

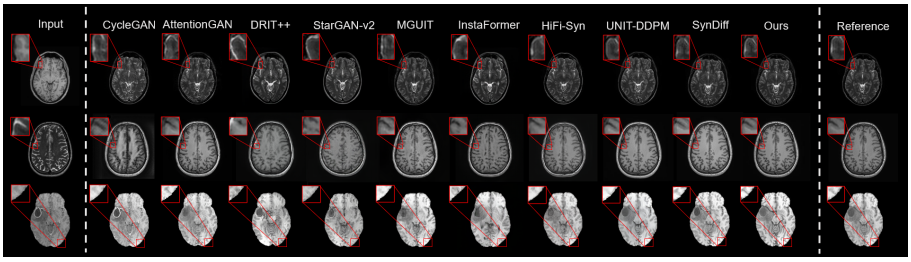

Fig. 2. Comparison of I2I translation methods on IXI and BraTS datasets. Compared to other SOTA methods, DisDiff shows superior structure-preserving translation results.

DisDiff consistently outperforms other methods across all image synthesis metrics. For T1w to T2w translation, it achieves the highest PSNR and SSIM scores in both the IXI (25.71 and 0.8445) and BraTS (29.49 and 0.9256) datasets, indicating superior structural similarity and perceptual quality. Visual comparisons (Fig. 2) show that DisDiff generates sharper, more realistic textures with better anatomical alignment. In the inverse translation task (T2w to T1w), DisDiff also leads with average SSIM and PSNR scores of 0.8661 and 27.05 for IXI, and 0.8648 and 26.31 for BraTS. In contrast, methods like CycleGAN and StarGAN-v2 exhibit noticeable errors in contrast and anatomical structures, especially in regions such as the cerebellar cortex. The outstanding performance of DisDiff is attributed to its disentangled learning framework, which separates content and attribute features, enabling more accurate modality translation. Regarding efficiency, we compared our method, DisDiff, with SynDiff. Both methods employ diffusion-based inference, therefore, we focus on training efficiency: under identical conditions with a batch size of 4, DisDiff achieves convergence in approximately 1.2 days, whereas SynDiff requires around 2.3 days to reach comparable performance.

Table 1. Performance for multi-contrast MRI translation tasks in IXI and BraTS dataset. PSNR (dB) and SSIM (%) are listed as mean±std across the test set. **Boldface** marks the top-performing model in each task.

Method	IXI				BraTS			
	T1 → T2		T2 → T1		T1 → T2		T2 → T1	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
CycleGAN	19.71±1.23	77.39±1.65	20.33±1.11	78.62±1.74	21.80±1.29	74.04±1.82	22.28±1.34	77.18±1.71
AttentionGAN	21.89±1.35	80.27±1.68	21.45±1.26	81.03±1.65	22.33±1.36	79.07±1.85	23.41±1.44	81.03±1.72
DRIT++	22.32±1.31	80.50±1.67	22.24±1.30	78.86±1.70	22.49±1.39	81.15±1.83	23.34±1.47	82.61±1.76
StarGAN-v2	22.16±1.30	79.75±1.73	22.55±1.36	80.53±1.68	22.94±1.40	81.82±1.86	23.77±1.51	82.55±1.80
MGUIT	24.37±1.40	79.79±1.71	23.27±1.45	83.21±1.67	23.28±1.41	82.97±1.89	24.55±1.53	83.51±1.84
InstaFormer	22.40±1.32	81.08±1.75	23.82±1.48	81.84±1.70	23.16±1.42	82.90±1.90	25.05±1.55	83.72±1.86
HiFi-Syn	23.84±1.39	82.40±1.69	24.18±1.50	83.71±1.66	24.11±1.45	83.31±1.88	24.98±1.56	84.59±1.85
UNIT-DDPM	22.44±1.26	81.64±3.06	24.01±0.72	86.59±2.16	23.71±1.50	88.75±2.49	19.84±1.54	85.92±2.28
SynDiff	24.92±4.90	83.89±1.68	26.07±4.35	85.47±1.26	27.97±2.23	90.43±1.46	24.33±3.80	84.96±1.42
DisDiff	**25.71±4.20**	**84.45±1.33**	**27.05±4.50**	**86.61±1.70**	**29.49±2.55**	**92.56±1.89**	**26.31±3.62**	**86.48±1.28**

To demonstrate how the disentanglement network improves the preservation of content features, we visualize the content features of T1w and T2w images extracted from the disentanglement Network (Fig. 3). As shown in Fig. 3, our disentanglement network effectively preserves the similar content features, regardless of the modality, which contributes to the high SSIM values significantly in our subsequent observed results.

Table 2. Ablation results of three different modules in DisDiff. **Boldface** marks the top-performing model in each task.

Task	Disentangled	Diffusion	D_c	PSNR	SSIM
T1w → T2w	✓	×	×	22.32±1.31	80.50±1.67
	×	✓	×	22.44±1.26	81.64±3.06
	✓	✓	×	23.52±3.56	79.22±2.28
	✓	✓	✓	**25.71±4.20**	**84.45±1.33**
T2w → T1w	✓	×	×	22.24±1.30	78.86±1.70
	×	✓	×	24.01±0.72	86.59±2.16
	✓	✓	×	24.31±2.89	81.12±1.74
	✓	✓	✓	**27.05±4.50**	**86.61±1.70**

4.2 Ablation Study

To evaluate the contributions of the three key components in our model – the disentanglement module, diffusion mechanism, and content discriminator, we conducted an ablation study. We could observe from Table 2 that the model incorporating all three components achieves the best performance. Specifically, the disentanglement network enhances the separation of content and style features, while the diffusion mechanism improves the smoothness and consistency of generated features. The content discriminator further refines the content preservation

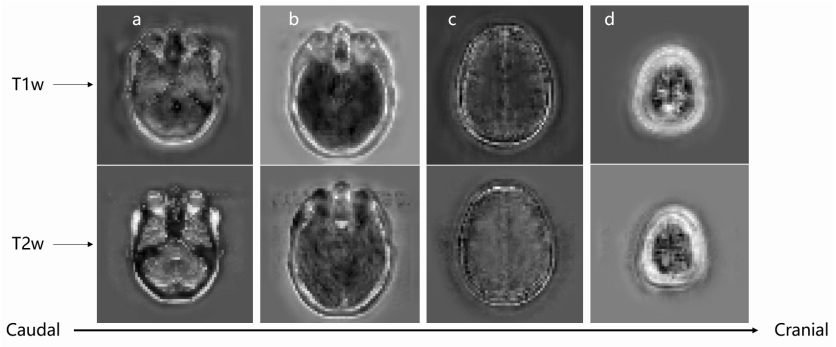

Fig. 3. Similar content-specific features are extracted from the same subject's corresponding T1w and T2w images by the disentanglement module.

by ensuring the consistency of content features throughout the transformation process. These findings underscore the importance of each component in improving the overall performance, with the combination of all three components leading to the best results.

5 Conclusion

We propose DisDiff, a novel disentangled diffusion model for MR images translation. DisDiff combines a disentanglement module, which preserves anatomical structures, and a rapid diffusion process for efficient generation of high-quality images. Trained in an unsupervised, cycle-consistent manner, DisDiff stands out in translating unpaired data while maintaining fidelity across modalities. Experimental results show that DisDiff outperforms existing methods in both image quality and translation performance, demonstrating its potential for high-fidelity medical image translation in clinical applications.

Acknowledgement. This work was supported by grants from the National Natural Science Foundation of China (82441016, 82471940), Shanghai Key Laboratory of Child Brain and Development (24dz2260100), and Natural Science Foundation of Shanghai (24TS1415000).

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

References

[1] Juan Eugenio Iglesias et al. "Is synthesizing MRI contrast useful for inter-modality analysis?" In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013: 16th International Conference, Nagoya, Japan, September 22-26, 2013, Proceedings, Part I 16*. Springer. 2013, pp. 631–638.

[2] Junghoon Lee et al. "Multi-atlas-based CT synthesis from conventional MRI with patch-based refinement for MRI-based radiotherapy planning". In: *Medical Imaging 2017: Image Processing*. Vol. 10133. SPIE. 2017, pp. 434–439.

[3] Thomas Joyce, Agisilaos Chartsias, and Sotirios A Tsaftaris. "Robust multi-modal MR image synthesis". In: *Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part III 20*. Springer. 2017, pp. 347–355.

[4] Snehashis Roy et al. "Atlas based intensity transformation of brain MR images". In: *Multimodal Brain Image Analysis: Third International Workshop, MBIA 2013, Held in Conjunction with MICCAI 2013, Nagoya, Japan, September 22, 2013, Proceedings 3*. Springer. 2013, pp. 51–62.

[5] Yawen Huang, Ling Shao, and Alejandro F Frangi. "Simultaneous super-resolution and cross-modality synthesis of 3D medical images using weakly-supervised joint convolutional sparse coding". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 6070–6079.

[6] Christopher Bowles et al. "Pseudo-healthy image synthesis for white matter lesion segmentation". In: *Simulation and Synthesis in Medical Imaging: First International Workshop, SASHIMI 2016, Held in Conjunction with MICCAI 2016, Athens, Greece, October 21, 2016, Proceedings 1*. Springer. 2016, pp. 87–96.

[7] Agisilaos Chartsias et al. "Multimodal MR synthesis via modality-invariant latent representation". In: *IEEE Transactions on Medical Imaging* 37.3 (2017), pp. 803–814.

[8] Wen Wei et al. "Fluid-attenuated inversion recovery MRI synthesis from multisequence MRI using three-dimensional fully convolutional networks for multiple sclerosis". In: *Journal of Medical Imaging* 6.1 (2019), pp. 014005–014005.

[9] Salman UH Dar et al. "Image synthesis in multi-contrast MRI with conditional generative adversarial networks". In: *IEEE Transactions on Medical Imaging* 38.10 (2019), pp. 2375–2388.

[10] Dong Nie et al. "Medical image synthesis with deep convolutional adversarial networks". In: *IEEE Transactions on Biomedical Engineering* 65.12 (2018), pp. 2720–2730.

[11] Karim Armanious et al. "MedGAN: Medical image translation using GANs". In: *Computerized Medical Imaging and Graphics* 79 (2020), p. 101684.

[12] Mahmut Yurt et al. "mustGAN: multi-stream generative adversarial networks for MR image synthesis". In: *Medical Image Analysis* 70 (2021), p. 101944.

[13] Onat Dalmaz, Mahmut Yurt, and Tolga Çukur. "ResViT: residual vision transformers for multimodal medical image synthesis". In: *IEEE Transactions on Medical Imaging* 41.10 (2022), pp. 2598–2614.

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 6840–6851.

[15] Muzaffer Özbey et al. "Unsupervised medical image translation with adversarial diffusion models". In: *IEEE Transactions on Medical Imaging* (2023).

[16] Hsin-Ying Lee et al. "Diverse image-to-image translation via disentangled representations". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 35–51.

[17] Ziqi Yu et al. "A generalizable brain extraction net (BEN) for multimodal MRI data from rodents, nonhuman primates, and humans". In: *Elife* 11 (2022), e81217.

[18] Brian B Avants et al. "A reproducible evaluation of ANTs similarity metric performance in brain image registration". In: *Neuroimage* 54.3 (2011), pp. 2033–2044.

[19] Bjoern H Menze et al. "The multimodal brain tumor image segmentation benchmark (BRATS)". In: *IEEE Transactions on Medical Imaging* 34.10 (2014), pp. 1993–2024.

[20] Jun-Yan Zhu et al. "Unpaired image-to-image translation using cycle-consistent adversarial networks". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 2223–2232.

[21] Hao Tang et al. "Attention-guided generative adversarial networks for unsupervised image-to-image translation". In: *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE. 2019, pp. 1–8.

[22] Yunjey Choi et al. "Stargan v2: Diverse image synthesis for multiple domains". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 8188–8197.

[23] Xun Huang et al. "Multimodal unsupervised image-to-image translation". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 172–189.

[24] Soohyun Kim et al. "InstaFormer: Instance-aware image-to-image translation with transformer". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 18321–18331.

[25] Ziqi Yu et al. "HiFi-Syn: Hierarchical granularity discrimination for high-fidelity synthesis of MR images with structure preservation". In: *Medical Image Analysis* 100 (2025), p. 103390.

[26] Hiroshi Sasaki, Chris G Willcocks, and Toby P Breckon. "Unit-ddpm: Unpaired image translation with denoising diffusion probabilistic models". In: *ArXiv Preprint arXiv:2104.05358* (2021).