D2Diff: A Dual-Domain Diffusion Model for Accurate Multi-Contrast MRI Synthesis

Sanuwani Dayarathna[1], Himashi Peiris[1], Kh Tohidul Islam[2], Tien-Tsin Wong[1], and Zhaolin Chen[1,2]

[1] Department of Data Science and AI, Faculty of IT, Monash University, Clayton, VIC, Australia.
[2] Monash Biomedical Imaging, Monash University, Clayton, VIC, Australia.
{Sanuwani.Hewamunasinghe, Himashi.Peiris, KhTohidul.Islam, TT.Wong, Zhaolin.Chen}@monash.edu

Abstract. Multi-contrast MRI synthesis is inherently challenging due to the complex and nonlinear relationships among different contrasts. Each MRI contrast highlights unique tissue properties, but their complementary information is difficult to exploit due to variations in intensity distributions and contrast-specific textures. Existing methods for multi-contrast MRI synthesis primarily utilize spatial domain features, which capture localized anatomical structures but struggle to model global intensity variations and distributed patterns. Conversely, frequency-domain features provide structured inter-contrast correlations but lack spatial precision, limiting their ability to retain finer details. To address this, we propose a dual-domain learning framework that integrates spatial and frequency domain information across multiple MRI contrasts for enhanced synthesis. Our method employs two mutually trained denoising networks, one conditioned on spatial domain and the other on frequency domain contrast features through a shared critic network. Additionally, an uncertainty-driven mask loss directs the model's focus toward more critical regions, further improving synthesis accuracy. Extensive experiments show that our method outperforms state-of-the-art (SOTA) baselines, and the downstream segmentation performance highlights the diagnostic value of the synthetic results. Code and model hyperparameters are available at https://github.com/sanuwanihewa/D2Diff

Keywords: MRI Synthesis · Dual-domain · Diffusion models.

1 Introduction

Magnetic Resonance Imaging (MRI) offers detailed anatomical and pathological insights through images of multiple contrasts [3]. However, acquiring multiple MRI contrasts poses significant challenges, including high imaging cost, prolonged scanning times, and potential safety concerns related to gadolinium-based contrast agents [5,11]. Medical image synthesis provides a powerful approach to address these challenges by reconstructing missing or corrupted image

contrasts from available contrasts [18]. However, synthesizing high-fidelity multi-contrast images remains challenging due to the complex, nonlinear, and often obscured relationships among contrasts, driven by intensity inconsistencies and modality-specific textures [5]. Therefore, capturing and aligning these intricate cross-contrast relationships is critical for an accurate image synthesis model.

Most of the existing methods [5,20,19] focus on fusing features from multiple contrasts, leveraging latent-level operations or hierarchical representations to model cross-contrast dependencies. Despite these advances, most approaches [15,20] rely heavily on rigid spatial domain representations and fusion strategies, which struggle to fully capture the complementary and distributed relationships across contrasts. Although spatial domain features excel at encoding localized structures and anatomical integrity, they often struggle to disentangle discerning intensity variations and overlapping distributions [12,6], particularly in scenarios with significant heterogeneity across contrasts such as brain lesions[5].

To address these limitations, we propose a dual-domain learning framework, D2Diff, for multi-contrast MRI synthesis, which employs two denoising networks that are mutually trained together. The first network is guided by frequency-domain representations [12,6] and captures structured inter-contrast correlations such as global intensity shifts and distributed intensity variations. Simultaneously, the second network is guided by spatial-domain features and ensures high-resolution, pixel-level detail fidelity. These networks are trained collaboratively through a shared critic network, which ensures adversarial consistency. Using a novel uncertainty-aware mask loss, the shared critic facilitates uncertainty estimation, guiding the synthesis process to focus on critical regions. By leveraging the complementary strengths of spatial and frequency domains, our framework effectively aligns complex cross-contrast correlations, providing a robust and accurate multi-contrast MRI synthesis. In summary, **our main contributions are**: **(1)** A dual-domain diffusion framework, simultaneously guided by multi-contrast MRI features in both frequency and spatial domains, and jointly trained using a shared critic network. **(2)** A multi-scale frequency feature integration module for adaptive inter-contrast feature combination to preserve subtle contrast-specific details. **(3)** A novel uncertainty-aware mask loss to enhance uncertainty-driven learning. **(4)** Comprehensive experiments confirm superior synthesis quality and further validation through downstream segmentation tasks.

2 Method

Problem Formulation. Let $\mathcal{X} = (\boldsymbol{X}_k, \boldsymbol{Y}_k)_{k=1}^{m}$, be a set of m co-registered MRI contrast image pairs, where $\boldsymbol{x_k}$ denotes the target contrast to be synthesized, and $\boldsymbol{y_k} = \{\boldsymbol{y_{k,i}}\}_{i=1}^{n}$, represents n source contrasts used as conditional inputs to generate the target contrast. We denote the denoising networks as $\mathcal{H}_j; j \in \{1, 2\}$ where encoder-decoder $\mathcal{F}_j$, and dual-domain feature extraction ϕ_j are their functional decompositions as,

$$f_j = \phi_j(\theta_j; \boldsymbol{Y}_k); \quad \mathcal{H}_j(\Theta^j; \mathcal{X}) = \mathcal{F}_j(\phi_j(\theta_j; \boldsymbol{Y}_k), \boldsymbol{X}_k) \quad j = 1 \text{ or } 2. \tag{1}$$

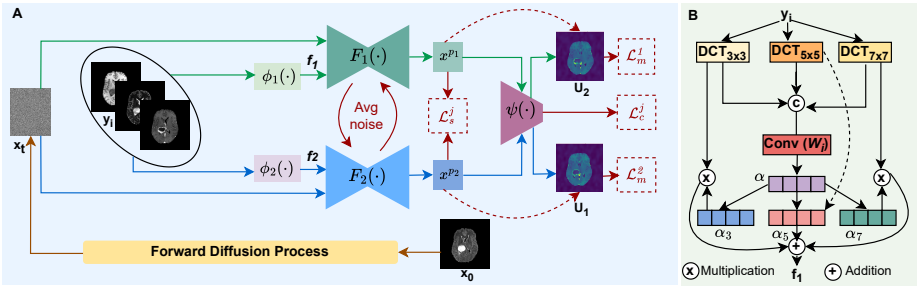

Fig. 1: Network architecture. **A**: Overall Architecture with frequency (ϕ_1) and spatial guidance (ϕ_2), **B**: Multi-scale adaptive frequency aggregation in ϕ_1.

Dual Domain Diffusion Model. Fig. 1(A) provides an overview of the D2Diff pipeline, which employs two denoising networks that collaboratively learn using frequency and spatial domain features from multi-contrast MRI. Diffusion models consist of two main processes: forward and reverse process [8]. In the forward process, random Gaussian noise is progressively added to the target MRI contrast ($\boldsymbol{x}_0$) to be synthesized as,

$$q(\boldsymbol{x}_\mathrm{t}|\boldsymbol{x}_{\mathrm{t}-1}) = \mathcal{N}(\boldsymbol{x}_\mathrm{t}; \sqrt{1-\beta_\mathrm{t}}\,\boldsymbol{x}_{\mathrm{t}-1}, \beta_\mathrm{t}\mathrm{I}), \tag{2}$$

where β_t is the noise variance schedule that is used to add noise to the data, $\mathcal{N}$ is the Gaussian distribution, and I is the identity covariance matrix. Utilizing the Markov property of the diffusion process, the marginal distribution of $\boldsymbol{x}_\mathrm{t}$ can be directly obtained as follows,

$$q(\boldsymbol{x}_\mathrm{t}|\boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_\mathrm{t}; \sqrt{\bar{\alpha}_\mathrm{t}}\,\boldsymbol{x}_0, (1-\bar{\alpha}_\mathrm{t})\mathrm{I}), \tag{3}$$

where $\alpha_\mathrm{t} := 1 - \beta_\mathrm{t}$ and $\bar{\alpha}_\mathrm{t} := \prod_{s=1}^{\mathrm{t}} \alpha_s$. The reverse diffusion process estimates the posterior distribution $p_\theta(x_{t-1}|x_t, \boldsymbol{y}_i)$ to generate a realistic x_0 guided by conditional contrasts $\boldsymbol{y}_i$,

$$p_\theta(\boldsymbol{x}_{\mathrm{t}-1}|\boldsymbol{x}_\mathrm{t}, \boldsymbol{y}_i) = \mathcal{N}(\boldsymbol{x}_{\mathrm{t}-1}; \mu_\theta(\boldsymbol{x}_\mathrm{t}, \mathrm{t}), \sigma_\mathrm{t}^2\mathrm{I}), \tag{4}$$

where $\mu_\theta(\boldsymbol{x}_\mathrm{t}, \mathrm{t})$ is the mean and σ_t^2 is the variance of the denoising network parameterized by θ. The noisy target contrast serves as input for both frequency- and spatial-guided synthesis models. Each denoising network then independently performs the reverse diffusion process on the perturbed data, leveraging multi-contrast features in their respective domains to approximate the posterior distribution parameterized [17] as follows,

$$p_\theta(\boldsymbol{x}_{\mathrm{t}-1}|\boldsymbol{x}_\mathrm{t}, \boldsymbol{y}_i) := q(\boldsymbol{x}_{\mathrm{t}-1}|\boldsymbol{x}_\mathrm{t}, \tilde{\boldsymbol{x}}_0 = H_j(\boldsymbol{x}_\mathrm{t}, \boldsymbol{y_i}, z, \mathrm{t})). \tag{5}$$

Frequency domain learning. Multi-contrast MRI exhibits significant variations in intensity and resolution across contrasts. The frequency domain structures spatial patterns into different frequency components, enabling better separation of global and local intensity variations[12]. To leverage this, we apply

Discrete Cosine Transform (DCT) [1] to convert spatial variations into frequency representations, aligning non-linear intensity differences. DCT employs real-valued cosine functions rather than complex exponentials, allowing efficient decomposition of MRI images into frequency components [16,21]. By transforming multi-contrast images into frequency coefficients, DCT effectively distributes intensity variability across distinct bands, enhancing feature consistency for synthesis. To achieve this, we used DCT with three different kernel sizes (k) in $\phi_1(\cdot)$, allowing the extraction of multi-scale frequency features as follows.

$$h_{k,\boldsymbol{y}_i} = \mathrm{DCT}_{k\times k}(\boldsymbol{y}_i), \quad \text{for } k \in \{3,5,7\} \tag{6}$$

where h_{k,y_i} is the the extracted frequency contrasts of each y_i.

To optimize frequency feature weighting, we introduce a novel adaptive feature aggregation using a lightweight attention mechanism that assigns importance scores via a learnable attention module and a convolutional layer. This refines and projects the combined representation into a common space. The adaptive frequency fusion (Fig. 1B) selectively emphasizes relevant frequency-specific information across MRI contrasts as follows,

$$f_1 = \Big[W_i \cdot \sum_{k\in\{3,5,7\}} \langle \alpha_k, h_{k,\boldsymbol{y}_i} \rangle \Big]_{i\in\{1,n\}} \tag{7}$$

where W_i is a learnable convolutional transformation layer and α_k represents the adaptive feature combination process as shown in Fig. 1 (B) where a softmax-based attention (α) mechanism is used to assign dynamic weights (α_3, α_5, α_7) to determine how much influence each contrast's frequency features should have in the final representation. The weighted sum of these features forms the fused representation f_1, to guide the first denoising network.

Spatial domain learning. To preserve fine anatomical details, the second denoising network is guided by spatial features from multi-contrast inputs in $\phi_2(\cdot)$. This enhances structural correlations, capturing finer details like edges and tissue boundaries to aid the denoising process as follows,

$$f_2 = [R_i(\boldsymbol{y}_i)]_{i\in\{1,n\}} \tag{8}$$

where R_i consists of separate residual blocks for each input contrast, which consist of a convolutional layer followed by a Group Normalization, ReLU activation.

Both denoising networks $\mathcal{F}_1$, $\mathcal{F}_2$ employs U-Net-based architecture as in [17] while sinusoidal positional embeddings [8] encode the timestep t with z serving as the latent vector for conditioning.

$$\mathcal{H}_j^{\Theta}(X) = \mathcal{F}_j(f_j, \mathrm{t}, z), \ j \in \{1,2\} \tag{9}$$

Alongside the denoising generators, we employ a shared time-dependent critic network ψ [17] to ensure collaborative training across them. ψ distinguish between $\boldsymbol{x}_{\mathrm{t}-1}$ and $\boldsymbol{x}_{\mathrm{t}}$ by assessing if $\boldsymbol{x}_{\mathrm{t}-1}$ is a plausible denoised version of $\boldsymbol{x}_{\mathrm{t}}$ using the critic loss $\mathcal{L}_c^j$.

$$\mathcal{L}_c^j\left(\theta_{\mathcal{H}}^j; \mathcal{X}\right) = \mathbb{E}_{q(\boldsymbol{x}_{\mathrm{t}}|\boldsymbol{x},\boldsymbol{y}_i),p_\theta(\boldsymbol{x}_{\mathrm{t}-1}|\boldsymbol{x}_{\mathrm{t}},\boldsymbol{y}_i)}[-log(\psi^\theta(\boldsymbol{x}_{\mathrm{t}-1}^{p_j}, \boldsymbol{x}_{\mathrm{t}}, \mathrm{t}))] \tag{10}$$

As each denoising network is trained on different feature domains of the same input contrasts, they can leverage the shared critic network to learn from each other, maintaining consistency in their predictions. To train the critic network against the actual ground truths, the predicted outputs ($\boldsymbol{x}^{p_1}$ and $\boldsymbol{x}^{p_2}$) from each denoising network are used as follows,

$$\begin{aligned}\mathcal{L}^j_{adv}(\theta_c;\mathcal{X}) = \mathbb{E}_{q(\boldsymbol{x}_\mathrm{t}|\boldsymbol{x},\boldsymbol{y}_i)}[\mathbb{E}_{q(\boldsymbol{x}_{\mathrm{t}-1}|\boldsymbol{x}_\mathrm{t},\boldsymbol{y}_i)}\eta[log(\psi^\theta(\boldsymbol{x}_{\mathrm{t}-1},\boldsymbol{x}_\mathrm{t},\mathrm{t}))] \\ +(1-\eta)\mathbb{E}_{p_\theta(\boldsymbol{x}_{\mathrm{t}-1}|\boldsymbol{x}_\mathrm{t},\boldsymbol{y}_i)}[log(1-\psi^\theta(\boldsymbol{x}^{p_j}_{\mathrm{t}-1},\boldsymbol{x}_\mathrm{t},\mathrm{t}))]],\end{aligned} \tag{11}$$

where $\eta = 0$ if $\boldsymbol{x}_{\mathrm{t}-1}$ is predicted by a denoising network, and $\eta = 1$ if $\boldsymbol{x}_{\mathrm{t}-1}$ is sampled from the actual target contrast distribution.

Uncertainty aware mask loss. We propose an uncertainty-aware mask loss that guides denoising networks to focus on high-uncertainty regions during the synthesis. This is achieved using spatial attention maps from the critic network, which identify reliable features from the target distribution. Specifically, we considered middle-layer features (f_m) from the critic network, which is sensitive to discriminative regions extracted via a sigmoid(σ) layer and interpolated (I) to match the output contrast dimension (dim) as,

$$U_j = I[\sigma[\psi^\theta(\boldsymbol{x}^{p_j}_{\mathrm{t}-1},\boldsymbol{x}_\mathrm{t},\mathrm{t})_{f_m}], dim(\boldsymbol{x})] \tag{12}$$

To enhance mutual learning across networks, each denoising network leverages the attention maps of other's output contrast from the shared critic to align individual predictions. Then, the Binary cross-entropy logistic criteria (BCE) is employed to quantify discrepancies, encouraging consistent probability estimations and refining focus on critical regions using mask loss $\mathcal{L}^j_m$ as,

$$\mathcal{L}^j_m(\theta^j_{\mathcal{H}};\mathcal{X}) = \langle U_2, BCE(\boldsymbol{x}^{p_1},\sigma(\boldsymbol{x}^{p_2}))\rangle + \langle U_1, BCE(\boldsymbol{x}^{p_2},\sigma(\boldsymbol{x}^{p_1}))\rangle \tag{13}$$

We also employed supervised loss $\mathcal{L}^j_s$ between individual predictions from each network and actual contrast as follows,

$$\mathcal{L}^j_s(\Theta^j_{\mathcal{H}};\mathcal{X}) = \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y}_i\epsilon\mathcal{X})}\left\|\boldsymbol{x}-\boldsymbol{x}^{p_j}\right\|_1 \tag{14}$$

Then, the two denoising networks are trained by minimizing the objective,

$$\mathcal{L}(\Theta^j_{\mathcal{H}};\mathcal{X}) = \sum_{j=1}^{2}[\lambda_s\ \mathcal{L}^j_s(\Theta^j_{\mathcal{H}};\mathcal{X}) + \lambda_m\ \mathcal{L}^j_m(\Theta^j_{\mathcal{H}};\mathcal{X}) + \lambda_c\ \mathcal{L}^j_c(\Theta^j_{\mathcal{H}};\mathcal{X})] \tag{15}$$

where $\lambda_s, \lambda_m, \lambda_c > 0$ control the contribution of each loss component.

Dual-domain consistency. During the inference process, we start at timestep T with random Gaussian noise as $\boldsymbol{x}_t$ and iteratively refine through T number of sampling steps. At each step we derive $\mathrm{t}-1^{th}$ sample using Markov property of forward process[8] as follows,

$$q(\boldsymbol{x}_{\mathrm{t}-1}|\boldsymbol{x}_\mathrm{t},\boldsymbol{x}_0=\boldsymbol{x}^{p_j}) = \mathcal{N}\Big(\boldsymbol{x}_{\mathrm{t}-1};\tilde{\mu}_\mathrm{t}(\boldsymbol{x}_\mathrm{t},\boldsymbol{x}^{p_j}),\tilde{\beta}_\mathrm{t}\mathrm{I}\Big) \tag{16}$$

where $\tilde{\mu_t}$ and $\tilde{\beta_t}$ is the mean and variance of the distribution.

To ensure mutual learning between two denoising networks, we derive the average mean noise predictions across two networks using eq.3 and eq.4 as below,

$$\tilde{\mu_t}_{avg}(\boldsymbol{x}_t, \boldsymbol{x}^{p_j}) := \frac{1}{2}\sum_{j=1}^{2}\Big[\frac{\sqrt{\bar{\alpha}_{t-1}}\,\tilde{\beta}_t}{1-\bar{\alpha}_t}\,\boldsymbol{x}^{p_j} \;+\; \frac{\sqrt{\alpha_t}\,(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\,\boldsymbol{x}_t\Big] \tag{17}$$

$$\tilde{\beta}_t := \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\,\beta_t \tag{18}$$

Then, we derive denoised contrast at each sampling step as follows,

$$\tilde{\boldsymbol{x}}_{t-1} = \tilde{\mu_t}_{avg} + \sqrt{\tilde{\beta}_t}\,\varepsilon;\; \varepsilon \sim \mathcal{N}(\varepsilon; 0, I). \tag{19}$$

3 Experiments

Datasets and Baselines. We evaluated our method on two datasets: the BraTS2019 brain tumour dataset [13] and a healthy dataset [9]. From the BraTS2019 dataset, we utilized 305 co-registered multi-contrast MR images, including T1w, T2w, T1CE, and FLAIR. For each scan, we selected 80 middle axial slices, resized them to (256×256), and split the data into 214 subjects for training, 61 for validation, and 30 for testing.

For our healthy dataset, we extracted 100 middle slices from 85 healthy brain MRI scans. We allocated 50, 20 and 15 subjects for training, validation, and testing. In both datasets, one contrast served as the synthesis target, while the remaining contrasts were used as source images to guide the denoising process.

We compared D2Diff with conventional generative networks, Pix2Pix[10], pGAN[4], DDPM[8], and SOTA MRI synthesis methods including, Hi-Net[20], MM-GAN[15], and SynDiff[14] adopted in a supervised manner.

4 Experimental Results

The qualitative performance of D2Diff is illustrated in Fig. 2 and 3 for both healthy and BraTS synthetic results. For healthy subjects, D2Diff better preserves anatomical structures, offering superior contrast details compared to other methods. In tumour datasets, it improves lesion synthesis, particularly in challenging tasks like T1CE, where other methods struggle with contrast enhancements. It produces sharper tumour boundaries that closely resemble ground truths. Additionally, the quantitative evaluation in Table 1 for PSNR, SSIM, MAE[7] confirms D2Diff's superiority, outperforming all methods across tasks.

Downstream segmentation task performance. To assess the diagnostic equivalence of our synthetic results, we conducted tumour segmentation using the BraTS dataset. A MONAI U-Net[2] was trained on all four contrasts to predict tumour masks with the same train-test split as synthesis tasks. Table 2

Dataset	Contrast	Metric	pGAN	Pix2Pix	DDPM	MMGAN	Hi-Net	SynDiff	D2Diff
BraTS	T1CE	PSNR	25.13±1.95	25.64±1.94	24.22±1.81	27.72±2.06	26.92±2.07	28.16±2.36	**28.58±2.69**
		SSIM %	87.34±3.28	88.06±3.15	56.05±7.37	87.47±3.74	90.03±2.94	91.15±2.88	**91.84±2.83**
		MAE %	3.29±1.43	2.93±1.15	25.97±5.61	11.83±6.83	2.66±1.31	2.08±1.14	**1.97±1.14**
	FLAIR	PSNR	25.37±1.90	24.98±1.71	25.60±2.03	24.50±1.73	25.20±1.82	27.13±2.11	**27.57±2.18**
		SSIM %	85.69±3.35	85.19±3.25	76.96±6.57	77.57±5.27	85.88±3.25	89.17±3.50	**89.56±3.42**
		MAE %	3.32±1.57	3.81±1.63	25.97±5.61	17.58±12.37	4.92±2.18	2.65±1.24	**2.55±1.30**
	T2	PSNR	25.70±2.01	25.52±1.89	24.05±1.86	26.01±2.12	26.26±2.04	28.24±2.64	**28.73±2.69**
		SSIM %	89.66±3.80	89.27±3.73	86.23±4.36	87.32±6.40	91.49±3.64	93.05±4.05	**93.51±3.97**
		MAE %	2.50±1.01	2.72±1.28	3.66±1.20	16.93±7.85	2.75±1.28	1.74±0.95	**1.65±0.09**
	T1	PSNR	26.23±1.89	26.71±1.82	24.78±1.70	25.26±1.92	27.19±2.05	29.36±2.83	**29.96±2.80**
		SSIM %	90.94±3.21	91.12±2.96	88.18±3.32	88.11±4.40	93.05±2.83	93.63±3.49	**94.13±3.33**
		MAE %	2.90±1.80	2.46±1.49	3.08±1.33	8.42±5.91	2.42±1.51	1.79±1.53	**1.71±1.52**
Healthy	FLAIR	PSNR	26.89±1.61	26.89±1.61	23.13±1.66	27.12±1.83	28.32±2.09	29.30±2.31	**29.65±2.25**
		SSIM %	92.03±2.38	91.45±2.94	42.77±6.81	91.48±3.54	94.26±2.15	95.01±2.11	**95.34±2.02**
		MAE %	10.47±8.05	1.81±0.54	3.45±0.70	6.93±3.47	1.50±0.51	1.28±0.44	**1.25±0.43**
	T2	PSNR	25.78±1.28	24.88±1.56	25.19±1.21	26.61±1.20	27.21±1.20	27.70±1.63	**28.54±1.67**
		SSIM %	89.07±3.76	87.26±4.80	79.24±5.80	88.88±4.52	92.00±2.95	92.72±2.90	**93.56±2.66**
		MAE %	25.90±6.70	2.23±0.66	2.35±0.54	16.76±5.96	2.08±0.69	1.51±0.46	**1.38±0.43**
	T1	PSNR	27.68±1.63	26.83±1.89	26.26±1.83	29.60±1.57	29.16±2.00	30.09±2.28	**30.82±2.33**
		SSIM %	93.49±2.49	92.22±3.19	86.85±4.00	93.84±3.08	95.32±2.17	95.76±2.03	**96.23±1.91**
		MAE %	12.43±1.70	1.78±0.56	2.09±0.61	6.85±4.99	1.33±0.52	1.19±0.42	**1.10±0.39**

Table 1: Performance comparison for healthy and BraTS datasets (mean±std) for different synthesis contrasts. The best performance is in bold with statistical significance $p < 0.05$ based on a paired mean t-test between D2Diff and the second best performed method.

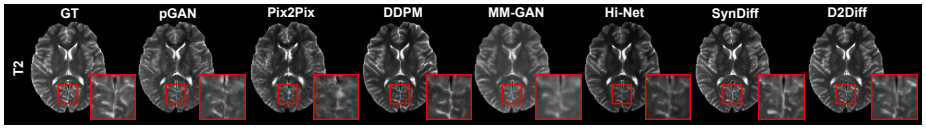

Fig. 2: Visualization of synthetic MRI results on healthy dataset.

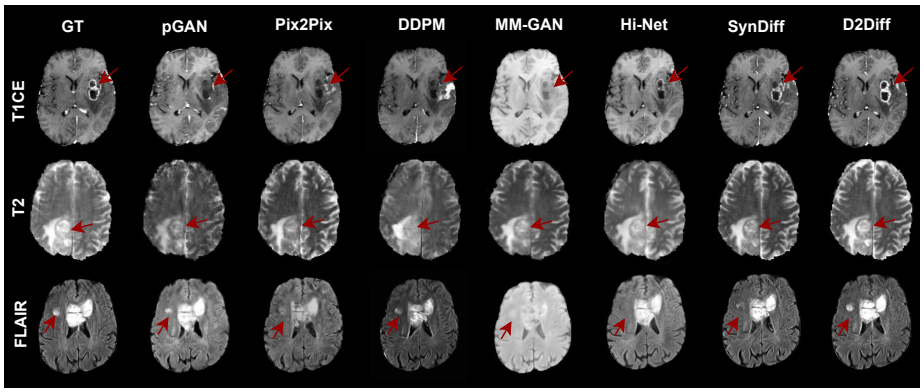

Fig. 3: Visualization of synthetic MRI results on BraTS dataset.

presents the Dice scores values for predicted masks with synthetic results, where x mark indicates that the corresponding contrast has been replaced by a synthetic contrast, while a ✓ denotes the use of the actual contrast from the test dataset. Fig. 4 presents qualitative comparisons for the predicted tumour masks. The "Complete" setup represents segmentation using all actual test contrasts without synthetic replacements. For comparison, we selected top-performing methods from each category. Results show that D2Diff closely matches the "Complete" setup, demonstrating clinically reliable and plausible synthesis quality. Notably, D2Diff achieves slightly higher Dice scores, likely due to the dataset's multi-site variability [13], including contrast differences and occasional artifacts. D2Diff's diffusion-based dual-domain architecture effectively mitigates these artifacts and handles contrast variations more robustly, leading to improved segmentation performance.

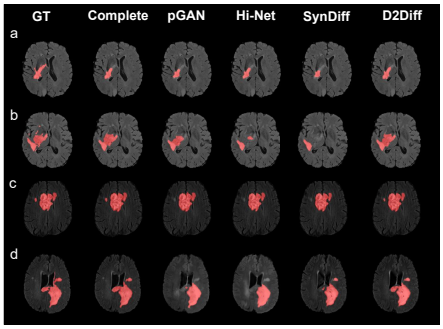

Fig. 4: Segmentation results.

Task	Contrasts				Dice Score% ↑			
	T1	T2	FLAIR	T1CE	pGAN	Hi-Net	SynDiff	D2Diff
a	✓	x	✓	✓	80.5	80.02	80.91	**81.05**
b	✓	x	✓	x	80.64	79.70	80.81	**81.19**
c	x	x	✓	x	79.96	78.93	80.66	**81.02**
d	x	x	x	x	76.50	73.49	79.31	**79.34**
complete	✓	✓	✓	✓		80.83		

Table 2: Segmentation performance.

	PSNR ↑	SSIM% ↑	MAE% ↓
Freq. guidance (H1)	28.01±2.41	91.24±2.88	2.09±1.19
Spatial guidance (H2)	28.39±2.45	91.82±2.87	2.02±1.30
w/o freq. feat. adaptation	28.54±2.32	91.34±2.82	2.07±1.27
w/o mask loss	28.29±2.40	91.57±2.80	2.04±1.25
D2Diff	**28.58±2.69**	**91.84±2.83**	**1.97±1.14**

Table 3: Ablation study.

Ablation Study. We conducted an ablation study to assess the impact of individual components and the dual-domain mutual learning approach. The T1CE synthesis task was selected to perform ablation as it is one of the most challenging tasks in tumour synthesis. As shown in Table 3, every component contributes to enhancing overall synthesis quality. The mask loss enhances lesion delineation—particularly at tumour boundaries where uncertainty is higher, as illustrated by sample uncertainty maps (U_1, U_2) in Fig. 1. In addition, the mutual learning between frequency- and spatial-domain networks leads to more effective representation learning, resulting in superior synthesis performance over their individual synthesis outputs. While D2Diff employs two denoising networks, its design leverages two shallow generators to maintain computational efficiency. The model contains 68.586M parameters, comparable to SOTA methods which use conventional diffusion-GAN based approaches like SynDiff with 67.465M, and achieves an average sampling time of 310.10 ms, only slightly above SynDiff's 303.70 ms. Thus, despite the dual-network structure, D2Diff does not impose a substantial computational burden.

5 Conclusion

In this work, we introduce a dual-domain diffusion framework for multi-contrast MRI synthesis, leveraging frequency and spatial features to capture both intensity variations and spatial differences across contrasts. Experimental results demonstrate that D2Diff outperforms baseline methods, producing more accurate synthetic images. Additionally, superior downstream tumour segmentation highlights the diagnostic value of the synthetic images. Despite these promising results, the current model is limited to 2D data, and further research is necessary to validate clinical reliability across multi-centre datasets, diverse clinical settings and imaging protocol variations. Overall, D2Diff offers a promising approach for high-fidelity multi-contrast MRI synthesis, contributing to the efficiency and safety of medical imaging.

Acknowledgments. This research was supported by Monash University's Faculty of IT International Postgraduate Research Scholarship, and Zhaolin Chen receives funding support from the Australian Research Council, including Mid-Career Fellowship Project (IM230100002) and Discovery Project (DP210101863).

Disclosure of Interests. The authors declare no competing interests.

References

1. Ahmed, N., Natarajan, T., Rao, K.: Discrete Cosine Transform. IEEE Transactions on Computers **C–23**(1), 90–93 (Jan 1974). https://doi.org/10.1109/t-c.1974.223784, http://dx.doi.org/10.1109/T-C.1974.223784
2. Cardoso, M.J., Li, W., Brown, R., et. al.: MONAI: An open-source framework for deep learning in healthcare (2022). https://doi.org/10.48550/ARXIV.2211.02701
3. Chen, Z., Pawar, K., Ekanayake, M., Pain, C., Zhong, S., Egan, G.F.: Deep Learning for Image Enhancement and Correction in Magnetic Resonance Imaging—State-of-the-Art and Challenges. Journal of Digital Imaging **36**(1), 204–230 (Nov 2022). https://doi.org/10.1007/s10278-022-00721-9
4. 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 Transactions on Medical Imaging **38**(10), 2375–2388 (Oct 2019). https://doi.org/10.1109/tmi.2019.2901750
5. Dayarathna, S., Islam, K.T., Uribe, S., Yang, G., Hayat, M., Chen, Z.: Deep learning based synthesis of MRI, CT and PET: Review and analysis. Medical Image Analysis **92**, 103046 (Feb 2024). https://doi.org/10.1016/j.media.2023.103046
6. Ding, H., Lu, J., Cai, J., Zhang, Y.: SLf-UNet: Improved UNet for Brain MRI Segmentation by Combining Spatial and Low-Frequency Domain Features (May 2023). https://doi.org/10.21203/rs.3.rs-2849524/v1
7. Dohmen, M., Klemens, M.A., Baltruschat, I.M., Truong, T., Lenga, M.: Similarity and quality metrics for MR image-to-image translation. Scientific Reports **15**(1) (Jan 2025). https://doi.org/10.1038/s41598-025-87358-0
8. Ho, J., Jain, A., Abbeel, P.: Denoising Diffusion Probabilistic Models. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems. vol. 33, pp. 6840–6851.

Curran Associates, Inc. (2020), https://proceedings.neurips.cc/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf
9. Islam, K.T., Zhong, S., Zakavi, P., Chen, Z., Kavnoudias, H., Farquharson, S., Durbridge, G., Barth, M., McMahon, K.L., Parizel, P.M., Dwyer, A., Egan, G.F., Law, M., Chen, Z.: Improving portable low-field mri image quality through image-to-image translation using paired low- and high-field images. Scientific Reports **13**(1) (Dec 2023). https://doi.org/10.1038/s41598-023-48438-1, http://dx.doi.org/10.1038/s41598-023-48438-1
10. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-Image Translation with Conditional Adversarial Networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). p. 5967–5976. IEEE (Jul 2017). https://doi.org/10.1109/cvpr.2017.632
11. Lohrke, J., Frenzel, T., Endrikat, J., Alves, F.C., Grist, T.M., Law, M., Lee, J.M., Leiner, T., Li, K.C., Nikolaou, K., Prince, M.R., Schild, H.H., Weinreb, J.C., Yoshikawa, K., Pietsch, H.: 25 Years of Contrast-Enhanced MRI: Developments, Current Challenges and Future Perspectives. Advances in Therapy **33**(1), 1–28 (Jan 2016). https://doi.org/10.1007/s12325-015-0275-4
12. Lou, Y., Zhang, J., Xu, D., Cao, Y., Wang, H., Huang, Y.: No-Reference MRI Quality Assessment via Contrastive Representation: Spatial and Frequency Domain Perspectives. In: 2024 IEEE International Conference on Multimedia and Expo (ICME). p. 1–6. IEEE (Jul 2024). https://doi.org/10.1109/icme57554.2024.10687481
13. Menze, B.H., Jakab, A., Bauer, S., et.al.: The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). IEEE Transactions on Medical Imaging **34**(10), 1993–2024 (Oct 2015). https://doi.org/10.1109/tmi.2014.2377694
14. Özbey, M., Dalmaz, O., Dar, S.U.H., Bedel, H.A., Özturk, c., Güngör, A., Çukur, T.: Unsupervised Medical Image Translation With Adversarial Diffusion Models. IEEE Transactions on Medical Imaging **42**(12), 3524–3539 (Dec 2023). https://doi.org/10.1109/TMI.2023.3290149
15. Sharma, A., Hamarneh, G.: Missing MRI Pulse Sequence Synthesis Using Multi-Modal Generative Adversarial Network. IEEE Transactions on Medical Imaging **39**(4), 1170–1183 (Apr 2020). https://doi.org/10.1109/tmi.2019.2945521
16. Wei, K., Kong, W., Liu, L., Wang, J., Li, B., Zhao, B., Li, Z., Zhu, J., Yu, G.: CT synthesis from MR images using frequency attention conditional generative adversarial network. Computers in Biology and Medicine **170**, 107983 (Mar 2024). https://doi.org/10.1016/j.compbiomed.2024.107983
17. Xiao, Z., Kreis, K., Vahdat, A.: Tackling the Generative Learning Trilemma with Denoising Diffusion GANs (2021). https://doi.org/10.48550/ARXIV.2112.07804
18. Yu, B., Wang, Y., Wang, L., Shen, D., Zhou, L.: Medical Image Synthesis via Deep Learning, p. 23–44. Springer International Publishing (2020). https://doi.org/10.1007/978-3-030-33128-3_2
19. Zhan, B., Li, D., Wu, X., Zhou, J., Wang, Y.: Multi-Modal MRI Image Synthesis via GAN With Multi-Scale Gate Mergence. IEEE Journal of Biomedical and Health Informatics **26**(1), 17–26 (Jan 2022). https://doi.org/10.1109/jbhi.2021.3088866
20. 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 (Sep 2020). https://doi.org/10.1109/tmi.2020.2975344
21. Ziashahabi, A., Buyukates, B., Sheshmani, A., You, Y.Z., Avestimehr, S.: Frequency Domain Diffusion Model with Scale-Dependent Noise Schedule. In: 2024 IEEE International Symposium on Information Theory (ISIT). p. 19–24. IEEE (Jul 2024). https://doi.org/10.1109/isit57864.2024.10619452