This article has been accepted for publication in IEEE Transactions on Biomedical Engineering. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TBME.2026.3698056

I2I-Mamba: Multi-modal medical image synthesis via selective state space modeling

Omer F. Atli, Bilal Kabas, Fuat Arslan, Arda C. Demirtas, Mahmut Yurt, Onat Dalmaz, and Tolga Çukur, *Senior Member*

***Abstract*— Multi-modal medical image synthesis involves nonlinear transformation of tissue signals between source and target modalities, where tissues exhibit contextual interactions across diverse spatial distances. As such, the utility of a network architecture in synthesis depends on its ability to express the broad set of contextual features in medical images. Convolutional neural networks (CNNs) offer high local precision at the expense of poor sensitivity to long-range context. While transformers promise to alleviate this issue, they suffer from an unfavorable trade-off between sensitivity to long- versus short-range context due to the intrinsic complexity of attention filters. To effectively capture contextual features while avoiding the complexity-driven trade-offs, here we introduce a novel multi-modal synthesis method, I2I-Mamba, based on the state space modeling (SSM) framework. Focusing on high-level representations across a hybrid residual architecture, I2I-Mamba leverages novel dual-domain Mamba (ddMamba) blocks for complementary contextual modeling in image and Fourier domains, while maintaining spatial precision with convolutional layers. Diverting from conventional raster-scan trajectories, ddMamba leverages novel SSM operators based on a spiral-scan trajectory to learn context with enhanced angular isotropy, and a channel-mixing layer to aggregate context across the channel dimension. Comprehensive demonstrations on multi-contrast MRI and MRI-CT protocols indicate that I2I-Mamba outperforms state-of-the-art CNNs, transformers and SSMs.**

***Index Terms*— medical image synthesis; modality; state space; Mamba**

I. Introduction

Multi-modal medical images with distinct tissue contrasts provide complementary information about underlying anatomy, boosting reliability in downstream analyses [1]. Multi-modal imaging is viable using different sequences on the same scanner or on entirely different scanners, albeit costs of running prolonged exams yield incomplete protocols under many scenarios [2]. As a remedy, target-modality images missing from a protocol can be synthesized based on the subset of source-modality images available [3], [4]. Key clinical applications of synthesis include imputing target modalities with diagnostically-relevant albeit redundant information omitted from imaging protocols to reduce scan time and improve efficiency; inferring invasive target modalities from non-invasive sources to avoid exposure to ionizing radiation or harmful contrast agents [5]; and recovering missing modalities to enhance consistency across participants in retrospective studies [6]. Yet, such synthesis tasks involve a challenging nonlinear transformation of signal levels between source and target images depending on tissue characteristics [7]. Although detailed tissue parameters that govern signal levels are generally difficult to infer from medical images, multi-modal synthesis can be guided via a rudimentary spatial prior on tissue composition that can be implicitly inferred from the signal distribution in source images [8]. Both healthy and pathological tissues can exhibit broad spatial distribution in the form of contiguous or segregated clusters across an anatomy [9], introducing not only local signal correlations in compact neighborhoods but also non-local correlations over extended distances. Thus, successful multi-modal synthesis inherently rests on the ability to exploit these short- to long-range contextual features.

Deep learning has recently emerged as the mainstream synthesis framework given its prowess in nonlinear function approximation [10]–[13]. In learning-based synthesis, a generative model attempts to map source onto target images through hierarchical nonlinear transformations on intermediate feature maps [14], [15]. Naturally, the fidelity of this mapping depends on the model's expressiveness for the diverse set of contextual features in medical images. While open questions remain about optimal training strategies (e.g., adversarial vs. diffusion) [16]–[19], architectural design is a critical factor that directly determines the expressiveness of generative models. Earlier studies have used CNN-based models with convolution operators for local filtering of feature maps [15], [20]–[23]. Popularity of CNNs has been fueled by linear model complexity with respect to image dimensionality, and high expressiveness for local context that can be critical in synthesis of detailed tissue structure [5], [24]–[33]. However, as convolution operators have strictly localized spatial footprints with poor sensitivity to long-range context [34], [35], CNNs can suffer from low accuracy in regions of heterogeneous tissue composition and uncommon pathology, where contextual relations are key in inferring the spatial distribution of tissue signals [36], [37].

Later studies have instead adopted transformer-based models based on self-attention operators that are capable of non-local filtering [34], [38]–[41]. Tokenizing an input image as a sequence of patches, transformers compute attention weights to measure inter-patch similarity and non-locally filter the input sequence. While the diffuse spatial footprint of self-attention helps increase sensitivity to long-range context, it induces quadratic model complexity with respect to sequence length (i.e., number of image patches), prohibiting use on small patches necessary for high spatial precision [40], [41]. Common mitigation strategies include efficient attention operators that compress contextual representations at the expense of limiting contextual sensitivity [34], [42], and tokenization

Corresponding author: Tolga Çukur (e-mail: cukur@ee.bilkent.edu.tr). This work was supported in part by TUBA GEBIP 2015 and BAGEP 2017 fellowships awarded to T. Çukur, and in part by The Scientific and Technological Research Council of Türkiye (TÜBİTAK) 1515 Frontier R&D Laboratories Support Program for Türk Telekom 6G R&D Lab under project number 5249902.

O.F. Atli, B. Kabas, F. Arslan, A.C. Demirtas and T. Çukur are with the Department of Electrical and Electronics Engineering, and National Magnetic Resonance Research Center, Bilkent University, Ankara 06800, Turkey. M. Yurt and O. Dalmaz are with the Department of Electrical Engineering, Stanford University, CA 94305, United States.

via large patches (e.g., a 16×16 patch as a single token) that compromises short-range contextual sensitivity [39], [40]. As such, transformers typically face an undesirable trade-off between spatial precision and global awareness. An emerging alternative to efficiently capture contextual representations is state-space models (SSM) [43], [44]. SSMs cast image pixels onto an input sequence via raster-scan trajectories in rectilinear orientations, and efficiently model this sequence via state-space operators that offer a refined trade-off between long- and short-range contextual sensitivity. Recent reports have already adopted selective SSMs (Mamba) for unimodal analysis and reconstruction tasks in medical imaging [45]–[47].

Given their success in other imaging tasks, SSMs hold significant promise for multi-modal medical image synthesis. However, several limitations hinder the effectiveness of existing SSMs in modality imputation. *Traditional SSMs perform recurrent sequence modeling directly in the image domain*, which often leads to an inevitable trade-off between capturing short- and long-range dependencies—*global context is attenuated* due to the difficulty of modeling interactions between distant sequence elements [45]. Moreover, multi-modal images of the same anatomy typically exhibit *signal correlations that vary with spatial frequency*: high-frequency edge structures are typically well-aligned across modalities, while low-frequency contrast information tends to be more modality-specific [48]. Conventional SSMs can *struggle to model this frequency-dependent behavior* effectively, limiting their ability to capture cross-modal interactions. Finally, SSMs typically impose *a rectilinear raster-scan trajectory (e.g., sweep, zig-zag)* to sequentialize images, resulting in anisotropic receptive fields and *reduced sensitivity to contextual interactions along non-axial directions* [47]. Note that diagnostically relevant structures such as vasculature, cortical folds, and tumor lesions often follow oblique orientations that are poorly captured by rectilinear scans. These limitations collectively constrain the potential of SSMs in medical image synthesis tasks.

In this study, we introduce I2I-Mamba, a novel learning-based model that, to our knowledge, represents the first SSM-based framework for multi-modal medical image synthesis (see [49] for an earlier conference version employing conventional image-domain SSM blocks with raster-scan operators and focusing exclusively on multi-contrast MRI synthesis). To achieve high contextual sensitivity while addressing key limitations of existing SSM approaches, I2I-Mamba incorporates dual-domain Mamba (ddMamba) blocks, residually fused with CNN blocks across a bottleneck that preserves high-resolution semantic representations (Fig. 1). To overcome the limitations of conventional image-domain SSMs in capturing global context and frequency-dependent interactions, ddMamba blocks feature independent SSM branches operating in the image and Fourier domains. This dual-domain design enables the state-space operators to leverage both spatial and spectral contextual cues. To mitigate the anisotropic spatial footprint introduced by traditional raster-scan trajectories, ddMamba blocks adopt a novel spiral-scan trajectory, enhancing angular isotropy in receptive fields (Fig. 2). Additionally, ddMamba blocks incorporate channel-mixing layers to enable context aggregation not only across spatial dimensions but also across channel dimensions of the feature maps.

The architectural design of I2I-Mamba is devised to capture rich cross-modal contextual interactions in medical images. We demonstrate the effectiveness of this design through comprehensive experiments on missing modality imputation in multi-contrast MRI and MRI-to-CT translation tasks. To systematically evaluate the impact of architecture, primary comparisons are performed in an adversarial training setup, providing a computationally efficient framework for controlled evaluation, while additional experiments benchmark I2I-Mamba against recent diffusion baselines. Our results show that I2I-Mamba consistently outperforms state-of-the-art CNN, transformer, and SSM-based models. The code is publicly available at: https://github.com/icon-lab/I2I-Mamba.

Contributions

- We develop, to our knowledge, the first application of SSMs for translating between multi-modal data to synthesize medical images from missing modalities.
- We propose a novel SSM model that operates on image and Fourier representations to enhance synthesis by capturing complementary spatial and spectral features.
- We design a novel ddMamba block that modulates SSM outputs with channel-mixing layers, enabling the model to jointly learn contextual dependencies across spatial-spectral and channel dimensions.
- We introduce a spiral-scan trajectory for image- and Fourier-domain SSM layers that enhances angular isotropy in receptive fields, improving the modeling of context in non-axial directions.

II. Related Work

A. Learning-based medical image synthesis

Learning-based synthesis relies on the choice of model architecture to infer mappings across modalities. CNNs use convolution operators to extract local image context, but struggle to generalize to atypical anatomy and model long-range dependencies [36]. Attention mechanisms have been considered in CNNs to emphasize semantically relevant regions during synthesis [37], [50]–[52]. Yet, because these mechanisms multiplicatively gate features derived via convolution operators, they provide only a modest change in sensitivity to global context [39]. Recurrent neural networks such as LSTM-based architectures have also been explore to enhance contextual modeling beyond local convolutions [53]–[56]. While these approaches introduce state propagation mechanisms, their formulation and computational characteristics differ from modern SSMs designed for efficient long-range dependency modeling at high spatial resolutions.

Transformers have been increasingly adopted to explicitly model long-range interactions [34]. Yet, pure transformer-based models incur high computational costs, which has led to the development of efficient variants using approximations such as windowed or low-rank attention [42]. These approximations, however, limit the model's ability to capture the full-scope of contextual interactions [40]. Alternatively, hybrid CNN-transformer architectures apply transformer modules only in low-resolution stages to reduce computation [39],

This article has been accepted for publication in IEEE Transactions on Biomedical Engineering. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TBME.2026.3698056

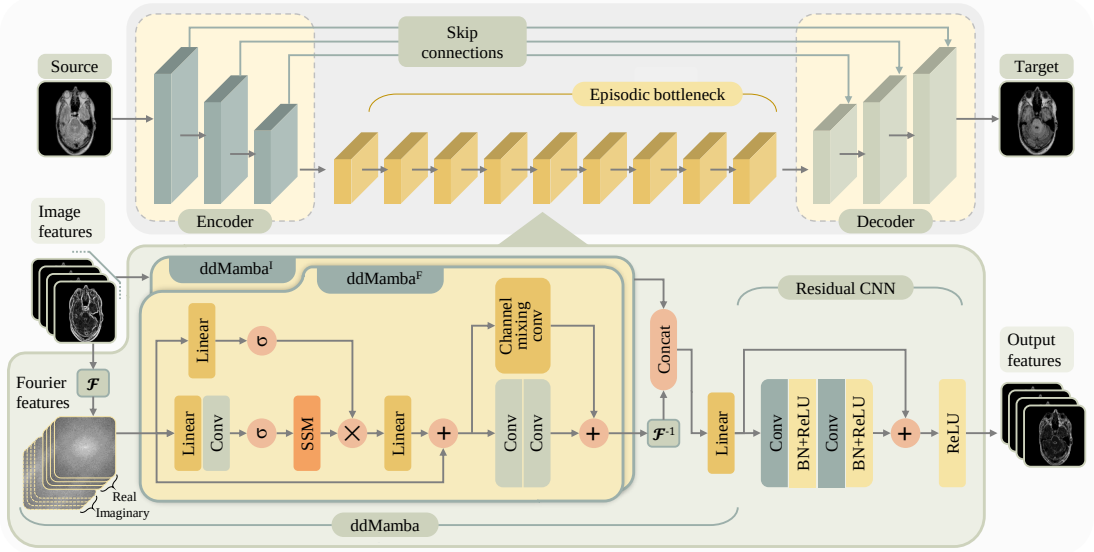

Fig. 1: Network architecture for I2I-Mamba. The proposed model comprises encoder, bottleneck, and decoder modules to synthesize target from source images. The encoder extracts high-level representations of the source image via convolutional layers. The bottleneck extracts task-relevant contextual information across spatial, frequency and channel dimensions with the aid of dual-domain Mamba (ddMamba) blocks (ddMamba^{I}: image domain, ddMamba^{F}: Fourier domain) comprising channel-mixing layers, and maintains high spatial precision with the aid of residual CNN blocks. The decoder back-projects the contextualized representations onto the target image via convolutional layers.

[57], compromising local precision. Consequently, transformers often trade off local precision for broader context due to the inherent computational limitations of attention mechanisms.

B. SSM models in medical imaging

SSMs have recently emerged as promising alternatives for achieving a refined balance between short- and long-range sensitivity, without incurring high model complexity. In medical imaging, SSM models have recently been proposed for unimodal tasks such as image segmentation [45], image reconstruction from undersampled measurements [46], [47], and image generation [58]. However, these studies have not addressed the distinct challenges posed by multi-modal synthesis. Here we introduce I2I-Mamba, which, to the best of our knowledge, is the first SSM-based framework designed for mapping between distinct imaging modalities, including both multi-contrast MRI and MRI-to-CT translation. While few independent studies [59] have appeared on this topic following our preprint [60], I2I-Mamba differs from these efforts through several architectural novelties, including (i) a residual SSM-CNN design —excluding transformer modules— that departs from the common UNet-style backbone to better preserve low-level spatial detail, (ii) dual-domain SSM blocks that operate jointly in image and frequency domains rather than conventional image-domain SSMs, and (iii) a spiral-scan tokenization strategy that replaces standard raster scans.

Common UNet-style SSMs pervasively adopted in the imaging literature typically reduce spatial resolution to coarse feature maps (e.g., 16×16) [44], thereby limiting spatial precision in contextual modeling [25]. Furthermore, conventional SSM models often operate solely in the image domain, making them less effective at capturing frequency-dependent patterns and long-range global structures. These models also use raster-scan trajectories to map images into 1D sequences, which introduces anisotropic biases in horizontal and vertical directions [44]. I2I-Mamba departs from existing SSM architectures in imaging in several ways. First, it employs a deep latent bottleneck at higher resolution (64×64), preserving local spatial detail while enabling efficient contextualization. Second, it uses dual-domain SSM processing, jointly operating in both image and Fourier domains. This dual representation enables state-space operators to simultaneously learn spatially localized features and globally coherent spectral patterns, offering improved sensitivity to multi-scale and frequency-aware features. Third, we propose a spiral-scan trajectory for SSM tokenization, yielding near-isotropic receptive fields and mitigating directional bias in contextual modeling. These technical advances position I2I-Mamba as a performant solution for modality translation in medical imaging.

A recent method for unimodal MRI reconstruction employs raster-scan SSMs in the image domain, and circular-scan SSMs in k-space to better preserve the energy spectrum across spatial frequencies [61]. In contrast, I2I-Mamba targets multi-modal synthesis tasks and introduces a unified spiral trajectory within its SSM operators in both image and Fourier domains, with the purpose of promoting isotropic receptive fields and reducing orientation bias. Furthermore, while [61] additively fuses image- and k-space features, I2I-Mamba adopts a learnable fusion strategy that enables adaptive integration of spatial and spectral context. Overall, the two methods differ substantially in the imaging tasks they target, the objectives of non-raster scans and the domains in which they are deployed, and the mechanisms used to aggregate image- and Fourier-domain representations.

C. Multi-domain processing in medical image synthesis

Multi-domain processing has remained relatively underexplored in medical image synthesis. Existing methods largely operate in the image domain, with limited attention to complementary transform-based representations. A recent study has investigated the multi-resolution nature of wavelets to capture hierarchical image features [62]. However, the wavelet-domain

This article has been accepted for publication in IEEE Transactions on Biomedical Engineering. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TBME.2026.3698056

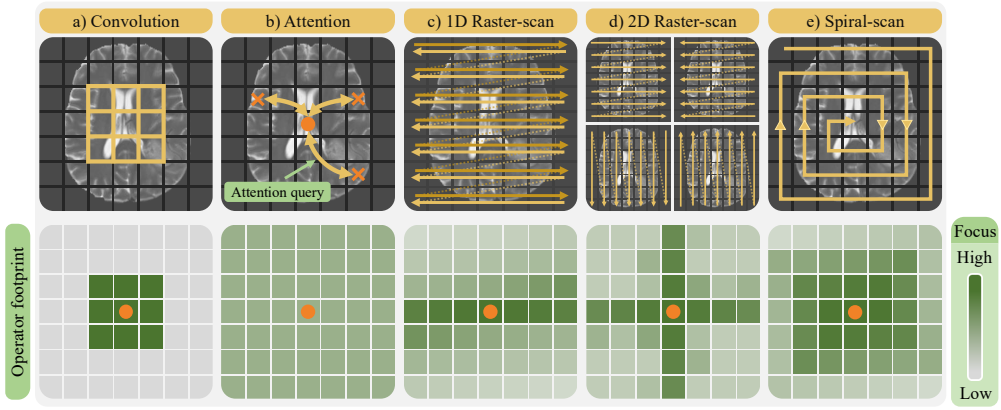

Fig. 2: Footprints illustrating the spatial distribution of focus that each learning operator deploys (see colorbar), while seeking contextual interactions of a central pixel (orange dots). **(a)** Convolution operators in CNNs have localized footprints with heavy focus over a restricted neighborhood, compromising sensitivity to long-range contextual interactions. **(b)** Attention operators in transformers have non-local footprints that diffusely distribute focus over the image, compromising local precision. **(c), (d)** Conventional state-space operators in SSMs are based on multiple raster-scan trajectories with anisotropic footprints biased towards rectangular image axes, limiting sensitivity to interactions in non-axial orientations. **(e)** I2I-Mamba's state-space operator leverages a novel spiral-scan trajectory that attains a near-isotropic footprint with more uniform focus across orientations, maintaining an improve balance between long- versus short-range contextual interactions.

method is based on a convolutional backbone, hence does not benefit from the contextual modeling capabilities of state-space operators in transform domains. To our knowledge, I2I-Mamba is the first to employ SSM in a dual-domain formulation based on image- and Fourier-domain representations. This dual-domain design enables context fusion across spatial and spectral representations, offering a principled way to integrate localized and global structural contents.

III. Methods

A. I2I-Mamba

I2I-Mamba is a novel medical image synthesis model that combines the contextual sensitivity of SSMs with the local precision of CNNs in a hybrid adversarial architecture. The generator of I2I-Mamba comprises encoder, high-resolution bottleneck, and decoder stages to synthesize target-modality images from source-modality images (Fig. 1).

Encoder. Receiving as input available source-modality images, the encoder extracts high-resolution semantic representations after concatenating input modalities into a tensor:

$$X = [x_1; x_2; \ldots; x_I], \tag{1}$$

where $I \in \mathbb{Z}^+$ is the number of source modalities, $x_i \in \mathbb{R}^{H,W}$ is the source image for the ith modality, and $X \in \mathbb{R}^{H,W,I}$ is the input tensor (H, W: image height and width). The input tensor is projected through multiple CNN blocks to derive latent representations at relatively high spatial resolution:

$$f_1 = \mathsf{CNN}(X), \tag{2}$$

where $f_1 \in \mathbb{R}^{H'=H/\alpha, W'=W/\alpha, C}$ (α: moderate downsampling rate, C: the number of feature channels).

Dual-domain Mamba bottleneck. Next, a deep bottleneck with J stages extracts task-relevant contextual representations. At each stage, *novel ddMamba blocks* are employed that contain SSM layers independently operating in image and Fourier domains to capture spatial and spectral context (Fig. 1), followed by channel-mixing layers to aggregate contextual interactions among the channel dimension. The contextualized feature maps from the ddMamba blocks are further processed via a residual CNN (rCNN) block to maintain high spatial precision. Receiving the feature map $f_j \in \mathbb{R}^{H',W',C}$, the jth bottleneck stage computes $f_{j+1} \in \mathbb{R}^{H',W',C}$ as:

$$f_{j+1} = \mathsf{rCNN}(\mathsf{ddMamba}(f_j)). \tag{3}$$

In each ddMamba block, a first branch $\mathsf{ddMamba}^I$ derives feature map $e^I_{cm} \in \mathbb{R}^{H',W',C}$ operating in the image domain:

$$e^I_{cm} = \mathsf{ddMamba}^{\mathsf{I}}(f_j). \tag{4}$$

A second branch $\mathsf{ddMamba}^F$ derives feature map $e^F_{cm} \in \mathbb{R}^{H',W',C}$ operating in the Fourier domain:

$$e^F_{cm} = \mathcal{F}^{-1}\{\mathsf{ddMamba}^{\mathsf{F}}(\mathcal{F}\{f_j\})\}, \tag{5}$$

where $\mathcal{F}, \mathcal{F}^{-1}$ denote forward and inverse Fourier transformation, respectively. To process complex-valued Fourier features in $\mathsf{ddMamba}^F$, real and imaginary components of $\mathcal{F}\{f_j\}$ are stacked across the channel dimension, and projected through an SSM layer with shared weights. The corresponding output channels are reorganized into a complex-valued feature map, inverse Fourier transformed, and any residual imaginary component is then discarded in the image domain. The contextualized feature maps are concatenated and two-fold compressed in the channel dimension via a linear projection:

$$\tilde{e}_{cm} = \mathsf{Lin}([e^I_{cm}; e^F_{cm}]), \tag{6}$$

which is forwarded to the rCNN block [63].

Within ddMamba branches, a gating variable $g \in \mathbb{R}^{H',W',C'}$ ($C' = \beta C$ for $\mathsf{ddMamba}^I$, $C' = 2\beta C$ for $\mathsf{ddMamba}^F$) is first computed as:

$$g = \sigma(\mathsf{Lin}_\beta(f_j)), \tag{7}$$

where σ is an activation function, Lin has expansion factor β. The input feature map f_j (or $\mathcal{F}(f_j)$) is then embedded via depth-wise convolution, and passed through an SSM layer:

$$e_j = \sigma(\mathsf{DWConv}(\mathsf{Lin}_\beta(f_j))), \tag{8}$$

$$M = \mathsf{SSM}(e_j). \tag{9}$$

Spiral-scan SSM operators. In conventional SSM layers, the input feature map is expanded along spatial dimensions onto a 1D sequence via multiple raster-scan trajectories (e.g., based on rectilinear sweep or zig-zag patterns) [44]. This causes state-space operators to have anisotropic footprints biased towards rectangular image axes, which can compromise

learning of contextual features in remaining orientations (Fig. 2c-d). To alleviate orientation bias in state-space operators, here we propose a *novel spiral-scan trajectory* that maps $e_j \in \mathbb{R}^{H',W',C'}$ onto a sequence z_{in} as (Fig. 2e):

$$z_{in}[n,c] = e_j(h[n], w[n], c), \tag{10}$$

where $n \in \mathbb{Z}^{[1\ H'W']}$ is the sequence index. ($h[n] \in \mathbb{Z}^{[1\ H']}$, $w[n] \in \mathbb{Z}^{[1\ W']}$) denote the row and column indices of the nth element, whose procedural derivation is given in Supp. Alg. 1 for an out-to-center, clockwise trajectory (see Supp. Fig. 1). For efficient implementation, we precompute a permutation matrix $\mathbf{S} \in \{0,1\}^{H'W' \times H'W'}$ that maps the row-major vectorized map $\mathrm{vec}(\mathbf{e}_j(:,:,c)) \in \mathbb{R}^{H'W'}$ to the spiral order as $z_{in}[:,c] = \mathbf{S}\ \mathrm{vec}(\mathbf{e}_j(:,:,c))$.

The sequence is then processed via the discretized state-space operator separately across channels:

$$\ell[n,c] = \bar{\mathbf{A}}\ell[n-1,c] + \bar{\mathbf{B}}[n] z_{in}[n,c], \tag{11}$$

$$z_{out}[n,c] = \bar{\mathbf{C}}[n]\ell[n,c], \tag{12}$$

where ℓ is the hidden state, $\bar{\mathbf{A}} \in \mathbb{R}^{N,N}$ is a learnable and $\bar{\mathbf{B}}[n] \in \mathbb{R}^{N,1}$, $\bar{\mathbf{C}}[n] \in \mathbb{R}^{1,N}$ are learnable, input-dependent parameters, N is the state dimensionality. The output sequence z_{out} is recast onto row-major order as $\mathrm{vec}(M[:,:,c]) = \mathbf{S}^T z_{out}[:,c]$, which is reshaped onto feature map $M \in \mathbb{R}^{H',W',C'}$. After gating M via a Hadamard product, it is linearly projected and residually combined with f_j:

$$e_{\mathrm{SSM}} = f_j + \mathsf{Lin}_{1/\beta}(g \odot M). \tag{13}$$

$e_{\mathrm{SSM}} \in \mathbb{R}^{H',W',C'}$ captures spatial/spectral relationships, while treating channels independently. Thus, we include a channel-mixing layer to aggregate context across channels:

$$e_{cm} = \mathsf{cmConv}(e_{\mathrm{SSM}}) + \mathsf{Conv}(e_{\mathrm{SSM}}), \tag{14}$$

where channel mixing is achieved via a 1×1 convolution operator [64], enabling efficient cross-channel interaction at each spatial location, while the parallel convolution branch preserves high spatial precision. Next, $\tilde{e}_{cm} \in \mathbb{R}^{H',W',C}$ fused across the ddMamba branches is projected through an rCNN block to compute the output feature map f_{j+1}.

Decoder. The decoder receives contextualized feature maps f_J from the bottleneck and projects them onto synthetic target image $y \in \mathbb{R}^{H,W}$ via transposed convolution blocks:

$$y = \mathsf{CNN}_{\mathsf{transposed}}(f_J). \tag{15}$$

I2I-Mamba is implemented as an adversarial model with a patch-based discriminator D [28], distinguishing actual (y_{act}) and synthetic (y_{syn}) target images. A combined pixel-wise and adversarial objective is used to train the generator G [39]:

$$L_G = \lambda_{pix}\mathbb{E}[||y_{syn} - y_{act}||_1] - \lambda_{adv}\{\mathbb{E}[D(y_{act}|X)^2] + \mathbb{E}[(D(y_{syn}|X) - 1)^2]\}, \tag{16}$$

where $\mathbb{E}$ denotes expectation, $y_{syn} = G(X)$, λ_{pix} and λ_{adv} are loss-term weightings. Meanwhile, an adversarial term is used to train the discriminator D [28]:

$$L_D = \mathbb{E}[D(y_{act}|X)^2] + \mathbb{E}[(D(y_{syn}|X) - 1)^2]. \tag{17}$$

B. Datasets

To mitigate potential biases due to demographic or protocol-related imbalances, we conducted evaluations on three distinct datasets with varying protocols and subject populations: two multi-contrast brain MRI datasets (IXI: https://brain-development.org/ixi-dataset/, BraTS [70]) and a multi-modal pelvic MRI-CT dataset [71]. All images were processed at a resolution of 256 × 256. When necessary, standard preprocessing steps (zero-padding in BraTS, resampling in MRI-CT) were applied to ensure consistent dimensions across datasets. Samples were selected sequentially as released in the original repositories to avoid manual filtering that could introduce bias. Experiments confirmed that model performance converged beyond the selected training set sizes.

IXI. T_1-, T_2-, PD-weighted MR images were analyzed with (7500, 3000, 5400) cross-sections reserved for (training, validation, test) sets, based on a subject-level split. Prior to modeling, T_2- and PD-weighted images were spatially registered onto T_1-weighted images in each subject via an affine transformation in FSL.

BraTS. T_1-, T_2-, FLAIR-weighted MR images from were analyzed with (7500, 3000, 7500) cross-sections reserved for (training, validation, test) sets, based on a subject-level split. As publicly shared, this dataset provides images that are co-registered onto T_1-weighted MRI scans.

MRI-CT. T_1-, T_2-weighted MRI, and CT images from were analyzed with (2430, 540, 1080) cross-sections reserved for (training, validation, test) sets, based on a subject-level split. As publicly shared, this dataset provides multi-modal images that are co-registered onto T_2-weighted MRI scans.

C. Architectural Design

The encoder module in I2I-Mamba had 3 stages, each containing a CNN block with a convolutional layer, batch normalization (BN), and ReLU activation. H=256, W=256, α=4, C=256 were used. The bottleneck had 9 stages, each containing a residual CNN block with 2 cascades of a convolutional layer, BN, and ReLU activation. Dual-domain ddMamba blocks were inserted in bottleneck stages j={1,5,9}, and used β=2, N=16, and SiLU activations. Channel-mixing blocks used a parallel combination of a channel-mixing convolutional layer and 2 regular convolutional layers. The decoder module had 3 stages, each containing a CNN block with a convolutional layer, BN, and ReLU activation, except for the final stage that used a Tanh activation. Long-range skip connections were employed between corresponding encoder-decoder stages to better preserve low-level spatial information.

D. Competing Methods

State-of-the-art baselines for medical image synthesis were considered including convolutional models (medSynth [28], pGAN [25], WDM [62]), transformer models (PTNet [34], ResViT [39], TransUNet [57], Swin-Unet [66]), and recent SSM models (U-Mamba [45], Mamba-Unet [67], Cunet-Mamba [68], CRAViM [69]). For systematic evaluations isolating the impact of architecture, these baselines were implemented using their originally proposed architectures, and as adversarial models that processed individual cross-sections identically to I2I-Mamba. Accordingly, they used a PatchGAN discriminator, and were trained using the losses

TABLE I: Performance for multi-contrast MRI synthesis tasks in IXI. PSNR (dB) and SSIM are listed as mean±std across the test set. Convolutional (Conv), diffusion (Diff), transformer (Trans), and state-space (SSM) baselines were considered. Boldface indicates the top-performing model for each task.

		$T_1, T_2 \rightarrow PD$		$T_1, PD \rightarrow T_2$		$T_2, PD \rightarrow T_1$		$T_2 \rightarrow PD$		$PD \rightarrow T_2$	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Conv	medSynth [28]	30.69±2.43	0.940±0.019	32.66±2.40	0.963±0.017	27.10±2.18	0.915±0.037	29.10±2.44	0.909±0.013	30.41±2.08	0.956±0.016
	pGAN [25]	31.40±2.40	0.948±0.013	31.96±1.95	0.955±0.015	27.14±2.45	0.928±0.031	30.86±2.39	0.957±0.015	30.11±2.34	0.949±0.019
	WDM [62]	32.50±2.48	0.966±0.011	33.30±1.75	0.964±0.012	28.01±2.48	0.934±0.033	31.27±2.48	0.960±0.014	32.34±1.55	0.961±0.012
Diff	DDPM [65]	31.77±5.31	0.893±0.076	32.50±2.17	0.948±0.049	26.41±4.37	0.921±0.077	29.68±6.39	0.720±0.255	31.87±2.97	0.964±0.024
	SynDiff [18]	30.86±5.21	0.934±0.093	29.35±8.45	0.890±0.087	26.23±2.66	0.918±0.038	29.85±4.86	0.930±0.092	28.42±6.09	0.925±0.111
Trans	PTNet [34]	30.33±2.59	0.935±0.018	32.62±2.09	0.954±0.013	27.73±2.79	0.931±0.029	29.75±2.66	0.932±0.012	32.46±2.19	0.956±0.022
	ResViT [39]	32.98±2.35	0.968±0.027	33.47±2.54	0.964±0.014	28.45±1.46	0.936±0.011	32.08±2.49	0.952±0.010	32.84±1.56	0.964±0.011
	TransUNet [57]	30.84±2.27	0.953±0.031	32.36±1.97	0.960±0.015	27.34±1.95	0.928±0.033	29.01±2.37	0.927±0.019	31.73±1.87	0.958±0.015
	Swin-Unet [66]	31.37±2.63	0.966±0.011	31.35±4.25	0.958±0.039	26.11±3.28	0.933±0.036	29.32±4.07	0.941±0.049	30.87±3.89	0.960±0.035
SSM	U-Mamba [45]	31.96±2.43	0.960±0.012	28.53±2.27	0.937±0.030	27.73±2.52	0.929±0.040	32.26±2.30	0.963±0.014	26.48±5.87	0.900±0.098
	Mamba-Unet [67]	29.56±4.48	0.967±0.016	32.39±3.90	0.966±0.035	26.93±3.40	0.940±0.037	29.54±3.88	0.948±0.046	30.33±3.97	0.955±0.035
	Cunet-Mamba [68]	31.50±2.34	0.960±0.012	32.15±1.85	0.958±0.013	28.03±2.32	0.933±0.033	30.89±2.54	0.959±0.022	31.10±2.05	0.954±0.014
	CRAViM [69]	31.94±2.34	0.958±0.014	32.64±1.83	0.958±0.014	27.68±2.03	0.930±0.030	30.79±2.31	0.953±0.017	31.83±1.54	0.957±0.014
	I2I-Mamba	**33.46±2.52**	**0.969±0.011**	**34.59±2.03**	**0.970±0.011**	**29.15±2.48**	**0.947±0.027**	**32.59±2.48**	**0.967±0.012**	**33.89±1.65**	**0.968±0.011**

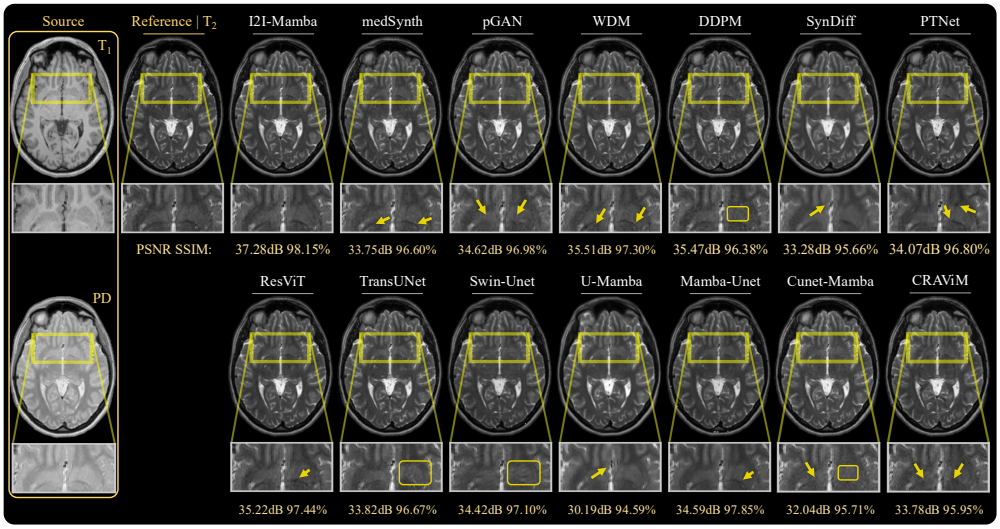

Fig. 3: Representative results for $T_1, PD \rightarrow T_2$ in IXI. Synthetic target images from competing methods are displayed along with source images and reference target images. Zoom-in windows, arrows, rectangular annotations, and performance metrics are also included to highlight differences among methods.

expressed in Eqs. 16-17. To benchmark I2I-Mamba against recent generative frameworks, diffusion baselines were also considered (DDPM [65], SynDiff [18]). Diffusion baselines were implemented using the architectures and loss functions described in their original papers. Key model hyperparameters for competing methods are listed in Supp. Table 1.

E. Modeling Procedures

Modeling was performed using the PyTorch framework on an Nvidia RTX 4090 GPU. All models were trained from scratch via the Adam optimizer with parameters $\beta_1 = 0.5$, $\beta_2 = 0.999$. Optimization hyperparameters were selected via stratified cross-validation. Data splitting was performed at the subject level, ensuring that no subject appeared in more than one fold. To avoid potential biases, a common set of hyperparameters observed to yield near-optimal validation performance across models was selected. Accordingly, all competing methods were trained using 2×10^{-4} learning rate, and 60 epochs (see Supp. Table 1). All adversarial models (convolutional, transformer and SSM baselines, and I2I-Mamba) prescribed $\lambda_{adv} = 1, \lambda_{pix} = 100$ (see Supp. Fig. 2). Among diffusion models, DDPM did not involve a loss-term weighting, and SynDiff prescribed $\lambda_{cyc} = 0.5$, $\lambda_{adv} = 1$. To evaluate performance, peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) metrics were measured between ground-truth and synthetic images. Significance levels were assessed with non-parametric signed-rank tests.

IV. Results

A. Multi-Contrast MRI Synthesis

We first conducted experiments for target-modality imputation in multi-contrast MRI. I2I-Mamba was comparatively demonstrated against convolutional (medSynth, pGAN, WDM), diffusion (DDPM, SynDiff), transformer (PTNet, ResViT, TransUNet, Swin-Unet), and SSM (U-Mamba, Mamba-Unet, Cunet-Mamba, CRAViM) models. Table I lists performance metrics for synthesis tasks in IXI that comprises data from healthy subjects. I2I-Mamba achieves the highest performance metrics consistently across tasks ($p<0.05$). On average, I2I-Mamba offers performance improvements of 2.1dB PSNR, 1.7% SSIM over convolutional baselines including the wavelet-domain WDM method; 3.0dB PSNR,

TABLE II: Performance for multi-contrast MRI synthesis tasks in BraTS. PSNR (dB) and SSIM are listed as mean±std across the test set. Boldface indicates the top-performing model for each task.

		$T_1, T_2 \rightarrow$ FLAIR		T_1, FLAIR $\rightarrow T_2$		T_2, FLAIR $\rightarrow T_1$		$T_2 \rightarrow$ FLAIR		FLAIR $\rightarrow T_2$	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Conv	medSynth	25.69±2.25	**0.911±0.031**	25.16±1.79	0.894±0.030	23.46±3.17	0.902±0.032	25.07±2.26	0.889±0.034	24.77±1.70	0.890±0.030
	pGAN	25.50±2.48	0.905±0.031	25.47±1.82	0.898±0.029	23.01±4.15	0.895±0.036	24.52±2.33	0.891±0.033	25.11±1.86	0.898±0.030
	WDM	25.70±2.36	0.901±0.032	25.70±1.85	0.903±0.028	23.37±3.53	0.900±0.033	25.03±2.40	0.893±0.034	25.06±1.92	0.896±0.029
Diff	DDPM	24.67±2.31	0.816±0.029	25.50±1.75	0.902±0.022	22.49±3.44	0.800±0.031	24.07±2.28	0.896±0.032	25.15±1.61	0.893±0.024
	SynDiff	25.70±1.68	0.903±0.022	25.14±1.53	0.903±0.024	23.01±3.77	0.891±0.037	23.59±2.22	0.901±0.032	24.82±2.26	0.892±0.024
Trans	PTNet	25.03±2.31	0.888±0.038	24.52±1.70	0.886±0.032	20.94±3.66	0.871±0.044	24.45±2.08	0.882±0.037	23.89±1.63	0.876±0.033
	ResViT	25.79±2.16	0.900±0.031	25.29±1.74	0.904±0.027	23.06±3.48	0.899±0.032	24.95±2.11	0.883±0.034	24.89±1.72	0.888±0.028
	TransUNet	25.54±2.24	0.899±0.032	25.59±1.86	0.909±0.024	23.26±3.52	**0.907±0.033**	25.28±2.26	0.896±0.032	25.04±1.84	0.899±0.029
	Swin-Unet	25.03±2.14	0.887±0.035	24.70±1.90	0.886±0.032	22.95±3.31	0.887±0.037	24.51±2.54	0.879±0.034	24.33±1.74	0.881±0.033
SSM	U-Mamba	25.37±2.39	0.891±0.037	25.58±1.70	0.909±0.026	23.36±3.59	0.895±0.035	24.86±1.55	0.882±0.022	25.03±1.33	0.894±0.024
	Mamba-Unet	25.17±1.93	0.892±0.033	25.16±1.83	0.893±0.030	23.12±2.28	0.889±0.036	24.83±2.10	0.883±0.037	24.51±1.93	0.889±0.027
	Cunet-Mamba	25.43±2.34	0.893±0.033	25.55±1.84	0.901±0.029	23.71±3.21	0.890±0.033	24.90±2.21	0.883±0.034	25.22±1.71	0.895±0.030
	CRAViM	25.70±2.25	0.896±0.033	24.99±1.82	0.890±0.031	24.12±3.44	0.895±0.031	24.90±2.10	0.880±0.035	24.88±1.66	0.890±0.029
	I2I-Mamba	**26.51±2.37**	**0.911±0.033**	**26.38±2.18**	**0.913±0.029**	**23.83±3.79**	0.904±0.034	**25.79±2.44**	**0.903±0.034**	**25.99±2.13**	**0.908±0.030**

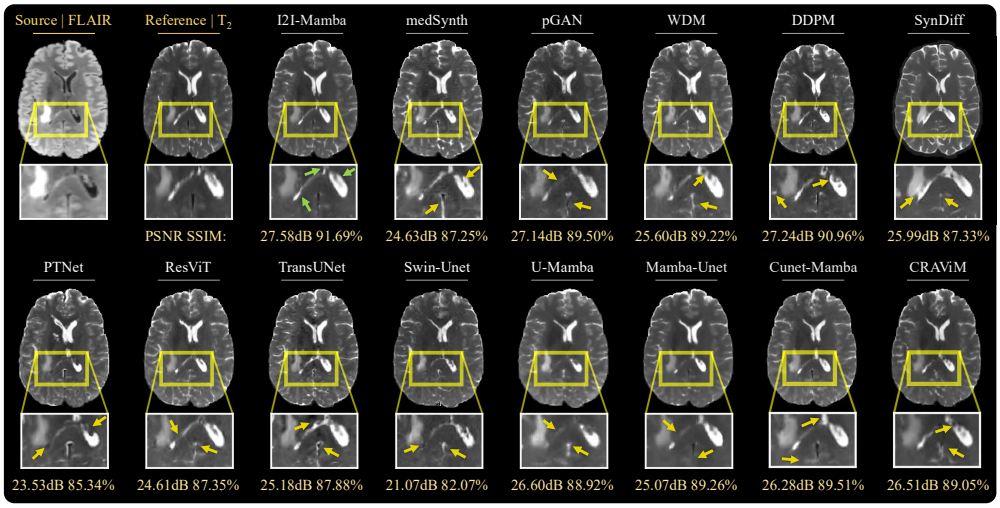

Fig. 4: Representative results for FLAIR $\rightarrow T_2$ in BraTS. Synthetic target images from competing methods are displayed along with source images and reference target images. Zoom-in windows, arrows, and performance metrics are also included to highlight differences.

6.0% SSIM over diffusion baselines; 2.1dB PSNR, 1.5% SSIM over transformer baselines; and 2.5dB PSNR, 1.5% SSIM over SSM baselines. Meanwhile, Table II lists performance metrics for synthesis tasks in BraTS that comprises data from glioma subjects. I2I-Mamba again achieves the highest performance metrics in all tasks ($p<0.05$), except for $T_1, T_2 \rightarrow$ FLAIR where medSynth yields similar SSIM and T_2, FLAIR $\rightarrow T_1$ where TransUNet yields modestly higher SSIM. On average, I2I-Mamba offers performance improvements of 0.9dB PSNR, 1.0% SSIM over convolutional baselines; 1.3dB PSNR, 2.8% SSIM over diffusion baselines; 1.3dB PSNR, 1.8% SSIM over transformer baselines; and 0.9dB PSNR, 1.6% SSIM over SSM baselines.

Representative synthetic images are displayed in Fig. 3 for IXI, and in Fig. 4 for BraTS. Convolutional baselines suffer from residual noise (e.g., pGAN) or structural inaccuracies (e.g., medSynth, WDM); diffusion baselines suffer from inaccurate depiction of low-to-moderate contrast tissue structures (e.g., DDPM) or a degree of spatial blur (e.g., SynDiff); and transformer and SSM baselines suffer either from a degree of spatial blur and contrast loss (e.g., PTNet, U-Mamba, Mamba-UNet, Cunet-Mamba) or pixel intensity artifacts (ResViT, TransUNet, Swin-Unet, CRAViM) that lead to inaccuracies in depiction of detailed anatomical structures. Particularly evident in BraTS images containing pathology, baselines generally suffer from hallucinatory features manifesting as hypo-intense or hyper-intense signals deviating from ground truth. In comparison, I2I-Mamba synthesizes target images with more accurate depiction of detailed structure and contrast in tissues, along with lower artifacts.

Performance benefits of I2I-Mamba over SSM baselines can be attributed to the fundamental architectural differences between methods. Note that U-Mamba, Mamba-Unet, Cunet-Mamba follow UNet-style architectures that substantially lower spatial resolution of encoded feature maps, except for CRAViM that follows a residual design. Furthermore, all SSM baselines rely on conventional image-domain SSM operators based on raster-scan trajectories (with CRAViM applying them on image patches). In contrast to these baselines, I2I-Mamba adopts a residual architecture where the bottleneck maintains relatively higher-resolution semantic representations, and it uses ddMamba blocks operating in image and Fourier domains, equipped with SSM operators based on spiral-scan trajectories. Our results indicate that these technical elements enable I2I-Mamba to sensitively capture a comprehensive set of contextual features to synthesize high-quality target images in multi-contrast MRI protocols.

TABLE III: Performance for the T_2-MRI → CT and T_1-MRI → CT synthesis tasks in the MRI-CT dataset.

		Conv			*Diff*		*Trans*				*SSM*				*Proposed*
		medSynth	pGAN	WDM	DDPM	SynDiff	PTNet	ResViT	TransUNet	Swin-Unet	U-Mamba	Mamba-Unet	Cunet-Mamba	CRAViM	I2I-Mamba
T_2 → CT	PSNR	26.87 ±1.79	24.64 ±1.59	26.83 ±1.80	26.83 ±1.80	26.78 ±1.95	27.35 ±2.23	27.87 ±2.06	28.06 ±2.20	24.14 ±1.87	26.32 ±1.77	24.66 ±2.69	27.54 ±2.39	27.49 ±1.99	**28.47 ±2.15**
	SSIM	0.886 ±0.026	0.867 ±0.030	0.906 ±0.021	0.906 ±0.021	0.901 ±0.023	0.911 ±0.026	0.913 ±0.027	0.915 ±0.021	0.874 ±0.028	0.902 ±0.022	0.880 ±0.031	0.909 ±0.031	0.910 ±0.023	**0.916 ±0.024**
T_1 → CT	PSNR	27.26 ±2.54	26.92 ±2.51	26.94 ±1.93	26.70 ±2.59	27.19 ±1.94	25.61 ±1.98	26.75 ±1.63	26.55 ±2.01	25.14 ±2.79	26.38 ±1.85	23.72 ±2.01	27.37 ±2.24	26.76 ±1.72	**28.01 ±1.09**
	SSIM	0.891 ±0.030	0.904 ±0.029	0.899 ±0.027	0.905 ±0.026	0.896 ±0.024	0.863 ±0.025	0.888 ±0.028	0.900 ±0.021	0.884 ±0.026	0.894 ±0.030	0.864 ±0.032	0.904 ±0.027	0.904 ±0.023	**0.909 ±0.020**

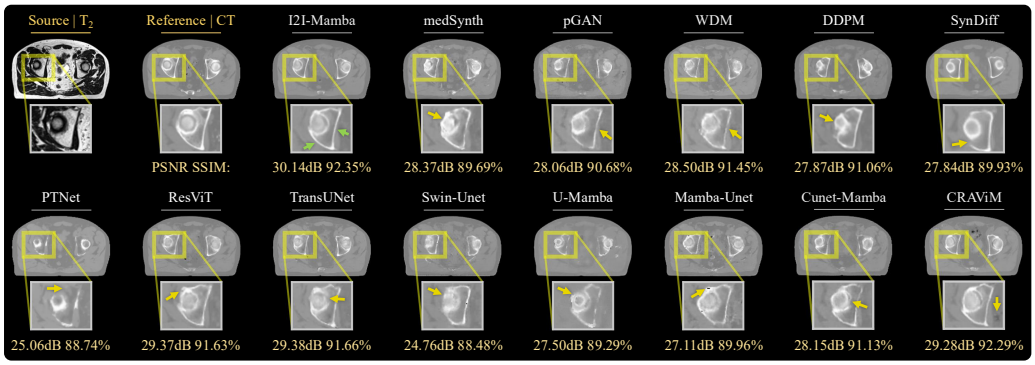

Fig. 5: Representative results for T_2→ CT in the MRI-CT dataset. Synthetic target images from competing methods are displayed along with source images and reference target images. Zoom-in windows, arrows, and performance metrics are included to highlight differences.

B. MRI-CT Synthesis

Next, we conducted experiments for target-modality imputation on the multi-modal MRI-CT dataset as listed in Table III. Among competing methods, I2I-Mamba achieves the highest performance metrics across tasks ($p<0.05$). On average, I2I-Mamba attains improvements of 1.7dB PSNR, 2.0% SSIM over convolutional baselines; 1.4dB PSNR, 1.1% SSIM over diffusion baselines; 1.8dB PSNR, 1.9% SSIM over transformer baselines; and 2.0dB PSNR, 1.7% SSIM over SSM baselines. Consistent with the findings in multi-contrast MRI synthesis, we observe that I2I-Mamba outperforms both the wavelet-domain convolutional baseline (WDM) and the SSM-based baselines (U-Mamba, Mamba-UNet), which rely on conventional image-domain SSM operators with raster-scan trajectories. These results collectively suggest that the dual-domain SSM operators in I2I-Mamba—operating jointly in the image and Fourier domains—enhance contextual sensitivity beyond what is achievable with wavelet-domain processing or standard image-domain SSMs. This elevated sensitivity to task-relevant contextual features contributes to the improved reliability of I2I-Mamba in multi-modal synthesis.

Representative synthetic images are displayed in Fig. 5. Convolutional baselines suffer from residual noise (e.g., pGAN, WDM) or structural degradations (e.g., medSynth); and diffusion baselines suffer from structural inaccuracies in depiction of low-to-moderate contrast tissue signals (e.g., DDPM) and a degree of spatial blur (e.g., SynDiff). Meanwhile, transformer and SSM baselines suffers from spatial smoothing (e.g., ResViT, TransUNet, Swin-Unet, U-Mamba, Cunet-Mamba), hypo- or hyper-intense contrast compared to ground truth in regions near bone tissue (e.g., PTNet, U-Mamba), or dark-pixel artifacts (e.g., Swin-Unet, Mamba-Unet, CRAViM). In comparison, I2I-Mamba synthesizes target images with lower artifacts and more accurate delineation of tissue structure and contrast, particularly across diagnostically relevant bone regions. Taken together, these results indicate that I2I-Mamba maintains higher sensitivity to contextual features in multi-modal medical images than baselines, thereby increasing fidelity in imputation of missing modalities.

C. Computational Complexity

Inference time, training time, memory load, and parameter counts for competing methods are summarized in Supp. Table 2. As expected, both diffusion and transformer models exhibit substantial computational overhead, reflected in longer inference and training times (due to iterative image generation in diffusion, and high model complexity in transformer models), along with higher memory usage, and increased parameter counts. This trend holds for hybrid CNN-transformer architectures (e.g., ResViT, TransUNet) as well as more efficient transformer variants employing approximate attention mechanisms (e.g., PTNet, Swin-Unet). In contrast, convolutional baselines offer high computational efficiency, with pure image-domain architectures (e.g., pGAN, medSynth) demonstrating the lowest inference times and memory demands. SSM-based methods, including I2I-Mamba, exhibit notably higher efficiency than transformer methods and show generally competitive efficiency relative to convolutional methods. Note that I2I-Mamba achieves inference time and memory usage comparable to the most efficient convolutional baselines, while its training time and parameter count are moderately higher. Among SSMs, CRAViM exhibits a smaller model footprint and correspondingly lower training cost, but does not reach the synthesis accuracy achieved by I2I-Mamba in our experiments. Taken together with its superior synthetic image quality, these results highlight I2I-Mamba as an efficient and effective solution for multi-modal medical image synthesis.

We further examined performance-efficiency trade-offs by varying the number of inference steps for the diffusion baselines (DDPM and SynDiff). As displayed in Supp. Fig. 3, the

selected number of steps for each baseline reflects a near-optimal balance between synthesis accuracy and inference cost, while further reductions in inference time lead to pronounced performance degradation. Across evaluated settings, I2I-Mamba consistently achieves higher performance at lower inference cost than diffusion baselines.

D. Ablation Studies

We performed a systematic set of ablation studies on I2I-Mamba to assess the contribution of its key design elements and configurations to synthesis performance.

Architectural components: First, we examined the importance of using SSM operators in image and Fourier domains, channel-mixing layers, and rCNN blocks. I2I-Mamba was compared against several variants for this purpose: 'w/o ddMamba' that ablated ddMamba blocks entirely, 'w/o ddMamba^F' that ablated the Fourier-domain branch from the ddMamba blocks, 'w/o ddMamba^I' that ablated the image-domain branch from the ddMamba blocks, 'w/o chan mix' that ablated channel-mixing layers from ddMamba blocks, and 'w/o rCNN' that ablated rCNN blocks. Table IV lists performance metrics for all ablated variants in representative synthesis tasks. I2I-Mamba consistently outperforms ablated variants ($p<0.05$), with improvements up to 1.0dB PSNR and 1.4% SSIM. These results indicate that each design element contributes significantly to model performance.

As listed in Table IV, we also considered a 'w/ Hermitian proj' variant, which applied a Hermitian projection on complex-valued feature maps at the output of the ddMamba^F block to explicitly enforce phase consistency prior to inverse Fourier transformation. We find that 'w/ Hermitian sym.' performs comparably to I2I-Mamba, with only modest differences. This result suggests that phase-inconsistent spectral configurations that might degrade image quality are implicitly discouraged during end-to-end training of I2I-Mamba. In addition, we evaluated 'w/ SE mix' that replaced the channel-mixing convolution with a squeeze-and-excitation layer, and 'w/ SA mix' that replaced it with a channel self-attention layer [68], [72]. We find that the proposed channel-mixing layer consistently outperforms both 'w/ SE mix' and 'w/ SA mix' ($p<0.05$), indicating that the convolutional channel-mixing design is effective.

Dual-domain processing: We then assessed the roles of image- and Fourier-domain branches in ddMamba blocks. First, we visualized branch-specific output feature maps for representative test samples after channel-wise aggregation (Supp. Fig. 4). We observe that Fourier-domain features tend to emphasize regions with rapid signal variation, such as tissue boundaries or locally low-intensity regions (e.g., near lesions), whereas image-domain features more closely follow high-intensity anatomical structures. Second, we visualized t-SNE embeddings of feature maps across the test set (Supp. Fig. 5), which reveal clear clustering and separation between image- and Fourier-domain representations. Finally, we quantified the relative contribution of each branch by measuring the aggregate feature intensity across spatial and channel dimensions (Supp. Fig. 6). While the relative intensity levels of the two branches vary across synthesis tasks, both branches consistently exhibit significant contributions. These results indicate that the image- and Fourier-domain branches encode distinct and complementary information within I2I-Mamba.

TABLE IV: Performance of I2I-Mamba variants built by ablating specific elements (w/o), or substituting alternative elements (w/). Results listed for representative synthesis tasks of $T_1, T_2 \rightarrow$ PD in IXI, FLAIR $\rightarrow T_2$ in BraTS, and $T_1 \rightarrow$ CT in the MRI-CT dataset.

	$T_1, T_2 \rightarrow$ PD		FLAIR $\rightarrow T_2$		$T_1 \rightarrow$ CT	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
I2I-Mamba	**33.46** ±**2.52**	0.969 ±0.011	25.99 ±2.13	**0.908** ±**0.030**	**28.01** ±**1.09**	**0.909** ±**0.020**
w/o ddMamba	32.65 ±2.56	0.966 ±0.010	24.86 ±2.05	0.899 ±0.029	27.18 ±1.98	0.902 ±0.021
w/o ddMamba^F	32.91 ±2.40	0.967 ±0.012	25.33 ±1.56	0.902 ±0.032	27.43 ±1.92	0.904 ±0.035
w/o ddMamba^I	33.09 ±2.58	0.968 ±0.012	25.80 ±2.13	0.907 ±0.030	27.59 ±2.06	0.906 ±0.024
w/o chan mix	33.14 ±2.63	0.968 ±0.012	25.19 ±1.91	0.898 ±0.028	27.53 ±2.41	0.897 ±0.028
w/o rCNN	32.69 ±2.53	0.966 ±0.011	25.07 ±1.83	0.894 ±0.024	26.60 ±2.05	0.885 ±0.030
w Hermitian proj	33.41 ±2.59	**0.970** ±**0.011**	**26.02** ±**2.04**	**0.908** ±**0.028**	27.95 ±1.97	0.908 ±0.024
w/ SE mix	33.28 ±2.52	0.963 ±0.011	25.66 ±2.09	0.905 ±0.029	27.68 ±2.26	0.907 ±0.025
w/ SA mix	33.25 ±2.50	0.968 ±0.011	25.72 ±2.08	0.906 ±0.029	27.74 ±2.11	0.907 ±0.027

TABLE V: Performance of I2I-Mamba variants built by using different scanning strategies.

	$T_1, T_2 \rightarrow$ PD		FLAIR $\rightarrow T_2$		$T_1 \rightarrow$ CT	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
I2I-Mamba	**33.46** ±**2.52**	**0.969** ±**0.011**	**25.99** ±**2.13**	**0.908** ±**0.030**	**28.01** ±**1.09**	**0.909** ±**0.020**
w sweep^a	32.72 ±2.48	0.964 0.013	24.97 ±1.86	0.891 ±0.031	27.31 ±2.11	0.890 ±0.027
w sweep^b	32.82 ±2.49	0.963 ±0.012	25.01 ±1.98	0.903 ±0.031	27.26 ±2.58	0.895 ±0.034
w zigzag^a	32.34 ±2.77	0.958 ±0.013	24.41 ±1.95	0.875 ±0.029	26.02 ±2.17	0.865 ±0.033
w zigzag^b	32.31 ±2.65	0.957 ±0.012	24.37 ±1.88	0.898 ±0.033	26.11 ±2.25	0.869 ±0.035
w sweep^a-zigzag^a	32.04 ±2.28	0.955 ±0.011	24.01 ±1.91	0.883 ±0.030	26.75 ±1.85	0.878 ±0.029

Spiral-scan SSM operator: Next, we examined the importance of implementing SSM operators based on the proposed spiral-scan trajectory. I2I-Mamba was compared against several variants based on raster-scan trajectories [44]: 'w sweep^a' that used 1D bidirectional sweep scans (2 separate scan trajectories), 'w sweep^b' that used 2D bidirectional sweep scans (4 trajectories), 'w zigzag^a' that used 1D bidirectional zigzag scans, 'w zigzag^b' that used 2D bidirectional zigzag scans, and 'w sweep^a-zigzag^a' that used an ensemble of sweep and zigzag scans. As listed in Table V, we find that I2I-Mamba outperforms all variants consistently across tasks ($p<0.05$), with improvements up to 1.6dB PSNR and 2.3% SSIM.

Assessment of receptive fields: To visually assess the influence of the spiral-scan trajectory on the effective receptive field of the SSM operator, we extracted per-sample activation maps for SSM layers [44]. Supp. Fig. 7 displays representative maps for I2I-Mamba (i.e., spiral scan), 'w sweep^a'

(i.e., 1D bidirectional raster scans), and 'w sweep^{b}' (i.e., 2D bidirectional raster scans). Since latent representations are processed in a sequence-dependent manner, regions of high activation may reflect abstract contextual dependencies rather than detailed anatomical structures. We observe that I2I-Mamba significantly improves angular isotropy of the SSM operator while maintaining coverage over a broad neighborhood of pixels. In contrast, 'w sweep^{a}', 'w sweep^{b}' suffer from notable bias across the primary scan directions (i.e., horizontal and/or vertical), diminishing focus over regions distanced obliquely to the scan directions.

For quantitative assessment, activation maps were transformed into a polar representation, partitioned into 10 radial bins and 20 angular wedges. Radial coverage (ρ) was measured as the percentage of radial distances with normalized activation values above 0.1 [73]. Angular isotropy (φ) was measured as $1 - CV$ across orientations where CV denotes coefficient of variation [73]. Supp. Table 3 lists ρ and φ across early-to-late bottleneck stages. Averaged across datasets and stages, I2I-Mamba attains comparable radial coverage ($\rho = 0.79$) to 'w sweep^{b}' ($\rho = 0.76$) and slightly higher coverage than 'w sweep^{a}' ($\rho = 0.74$). In contrast, I2I-Mamba consistently achieves substantially improved angular isotropy ($\varphi = 0.86$) relative to 'w sweep^{a}' ($\varphi = 0.52$) and 'w sweep^{b}' ($\varphi = 0.61$) ($p<0.05$). These findings indicate that the spiral-scan SSM operator maintains broad spatial coverage while markedly reducing directional bias compared to raster-based sequentialization strategies.

Other Design Parameters: Lastly, we examined the influence of several design parameters. To assess the importance of injecting ddMamba blocks in select bottleneck stages (i.e., $S = \{b1, b5, b9\}$), we compared I2I-Mamba against variants built by inserting ddMamba blocks in different configurations across the network as listed in Supp. Table 4. While all models generally yield similar SSIM values, we find that I2I-Mamba consistently yields superior PSNR against variants ($p<0.05$). Lower performance in variants with fewer ddMamba blocks can be attributed to a lowered ability to extract contextual features, whereas lower performance in variants with a greater number of ddMamba blocks is best attributed to elevated model complexity. These results suggest that I2I-Mamba attains a favorable trade-off between contextual sensitivity and model complexity by avoiding indiscriminate placement.

To assess the influence of the SSM state dimensionality N, we constructed I2I-Mamba variants with different values of N. As summarized in Supp. Table 5, the selected $N = 16$ achieves near-optimal performance, with results generally comparable across variants. The variants also exhibit similar computational characteristics, with only modest increases in model size and training time towards higher N.

Model Reliability: To assess the reliability of I2I-Mamba with respect to data partitioning, we conducted 5-fold cross-validation using non-overlapping subject-level splits. Supp. Table 6 shows that performance remains highly consistent, with standard deviations across folds below 0.09dB PSNR and 0.001% SSIM. These results suggest that I2I-Mamba generalizes reliably across different sets of subjects.

V. Discussion

A. Scope and Implications

In this work, we introduced a novel SSM model for imputing missing imaging modalities. CNNs lack sensitivity to long-range dependencies across distant anatomical regions [34], while transformers suffer from quadratic complexity that restricts attention to low-resolution or coarse features [39], [57]. SSMs have recently emerged as an efficient alternative, yet existing variants typically operate in the image domain with raster-scan trajectories, which constrain their capture of global and frequency-specific context and introduce directional bias in their receptive fields [74]. I2I-Mamba addresses these limitations via spiral-scan SSM operators that jointly process features in the image and Fourier domains, enhancing contextual sensitivity without incurring high computational costs.

I2I-Mamba's performance benefits might be of value in several clinical scenarios where efficient, high-fidelity synthesis of missing modalities is needed: (i) imputing redundant but diagnostically useful target modalities to reduce total scan time; (ii) inferring invasive or high-risk modalities—such as those requiring contrast agents or ionizing radiation—using safer, non-invasive scans [5]; and (iii) retrospectively recovering missing modalities to enhance harmonization across large-scale imaging datasets in population studies [6].

B. Limitations and Future Work

Several lines of limitations can be addressed to further improve performance. A first group concerns learning strategies. Here we examined one-to-one and many-to-one tasks in multi-contrast MRI and MRI-CT protocols. For optimal performance, a separate model was built for each individual task. In certain scenarios, missing and acquired modalities in a multi-modal protocol may vary sporadically across the imaging cohort [75]. To improve practicality, a unified I2I-Mamba can be built by adopting a masked training strategy so as to perform many-to-many synthesis [39], [75]. When needed, reliability of such multi-tasking models might be boosted via large multi-site datasets containing diverse training samples [72], [76]. Here, we performed supervised learning by assuming that paired sets of source and target images are available in each subject [25]. When it is difficult to curate paired sets, unsupervised learning based on cycle-consistency could be adopted [17], [77].

A second group concerns synthesis tasks. Here we observed high performance in translation among endogenous MRI contrasts and in translation from MRI to CT. Literature suggests that synthesizing exogenous MRI contrasts that involve external contrast agents, or synthesis of MRI images from CT are rather ill-posed [5], [39]. Incorporating regularization priors regarding the distribution of the target modality might improve fidelity in challenging tasks [7], [27]. Consistent with common practice in literature, here we considered synthesis tasks involving 2D cross-sections rather than full 3D volumes to maintain high computational efficiency [25]. While we did not observe noticeable inter-slice intensity inconsistencies or misalignment, 2D processing naturally limits the exploitation of contextual dependencies along the through-plane axis of

volumetric imaging data. Several practical strategies can be considered to extend I2I-Mamba to volumetric settings while controlling computational cost. One option is to reduce feature map resolution in the bottleneck when moving to 3D; however, such compression may compromise the preservation of fine anatomical detail. A more favorable compromise is 2.5D processing, where multiple adjacent slices are processed jointly, enabling limited through-plane contextual modeling while largely preserving the efficiency of 2D architectures. In this case, the proposed in-plane spiral trajectory could be combined with a linear sweep along the through-plane dimension. Alternatively, volumetric processing over moderately sized 3D patches could be considered, where non-raster scan trajectories such as outside-to-center traversal across concentric volumetric shells may be used. Systematic evaluation of these design choices and their accuracy-efficiency trade-offs remains an important direction for future work.

A third group concerns the employed loss functions. For systematic and efficient evaluations, we implemented primary competing methods based on a combined pixel-wise/adversarial loss [39]. This adversarial approach elicited high-quality synthetic images, and in preliminary studies we did not observe a notable benefit from a diffusion-based implementation. That said, further improvements might be viable with advanced loss terms including gradient-based, difficulty-aware, or cross-entropy losses [28], [78]–[80]. A potential limitation of adversarial learning is training instabilities that can hamper image fidelity [18]. While we did not observe signs of instability, diffusion learning could be adopted to improve reliability when necessary [52], [58]. Additional improvements for a diffusion-based implementation include bridge formulations to boost task-relevant information [35]. Lastly, synthesis performance might be further improved by adopting pre-training procedures.

Lastly, architectural variations can be considered within the ddMamba design. Here we adopted the same architectural template for the image-domain and Fourier-domain branches, while learning their parameters independently. This choice avoids imposing hand-crafted structural assumptions, and instead allows specialization to emerge naturally from data through learned parameters and input-adaptive state-space dynamics of SSM layers. Despite sharing the same macro-architecture, the two branches exhibit distinct behaviors when operating on spatial versus spectral representations. As demonstrated by our ablation and feature analysis results, this formulation is sufficient to induce complementary representations across the image and Fourier domains. Exploring more explicitly domain-specialized architectural variants remains an interesting direction for future investigation.

VI. Conclusion

In this study, we proposed a novel learning-based method for imputing missing modalities in multi-modal medical imaging protocols. To improve fidelity in synthetic images, I2I-Mamba leverages a novel SSM-based architecture with dual-domain Mamba blocks operating in image and Fourier domains to capture spatial and spectral contextual features. These ddMamba blocks are further equipped with spiral-scan trajectories to improve angular isotropy in the receptive field of SSM operators. With its superior performance and competitive computational efficiency with respect to state-of-the-art baselines, I2I-Mamba holds great potential for multi-modal medical image synthesis.

References

[1] B. J. Pichler, M. S. Judenhofer, and C. Pfannenberg, *Multimodal Imaging Approaches: PET/CT and PET/MRI*. Springer, 2008, pp. 109–132.

[2] B. Thukral, "Problems and preferences in pediatric imaging," *Ind J Rad Imaging*, vol. 25, pp. 359–364, 2015.

[3] J. E. Iglesias *et al.*, "Is synthesizing MRI contrast useful for inter-modality analysis?" in *MICCAI*, 2013, pp. 631–638.

[4] Y. Huo *et al.*, "Adversarial synthesis learning enables segmentation without target modality ground truth," in *ISBI*, 2018, pp. 1217–1220.

[5] D. Lee *et al.*, "CollaGAN: Collaborative GAN for missing image data imputation," in *CVPR*, 2019, pp. 2487–2496.

[6] L. T. Clark *et al.*, "Increasing Diversity in Clinical Trials: Overcoming Critical Barriers," *Cur. Prob. Cardiol.*, vol. 44, no. 5, pp. 148–172, 2019.

[7] J. Lee *et al.*, "Multi-atlas-based CT synthesis from conventional MRI with patch-based refinement for MRI-based radiotherapy planning," in *SPIE Med. Imag.*, 2017, p. 101331I.

[8] Y. Huang, L. Shao, and A. F. Frangi, "Simultaneous super-resolution and cross-modality synthesis of 3D medical images using weakly-supervised joint convolutional sparse coding," in *CVPR*, 2017, pp. 5787–5796.

[9] A. Adam *et al.*, *Grainger & Allison's Diagnostic Radiology*. Elsevier, 2014.

[10] R. Vemulapalli, H. Van Nguyen, and S. K. Zhou, "Unsupervised cross-modal synthesis of subject-specific scans," in *ICCV*, 2015, pp. 630–638.

[11] Y. Wu *et al.*, "Prediction of CT substitutes from MR images based on local diffeomorphic mapping for brain PET attenuation correction," *J Nuc Med*, vol. 57, no. 10, pp. 1635–1641, 2016.

[12] D. C. Alexander *et al.*, "Image quality transfer via random forest regression: Applications in diffusion MRI," in *MICCAI*, 2014, pp. 225–232.

[13] T. Huynh *et al.*, "Estimating CT image from MRI data using structured random forest and auto-context model," *IEEE Trans Med Imaging*, vol. 35, no. 1, pp. 174–183, 2016.

[14] C. Zhao *et al.*, "Whole brain segmentation and labeling from CT using synthetic MR images," in *MLMI*, 2017, pp. 291–298.

[15] A. Chartsias *et al.*, "Multimodal MR synthesis via modality-invariant latent representation," *IEEE Trans Med Imaging*, vol. 37, no. 3, pp. 803–814, 2018.

[16] ——, "Adversarial image synthesis for unpaired multi-modal cardiac data," in *SSMI*, 2017, pp. 3–13.

[17] J. Wolterink *et al.*, "Deep MR to CT synthesis using unpaired data," in *SSMI*, 2017, pp. 14–23.

[18] M. Özbey *et al.*, "Unsupervised medical image translation with adversarial diffusion models," *IEEE Trans Med Imaging*, vol. 42, no. 12, pp. 3524–3539, 2023.

[19] W. H. L. Pinaya *et al.*, "Generative AI for Medical Imaging: extending the MONAI Framework," *arXiv:2307.15208*, 2023.

[20] C. Bowles *et al.*, "Pseudo-healthy image synthesis for white matter lesion segmentation," in *SSMI*, 2016, pp. 87–96.

[21] T. Joyce, A. Chartsias, and S. A. Tsaftaris, "Robust multi-modal MR image synthesis," in *MICCAI*, 2017, pp. 347–355.

[22] H. Yang *et al.*, "Unpaired brain MR-to-CT synthesis using a structure-constrained cycleGAN," *IEEE Trans Med Imaging*, vol. 39, no. 12, pp. 4249–4261, 2020.

[23] W. Wei *et al.*, "Fluid-attenuated inversion recovery MRI synthesis from multisequence MRI using three-dimensional fully convolutional networks for multiple sclerosis," *J Med Imaging*, vol. 6, no. 1, p. 014005, 2019.

[24] A. Beers *et al.*, "High-resolution medical image synthesis using progressively grown generative adversarial networks," *arXiv:1805.03144*, 2018.

[25] S. U. Dar *et al.*, "Image synthesis in multi-contrast MRI with conditional generative adversarial networks," *IEEE Trans Med Imaging*, vol. 38, no. 10, pp. 2375–2388, 2019.

[26] B. Yu *et al.*, "3D cGAN based cross-modality MR image synthesis for brain tumor segmentation," in *ISBI*, 2018, pp. 626–630.

[27] ——, "Ea-gans: edge-aware generative adversarial networks for cross-modality mr image synthesis," *IEEE Trans Med Imaging*, vol. 38, no. 7, pp. 1750–1762, 2019.

[28] D. Nie *et al.*, "Medical image synthesis with deep convolutional adversarial networks," *IEEE Trans Biomed Eng*, vol. 65, no. 12, pp. 2720–2730, 2018.

[29] K. Armanious *et al.*, "MedGAN: Medical image translation using GANs," *Comput Med Imaging Graph*, vol. 79, p. 101684, 2019.

[30] H. Li *et al.*, "DiamondGAN: Unified multi-modal generative adversarial networks for MRI sequences synthesis," in *MICCAI*, 2019, pp. 795–803.

[31] T. Zhou *et al.*, "Hi-Net: Hybrid-fusion network for multi-modal MR image synthesis," *IEEE Trans Med Imaging*, vol. 39, no. 9, pp. 2772–2781, 2020.

[32] G. Wang *et al.*, "Synthesize high-quality multi-contrast magnetic resonance imaging from multi-echo acquisition using multi-task deep generative model," *IEEE Trans Med Imaging*, vol. 39, no. 10, pp. 3089–3099, 2020.

[33] M. Yurt *et al.*, "mustGAN: multi-stream generative adversarial networks for MR image synthesis," *Med Image Anal*, vol. 70, p. 101944, 2021.

[34] X. Zhang *et al.*, "PTNet3D: A 3D High-Resolution Longitudinal Infant Brain MRI Synthesizer Based on Transformers," *IEEE Trans Med Imaging*, vol. 41, no. 10, pp. 2925–2940, 2022.

[35] Y. Korkmaz, T. Cukur, and V. M. Patel, "Self-supervised mri reconstruction with unrolled diffusion models," in *MICCAI*, 2023, pp. 491–501.

[36] O. Oktay *et al.*, "Attention U-Net: Learning where to look for the pancreas," *arXiv:1804.03999*, 2018.

[37] H. Lan, A. Toga, and F. Sepehrband, "SC-GAN: 3D self-attention conditional GAN with spectral normalization for multi-modal neuroimaging synthesis," *bioRxiv:2020.06.09.143297*, 2020.

[38] H.-C. Shin *et al.*, "GANBERT: Generative adversarial networks with bidirectional encoder representations from transformers for MRI to PET synthesis," *arXiv:2008.04393*, 2020.

[39] O. Dalmaz, M. Yurt, and T. Çukur, "ResViT: Residual vision transformers for multi-modal medical image synthesis," *IEEE Trans Med Imaging*, vol. 44, no. 10, pp. 2598–2614, 2022.

[40] J. Liu *et al.*, "One Model to Synthesize Them All: Multi-Contrast Multi-Scale Transformer for Missing Data Imputation," *IEEE Trans Med Imaging*, vol. 42, no. 9, pp. 2577–2591, 2023.

[41] Y. Li *et al.*, "Multi-scale transformer network with edge-aware pre-training for cross-modality mr image synthesis," *IEEE Trans Med Imaging*, vol. 42, no. 11, pp. 3395–3407, 2023.

[42] H. A. Bedel *et al.*, "Bolt: Fused window transformers for fmri time series analysis," *Med Image Anal*, vol. 88, p. 102841, 2023.

[43] L. Zhu *et al.*, "Vision mamba: Efficient visual representation learning with bidirectional state space model," in *ICML*, 2024.

[44] Y. Liu *et al.*, "Vmamba: Visual state space model," in *NeurIPS*, 2025.

[45] J. Ma, F. Li, and B. Wang, "U-mamba: Enhancing long-range dependency for biomedical image segmentation," *arXiv:2401.04722*, 2024.

[46] J. Huang *et al.*, "Enhancing global sensitivity and uncertainty quantification in medical image reconstruction with monte carlo arbitrary-masked mamba," *Med Image Anal*, vol. 99, p. 103334, 2025.

[47] Y. Korkmaz and V. M. Patel, "Mambarecon: Mri reconstruction with structured state space models," in *IEEE/CVF WACV*, 2025, pp. 4142–4152.

[48] B. Bilgic, V. K. Goyal, and E. Adalsteinsson, "Multi-contrast reconstruction with Bayesian compressed sensing." *Magn Reson Med*, vol. 66, no. 6, pp. 1601–1615, 2011.

[49] O. F. Atli *et al.*, "Multi-contrast MR image synthesis with episodic state-space modeling," in *ISMRM*, 2025, p. 1119.

[50] J. Zhao *et al.*, "Tripartite-GAN: Synthesizing liver contrast-enhanced MRI to improve tumor detection," *Med Image Anal*, vol. 63, p. 101667, 2020.

[51] Z. Yuan *et al.*, "SARA-GAN: Self-attention and relative average discriminator based generative adversarial networks for fast compressed sensing MRI reconstruction," *Front Neuroinform*, vol. 14, p. 58, 2020.

[52] F. Arslan *et al.*, "Self-consistent recursive diffusion bridge for medical image translation," *Med Image Anal*, vol. 106, p. 103747, 2025.

[53] D. Yousaf *et al.*, "AI-driven framework for text neck syndrome detection using non-contact software-defined rf sensing and sequential deep learning," *Telecom Sys*, vol. 88, 07 2025.

[54] H. Bilal *et al.*, "Identification and diagnosis of chronic heart disease: A deep learning-based hybrid approach," *Alex Eng J*, vol. 124, pp. 470–483, 2025.

[55] ——, "An intelligent approach for early and accurate predication of cardiac disease using hybrid artificial intelligence techniques," *Bioeng*, vol. 11, no. 12, 2024.

[56] ——, "M-CNN-RF: A hybrid deep learning model for accurate pediatric skeletal age estimation using hand bone radiographs," *Alex Eng J*, vol. 129, pp. 289–301, 2025.

[57] J. Chen *et al.*, "Transunet: Rethinking the u-net architecture design for medical image segmentation through the lens of transformers," *Med Image Anal*, vol. 97, p. 103280, 2024.

[58] Z. Ju and W. Zhou, "Vm-ddpm: Vision mamba diffusion for medical image synthesis," *arXiv:2405.05667*, 2024.

[59] S. Chen *et al.*, "Coupling of state space modules and attention mechanisms: An input-aware multi-contrast mri synthesis method," *Med Phys*, vol. 52, no. 4, pp. 2269–2278, 2025.

[60] O. F. Atli *et al.*, "I2I-Mamba: Multi-modal medical image synthesis via selective state space modeling," *arXiv:2405.14022*, 2024.

[61] Y. Meng *et al.*, "DH-Mamba: Exploring dual-domain hierarchical state space models for MRI reconstruction," *arXiv:2501.08163*, 2025.

[62] P. Friedrich *et al.*, "Wdm: 3d wavelet diffusion models for high-resolution medical image synthesis," in *DGM4MICCAI*, 2025, pp. 11–21.

[63] K. He *et al.*, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[64] I. Tolstikhin *et al.*, "Mlp-mixer: An all-mlp architecture for vision," in *NeurIPS*, 2021.

[65] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *NeurIPS*, H. Larochelle *et al.*, Eds., vol. 33, 2020, pp. 6840–6851.

[66] H. Cao *et al.*, "Swin-unet: Unet-like pure transformer for medical image segmentation," in *Proc ECCV*, 2022.

[67] C. Ma and Z. Wang, "Semi-mamba-unet: Pixel-level contrastive and cross-supervised visual mamba-based unet for semi-supervised medical image segmentation," *Knowled Syst*, vol. 300, p. 112203, 2024.

[68] W. Lin *et al.*, "Mamba-convolutional UNet for multi-modal medical image synthesis," *Med Phys*, vol. 52, no. 10, p. e70029, 2025.

[69] J. Zhang and S. Jiang, "Hybrid State-Space Models and Denoising Training for Unpaired Medical Image Synthesis ," in *MICCAI*, vol. 15961, 2025, pp. 239 – 248.

[70] U. Baid *et al.*, "The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification," *arXiv:2107.02314*, 2021.

[71] T. Nyholm *et al.*, "MR and CT data with multiobserver delineations of organs in the pelvic area—part of the gold atlas project," *Med Phys*, vol. 45, no. 3, pp. 1295–1300, 2018.

[72] O. Dalmaz *et al.*, "One model to unite them all: Personalized federated learning of multi-contrast MRI synthesis," *Med Image Anal*, vol. 94, p. 103121, 2024.

[73] N. I. Fisher, *Statistical Analysis of Circular Data*. Cambridge Univ-Press, 1993.

[74] S. Ozturk, O. C. Duran, and T. Çukur, "DenoMamba: A fused state-space model for low-dose CT denoising," *IEEE J Biomed Health Inf*, pp. 1–14, 2025.

[75] A. Sharma and G. Hamarneh, "Missing MRI pulse sequence synthesis using multi-modal generative adversarial network," *IEEE Trans Med Imaging*, vol. 39, pp. 1170–1183, 2020.

[76] V. A. Nezhad *et al.*, "Generative autoregressive transformers for model-agnostic federated MRI reconstruction," *arXiv:2502.04521*, 2025.

[77] Y. Ge *et al.*, "Unpaired MR to CT synthesis with explicit structural constrained adversarial learning," in *ISBI*, 2019, pp. 1096–1099.

[78] B. Kabas *et al.*, "Physics-driven autoregressive state space models for medical image reconstruction," *arXiv:2412.09331*, 2024.

[79] B. Zhan *et al.*, "LR-cGAN: Latent representation based conditional generative adversarial network for multi-modality MRI synthesis," *Biomed. Signal Process. Control*, vol. 66, p. 102457, 2021.

[80] D. Nie and D. Shen, "Adversarial Confidence Learning for Medical Image Segmentation and Synthesis," *Int J Comput. Vision*, vol. 128, no. 10, pp. 2494–2513, 2020.