Physics-Driven Autoregressive State Space Models for Medical Image Reconstruction

Bilal Kabas, Fuat Arslan, Valiyeh A. Nezhad, *Graduate Student Member, IEEE*, Kader K. Oguz, Saban Ozturk, Emine U. Saritas, *Member, IEEE*, and Tolga Çukur, *Senior Member, IEEE*

***Abstract*— Medical image reconstruction from undersampled acquisitions is an ill-posed inverse problem requiring accurate recovery of anatomical structures from incomplete measurements. Physics-driven (PD) network models have gained prominence for this task by integrating data-consistency mechanisms with learned priors, enabling improved performance over purely data-driven approaches. However, reconstruction quality still hinges on the network's ability to disentangle artifacts from true anatomical signals—both of which exhibit complex, multi-scale contextual structure. Convolutional neural networks (CNNs) capture local correlations but often struggle with non-local dependencies. While transformers aim to alleviate this limitation, practical implementations involve design compromises to reduce computational cost by balancing local and non-local sensitivity, occasionally resulting in performance comparable to CNNs. To address these challenges, we propose MambaRoll, a novel physics-driven autoregressive state space model (SSM) for high-fidelity and efficient image reconstruction. MambaRoll employs an unrolled architecture where each cascade autoregressively predicts finer-scale feature maps conditioned on coarser-scale representations, enabling consistent multi-scale context propagation. Each stage is built on a hierarchy of scale-specific PD-SSM modules that capture spatial dependencies while enforcing data consistency through residual correction. To further improve scale-aware learning, we introduce a Deep Multi-Scale Decoding (DMSD) loss, which provides supervision at intermediate spatial scales in alignment with the autoregressive design. Demonstrations on accelerated MRI and sparse-view CT reconstructions show that MambaRoll provides improved performance relative to state-of-the-art CNN-, transformer-, and SSM-based methods.**

***Index Terms*— Medical image reconstruction, state space, autoregressive, physics-driven.**

Received 30 April 2026; revised 7 July 2026; accepted 17 July 2026. This work was supported in part by the Fellowships TUBA GEBIP 2015 and BAGEP 2017.The work of Tolga Çukur was supported by TUBITAK 1001 under Grant 121E488. Recommended by Associate Editor F. Lam. *(Corresponding author: Tolga Çukur.)*

Bilal Kabas, Fuat Arslan, Valiyeh A. Nezhad, Emine U. Saritas, and Tolga Çukur are with the Department of Electrical-Electronics Engineering, Bilkent University, TR-06800 Ankara, Türkiye (e-mail: bilal.kabas@bilkent.edu.tr; fuat.arslan@bilkent.edu.tr; valiyeh.ansarian@bilkent.edu.tr; saritas@ee.bilkent.edu.tr; cukur@ee.bilkent.edu.tr).

Kader K. Oguz is with the National Magnetic Resonance Research Center (UMRAM), Bilkent University, TR-06800 Ankara, Turkiye, also with the Department of Radiology, UC Davis Medical Center, California, CA 95817 USA, and also with Ankara Haci Bayram Veli University, TR-06570 Ankara, Turkiye (e-mail: kkarlioguz@gmail.com).

Saban Ozturk is with the National Magnetic Resonance Research Center (UMRAM), Bilkent University, TR-06800 Ankara, Turkiye (e-mail: sabn.ozturk@gmail.com).

Digital Object Identifier 10.1109/TMI.2026.3716153

I. Introduction

In MANY tomographic modalities, tissue structure is spatially encoded prior to acquisition, placing raw data in a distinct measurement domain (e.g., k-space in MRI, projection domain in CT). Higher spatial resolution requires collecting a greater number of encodings, but this increases scan time, patient discomfort, and/or radiation dose. A common strategy to improve scan efficiency is undersampling, i.e., acquiring only a subset of k-space points or projection views [1]. However, reconstructing images from undersampled data is inherently ill-posed, requiring inversion of the imaging operator linking measurement and image domains [2], [3]. Traditional methods often yield residual aliasing artifacts and noise during inversion [4], [5], highlighting the need for robust reconstruction techniques.

Among learning-based reconstruction methods, data-driven (DD) approaches train networks to directly map input images linearly recovered from undersampled data to ground-truth images from fully-sampled acquisitions [6], [7], [8], [9], [10]. While DD methods build conditional priors that benefit from task-specific feature extraction, ignoring the imaging operator can impair performance [11]. In comparison, physics-driven (PD) approaches employ joint objectives based on learned network modules for artifact suppression and data-consistency steps guided by the imaging operator, and they often demonstrate stronger performance [12], [13], [14], [15], [16], [17]. Unconditional PD variants (e.g., generative models) typically require intensive sampling at inference and are relatively difficult to deploy in resource-constrained clinical settings [18]. Instead, conditional PD approaches that train unrolled architectures kept frozen at test time offer a practical alternative [19], [20], [21], [22].

While the PD framework is analytically well established in image reconstruction, its empirical success hinges on the specific design of network modules used to suppress aliasing artifacts. Critically, accurate separation of artifacts from true anatomical features demands sensitivity to contextual relationships in medical images across multiple spatial scales [23], [24]. Convolutional neural networks (CNNs) are adept at capturing local patterns but often struggle with long-range dependencies [12], [25], [26]. Transformers offer improved modeling of non-local context, yet their substantial

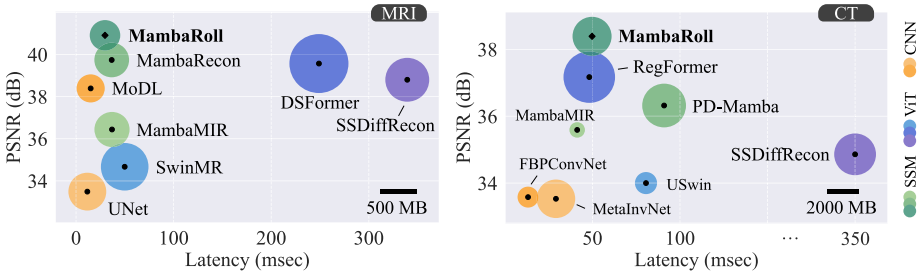

Fig. 1. Performance versus efficiency in MRI (left panel) and CT (right panel) reconstructions. Each method is represented by a circular marker whose center denotes the performance-latency trade-off (where latency is taken as inference time), and whose diameter indicates the memory load (see scale bars).

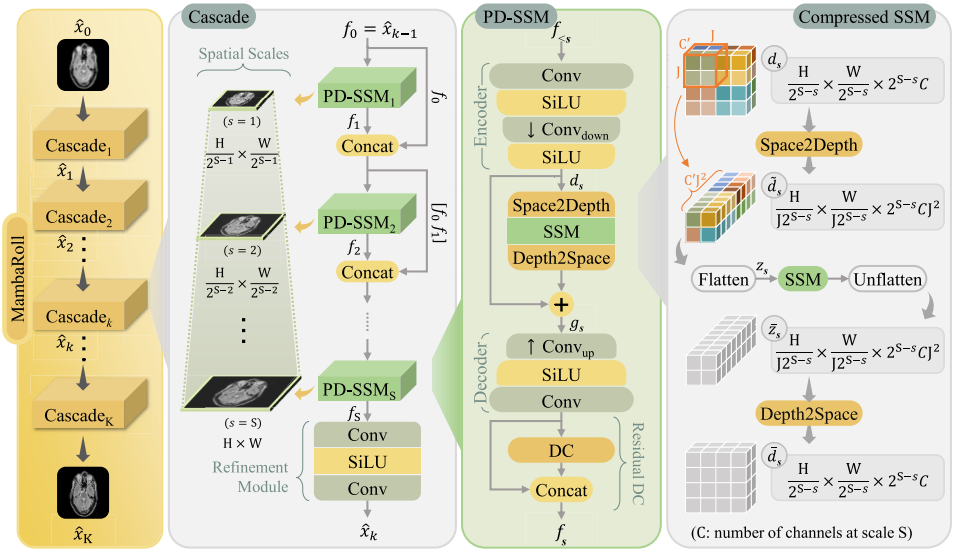

Fig. 2. MambaRoll employs an unrolled architecture with K cascades to map a linear reconstruction $\hat{x}_0$ to the learned output $\hat{x}_K$. Each cascade progressively recovers high-resolution contextualized features g_S across S spatial scales via autoregressive prediction (Fig. 3). At scale s, the proposed PD-SSM module convolutionally encodes input feature maps $f_{<s}$, captures contextualized feature representations through a compressed SSM block, decodes high-resolution feature maps, and residually enforces data fidelity. Following the final PD-SSM module within a cascade, a refinement module further restores fine textures.

computational cost frequently necessitates simplified designs that compromise either spatial resolution or contextual expressiveness [23], [27], [28], [29], [30], [31]. These trade-offs underscore the continued need for architectures that more effectively balance contextual sensitivity and computational efficiency [32].

State space models (SSMs) emerge as a promising alternative in this regard [33]. SSMs use rectilinear scans to rasterize images into one-dimensional (1D) sequences and model inter-pixel dependencies as a linear dynamical system with learnable parameters. This formulation promises stronger long-range sensitivity than CNNs and higher efficiency than transformers. Initial works have introduced SSMs in DD pipelines for medical image reconstruction [34], [35], where neglect of the imaging physics can limit generalization across different operators. Only very recently, a PD architecture has been explored for MRI reconstruction [36]. However, existing SSM methods rely on single-spatial-scale processing that can exhibit suboptimal sensitivity to distant pixel interactions, as such pixels may be separated by many intermediate elements when converted to 1D sequences [37], compromising comprehensive contextual feature capture across spatial scales.

Given that many existing reconstruction models use network backbones that process multi-scale feature representations (e.g., UNet, Swin Transformer), extending SSMs to operate on multi-scale feature maps could alleviate limitations elicited by single-scale processing. However, without explicit cross-scale linking mechanisms and supervision of intermediate representations, multi-scale feature maps alone may be insufficient for high-fidelity reconstruction. Naive linking strategies can lead to inconsistencies and artifacts degrading image quality [38], while conventional supervision focused solely on final reconstructed images might provide suboptimal guidance for learning accurate intermediate-scale representations [39].

To address these challenges, we introduce MambaRoll, a physics-driven autoregressive SSM for high-fidelity and efficient medical image reconstruction (Fig. 1). For this purpose, MambaRoll uniquely integrates three key advances: (i) Multi-scale state space modeling, where SSMs operate across hierarchical feature resolutions to overcome the limitations of single-scale modeling; (ii) Autoregressive linking across scales, where finer features conditionally refine coarse-scale predictions, encouraging consistent anatomical structure; and (iii) Physics-driven unrolled optimization, where each stage combines a learned PD-SSM module with operator-based data consistency steps and maintains high efficiency through compressed SSM blocks using space-to-depth transformation (Fig. 2). *To our knowledge, MambaRoll is the first medical*

image reconstruction framework to unify multi-scale SSMs, autoregressive modeling, and PD unrolled optimization in a single architecture.

To further enhance multi-scale learning, we introduce a Deep Multi-Scale Decoding (DMSD) loss that supervises scale-specific reconstructions at the last cascade of the unrolled network. Unlike conventional training strategies that only supervise the final output, DMSD provides auxiliary supervision at multiple scales. Comprehensive experiments on accelerated MRI and sparse-view CT demonstrate that MambaRoll yields performance benefits relative to state-of-the-art DD and PD methods based on CNN, transformer, or SSM backbones. Code for MambaRoll is publicly available at https://github.com/icon-lab/MambaRoll

A. Contributions

- We introduce a novel physics-driven SSM for medical image reconstruction that enhances contextual sensitivity and image fidelity through multi-scale autoregressive inference.
- We design an unrolled architecture that predicts feature maps at progressively finer spatial scales, with each level conditioned on coarser-scale outputs to support hierarchical context integration.
- The architecture comprises scale-specific PD-SSM modules that maintain efficiency via compressed SSM blocks, and data consistency via corrections at the image scale.
- We propose a DMSD loss that supervises intermediate representations at each spatial scale to improve overall reconstruction accuracy.

II. Related Work

A. SSM Reconstruction Models

SSMs have emerged as an efficient alternative to transformers with refined balance between local and non-local contextual sensitivity [34]. Several earlier studies have explored SSMs for MRI reconstruction [34], [35], [36], with only MambaMIR extending to CT and PET [34]. These works demonstrate SSM's potential in image reconstruction but typically employ either DD (MambaMIR and MMR-Mamba) [34], [35] or PD (MambaRecon) [36] designs based on single-spatial-scale SSM blocks, potentially causing suboptimal capture of multi-scale contextual features in medical images. In contrast, MambaRoll extends SSM-based reconstruction to operate across spatial scales with explicit cross-scale interactions.

After our preprint release [40], we became aware of two concurrent efforts exploring SSMs for MRI reconstruction. While these highlight growing interest in SSMs, the proposed approach differs in both architectural design and application scope. DH-Mamba proposes a DD SSM method with hierarchically tokenized scan trajectories to enhance contextual modeling [41]. Unlike DH-Mamba, MambaRoll employs a PD approach and explicitly integrates autoregressive priors across spatial scales to link multi-scale feature maps. CIMP-Net devises a PD method with spatial and spectral SSM modules for improved global context modeling [42]. Yet, both SSM modules are confined to single spatial scales, potentially limiting contextual reach. The proposed method instead adopts multi-scale processing with additional scale-aware supervision to improve contextual feature modeling. Finally, while most prior SSM-based methods are limited to MRI, except MambaMIR, MambaRoll is designed to generalize to other medical image reconstruction tasks under a unified framework.

B. Autoregressive Reconstruction Models

Autoregressive modeling decomposes joint probability distributions of element sequences into products of conditional distributions, such that each element is predicted conditioned on the preceding subset. Several previous CNN-based reconstruction studies have applied autoregressive modeling to capture dependencies between elements in raw pixel sequences [43], [44] or between temporal steps in diffusion generative models [45]. In contrast, the proposed approach models dependencies across spatial scales within a PD reconstruction framework.

C. Multi-Scale Models in Machine Learning

Recent work in the broader machine learning literature has explored multi-scale architectures based on SSMs as well as coarse-to-fine autoregressive prediction strategies. Several studies have proposed multi-scale SSM backbones for computer vision tasks to improve the modeling of spatial context across resolutions [46], [47], while autoregressive frameworks such as VAR generate images progressively across spatial scales using transformer-based representations [48]. These approaches highlight the benefits of combining hierarchical representations with scale-aware prediction strategies. While MambaRoll follows a similar high-level motivation of leveraging cross-scale-conditioned hierarchical representations, these prior approaches are primarily designed for computer vision tasks and typically do not incorporate physics-based constraints arising in inverse imaging problems. In contrast, the proposed method incorporates explicit data-fidelity constraints through the imaging forward operator for medical image reconstruction.

III. Theory

A. Medical Image Reconstruction

In many modalities, acquired data and the underlying image reflecting the spatial distribution of tissue structure are associated through a linear system [1]:

$$\mathcal{A}x + \varepsilon = y, \tag{1}$$

where $x \in \mathbb{C}^N$ denotes the underlying complex image in vector form (N: number of pixels), $y \in \mathbb{C}^M$ are acquired complex data in vector form (M: number of measurements), ε is measurement noise, and $\mathcal{A} \in \mathbb{C}^{M\times N}$ is the imaging operator that describes the relationship between image and measurement domains. For MRI scans, $\mathcal{A} = \mathcal{P}\mathcal{F}\mathcal{C}$ where $\mathcal{P}$ denotes the k-space sampling pattern, $\mathcal{F}$ denotes Fourier transformation, and $\mathcal{C}$ denotes coil sensitivities. Accordingly, $\mathcal{A}$ can be computed based on $\mathcal{P}$ and $\mathcal{F}$ that depend on

known scan parameters, and $\mathcal{C}$ can be estimated from data in a central calibration region [13]. For CT scans, $\mathcal{A}$ can be taken as a Radon transformation $\mathcal{R}$ [19]. Image reconstruction from acquired measurements involves solution of Eq. 1 by inverting $\mathcal{A}$, which becomes increasingly ill-conditioned as the number of measurements decline with respect to the number of image pixels [1]. Thus, image priors are employed to improve problem conditioning by regularizing reconstructions:

$$\hat{x} = \arg\min_{x} \; \|\mathcal{A}x - y\|_2^2 + R(x), \tag{2}$$

where $R(x)$ is an image prior that regularizes reconstructions based on the distribution of high-quality medical images.

In PD models, the solution of Eq. 2 is typically operationalized as iterations through cascades of an unrolled architecture G_{θ_k}, such that the output of cascade $k \in [1\ K]$ depends on the output of the previous cascade [22]:

$$\hat{x}_k = G_{\theta_k}(\hat{x}_{k-1}; \mathcal{A}, y), \tag{3}$$

where θ_k denotes the parameters of the kth cascade, and $\hat{x}_0 = \mathcal{A}^\dagger y$ is a linear reconstruction of undersampled data with $\dagger$ denoting the Hermitian adjoint. For a given cascade, $G_{\theta_k}(\hat{x}_{k-1}; \mathcal{A}, y) := \Psi_{dc}(\Psi_{nm}(\hat{x}_{k-1}); \mathcal{A}, y)$ interleaves projections through network and data-consistency (DC) modules:

$$v_{out} = \Psi_{nm}(v_{in}) = \arg\min_{v_{out}} R_{\theta_k}(v_{out}|v_{in}), \tag{4}$$

$$v_{out} = \Psi_{dc}(v_{in}; A, y) = v_{in} + \mathcal{A}^\dagger(y - \mathcal{A}v_{in}), \tag{5}$$

where the projection in Eq. 4 is a forward pass through the network module, v_{in} and v_{out} denote the projection input and output, respectively. Performance in PD models depends critically on the ability of Ψ_{nm} in separating undersampling-related artifacts from underlying tissue structure.

B. MambaRoll

MambaRoll employs PD-SSM modules to recover high-resolution medical images gradually across spatial scales, and a refinement module (RM) to enhance recovery of fine-grained tissue structure. Inspired by recent advances in autoregressive modeling for computer vision [48], we design PD-SSMs to autoregressively predict feature maps at finer spatial scales conditioned on coarser-scale representations, enhancing contextual representation by progressively aggregating multi-scale features while remaining consistent with acquired data.

1) Network Architecture: Receiving a linear reconstruction of undersampled data as a two-channel input $\hat{x}_0 = \mathcal{A}^\dagger y \in \mathbb{R}^{H\times W\times 2}$ with channels storing real and imaginary components (H: image height, W: image width), MambaRoll projects its input through K cascades comprising *novel PD-SSM modules* and refinement modules to compute the reconstruction $\hat{x}_K \in \mathbb{R}^{H\times W\times 2}$. At cascade k, the input image $f_0 = \hat{x}_{k-1} \in \mathbb{R}^{H\times W\times 2}$ is projected through S consecutive PD-SSM modules that autoregressively process feature maps across multiple spatial scales (Fig. 2):

$$f_s = \text{PD} - \text{SSM}_s(f_{<s}), \quad \text{for} \;\; s \in \{1, 2, \ldots, S\}, \tag{6}$$

where s denotes the scale index, $f_s \in \mathbb{R}^{H\times W\times 4}$ for $s \geq 1$ (4 channels due to residual DC blocks expressed in Eq. 13),

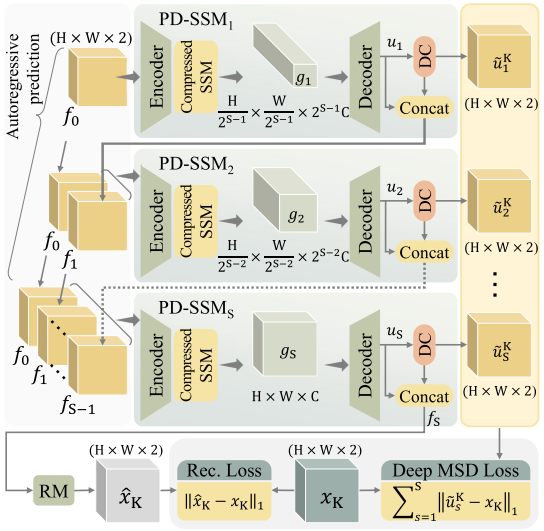

Fig. 3. To explicitly link multi-scale representations via an autoregressive prior, the PD-SSM module at a given scale s predicts the finer-scale feature map (f_s) conditioned on the preceding set of feature maps obtained at coarser scales ($f_{<s} := [f_0\ f_1\ \ldots\ f_{s-1}]$). For model supervision, a deep multi-scale decoding (DMSD) loss.

and $f_{<s} := [f_0\ f_1\ \ldots\ f_{s-1}] \in \mathbb{R}^{H\times W\times(4s-2)}$ is formed by concatenating feature maps from PD-SSMs at earlier scales. To ensure PD-SSMs process feature maps while maintaining data fidelity, f_s for all s carry the same spatial dimensionality as the input image $\hat{x}_0$. *Thus, processing at different spatial scales is achieved by adjusting resolution internally within PD-SSMs (Fig. 3).* Afterwards, the RM module enhances recovery of structural details via convolutional layers:

$$\hat{x}_k = \text{RM}(f_S) := \text{Conv}(\sigma(\text{Conv}(f_S))), \tag{7}$$

where $\hat{x}_k \in \mathbb{R}^{H\times W\times 2}$ is the output of the kth cascade, and σ is a SiLU activation function.

PD-SSM modules. Diverging from conventional SSM modules, PD-SSMs enable progressive image recovery across multiple spatial scales while maintaining data fidelity at each scale. A given PD-SSM$_s$ module at scale s first projects its input $f_{<s} \in \mathbb{R}^{H\times W\times 4s-2}$ through a convolutional encoder to map the high-resolution input onto the desired scale:

$$d_s = \text{Enc}_s(f_{<s}) := \sigma(\text{Conv}_{\text{down}}(\sigma(\text{Conv}(f_{<s})))), \tag{8}$$

where $d_s \in \mathbb{R}^{H_s\times W_s\times C_s}$ with $H_s = H/2^{(S-s)}$, $W_s = W/2^{(S-s)}$, $C_s = 2^{(S-s)}C$ (C: number of channels at scale S), and $\text{Conv}_{\text{down}}$ performs spatial downsampling while Conv expands channel dimensionality to C_s.

A compressed SSM block with residual connection extracts contextual features at the sth scale as $g_s = \text{SSM}(d_s)$, while improving efficiency in state-space modeling. Since the computational complexity of SSMs scales with sequence length (number of image pixels), the compressed SSM block first applies a space-to-depth transformation to the input feature map d_s. This operation rearranges local $J \times J$ spatial neighborhoods into the channel dimension, effectively reducing the spatial resolution by a factor of J in each dimension while increasing the number of channels by J^2 [50]. As a result, the sequence length is reduced by a factor of J^2. The transformed

map is then reshaped into a sequence $z_s \in \mathbb{R}^{H_s W_s / J^2 \times C_s J^2}$ via a two-dimensional sweep scan, and independently processed across channels with a discretized SSM:

$$h_s[n,c] = \mathbf{A} h_s[n-1,c] + \mathbf{B} z_s[n,c], \tag{9}$$

$$\bar{z}_s[n,c] = \mathbf{C} h_s[n,c], \tag{10}$$

where $n \in [1\ H_s W_s / J^2]$ denotes the sequence index, $h_s[n,c] \in \mathbb{R}^{D\times 1}$ denotes the scale-specific hidden state at step n, and $\mathbf{A} \in \mathbb{R}^{D\times D}$, $\mathbf{B} \in \mathbb{R}^{D\times 1}$, $\mathbf{C} \in \mathbb{R}^{1\times D}$ are learnable parameters (D: state dimensionality). After the SSM layer, the output sequence $\bar{z}_s$ is reshaped back to the spatial domain and a depth-to-space transformation is applied to restore the original resolution by redistributing channel information into spatial locations [50]. The restored feature map $\bar{d}_s \in \mathbb{R}^{H_s \times W_s \times C_s}$ is then residually added onto the input:

$$g_s = \bar{d}_s + d_s. \tag{11}$$

To enable PD-SSM modules to enforce fidelity to acquired data, the contextualized feature map g_s is projected through a convolutional decoder to recollect a high-resolution image:

$$u_s = \mathrm{Dec}_s(g_s) := \mathrm{Conv}(\sigma(\mathrm{Conv}_{\mathrm{up}}(g_s))), \tag{12}$$

where $u_s \in \mathbb{R}^{H\times W\times 2}$, $\mathrm{Conv}_{\mathrm{up}}$ performs spatial upsampling while Conv reduces channel dimensionality to 2 to store real and imaginary components. Finally, the recollected image can be projected through a residual DC block that concatenates u_s with a data-consistent version of itself:

$$\tilde{u}_s = \mathrm{DC}(u_s) := u_s + \mathcal{A}^{\dagger}(y - \mathcal{A}u_s), \tag{13}$$

$$f_s = \mathrm{resDC}(u_s) := [u_s \quad \tilde{u}_s]. \tag{14}$$

DC enforcement is applied once per scale within each cascade through the adjoint residual correction in Eq. 13, yielding a closed-form update for $\tilde{u}_s$ without inner iterations or additional parameters. Rather than replacing u_s with $\tilde{u}_s$, f_s concatenates both representations so that subsequent layers can adaptively balance the data-consistent estimate $\tilde{u}_s$ with the data-driven prediction u_s. This allows the network to adaptively modulate the strength of DC enforcement across stages and scales, which can help compensate for inaccuracies in the imaging operator.

2) Deep Multi-Scale Decoding (DMSD) Loss: MambaRoll leverages autoregressive processing across spatial scales to capture multi-scale contextual features. To promote enhanced learning at intermediate scales, we introduce a DMSD loss on representations at the final cascade so that earlier cascades remain primarily governed by autoregressive interactions and data-consistency updates. Accordingly, the feature map g_s^K at scale s within the final cascade (i.e., Kth) is decoded to a full-resolution image $u_s^K = \mathrm{Dec}_s(g_s^K)$ through a learned decoder (Fig. 3). These decoded images are then projected through the DC block, and supervised against the ground-truth image x_K to enable direct learning at multiple scales:

$$\mathcal{L}_{\mathrm{DMSD}} = \sum_{s=1}^{S} \|\tilde{u}_s^K - x_K\|_1 = \sum_{s=1}^{S} \|\mathrm{DC}(\mathrm{Dec}_s(g_s^K)) - x_K\|_1. \tag{15}$$

The decoding operators $\mathrm{Dec}_s(\cdot)$ used to compute the DMSD loss share weights with the corresponding decoding modules already present in the PD-SSM blocks. As such, DMSD does not introduce additional learnable parameters and instead provides auxiliary supervision for the existing scale-specific representations. By supervising a strictly data-consistency enforced version of decoded images (i.e., $\tilde{u}_s^K$) rather than the residually-corrected images (i.e., f_s^K), the DMSD loss provides auxiliary supervision for scale-specific predictions that may help improve reconstruction quality. Meanwhile, the decoded outputs u_s^K (prior to the DC projection) continue to participate in the autoregressive processing across scales, allowing intermediate representations to evolve without being directly constrained by the auxiliary supervision.

The overall loss then combines the reconstruction loss on the predicted image with Eq. 15:

$$\mathcal{L}_{\mathrm{total}} = \|\hat{x}_K - x_K\|_1 + \mathcal{L}_{\mathrm{DMSD}}, \tag{16}$$

where $\hat{x}_K$ denotes the output image predicted by the model. A fixed unit weighting was used for $\mathcal{L}_{\mathrm{DMSD}}$ in all experiments. We did not tune an additional loss weight, as this setting provided consistently strong validation performance while avoiding an extra hyperparameter. DMSD loss departs from conventional supervision by treating scale-wise predictions not merely as intermediate latent steps, but as independently decodable image estimates. This design is intended to reduce residual artifacts within feature maps at coarser spatial scales.

IV. Methods

A. Datasets

We conducted evaluations on public datasets with random subject-level splits into independent training, validation and test sets. **Accelerated MRI** experiments used brain data from fastMRI [52]. T_1, T_2 and FLAIR contrasts were analyzed as multi-coil complex k-space data, each volume comprising 10 cross-sections of size 320×320. Multi-coil data were compressed onto 5 virtual coils preserving 90% of variance [53], with coil sensitivities estimated via ESPIRiT [54]. Data were split into (240, 60, 120) subjects for training, validation, test sets, corresponding to (7200, 1800, 3600) cross-sectional samples. Ground-truth k-space data were undersampled at rates $R = 4-12$ via two-dimensional variable-density patterns following a Gaussian probability density function [55]. Linear reconstructions were obtained by zero-filling missing k-space samples followed by inverse Fourier transformation (ZF).

Sparse-view CT experiments used lung data from LoDoPaB-CT [56]. Each volume comprised 90 cross-sections of size 352×352. Data were split into (30, 6, 12) subjects for training, validation, test sets, corresponding to (2700, 540, 1080) cross-sectional samples. For sparse-view CT scans, ground-truth reconstructions were Radon transformed using two-dimensional parallel-beam geometry. Ground-truth sinograms were undersampled at rates $R = 4-8$ via uniform patterns across the angular dimension. Linear reconstructions used filtered back projection (FBP).

All experiments were conducted on the fastMRI and LoDoPaB-CT datasets using the splits described above, with the exception of cross-population generalization experiments. These were conducted separately for MRI and CT using

TABLE I
MRI RECONSTRUCTION PERFORMANCE AT $R = 4 - 12$ ON FASTMRI. PSNR (DB) AND SSIM (%) LISTED AS MEAN±STD ACROSS THE TEST SET. BOLDFACE MARKS THE TOP PERFORMING METHOD. (DD: DATA-DRIVEN, PD: PHYSICS-DRIVEN, CNN: CONVOLUTIONAL, VIT: TRANSFORMER, SSM: STATE-SPACE)

	Framework		Architecture			$R = 4$		$R = 8$		$R = 12$		Average	
	DD	PD	CNN	ViT	SSM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
UNet [49]	✓		✓			36.25±2.42	94.27±3.53	33.93±2.38	91.63±4.98	30.29±5.43	89.63±5.69	33.49	91.84
MoDL [22]		✓	✓			41.76±2.55	97.64±2.08	37.56±2.60	95.03±3.69	35.87±2.41	93.19±4.48	38.39	95.28
SwinMR [29]	✓			✓		37.31±2.38	94.63±2.61	34.10±2.18	91.58±3.90	32.60±2.06	89.65±4.47	34.67	91.95
SSDiffRecon [32]		✓		✓		42.47±2.99	97.63±2.14	37.99±2.66	94.79±3.71	35.95±2.45	93.12±4.54	38.80	95.18
DSFormer [30]		✓		✓		43.13±3.12	97.91±2.12	38.69±2.84	95.04±3.74	36.90±2.71	93.23±4.62	39.57	95.39
MambaMIR [34]	✓				✓	38.58±2.60	95.12±3.23	36.05±2.56	92.66±4.88	34.70±2.47	91.38±5.61	36.44	93.05
MambaRecon [36]		✓			✓	43.45±3.31	98.16±2.31	38.74±2.85	95.57±3.94	37.03±2.72	93.99±4.69	39.74	95.91
MambaRoll		✓			✓	**45.55±3.71**	**98.57±3.57**	**39.51±3.04**	**96.12±3.71**	**37.68±2.77**	**94.61±4.37**	**40.91**	**96.43**

fastMRI+, containing annotated structural brain abnormalities [57], and LIDC-IDRI, containing annotated lung nodules and related lesions [58], respectively. The training, validation, and test split sizes were chosen to match those used in the primary experiments. In this setting, the training and validation sets contained only healthy cases (i.e., without reported findings), while the test set contained only patient cases (i.e., with pathology findings).

B. Architectural Details

MambaRoll used $S = 3$ spatial scales (at 0.25, 0.5, 1 scale of the original image resolution). In PD-SSM modules, the encoder used expansion factors across the channel dimension of (4, 2, 1), whereas the decoder used expansion factors of (0.25, 0.5, 1) across scales. The SSM block used a compression factor of $J = 4$ and a sweep-scan trajectory to project feature maps onto a sequence, and a state-dimensionality of $D = 16$. The refinement module comprised a convolutional block. All convolution layers used 3×3 kernels and SiLU activation.

C. Competing Methods

MambaRoll was demonstrated against state-of-the-art baselines in MRI and CT reconstruction, including both DD and PD methods. **For accelerated MRI reconstruction**, competing methods included CNNs (data-driven UNet [49] and physics-driven MoDL [22]); transformers (data-driven SwinMR [29], physics-driven SSDiffRecon [32] and DSFormer [30]); and SSMs (data-driven MambaMIR [34], physics-driven MambaRecon [36]). **For sparse-view CT reconstruction**, competing methods included CNNs (data-driven FBPConvNet [19], physics-driven MetaInv-Net [15]); transformers (data-driven USwin [51], physics-driven SSDiffRecon [32] and RegFormer [28]); and SSMs (data-driven MambaMIR [34], physics-driven PD-Mamba based on [37]).

D. Modeling Procedures

All experiments were conducted on individual cross-sections for efficient implementation; accordingly, two-dimensional (2D) network models were used. The models used two separate input-output channels to represent real and imaginary components of MR images, and a single input-output channel for CT images. To ensure fair comparisons, baseline models were implemented using the network architectures described in their original publications. Baseline methods were trained using an ℓ_1 reconstruction loss as a common objective across methods to enable controlled comparison of architectural differences. Note that MambaRoll employs the same reconstruction loss, with the proposed DMSD loss added as an auxiliary supervision term aligned with its multi-scale design. SSDiffRecon was adapted from its original self-supervised formulation by replacing the masking-based k-space loss with a supervised image-domain reconstruction loss to ensure consistency with the training setting of other methods. Key hyperparameters, including the number of cascades in unrolled architectures (K) and learning rate, were tuned for each method based on performance in the validation set. Accordingly, $K = 5$ in MRI reconstruction and $K = 3$ in CT reconstruction yielded near-optimal performance consistently across methods, whereas learning rates in the range 10^{-6} to 10^{-3} were selected per method based on validation performance. Modeling was performed via the PyTorch framework on an Nvidia RTX 4000 (Ada) GPU. Models were trained and tested on two-dimensional cross sections. Training was conducted using the Adam optimizer for 50 epochs, $(\beta_1, \beta_2) = (0.9, 0.999)$. Performance was assessed by peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) between recovered and reference images. Reference images were derived via Fourier reconstruction of fully-sampled acquisitions in MRI, and via filtered back-projection of fully-sampled sinograms in CT. Significance of performance differences was assessed via Wilcoxon signed-rank tests.

V. RESULTS

A. Accelerated MRI Reconstruction

Demonstrations were first performed for accelerated MRI reconstruction. MambaRoll was compared against state-of-the-art models based on convolutional (UNet, MoDL), transformer (SwinMR, SSDiffRecon, DSFormer), and SSM (MambaMIR, MambaRecon) backbones. Reconstruction performances on fastMRI at undersampling rates $R = 4-12$ are listed in Table I. MambaRoll yields higher performance metrics than competing methods across reconstruction tasks ($p < 0.05$). On average across tasks, MambaRoll provides improvements of 2.82dB PSNR and 1.95% SSIM relative to SSM baselines. Note that PD methods generally yield improved performance against

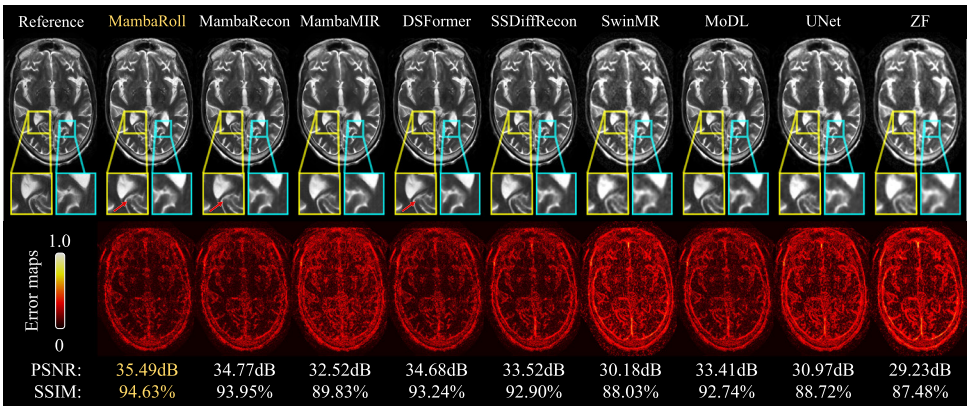

Fig. 4. Accelerated MRI reconstructions of a representative cross-section from a T_2-weighted acquisition in fastMRI at R = 12. Reconstructed images from competing methods are shown along with a linear reconstruction of undersampled data (ZF), and the reference image derived from fully-sampled data. Magnified views (demarcated by colored squares), arrows, and error maps are provided to highlight comparative differences in tissue depiction. PSNR and SSIM values are listed below each method.

TABLE II
CT RECONSTRUCTION PERFORMANCE AT $R = 4 - 8$ ON LODOPAB-CT. BOLDFACE MARKS THE TOP PERFORMING METHOD

	Framework		Architecture			$R = 4$		$R = 6$		$R = 8$		Average	
	DD	PD	CNN	ViT	SSM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
FBPConvNet [19]	✓		✓			34.41±1.65	85.68±4.92	33.85±1.73	84.86±5.75	32.49±1.57	81.76±5.94	33.58	84.10
MetaInv-Net [15]		✓	✓			36.31±2.23	90.03±5.72	33.10±1.68	84.82±6.61	31.17±1.38	80.90±6.82	33.53	85.25
USwin [51]	✓			✓		36.41±2.58	91.96±5.33	33.48±1.94	87.41±6.24	32.11±1.68	84.21±6.48	34.00	87.86
SSDiffRecon [32]		✓		✓		38.25±2.75	91.95±4.96	34.92±2.11	86.97±6.01	31.40±1.54	80.99±7.19	34.86	86.64
RegFormer [28]		✓		✓		39.88±3.23	93.33±4.89	36.85±2.58	89.49±5.99	34.79±2.17	86.29±6.39	37.17	89.70
MambaMIR [34]	✓				✓	37.42±2.40	89.57±4.77	35.42±2.20	86.73±5.89	33.92±1.99	83.74±6.19	35.59	86.68
PD-Mamba [37]		✓			✓	38.99±2.97	92.64±5.08	35.90±2.33	88.35±6.18	34.06±1.99	85.27±6.65	36.32	88.75
MambaRoll		✓			✓	**40.72±3.56**	**94.05±4.86**	**38.00±2.93**	**91.15±5.97**	**36.44±2.62**	**89.06±6.49**	**38.39**	**91.42**

DD methods (MambaMIR, SwinMR, UNet). Collectively, these results suggest that incorporating PD-SSM modules can enhance reliability in MRI reconstruction.

Representative reconstructions are shown in Fig. 4, along with annotations and error maps to highlight differences in tissue depiction. Among DD baselines, MambaMIR shows a degree of blur and pixelation artifacts, while SwinMR and UNet exhibit visible blur, noise and residual artifacts. Meanwhile, PD methods tend to reduce noise and artifacts, and among these methods, reconstructions are generally comparable with subtle qualitative differences. MambaRecon and DSFormer exhibit variations in tissue contrast that can occasionally lead to reduced boundary definition near small structures (annotated with arrows in Fig. 4), and SSDiffRecon and MoDL show a degree of edge blurring. MambaRoll tends to offer improved tissue delineation with reduced artifacts, which may reflect improved capture of multi-scale contextual features.

B. Sparse-View CT Reconstruction

Next, demonstrations were performed for sparse-view CT reconstruction. MambaRoll was compared against state-of-the-art models based on convolutional (FBPConvNet, MetaInv-Net), transformer (USwin, SSDiffRecon, RegFormer), and SSM (MambaMIR, PD-Mamba) backbones. Performances on LoDoPaB-CT at $R = 4 - 8$ are listed in Table II. MambaRoll achieves higher performance metrics than competing methods across reconstruction tasks ($p < 0.05$). On average across tasks, MambaRoll provides improvements of 2.44dB PSNR and 3.70% SSIM relative to SSM baselines. While PD methods generally offer performance benefits over DD counterparts (e.g., PD-Mamba vs MambaMIR, RegFormer vs USwin), performance differences are moderate compared to MRI. MambaRoll's benefits over convolutional baselines are notably stronger in CT, which might be attributed to the streaking-like aliasing artifacts spanning long distances, compromising artifact suppression in convolutional methods with locality biases.

Representative images are displayed in Fig. 5. DD baselines show notable reconstruction errors, including streaking artifacts (USwin, FBPConvNet) and grid-like pixelation artifacts (MambaMIR). PD baselines (PD-Mamba, SSDiffRecon, RegFormer, MetaInvNet) generally reduce these artifacts but still exhibit residual streaking that can lead to contrast inaccuracies near bright tissue structures. In comparison, MambaRoll tends to provide more effective artifact mitigation and improved fidelity in tissue depiction.

C. Generalization Under Acquisition and Population Variability

To investigate the robustness of reconstruction models under distributional shifts, we evaluated generalization performance across varying undersampling rates, and across subject populations. As representative cross-rate scenarios, we considered MRI models trained at $R = 8$ and tested at $R = 12$ on fastMRI, and CT models trained at $R = 6$ and tested at $R = 8$ on

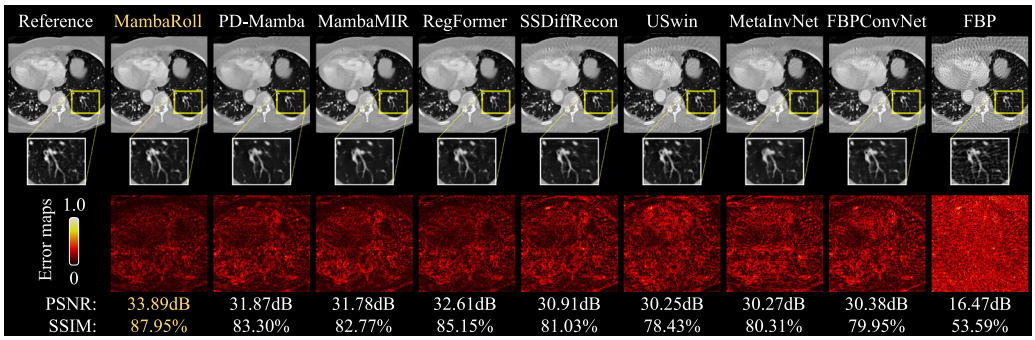

Fig. 5. Sparse-view CT reconstructions of a representative cross-section in LoDoPaB-CT at $R = 8$. Reconstructed images from competing methods are shown along with a linear reconstruction of undersampled data (FBP), and the reference image derived from fully-sampled data.

TABLE III
GENERALIZATION PERFORMANCE FOR MRI (FASTMRI) AND CT (LODOPAB-CT) RECONSTRUCTIONS UNDER SHIFTS IN UNDERSAMPLING RATES ($R_{TRAINING} \rightarrow R_{TEST}$)

	MRI ($R: 8 \rightarrow 12$)	
	PSNR	SSIM
UNet	31.31±1.98	86.84±6.07
MoDL	34.09±2.18	90.85±4.72
SwinMR	32.01±1.90	87.97±5.16
SSDiffRecon	34.44±2.29	90.69±4.84
DSFormer	35.00±2.44	91.02±4.76
MambaMIR	33.01±2.20	87.89±5.70
MambaRecon	35.45±2.49	91.99±4.60
MambaRoll	**35.93±2.55**	**92.68±4.38**

	CT ($R: 6 \rightarrow 8$)	
	PSNR	SSIM
FBPConvNet	31.29±1.48	79.44±6.21
MetaInvNet	30.72±1.46	79.96±6.68
USwin	28.91±1.94	78.04±6.55
SSDiffRecon	29.36±1.33	80.57±6.28
RegFormer	33.90±1.93	85.03±6.29
MambaMIR	32.80±1.67	82.29±5.87
PD-Mamba	33.38±1.84	84.05±6.42
MambaRoll	**34.68±2.37**	**85.81±6.34**

TABLE IV
GENERALIZATION PERFORMANCE FOR MRI (FASTMRI+, $R = 4$) AND CT (LIDC-IDRI, $R = 6$) RECONSTRUCTIONS UNDER SHIFTS IN SUBJECT POPULATION (TRAINING ON HEALTHY SUBJECTS AND TESTING ON PATIENTS)

	MRI	
	PSNR	SSIM
UNet	33.88±1.38	77.98±3.15
MoDL	38.78±2.61	96.19±2.28
SwinMR	36.77±2.15	94.38±2.80
SSDiffRecon	40.27±2.63	96.90±2.28
DSFormer	40.88±2.78	97.14±2.19
MambaMIR	35.22±1.53	76.95±2.94
MambaRecon	41.43±2.72	97.57±1.97
MambaRoll	**43.51±3.11**	**98.21±1.86**

	CT	
	PSNR	SSIM
FBPConvNet	33.68±0.59	84.17±3.66
MetaInvNet	34.02±0.65	87.58±1.78
USwin	35.04±1.43	89.63±1.73
SSDiffRecon	35.61±0.56	89.00±1.70
RegFormer	37.56±0.73	91.50±1.77
MambaMIR	35.91±0.56	88.02±1.79
PD-Mamba	36.63±0.58	90.43±1.71
MambaRoll	**38.54±0.82**	**92.86±1.79**

LoDoPaB-CT. As representative cross-population scenarios, we considered models trained on healthy subjects and tested on patients, while using fastMRI+ for MRI ($R = 4$) and LIDC-IDRI for CT ($R = 6$). Performance metrics are listed in Table III for cross-rate scenarios, and in Table IV for cross-population scenarios. We find that MambaRoll yields higher performance metrics than respective baselines in all cases ($p < 0.05$). Compared to SSM baselines, MambaRoll offers average improvements of 1.65dB PSNR, 2.69% SSIM in cross-rate tasks, and 3.73dB PSNR, 7.29% SSIM in cross-population tasks.

Representative images for cross-rate and cross-population settings are depicted in Figs. 6 and 7, respectively. Baseline methods exhibit varying degrees of residual streaking artifacts, which can affect structural fidelity. In comparison, MambaRoll shows relatively more effective suppression of residual artifacts. Consistent with the quantitative results, these observations suggest that MambaRoll can offer improvements in reconstruction quality relative to baselines, under shifts in undersampling-rate and subject population.

D. Radiological Evaluation

To assess perceived image quality beyond quantitative metrics (i.e., PSNR, SSIM), we conducted an observer study in which a board-certified radiologist (> 25 year experience) evaluated the similarity of reconstructed images to reference images obtained from fully-sampled data. Assessments were performed using a 5-point Likert scale (1: lowest score, 5: highest score), while being blinded to method identity. Radiological opinion scores of competing methods for $R = 12$ in MRI (fastMRI) and $R = 8$ in CT (LoDoPaB-CT), across twenty randomly selected samples spanning the test set, are summarized in Fig. 8. MambaRoll achieves higher scores than all baseline methods ($p < 0.05$), suggesting consistent improvements in perceived image quality. The score improvements are relatively moderate in MRI (0.20 higher than the second-best method), with reader feedback indicating subtle differences in depiction of tissue-to-fluid boundaries (e.g., along the ventricular walls and cortical sulci), and in edge blurring and residual artifacts. In contrast, score improvements are more pronounced in CT (0.65 higher than the second-best method), with reader feedback indicating differences in depiction of lung parenchyma (e.g., peripheral bronchovascular bundles), and in smoothing and streaking artifacts across hilar and mediastinal regions. These results suggest that MambaRoll can improve perceived image quality under radiological assessment. Given the limited scope of the observer study, however, future studies with multiple readers and larger cohorts are warranted to further evaluate its radiological utility.

E. Contextual Awareness

To assess spatial context utilization during reconstruction, we examined contextual awareness—the model's ability to integrate information from spatially distant but contextually dependent regions. In medical image reconstruction, such contextual integration is critical for capturing extended anatomical structures and reducing local ambiguities under undersampled

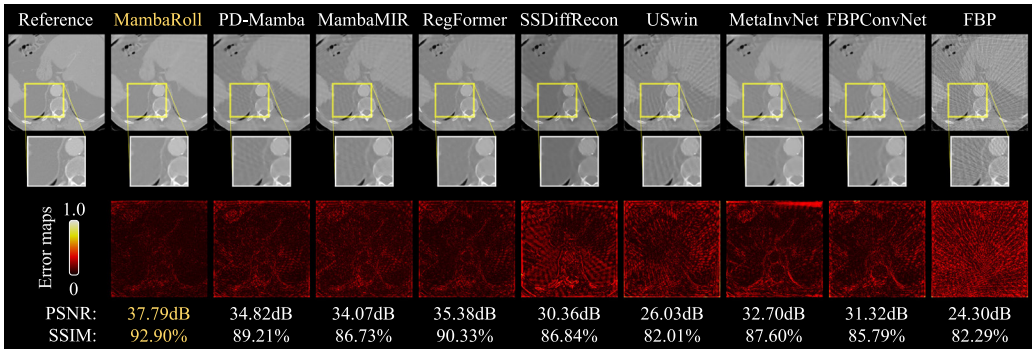

Fig. 6. Sparse-view CT reconstructions of a representative cross-section in LoDoPaB-CT at $R = 8$, with models trained at $R = 6$.

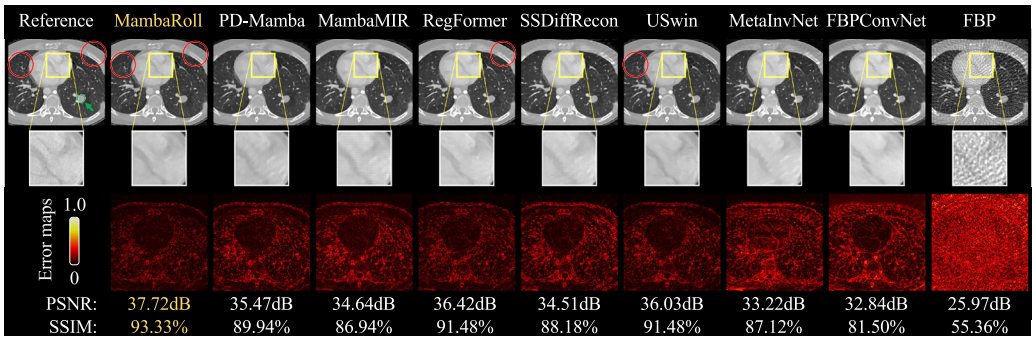

Fig. 7. Sparse-view CT reconstructions of a representative cross-section in LIDC-IDRI at $R = 6$ with models trained on healthy subjects and tested on patients. Circular annotations are provided to highlight subtle comparative differences in tissue depiction. The green contour on the reference image delineates a malignant nodule in this patient.

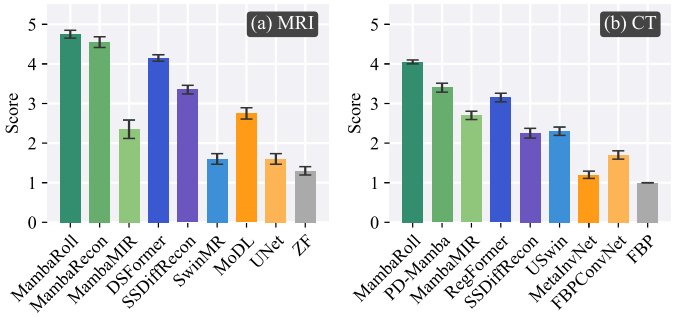

Fig. 8. Radiological opinion scores of competing methods for **(a)** MRI reconstruction at $R = 12$, **(b)** CT reconstruction at $R = 8$. Bar plots show mean±se of opinion scores.

or corrupted measurements. As a proxy for contextual awareness, we computed effective receptive fields (ERFs) that indicate how strongly different input pixels influence predictions at a given output pixel. ERF maps were obtained by computing gradients of the central pixel in the reconstructed image with respect to the input image, aggregating gradients across channels and averaging over test samples. The resulting maps were normalized by their maximum value, so that ERF intensity reflects the influence of each spatial location relative to the most influential input pixel. Thus, consistent with common practice in ERF analyses, interpretation focuses on the spatial spread of contextual influence away from the dominant central region rather than absolute gradient magnitudes, which can vary with parameter scaling and activation levels and may not be directly comparable across models.

As depicted in Fig. 9, MambaRoll exhibits the broadest and most coherent ERF coverage, indicating consistent multi-scale contextual integration. Conventional SSMs show moderately expansive but focalized ERFs along horizontal and vertical lines, reflecting raster-scan biases used to sequentialize image pixels. Transformer-based models display relatively compact or fragmented ERFs, especially those using local window attention (e.g., SwinMR, USwin), suggesting that despite their theoretical capacity for global attention, practical implementations may underutilize broad spatial cues. CNN-based models show variable ERFs: deeper models with UNet-like structures (e.g., UNet, FBPConvNet) achieve moderately spread fields through multi-resolution processing and skip connections, while uniform designs maintaining consistent feature map resolution (e.g., MoDL, MetaInvNet) remain highly localized. These findings highlight that by combining autoregressive modeling with multi-scale state space processing, MambaRoll more coherently captures long-range dependencies, providing broad spatial sensitivity valuable for anatomical coherence and global structure recovery in medical image reconstruction.

F. Computational Complexity

Computational complexity is an important consideration for adoption of learning-based models. Run times and memory loads for competing methods are listed in Table V for MRI and Table VI for CT. CNN-based methods generally demonstrate fast run times and low memory demand, though with exceptions. In CT reconstruction, MetaInvNet shows higher compute load as a PD architecture due to intrinsic complexity of Radon transformation within data-consistency modules. Transformer-based methods typically incur elevated compute burden with notably higher run times and memory demand

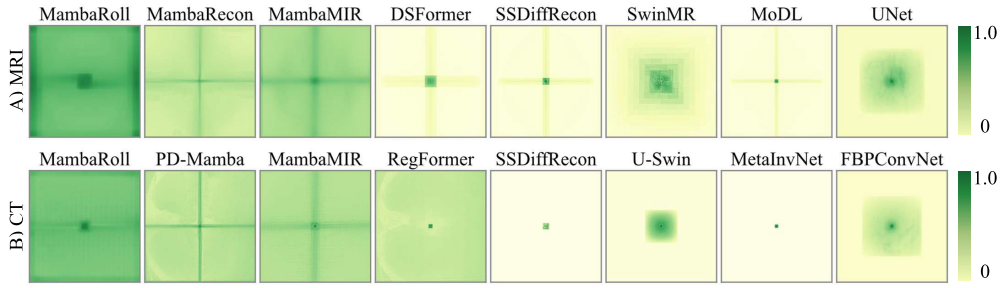

Fig. 9. Effective receptive field (ERF) visualizations for MRI (top row) and CT (bottom row) reconstruction. For each method, the corresponding ERF map highlights relative influences of input image regions on the central pixel of the reconstructed image. Green/yellow areas indicate higher/lower contribution, respectively (see colorbar).

TABLE V
AVERAGE RUN TIME (MSEC) AND MEMORY LOAD (MB) PER CROSS-SECTION FOR MRI RECONSTRUCTION DURING TRAINING AND INFERENCE PHASES

	Run time (msec)		Memory (MB)	
	Training	Inference	Training	Inference
UNet	42.37	11.56	1516	550
MoDL	50.60	14.99	1398	304
SwinMR	149.60	49.85	3864	870
SSDiffRecon	196.73	339.40	3858	740
DSFormer	750.21	248.89	16090	1338
MambaMIR	121.34	36.84	2382	474
MambaRecon	139.89	36.55	2304	462
MambaRoll	96.09	29.67	2568	358

TABLE VI
AVERAGE RUN TIME (MSEC) AND MEMORY LOAD (MB) PER CROSS-SECTION FOR CT RECONSTRUCTION DURING TRAINING AND INFERENCE PHASES

	Run time (msec)		Memory (MB)	
	Training	Inference	Training	Inference
FBPConvNet	84.93	13.23	1962	974
MetaInvNet	77.57	29.16	5144	3288
USwin	246.06	80.59	5926	1082
SSDiffRecon	193.19	346.74	7076	3866
RegFormer	108.64	48.20	7890	5884
MambaMIR	143.05	41.29	2830	532
PD-Mamba	303.37	90.93	9264	4244
MambaRoll	134.23	49.79	6144	3374

despite efficiency-oriented implementations (e.g., windowed attention mechanisms). While RegFormer has competitive run times to CNNs in CT reconstruction, it has high memory demand as a PD architecture. SSM-based methods generally demonstrate efficiency levels between CNNs and transformers. In MRI, MambaRoll achieves the highest computational efficiency among evaluated SSM methods, with run times and memory demand approaching those of CNN models. In CT, MambaMIR has lower inference latency and memory demand, whereas MambaRoll provides a favorable efficiency-performance trade-off.

Comparative visualizations of performance and efficiency are presented in Fig. 1. While some CNN-based methods achieve higher computational efficiency (lower latency and/or memory load), this comes at the expense of reconstruction quality (lower PSNR and SSIM). MambaRoll offers an attractive balance between reconstruction performance and computational efficiency. For CT reconstructions, PD models (PD-Mamba, RegFormer, SSDiffRecon, MetaInvNet) incur additional computational costs due to forward and inverse Radon transformations in data-consistency modules, effectively addressed by MambaRoll's efficient design. These results suggest particular promise for MambaRoll in achieving high fidelity and efficiency in medical image reconstruction.

G. Ablation Studies

1) Network Architecture: We conducted ablation studies to evaluate major architectural elements of MambaRoll. To assess PD-SSM modules, we trained a 'w/o PD-SSM' variant that removed the PD-SSM modules albeit retained the refinement modules so as to create an unrolled convolutional architecture. To assess autoregressive modeling across spatial scales, we trained a 'w/o AR' variant that retained the PD-SSM modules operating across S spatial scales, while removing cross-scale conditioning so that each module receives only the base input f_0. The resulting scale-specific outputs $(f_1, \ldots, f_S)$ were then combined via a learnable fusion layer to form an aggregated map that was propagated to the RM module. To assess multi-scale modeling, we trained a 'w/o multi-scale' variant that prescribed $S = 1$ (see Fig. 11 for a visual comparison). To assess SSM layers in capturing contextual representations, we trained a 'w/o comp. SSM' variant that ablated compressed SSM blocks from PD-SSM modules. To assess DC blocks, we trained a 'w/o DC' variant that ablated DC blocks albeit retaining their residual connections to create a DD SSM model. Table VII lists performance metrics for variants. MambaRoll yields superior performance against all variants, corroborating the importance of key architectural design elements.

2) Loss Function: We evaluated the importance of DMSD loss on reconstruction performance by training MambaRoll variants with alternative supervision. A 'Shallow SS' variant only used a single-scale loss term on the network output. A 'Shallow MS' variant used the single-scale loss term on the network output alongside multi-scale loss terms on spatially downsampled versions of the output (with a total of S scales). A 'Deep MSL' variant used a deep multi-scale loss on the latent feature maps extracted by PD-SSM modules (i.e., g_s), instead of the decoded images as in MambaRoll. Figure 10 shows that the proposed DMSD loss yields improved reconstruction performance relative to these variants, indicating the benefit of auxiliary multi-scale supervision in MambaRoll. A

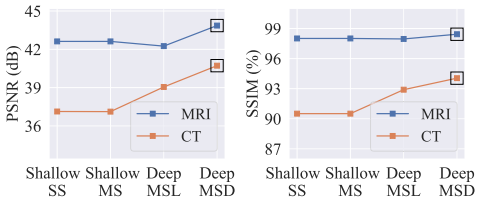

Fig. 10. Performance (FLAIR, R = 4 for MRI and R = 4 for CT) with the proposed DMSD loss (marked with squares), along with Shallow SS, Shallow MS, and Deep MSL variants.

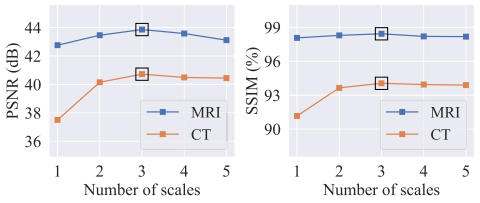

Fig. 11. Performance of MambaRoll variants that prescribed varying numbers of spatial scales (selected values marked with squares).

TABLE VII
PERFORMANCE OF MAMBAROLL VARIANTS, WHERE 'W/O PD-SSM' ABLATES PD-SSM MODULES, 'W/O AR' ABLATES AUTOREGRESSIVE PROCESSING, 'W/O MULTI-SCALE' EMPLOYS A SINGLE-SCALE (S = 1), 'W/O COMP. SSM' ABLATES COMPRESSED SSM BLOCKS, AND 'W/O DC' ABLATES DC MODULES

	fastMRI - FLAIR, $R = 4$		CT, $R = 4$	
	PSNR	SSIM	PSNR	SSIM
MambaRoll	**43.86±3.32**	**98.43±1.43**	**40.72±3.56**	**94.05±4.86**
w/o PD-SSM	40.49±2.70	97.08±1.98	31.00±1.27	83.50±5.00
w/o AR	42.97±3.23	98.10±1.65	39.26±3.09	92.86±5.06
w/o multi-scale	42.62±3.11	98.01±1.66	37.11±2.56	90.53±5.33
w/o comp. SSM	41.43±3.11	97.44±2.13	35.65±2.41	88.51±6.17
w/o DC	36.88±2.50	93.44±3.38	38.66±2.99	92.14±5.15

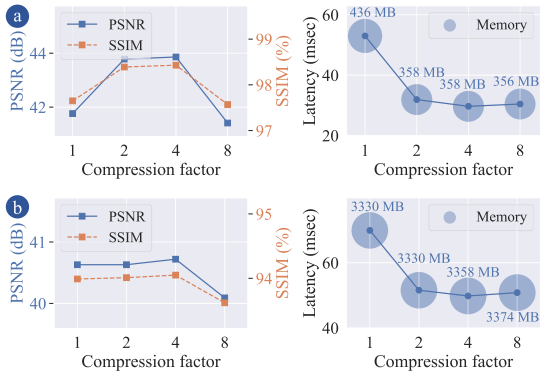

Fig. 12. Performance of MambaRoll variants that prescribed varying compression factors **(a)** on FLAIR, R = 4 for MRI; **(b)** on R = 4 for CT. PSNR, SSIM are displayed in the left panels; memory load, latency are displayed in the right panels.

potential explanation for these improvements is that DMSD enhances the fidelity of scale-specific representations used in autoregressive processing, although verifying this effect requires further targeted analysis.

3) Design Parameters: We assessed the influence of spatial scales for autoregressive modeling and compression factors in PD-SSM modules. As shown in Fig. 11, performance increases with growing scales until S = 3, without apparent benefits beyond this point, indicating a near-optimal performance-efficiency trade-off. Meanwhile, Fig. 12 shows performance and efficiency across different compression factors in compressed SSM blocks. While larger compression factors enhance contextual information exchange across distant regions by increasing spatial sampling steps, they expand feature maps along channel dimension, elevating computational overhead. A factor of J = 4 provides optimal performance while maintaining efficiency. Factors beyond 4 degrade performance without efficiency benefits, likely due to excessive information dispersion disrupting local spatial coherence necessary for effective medical image reconstruction.

VI. DISCUSSION

The utility of MambaRoll could be further improved by addressing several technical limitations. A first group of improvements concern the learning setup. Here, a supervised setup using fully-sampled acquisitions as ground truth was considered. To enable training on datasets with only undersampled data, self-supervised learning could be adopted [59]. When curating any training data is difficult altogether, test-time training could be used for scan-specific reconstruction at increased inference cost [24], [60].

A second group of improvements concern learning objectives. Here we observed high-quality reconstructions using models trained for autoregressive image prediction guided via DMSD loss. In the current design, DMSD acts as auxiliary supervision applied to late-stage scale-specific outputs after repeated autoregressive interactions and data-consistency updates, which helps guide multi-scale representations without directly constraining early-stage representations. While deep supervision across scales may in principle reduce the flexibility of representations at coarser scales, our ablation studies indicate that combining autoregressive conditioning with DMSD improves final reconstruction quality in the evaluated settings. Future work could explore multi-path supervision strategies in which models trained under different scale-supervision regimes are ensembled, allowing complementary representations to be learned without forcing a single supervision pathway to dominate intermediate features. Our results suggest that autoregressive modeling might be a promising alternative to generative methods based on adversarial and diffusive losses [48]. Yet, combining autoregressive loss with other generative losses might help further improve sensitivity to fine details. While our primary focus here was on architectural design of reconstruction models, MambaRoll can be combined with generative learning objectives to examine their potential benefits in future work [18], [61], [62].

A third group of improvements concern reconstruction tasks. Here we built models that operate on single-modality acquisitions (a single MRI contrast or a single CT contrast). Recent work suggests SSMs can also aid in joint reconstruction of multiple modalities [35]. A basic approach to extend MambaRoll for multi-modal reconstruction would be to assign separate channels per modality in the input-output layers and across DC layers in PD-SSM modules. Separate model branches could also be employed to process each modality, while cross-modality interactions are mediated by attentional or state-space modules for fine-grained control of feature representations shared between modalities [63].

For systematic assessment of CT reconstruction performance, we considered sparse-view measurements under a parallel-beam geometry obtained by retaining a subset of projection angles from fully-sampled sinograms. The resulting data inherit the measurement noise present in the acquired projections, although dose-dependent noise variations arising from changes in photon counts were not explicitly modeled. In clinical CT systems, however, more complex acquisition settings are typically encountered: projections are commonly acquired using fan-beam or helical geometries, and measurements follow signal-dependent photon-counting statistics [64]. While parallel-beam formulations are widely used for controlled evaluation and algorithm development, these clinical settings introduce additional challenges. Extending the proposed method to such scenarios would therefore require incorporating the appropriate forward and adjoint imaging operators within the data-consistency modules, as well as accounting for realistic noise behavior. Systematic evaluation under such clinical acquisition conditions remains an important direction for future work.

A final group of improvements concerns architectural design. Each PD-SSM module integrates a linear SSM layer between scale-specific convolutional layers that downsample/upsample feature maps, enabling autoregressive modeling across spatial scales. Our experiments suggest that this design outperforms existing architectures based on conventional Mamba modules (e.g., MambaMIR, MambaRecon), which use a popular gating mechanism to modulate SSM outputs [37]. A hybrid architecture that pools linear and gated SSMs might offer performance benefits. Dual-domain approaches that combine spatial and spectral processing can also enhance performance [65]. In addition, hybrid SSM-attention designs may offer a useful compromise between representational fidelity and efficiency, where predominantly SSM-based backbones provide efficient long-range aggregation and occasional attention modules help capture finer pairwise contextual relationships [66]. Beyond these directions, the contextual sensitivity of MambaRoll could be further extended toward volumetric settings. While the current implementation operates on individual 2D cross-sections, efficient volumetric variants can be considered. In particular, 2.5D processing, where small stacks of adjacent slices are jointly processed, or reconstruction over moderately sized 3D patches could enable the model to capture through-plane dependencies while maintaining manageable computational cost. In such extensions, volumetric spatial grids can be mapped to vectorized sequences within SSM blocks using locality-preserving traversals, such as slice-wise raster, centric spiral or space-filling trajectories [67], [68]. Further work is needed to systematically explore the benefits of hybrid architectures and volumetric processing in medical image reconstruction.

VII. Conclusion

We introduced a novel deep learning method for medical image reconstruction based on physics-driven autoregressive state-space modeling. MambaRoll leverages autoregressive prediction across spatial scales along with physics-driven SSM modules to aggregate contextual features. To enhance learning efficacy, MambaRoll employs a deep multi-scale decoding loss tailored to the autoregressive design. Demonstrations on MRI and CT reconstructions indicate that the proposed model provides improved performance relative to state-of-the-art methods in terms of image quality, without compromising computational efficiency. Therefore, MambaRoll holds great potential for improving the utility of learning-based medical image reconstruction.

References

[1] G. Wang, J. C. Ye, K. Mueller, and J. A. Fessler, "Image reconstruction is a new frontier of machine learning," *IEEE Trans. Med. Imag.*, vol. 37, no. 6, pp. 1289–1296, Jun. 2018.

[2] Y. Du, R. Dharmakumar, and S. A. Tsaftaris, "The MRI scanner as a diagnostic: Image-less active sampling," in *Proc. MICCAI*, 2024, pp. 467–476.

[3] T. Çukur et al., "A tutorial on MRI reconstruction: From modern methods to clinical implications," *IEEE Trans. Biomed. Eng.*, vol. 73, no. 5, pp. 1900–1920, May 2026.

[4] S. K. Zhou et al., "A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises," *Proc. IEEE*, vol. 109, no. 5, pp. 820–838, May 2021.

[5] F. Lam, X. Peng, and Z.-P. Liang, "High-dimensional MR spatiospectral imaging by integrating physics-based modeling and data-driven machine learning: Current progress and future directions," *IEEE Signal Process. Mag.*, vol. 40, no. 2, pp. 101–115, Mar. 2023.

[6] S. Wang et al., "Accelerating magnetic resonance imaging via deep learning," in *Proc. IEEE 13th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2016, pp. 514–517.

[7] D. Lee, J. Yoo, S. Tak, and J. C. Ye, "Deep residual learning for accelerated MRI using magnitude and phase networks," *IEEE Trans. Biomed. Eng.*, vol. 65, no. 9, pp. 1985–1995, Sep. 2018.

[8] Y. Han and J. C. Ye, "Framing U-Net via deep convolutional framelets: Application to sparse-view CT," *IEEE Trans. Med. Imag.*, vol. 37, no. 6, pp. 1418–1429, Jun. 2018.

[9] W. Wu, D. Hu, C. Niu, H. Yu, V. Vardhanabhuti, and G. Wang, "DRONE: Dual-domain residual-based optimization network for sparse-view CT reconstruction," *IEEE Trans. Med. Imag.*, vol. 40, no. 11, pp. 3002–3014, Nov. 2021.

[10] M. Lee, H. Kim, and H.-J. Kim, "Sparse-view CT reconstruction based on multi-level wavelet convolution neural network," *Phys. Medica*, vol. 80, pp. 352–362, Dec. 2020.

[11] Z. Zhang, X. Liang, X. Dong, Y. Xie, and G. Cao, "A sparse-view CT reconstruction method based on combination of DenseNet and deconvolution," *IEEE Trans. Med. Imag.*, vol. 37, no. 6, pp. 1407–1417, Jun. 2018.

[12] J. Schlemper, J. Caballero, J. V. Hajnal, A. Price, and D. Rueckert, "A deep cascade of convolutional neural networks for MR image reconstruction," in *Proc. Int. Conf. Info. Process. Med. Imag. (IPMI)*, 2017, pp. 647–658.

[13] K. Hammernik et al., "Learning a variational network for reconstruction of accelerated MRI data," *Magn. Reson. Med.*, vol. 79, no. 6, pp. 3055–3071, Jun. 2018.

[14] Y. Zhang et al., "DREAM-Net: Deep residual error iterative minimization network for sparse-view CT reconstruction," *IEEE J. Biomed. Health Informat.*, vol. 27, no. 1, pp. 480–491, Jan. 2023.

[15] H. Zhang, B. Liu, H. Yu, and B. Dong, "MetaInv-Net: Meta inversion network for sparse view CT image reconstruction," *IEEE Trans. Med. Imag.*, vol. 40, no. 2, pp. 621–634, Feb. 2021.

[16] B. Zhou and S. K. Zhou, "DuDoRNet: Learning a dual-domain recurrent network for fast MRI reconstruction with deep T1 prior," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4272–4281.

[17] S. A. H. Hosseini, B. Yaman, S. Moeller, M. Hong, and M. Akçakaya, "Dense recurrent neural networks for accelerated MRI: History-cognizant unrolling of optimization algorithms," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 6, pp. 1280–1291, Oct. 2020.

[18] A. Jalal, M. Arvinte, G. Daras, E. Price, A. G. Dimakis, and J. Tamir, "Robust compressed sensing MRI with deep generative priors," in *Proc. NeurIPS*, vol. 34, 2021, pp. 14938–14954.

[19] K. H. Jin, M. T. McCann, E. Froustey, and M. Unser, "Deep convolutional neural network for inverse problems in imaging," *IEEE Trans. Image Process.*, vol. 26, no. 9, pp. 4509–4522, Sep. 2017.

[20] G. Yang et al., "DAGAN: Deep de-aliasing generative adversarial networks for fast compressed sensing MRI reconstruction," *IEEE Trans. Med. Imag.*, vol. 37, no. 6, pp. 1310–1321, Jun. 2018.

[21] M. Mardani et al., "Deep generative adversarial neural networks for compressive sensing MRI," *IEEE Trans. Med. Imag.*, vol. 38, no. 1, pp. 167–179, Jan. 2019.

[22] H. K. Aggarwal, M. P. Mani, and M. Jacob, "MoDL: Model-based deep learning architecture for inverse problems," *IEEE Trans. Med. Imag.*, vol. 38, no. 2, pp. 394–405, Feb. 2019.

[23] P. Guo, Y. Mei, J. Zhou, S. Jiang, and V. M. Patel, "ReconFormer: Accelerated MRI reconstruction using recurrent transformer," *IEEE Trans. Med. Imag.*, vol. 43, no. 1, pp. 582–593, Jan. 2024.

[24] Y. Korkmaz, S. U. H. Dar, M. Yurt, M. Özbey, and T. Çukur, "Unsupervised MRI reconstruction via zero-shot learned adversarial transformers," *IEEE Trans. Med. Imag.*, vol. 41, no. 7, pp. 1747–1763, Jul. 2022.

[25] S. U. H. Dar, M. Özbey, A. B. Çatlı, and T. Çukur, "A transfer-learning approach for accelerated MRI using deep neural networks," *Magn. Reson. Med.*, vol. 84, no. 2, pp. 663–685, Aug. 2020.

[26] J. Xiang, Y. Dong, and Y. Yang, "FISTA-Net: Learning a fast iterative shrinkage thresholding network for inverse problems in imaging," *IEEE Trans. Med. Imag.*, vol. 40, no. 5, pp. 1329–1339, May 2021.

[27] C. Wang, K. Shang, H. Zhang, Q. Li, and S. K. Zhou, "DuDoTrans: Dual-domain transformer for sparse-view CT reconstruction," in *Proc. MLMIR*, 2022, pp. 84–94.

[28] W. Xia, Z. Yang, Z. Lu, Z. Wang, and Y. Zhang, "RegFormer: A local–nonlocal regularization-based model for sparse-view CT reconstruction," *IEEE Trans. Radiat. Plasma Med. Sci.*, vol. 8, no. 2, pp. 184–194, Feb. 2024.

[29] J. Huang et al., "Swin transformer for fast MRI," *Neurocomputing*, vol. 493, pp. 281–304, Jul. 2022.

[30] B. Zhou et al., "DSFormer: A dual-domain self-supervised transformer for accelerated multi-contrast MRI reconstruction," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 4966–4975.

[31] C.-M. Feng et al., "Multimodal transformer for accelerated MR imaging," *IEEE Trans. Med. Imag.*, vol. 42, no. 10, pp. 2804–2816, Oct. 2023.

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

[33] L. Zhu, B. Liao, Q. Zhang, X. Wang, W. Liu, and X. Wang, "Vision Mamba: Efficient visual representation learning with bidirectional state space model," in *Proc. ICML*, 2024, pp. 1–14.

[34] 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, Jan. 2025, Art. no. 103334.

[35] J. Zou et al., "MMR-Mamba: Multi-modal MRI reconstruction with mamba and spatial-frequency information fusion," *Med. Image Anal.*, vol. 102, May 2025, Art. no. 103549.

[36] Y. Korkmaz and V. M. Patel, "MambaRecon: MRI reconstruction with structured state space models," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Feb. 2025, pp. 4142–4152.

[37] Y. Liu et al., "VMamba: Visual state space model," in *Proc. Adv. Neural Inf. Process. Syst.* 37, 2024, pp. 103031–103063.

[38] M. Yurt, S. U. Dar, A. Erdem, E. Erdem, K. K. Oguz, and T. Çukur, "MustGAN: Multi-stream generative adversarial networks for MR image synthesis," *Med. Image Anal.*, vol. 70, May 2021, Art. no. 101944.

[39] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[40] B. Kabas, F. Arslan, V. A. Nezhad, S. Ozturk, E. U. Saritas, and T. Çukur, "Physics-driven autoregressive state space models for medical image reconstruction," 2024, *arXiv:2412.09331*.

[41] Y. Meng, Z. Yang, Z. Song, and Y. Shi, "DH-Mamba: Exploring dual-domain hierarchical state space models for MRI reconstruction," 2025, *arXiv:2501.08163*.

[42] T. Hou, H. Zhu, B. Huang, K. Chen, and Z. Zheng, "Causal inertia proximal mamba network for magnetic resonance image reconstruction," *Med. Image Anal.*, vol. 105, Oct. 2025, Art. no. 103649.

[43] G. Luo, N. Zhao, W. Jiang, E. S. Hui, and P. Cao, "MRI reconstruction using deep Bayesian estimation," *Magn. Reson. Med.*, vol. 84, no. 4, pp. 2246–2261, Oct. 2020.

[44] T. H. Kim, P. Garg, and J. P. Haldar, "LORAKI: Autocalibrated recurrent neural networks for autoregressive MRI reconstruction in k-Space," 2019, *arXiv:1904.09390*.

[45] G. Luo, S. Huang, and M. Uecker, "Autoregressive image diffusion: Generation of image sequence and application in MRI," in *Proc. Adv. Neural Inf. Process. Syst.* 37, 2024, pp. 129094–129119.

[46] Y. Shi, M. Dong, and C. Xu, "Multi-scale VMamba: Hierarchy in hierarchy visual state space model," in *Proc. Adv. Neural Inf. Process. Syst.*, 2024, pp. 25687–25708.

[47] J. Qiao et al., "Hi-Mamba: Hierarchical Mamba for efficient image super-resolution," 2024, *arXiv:2410.10140*.

[48] K. Tian, Y. Jiang, Z. Yuan, B. Peng, and L. Wang, "Visual autoregressive modeling: Scalable image generation via next-scale prediction," in *Proc. Adv. Neural Inf. Process. Syst.* 37, 2024, pp. 84839–84865.

[49] P. Guo, P. Wang, J. Zhou, S. Jiang, and V. M. Patel, "Multi-institutional collaborations for improving deep learning-based magnetic resonance image reconstruction using federated learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2423–2432.

[50] W. Shi et al., "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1874–1883.

[51] Y. Xu, S. Han, D. Wang, G. Wang, J. S. Maltz, and H. Yu, "Hybrid U-Net and Swin-transformer network for limited-angle cardiac computed tomography," *Phys. Med. Biol.*, vol. 69, no. 10, May 2024, Art. no. 105012.

[52] F. Knoll et al., "FastMRI: A publicly available raw k-space and DICOM dataset of knee images for accelerated MR image reconstruction using machine learning," *Radiol., Artif. Intell.*, vol. 2, no. 1, Jan. 2020, Art. no. e190007.

[53] T. Zhang, J. M. Pauly, S. S. Vasanawala, and M. Lustig, "Coil compression for accelerated imaging with Cartesian sampling," *Magn. Reson. Med.*, vol. 69, no. 2, pp. 571–582, Feb. 2013.

[54] M. Uecker et al., "ESPIRiT—An eigenvalue approach to autocalibrating parallel MRI: Where SENSE meets GRAPPA," *Magn. Reson. Med.*, vol. 71, no. 3, pp. 990–1001, Mar. 2014.

[55] M. Lustig, D. Donoho, and J. M. Pauly, "Sparse MRI: The application of compressed sensing for rapid MR imaging," *Magn. Reson. Med.*, vol. 58, no. 6, pp. 1182–1195, Dec. 2007.

[56] J. Leuschner, M. Schmidt, D. O. Baguer, and P. Maass, "LoDoPaB-CT, a benchmark dataset for low-dose computed tomography reconstruction," *Sci. Data*, vol. 8, no. 1, p. 109, Apr. 2021.

[57] R. Zhao et al., "FastMRI+, clinical pathology annotations for knee and brain fully sampled magnetic resonance imaging data," *Sci. Data*, vol. 9, no. 1, p. 152, Apr. 2022.

[58] S. G. Armato III et al., "The lung image database consortium (LIDC) and image database resource initiative (IDRI): A completed reference database of lung nodules on CT scans," *Med. Phys.*, vol. 38, no. 2, pp. 915–931, 2011.

[59] B. Yaman, S. A. H. Hosseini, S. Moeller, J. Ellermann, K. Uğurbil, and M. Akçakaya, "Self-supervised learning of physics-guided reconstruction neural networks without fully sampled reference data," *Magn. Reson. Med.*, vol. 84, no. 6, pp. 3172–3191, Dec. 2020.

[60] M. Z. Darestani and R. Heckel, "Accelerated MRI with un-trained neural networks," *IEEE Trans. Comput. Imag.*, vol. 7, pp. 724–733, 2021.

[61] A. Güngör et al., "Adaptive diffusion priors for accelerated MRI reconstruction," *Med. Image Anal.*, vol. 88, Aug. 2023, Art. no. 102872.

[62] M. U. Mirza et al., "Learning Fourier-constrained diffusion bridges for MRI reconstruction," *IEEE Trans. Med. Imag.*, vol. 45, no. 6, pp. 3229–3245, Jun. 2026.

[63] F. Genç, B. İ. Macun, S. S. Özaslan, E. U. Saritas, and T. Çukur, "NeuroSSM: Multiscale differential state-space modeling for context-aware fMRI analysis," 2026, *arXiv:2601.01229*.

[64] Ş. Öztürk, O. C. Duran, and T. Çukur, "DenoMamba: A fused state-space model for low-dose CT denoising," *IEEE J. Biomed. Health Informat.*, vol. 30, no. 6, pp. 5353–5366, Jun. 2026.

[65] T. Eo, Y. Jun, T. Kim, J. Jang, H.-J. Lee, and D. Hwang, "KIKI-Net: Cross-domain convolutional neural networks for reconstructing undersampled magnetic resonance images," *Magn. Reson. Med.*, vol. 80, no. 5, pp. 2188–2201, Nov. 2018.

[66] L. Ren, Y. Liu, Y. Lu, Y. Shen, C. Liang, and W. Chen, "Samba: Simple hybrid state space models for efficient unlimited context language modeling," in *Proc. ICLR*, 2025, pp. 53551–53575.

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

[68] X. Liu, C. Zhang, and L. Zhang, "Vision Mamba: A comprehensive survey and taxonomy," 2024, *arXiv:2405.04404*.