This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/TCI.2025.3536078, IEEE Transactions on Computational Imaging

Cascaded Frequency-Encoded Multi-Scale Neural Fields for Sparse-View CT Reconstruction

Jia Wu, Jinzhao Lin, Yu Pan, Xiaoming Jiang, Xinwei Li, Hongying Meng, *Senior Member, IEEE*, Yamei Luo, Lu Yang, and Zhangyong Li

***Abstract*— Sparse-view Computed Tomography (CT) reconstruction has attracted considerable attention as a method for reducing radiation exposure and acquisition time in CT imaging. However, the ill-posed nature of the sparse-view reconstruction problem poses challenges for image quality and computational efficiency. In this study, we propose a novel cascaded framework for sparse-view CT reconstruction, designated Cascaded Frequency-encoded Multi-scale Neural Field (Ca-FMNF). This framework combines an iterative unfolding network based on state space models (SSMs) with a frequency-encoded multi-scale neural field (FMNF) representation. The SSM-based iterative unfolding network generates an effective initial reconstruction, which is subsequently refined by the FMNF network through a continuous optimization process in the image space. The FMNF network utilizes a multi-scale grid structure for spatial decomposition and associates each scale with specific frequency bands through Fourier feature encoding, enabling efficient and precise local feature modeling. Additionally, a hybrid loss function, incorporating data fidelity, wavelet sparsity, and total variation regularization, is implemented to enhance the stability and robustness of the reconstruction process. Extensive experiments on the AAPM dataset demonstrate that our Ca-FMNF method outperforms state-of-the-art approaches in terms of both quantitative metrics and visual quality, yielding superior reconstruction results with preserved edges and structural features.**

***Index Terms*— Sparse-View CT Reconstruction, Iterative Unfolding, State Space Models, Neural Fields, Multi-Scale Representation**

I. Introduction

Computed Tomography (CT) is an essential imaging tool widely employed in medical and industrial fields for non-destructive, non-invasive internal examinations. CT imaging involves capturing 2D X-ray projections from various angles, which are subsequently reconstructed into a 3D volume utilizing algorithms such as Filtered Back Projection (FBP) or Feldkamp-Davis-Kress (FDK) [1], [2]. Although a higher quantity and optimal angular distribution of projections can enhance image accuracy, they also increase radiation exposure, presenting significant risks, particularly for frequent CT users. To address this issue, sparse-view CT has been developed, which reduces the number of projection angles to decrease radiation doses. However, this reduction results in an ill-posed problem, substantially compromising the efficacy of traditional algorithms like FBP and resulting in images with significant noise and artifacts.

Model-based optimization methods address the ill-posed problem of sparse-view CT reconstruction by incorporating nonlinear regularizers (priors) to account for the missing information [3]. These approaches have demonstrated adaptability and produce high-quality reconstructions by leveraging prior knowledge about the underlying image structure. Examples of such methods include Total Variation (TV) [4], dictionary learning [5], nonlocal means filtering [6], tight wavelet frames [7], low-rank models [8], transform learning [9], and convolutional sparse coding [10]. By incorporating prior knowledge, these methods effectively mitigate the artifacts and noise typically associated with under-sampled data. However, the high-quality reconstructions achieved by these methods often come at the expense of computational efficiency. Moreover,

This work was supported in part by the National Natural Science Foundation of China under Grants U21A20447 and 62171073, the Chongqing Natural Science Foundation under Grant CSTB2023NSCQ-LZX0064, the Chunhui Plan of the China Education Ministry under Grant HZKY20220209, the Southwest Medical University Natural Science Foundation under Grant 2023ZD004, and the Sichuan Science and Technology Program under Grant 2022YFS0616. (Corresponding authors: Zhangyong Li; Jinzhao Lin)

Jia Wu is with the School of Communications and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing, 400065, China, and also with the School of Medical Information and Engineering, Southwest Medical University, Luzhou, 646000, China (e-mail: wujiahj@126.com).

Jinzhao Lin is with the School of Communications and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing, 400065, China (e-mail: lin.jz@cqupt.edu.cn).

Yu Pan is with the School of Optoelectronic Engineering, Chongqing University of Posts and Telecommunications, Chongqing, China (e-mail: pangyu@cqupt.edu.cn).

Hongying Meng is with the Department of Electronic and Electrical Engineering, College of Engineering Design and Physical Sciences, Brunel University London, Uxbridge, UB8 3PH, UK (e-mail: hongying.meng@brunel.ac.uk).

Yamei Luo is with the School of Medical Information and Engineering, Southwest Medical University, Luzhou, 646000, China (e-mail: luoluoeryan@126.com).

Lu Yang is with the Department of Radiology, The Affiliated Hospital of Southwest Medical University, Luzhou, 646000, China (e-mail: yanglu@swmu.edu.cn).

Zhangyong Li, Xiaoming Jiang, and Xinwei Li are with the Chongqing Engineering Research Center of Medical Electronics and Information Technology, Chongqing University of Posts and Telecommunications, Chongqing, 400065, China (e-mail: lizy@cqupt.edu.cn; jiangxm@cqupt.edu.cn; lixinwei@cqupt.edu.cn).

the performance of these methods relies on the choice of modeling assumptions and regularization parameters, which can be challenging to optimize.

Deep learning techniques have significantly enhanced CT image reconstruction, utilizing large-scale data to achieve notable improvements. Instead of directly addressing the ill-posed inverse problem, deep learning approaches utilize convolutional neural networks (CNNs) trained on extensive datasets to learn the end-to-end mapping from raw measurement data to reconstructed images [11]–[14]. This data-driven training process unveils the complex transformation patterns inherent in the data. Further innovations have enhanced deep learning reconstruction by incorporating priors, such as plug-and-play (PnP) [15] and regularization by denoising (RED) [16], which incorporate additional prior knowledge to refine the output. Additionally, generative adversarial networks (GANs) [17], diffusion score models [18]–[20], and geometric deep learning frameworks [21] have emerged, expanding the model's ability to generalize across various imaging scenarios. Hybrid approaches, combining model-driven and data-driven paradigms, have also been developed in the form of deep iterative unrolling models [22], [23]. These models interpret iterative steps of model-based optimization as CNN layers, facilitating end-to-end training in a supervised manner. Despite the clear superiority of deep learning in CT image reconstruction, challenges persist, especially in acquiring large-scale training datasets and handling variability across different datasets. These challenges can impact the robustness of reconstructions, particularly when dealing with subtle but critical anatomical changes, such as tumor growth. Moreover, trained models may encounter difficulties in adapting to patients with varying anatomical structures, highlighting the need for models that generalize effectively across different patient scans [24].

Neural fields [25] represent a significant development in computational modeling, employing coordinate-based neural networks to parameterize physical attributes across spatial and temporal dimensions. Unlike traditional CNNs that rely on discrete data representations, neural fields define a continuous function, offering a detailed and continuous representation of the scene's dynamics. This continuous nature enables the precise modeling of complex and intricate patterns while also enhancing the representation of high-frequency details, thereby overcoming the 'spectral bias' that is commonly encountered in CNNs. As a result, neural fields demonstrate exceptional performance across various visual tasks, such as surface reconstruction [26], view synthesis [27], and image super-resolution [28]. The ability of neural fields to capture and represent detailed information makes them a promising candidate for addressing the challenges faced by deep learning methods in CT image reconstruction.

In the sparse-view CT imaging, the potential of neural fields has gained increasing recognition. Early explorations, like those by Tancik et al. [29], showcased the capability of neural fields to reconstruct CT images from sparse data sets without relying on extensive external inputs. Building on these initial findings, a variety of neural field-based approaches have been developed for CT reconstruction. These methods, such as those implemented in the CoIL framework by Sun et al. [30], utilize the continuous modeling capability of neural fields to predict dense-view sinograms from sparse observations. However, the discord between the coordinate space of sinograms and the inherently continuous neural field models can affect the efficacy of CoIL. To address these complexities, newer methodologies, such as NeRP and IntroTomo, proposed by Shen et al. [24] and Zang et al. [1], respectively, were developed. These approaches combine longitudinal scan data and neural field modeling to refine CT image reconstruction. They demonstrate the power of neural fields in enhancing image quality through iterative and prior-integrated training processes. Further advancing the field, Wu et al. [31] introduced SCOPE, which employs neural field principles and a novel re-projection strategy to enhance the solution space and stability of the CT image reconstruction process. However, methods like CoIL, IntroTomo, and SCOPE primarily focus on using multi-layer perceptron (MLP) to represent the measurement field rather than directly reconstructing the image. They subsequently rely on other existing methods for the final image reconstruction. This indirect approach can propagate measurement inaccuracies, potentially leading to secondary artifacts in the reconstructed images. For instance, CoIL's performance might be compromised by the misalignment between the measured sinogram's coordinate space and the neural field's continuous model. In particular, the IntroTomo method integrates explicit priors such as TV and non-local means within an optimization framework to refine CT images. This approach can improve reconstruction fidelity but at the cost of extended reconstruction times. While methods like NeRF can directly yield reconstructed images, their effectiveness heavily depends on the accuracy of the prior image. A significant disparity between the prior and the target can impede optimal reconstruction results.

To leverage the strengths of neural fields while addressing their limitations, we propose a cascaded method that integrates a pre-trained iterative unfolding network with a neural field technique. Our framework initially employs an iterative unfolding network to quickly generate an initial reconstruction, providing an efficient starting point and effectively narrowing the solution space. Subsequently, we introduce a neural field technique to enhance the reconstruction, leveraging its superior continuous representation capabilities to capture and optimize fine image details. This neural field approach incorporates the forward model of the CT imaging system into the network architecture, transforming the image reconstruction challenge into a network optimization problem. This strategy maintains computational efficiency while leveraging the advantages of neural fields in precise modeling. It also enables accurate reconstruction from highly sparse sinograms while significantly reducing computational time.

In the initial phase of our hybrid framework, we utilize an iterative unfolding network based on State Space Models (SSMs) [32]–[34]. This network employs Vision SSM (VSSM) to construct feature extractors within each submodule, improving its capacity to learn concurrently from projection data and image priors, thereby effectively capturing long-range dependencies. We explicitly incorporate measurement fidelity terms into each submodule to maintain consistency

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/TCI.2025.3536078, IEEE Transactions on Computational Imaging

between reconstructed results and measured data. Compared to traditional iterative unfolding networks, this architecture reduces the required number of sub-blocks, shortens inference times for initial image intensity prediction, and effectively narrows the search space for subsequent optimization. Building upon the iterative unfolding network, we introduce a neural field-based method to refine the reconstruction results. This approach encodes the entire image as a continuous implicit representation within the neural network's weights, aiming to capture the complete image space. Inspired by Neural Fourier Filter Banks [35], we encode coordinate signals simultaneously in both spatial and frequency domains. We implement a multi-scale grid structure and apply Fourier feature encoding to the grid features before network input. This method effectively transforms linear variations of grid features into learnable frequencies at each scale level. An MLP with sinusoidal activation functions then processes these Fourier-encoded features, forming a pipeline that progressively accumulates high-frequency information. The summation of all intermediate outputs generates the final estimated intensity values, specifically optimizing residual intensity information.

To summarize, our contributions are as follows:

- We present an implicit neural representation approach for sparse-view CT reconstruction, utilizing a neural field to encode the CT image as a continuous function. Employing a multi-scale grid structure for spatial decomposition and Fourier feature encoding for frequency correlation, our method optimizes a continuous CT intensity field.
- We propose a cascading framework that integrates a learning-based method with implicit neural fields. Initially, we train an iterative unfolding network based on SSMs to predict the initial CT image. Subsequently, implicit neural fields are employed to refine the residual intensity values, optimizing a continuous CT intensity field for higher-quality reconstruction outcomes.
- Experiments on AAPM datasets show that our proposed method achieves state-of-the-art performance while preserving the edges and structural features of the CT images.

The remainder of this paper is organized as follows. Section II introduces the CT imaging problem and details our proposed Ca-FMNF framework, including the SSM-based iterative unfolding network and the frequency-encoded multi-scale neural field representation. Section III presents our experimental results, including simulated and real data evaluations, comparisons with state-of-the-art methods, and an ablation study. Section IV concludes the paper with a summary of our findings. It also discusses potential future directions.

II. Method

A. CT Imaging Problem and Ca-FMNF Framework

The CT imaging system is modeled as $\mathbf{y} = \mathbf{A}\mathbf{x} + \boldsymbol{\varepsilon}$, where $\mathbf{x} \in \mathbb{R}^N$ represents the target image, $\mathbf{y} \in \mathbb{R}^M$ is the sampled sensor measurement of the projection sinogram, $\boldsymbol{\varepsilon} \in \mathbb{R}^M$ represents the measurement noise or error, and $\mathbf{A} \in \mathbb{R}^{M \times N}$ denotes the system's projection matrix. The aim of CT reconstruction is to recover the target CT image $\mathbf{x}$ given the projection sinogram $\mathbf{y}$. In sparse-view CT, under-sampling in the projection domain reduces radiation exposure but results in an ill-posed problem. Coordinate-based neural networks have emerged as a promising approach to address ill-posed problems in image reconstruction [24]. These implicit neural representations offer continuous parameterization of image properties, with sinusoidal activation functions being particularly effective in modeling 2D images. In CT reconstruction, these networks can be optimized to map spatial coordinates to image intensities, incorporating data consistency constraints from projection measurements [30]. However, applying implicit neural networks to sparse-view CT reconstruction presents significant challenges. The presence of noise $\boldsymbol{\varepsilon}$ and the incompleteness of $\mathbf{y}$ in sparse sampling scenarios can lead to overfitting of the implicit neural field to the projection data, potentially compromising the accuracy of the reconstructed images.

To address these limitations and enhance reconstruction quality in sparse-view CT, we propose Ca-FMNF, as illustrated in Fig. 1. Our method integrates a learning-based iterative unfolding network with Frequency-Encoded Multi-Scale Neural Fields (FMNF) to achieve high-quality image reconstruction. Initially, we pre-train an iterative unfolding network to predict an initial CT image. This pre-training provides a well-defined starting point and narrows the search space for the optimal solution in subsequent steps. We then shift our focus to learning the neural representation of the target reconstruction image from sparse projection measurements. By integrating a differentiable forward model (Radon transform), the FMNF network constructs a projection loss between the image space and the measurement space, thereby refining the continuous CT image intensity field. To enable precise and efficient modeling of local features, we employ a multi-scale grid structure for spatial decomposition and Fourier feature encoding at each scale to model specific frequency bands. By optimizing within a continuous functional space of parameters, constrained by sparse projection measurements, the network effectively enhances the quality of the reconstruction. Finally, the trained network's CT intensity field is inferred across all spatial coordinates to produce the final reconstructed image.

B. SSM-based Iterative Unfolding Network

In our cascaded framework, we introduce an iterative unfolding network based on SSMs. SSMs, originating from classical control theory, have recently gained prominence in deep learning due to their ability to efficiently model long-range dependencies [33], [34]. We utilize Mamba [32] as the underlying SSM implementation, enhancing the network's capacity to capture complex temporal and spatial relationships in the data.

Fig. 2 illustrates the network architecture. Given sparse projections $\mathbf{y}$ and an initial reconstruction $\mathbf{x}_0$, the network learns a mapping function that estimates the CT image, effectively transforming data from the measurement domain to the image domain. The network consists of multiple cascaded iteration blocks, each comprising a Measurement Consistency

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/TCI.2025.3536078, IEEE Transactions on Computational Imaging

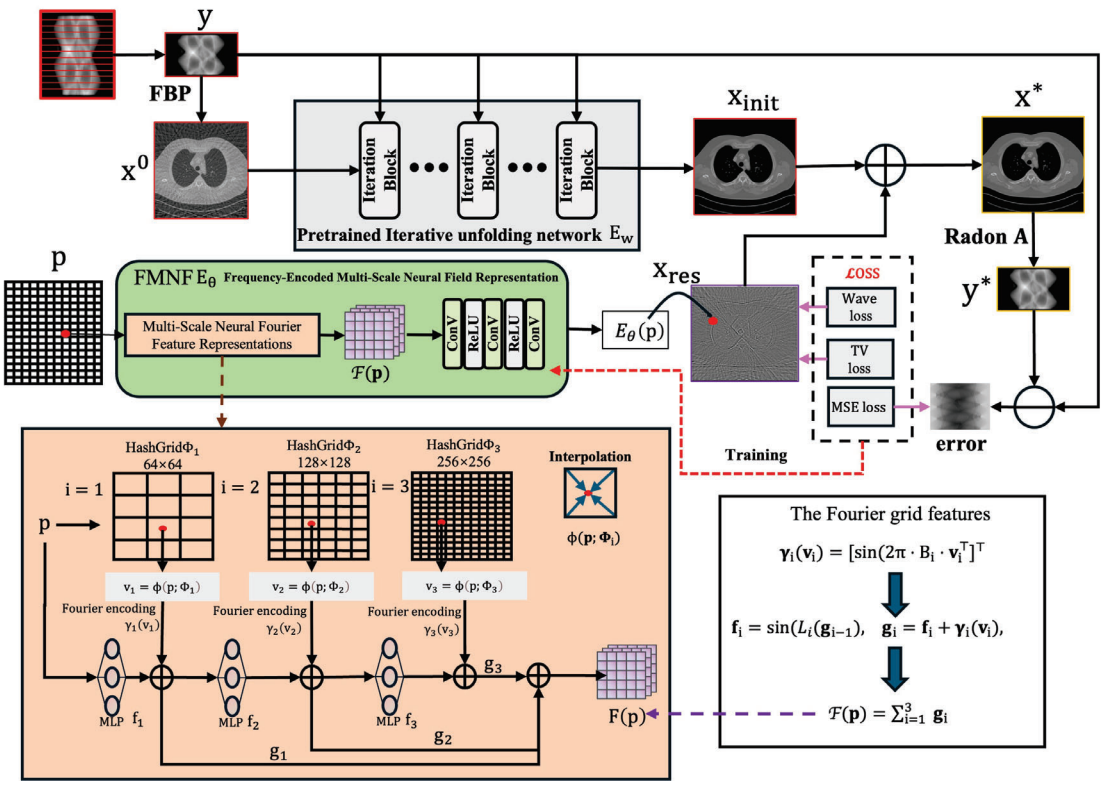

Fig. 1. The proposed cascaded Ca-FMNF framework for sparse-view CT reconstruction. The framework consists of two main stages: (1) a pre-trained iterative unfolding network based on SSM for initial CT image prediction, and (2) a FMNF network for refining the continuous CT intensity field.

Fidelity Block (MCFB) and a Residual State Space Block (RSSB). The fidelity module ensures consistency between the reconstructed image and the measured sinogram data. Meanwhile, the RSSB captures long-range dependencies and high-frequency details by integrating Mamba SSMs with local convolutional operations.

Our proposed iterative unfolding network adopts LEARN [23] as its backbone, integrating SSM principles to enhance its feature extraction and representation capabilities. We reformulate the reconstruction process as:

$$\mathbf{x}^{t+1} = \mathbf{x}^t - \Psi\left(\mathbf{A}^{\mathrm{T}}\left(\mathbf{A}\mathbf{x}^t - \mathbf{y}\right)\right) + \mathcal{M}\left(\mathbf{x}^t\right), \qquad (1)$$

where Ψ represents a three-layer CNN module that, in conjunction with the CNN spatial module, adaptively refines the image reconstruction process by balancing contributions from both data fidelity and feature extraction mechanisms. $\mathcal{M}$ denotes the RSSB, which emphasizes the capture of detailed image features and long-range dependencies.

Inspired by the success of Mamba [32] in long-range modeling with linear complexity, we introduce the RSSB as an implicit regularization term $\mathcal{M}$. Fig. 2 illustrates the RSSB, which combines the Visual State Space Model (VSSM) with local convolutional operations to effectively model both global and local image features [33]. Given input features $\mathbf{X} \in \mathbb{R}^{B\times H\times W\times 1}$, we first process them through a convolutional layer with a Switch activation function to extract feature embeddings $\mathbf{X} \in \mathbb{R}^{B\times H\times W\times C}$, where C represents the dimensionality of the feature embeddings. These features are then layer-normalized to obtain $\mathbf{Y} = \mathrm{LN}\left(\mathrm{Switch}(\mathrm{Conv}(\mathbf{X}))\right)$. The normalized features $\mathbf{Y}$ are input into the VSSM to model long-range dependencies among features. To effectively combine the original input information with the long-range dependencies captured by the VSSM, we employ a learnable skip connection. This allows the network to maintain a balance between preserving low-level details and incorporating global context. The output is expressed as $\mathbf{Z} = \mathbf{s}_l \odot \mathbf{Y} + \mathrm{VSSM}(\mathbf{Y})$, where $\mathbf{s}_l$ represents the learnable parameter. Finally, $\mathbf{Z}$ is processed through additional convolutional layers, followed by a Switch activation function and another layer normalization to generate the final feature representation:

$$\mathbf{X}_{\mathrm{out}} = \mathrm{LN}\left(\mathrm{Switch}(\mathrm{Conv}(\mathbf{Z}))\right). \qquad (2)$$

As illustrated in Fig. 2d, the VSSM comprises a linear layer, depthwise convolutions, a 2D Selective Scan Module (2D-SSM), and additional depthwise convolutions. The input features $\mathbf{Y} \in \mathbb{R}^{B\times H\times W\times C}$ are initially processed through a projection layer, expanding their dimensionality to $\mathbf{Y}_{\mathrm{proj}} \in \mathbb{R}^{B\times H\times W\times D_{\mathrm{inner}}}$, where D_{inner} represents the expanded feature dimension.

$\mathbf{Y}_{\mathrm{proj}}$ is reshaped to $\mathbb{R}^{B\times D_{\mathrm{inner}}\times L}$, where $L = H \times W$, and then subjected to a two-dimensional depthwise convolution followed by a SiLU activation function. To enhance the capture of spatial relationships, we incorporate the 2D-SSM as proposed by [33]. The resulting features are then expanded into four orientations: the original, transposed, and their respective spatially reversed counterparts.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/TCI.2025.3536078, IEEE Transactions on Computational Imaging

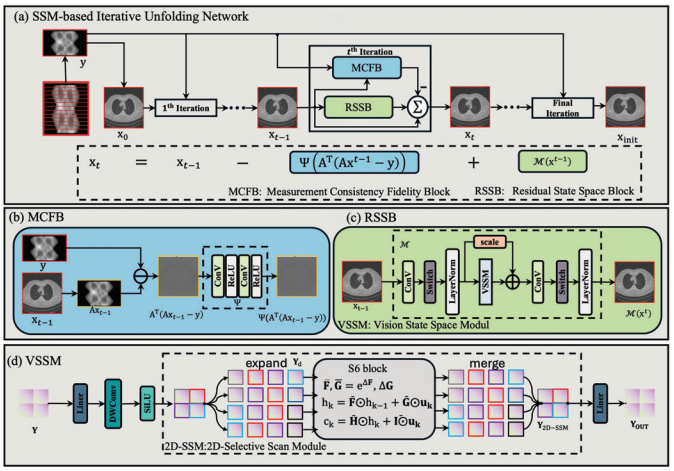

Fig. 2. The architecture and key components of the proposed SSM-based Iterative Unfolding Network. (a) The overall structure of the network, consisting of multiple cascaded iteration blocks. Each iteration comprises two main components: the MCFB and the RSSB. The equation at the bottom describes the iterative process. (b) The detailed structure of the MCFB, which enforces consistency between the reconstructed image and the measured sinogram data. (c) The architecture of the RSSB, which integrates the VSSM with local convolutional operations to capture long-range dependencies and high-frequency details. (d) The internal structure of the VSSM, highlighting the 2D-SSM that enables adaptive feature extraction from various directions, enhancing the overall representation of the input data.

For each orientation, the module applies state space equations. Let $\mathbf{u}_k \in \mathbb{R}^{B\times D_{\text{inner}}}$ denote the k-th slice along the spatial dimension. The state space equations are:

$$\begin{aligned}\hat{\mathbf{h}}_k &= \hat{\mathbf{F}} \odot \hat{\mathbf{h}}_{k-1} + \hat{\mathbf{G}} \odot \mathbf{u}_k \\ \hat{\mathbf{c}}_k &= \hat{\mathbf{H}} \odot \hat{\mathbf{h}}_k + \hat{\mathbf{I}} \odot \mathbf{u}_k,\end{aligned} \tag{3}$$

where $\hat{\mathbf{h}}_k \in \mathbb{R}^{B\times D_{\text{inner}}\times J}$ is the hidden state, $\hat{\mathbf{c}}_k \in \mathbb{R}^{B\times D_{\text{inner}}}$ is the output at position k, and J is the state size.

The parameters $\hat{\mathbf{F}}, \hat{\mathbf{G}}, \hat{\mathbf{H}}$, and $\hat{\mathbf{I}}$ are derived from learnable components, enabling adaptive processing. Specifically: $\hat{\mathbf{F}} = e^{\Delta\mathbf{F}}$, $\hat{\mathbf{G}} = \Delta\mathbf{G}$, $\hat{\mathbf{H}} = \mathbf{H}$, and $\hat{\mathbf{I}} = \mathbf{I}$, where $\mathbf{F} \in \mathbb{R}^{D_{\text{inner}}\times J}$ represents state transition, $\mathbf{G} \in \mathbb{R}^{B\times 1\times J\times L}$ is input projection, $\mathbf{H} \in \mathbb{R}^{B\times 1\times J\times L}$ is output projection, $\mathbf{I} \in \mathbb{R}^{B\times D_{\text{inner}}\times D_{\text{inner}}}$ is the skip connection, and $\Delta \in \mathbb{R}^{B\times D_{\text{inner}}\times L}$ is the learned time step. The Δ parameter is dynamically computed based on input features using the following equation:

$$\Delta = \mathcal{S}(\mathbf{W}_\delta \cdot \text{SiLU}(\text{Conv}(\mathbf{Y}_{\text{proj}})) + \mathbf{b}_\delta), \tag{4}$$

where $\mathbf{W}_\delta$ and $\mathbf{b}_\delta$ are learnable parameters, and $\mathcal{S}$ is the softplus activation function. This formulation allows Δ to adapt to input content, dynamically adjusting temporal dynamics for different input regions.

After completing state space computations across all four orientations (original, transposed, and their spatially reversed counterparts), the module aggregates the outputs through element-wise addition and reshapes the result to $\mathbb{R}^{B\times H\times W\times D_{\text{inner}}}$. Finally, the module projects these features back to the original input dimension C. This multi-orientation processing and aggregation effectively captures long-range dependencies in various spatial directions while maintaining computational efficiency, significantly enhancing the model's capacity to analyze complex spatial relationships and global context in the input data.

To train the iterative unfolding network, we use the DeepLesion [36] dataset with paired data $\{y^n, x_{gt}^n\}_{n=1}^N$, where y^n represents sparse-view projection and x_{gt}^n represents corresponding ground truth image of size 256×256. The sparse-view projections are obtained through simulated fan-beam geometry using the Operator Discretization Library (ODL) [37]. The network E_w, parameterized by weights w, is trained to optimize the mapping from sparse projections to ground truth images using the Mean Squared Error (MSE) loss function:

$$\mathcal{L}_w = \frac{1}{N}\sum_{n=1}^{N} \mathcal{L}_{mse}(E_w(y^n, x_o^n), x_{gt}^n), \tag{5}$$

where $E_w(y^n, x_o^n)$ is the estimated image from the network for the n-th sample, with x_o^n representing the input FBP reconstruction.

In the subsequent reconstruction process, we employ the trained network E_w^* to generate an initial reconstruction for each sample, defined as $x_{init}^n = E_w^*(y^n, x_o^n)$. This initial reconstruction serves as an efficient starting point, effectively narrowing the solution space for subsequent FMNF optimization. Moreover, it accelerates convergence in the FMNF by providing a robust initial estimate of the image intensity distribution.

C. Frequency-Encoded Multi-Scale Neural Field Representation

Building upon the initial reconstruction from the iterative unfolding network, our FMNF combines the advantages of learning-based methods and neural field representation to optimize the residual image and ensure data consistency. Inspired by the Neural Fourier Filter Bank [35], the FMNF employs a multi-scale grid structure that captures distinct frequency bands of the residual CT image field, effectively addressing both low and high-frequency information. This approach distinguishes itself from previous methods [1], [30], [31] by offering a more comprehensive representation of the residual components, rather than encoding coordinates directly in spatial or frequency domains.

As illustrated in Fig. 1, the FMNF architecture comprises a multi-scale grid structure with neural Fourier feature representations and a three-layer CNN for final signal construction. This structure enables the mapping of spatial 2D points to their corresponding intensity values in the residual CT image. By optimizing this neural field through a projection-based loss function, we maintain consistency with the observed data while leveraging the learning capabilities of neural networks, thus synergizing the strengths of data-driven and model-based approaches.

1) Neural Fourier Feature Representations: Inspired by the Neural Fourier Filter Bank [35], we propose a multi-scale grid feature extraction method that maps two-dimensional coordinates to a high-dimensional feature space. Our approach defines a grid feature function $\kappa_i : \mathbb{R}^2 \to \mathbb{R}^M$ at each level i, where input coordinates are mapped to an M-dimensional feature space. Following [35], we implement the lookup table

$\mathbf{\Phi}_i$ using a trainable hash table, facilitating efficient storage and retrieval of feature vectors associated with grid vertices at the i-th level. The resolution of this grid at each level is determined by a base resolution $\mathbf{P_0}$ and a per-level scaling factor s. In our implementation, we set the base resolution to 64×64 and the per-level scale s to 2. This scaling strategy progressively increases the resolution across levels, enabling the capture of features at various scales. We utilize three levels ($i = 1, 2, 3$), effectively encompassing a range from coarse to fine details.

In our feature extraction process, we begin with an input coordinate $\mathbf{p}$. To determine the appropriate grid features, we first identify the four grid vertices that encompass this point. These vertices are determined by computing the floor and ceiling functions of the input coordinate. Specifically, we define the lower bound vertex as $\mathbf{p}_{\text{lower}} = \lfloor \mathbf{p} \rfloor$ and the upper bound vertex as $\mathbf{p}_{\text{upper}} = \lceil \mathbf{p} \rceil$. This procedure yields a set of four vertices $\{\mathbf{p}_j\}_{j=1}^4$, which form the corners of the grid cell containing $\mathbf{p}$.

To map each vertex to an index in the hash table $\mathbf{\Phi}_i$, we employ a spatial hashing function. This mapping is achieved through the following equation:

$$h(\mathbf{p}_j) = (\bigoplus_{k=1}^{2} (p_{j,k} \cdot \Pi_k)) \bmod T_i, \tag{6}$$

where $p_{j,k}$ denotes the k-th component (x or y coordinate) of the j-th vertex coordinate. The $\bigoplus$ symbol denotes the bitwise XOR operation, applied cumulatively over both dimensions. Π_k represents predefined large prime numbers, one for each dimension. Specifically, we choose $\Pi_1 = 1$ and $\Pi_2 = 2654435761$ [35]. Finally, T_i is the size of the hash table for the i-th level of our multi-scale grid structure.

Utilizing this hashing mechanism, we retrieve the corresponding feature vectors from $\mathbf{\Phi}_i$ for each of the four vertices. This results in a set of vectors $\{\mathbf{q}_j\}_{j=1}^4$, each encapsulating the local characteristics of the grid at the respective vertex. To complete the feature extraction process, we calculate the relative position of $\mathbf{p}$ within the grid cell, denoted as $\mathbf{w} = \mathbf{p} - \mathbf{p}_{\text{lower}}$. This relative position is then used to perform bilinear interpolation on the retrieved vertex features, yielding the final feature vector:

$$\begin{aligned}\mathbf{v}_i = \phi(\mathbf{p}; \mathbf{\Phi}_i) =& (1 - w_x)(1 - w_y)\mathbf{q}_1 + w_x(1 - w_y)\mathbf{q}_2 \\ &+ (1 - w_x)w_y\mathbf{q}_3 + w_x w_y \mathbf{q}_4\end{aligned} \tag{7}$$

where w_x and w_y are the components of $\mathbf{w}$, representing the relative position of $\mathbf{p}$ within the grid cell along the x and y axes, respectively.

To enhance the learning of high-frequency functions, promote faster convergence, and improve generalization capabilities, we incorporate frequency information into our feature representation. Specifically, we apply Fourier feature encoding to the interpolated grid features. This encoding is achieved through the following equation:

$$\gamma_i(\mathbf{v}_i) = \left[\sin\left(2\pi \cdot \mathbf{B}_i \cdot \mathbf{v}_i^\top\right)\right]^\top, \tag{8}$$

where $\gamma_i(\mathbf{v}_i)$ represents the encoded features for level i, and $\mathbf{B}_i$ denotes a learnable frequency transform matrix specific to that level. The elements of $\mathbf{B}_i$ are initialized using a normal distribution $\mathcal{N}(0, \sigma_i^2)$, where the standard deviation σ_i for the i-th level is calculated as $\sigma_i = 5.0 \times 2^i$. This exponential scaling of the standard deviation across levels allows for the capture of features at progressively higher frequencies as i increases.

The Fourier-encoded features are subsequently processed by an MLP incorporating sinusoidal activation functions, as depicted in Fig. 1. This MLP architecture employs a multi-scale approach, with each layer operating on features derived from different spatial resolutions. Specifically, the i-th layer, denoted as L_i, takes as input the output $\mathbf{g}_{i-1}$ from the preceding layer, except for the first layer which directly processes the input position $\mathbf{p}$. The output of each layer is then combined with the corresponding Fourier-encoded features $\gamma_i(\mathbf{v}_i)$ to produce an updated feature representation $\mathbf{g}_i$, as formalized in the following equations:

$$\mathbf{f}_i = \sin\left(\mathrm{L}_i(\mathbf{g}_{i-1})\right), \quad \mathbf{g}_i = \mathbf{f}_i + \gamma_i(\mathbf{v}_i), \tag{9}$$

where $\sin(\cdot)$ is applied element-wise to the output of L_i.

Our network comprises multiple hidden layers, each consisting of 128 units. This hierarchical structure generates a set of intermediate outputs $G = \{\mathbf{g}_1, \mathbf{g}_2, \mathbf{g}_3\}$, effectively integrating information across different spatial scales.

2) Multi-scale Feature Integration: The multi-scale features obtained through Fourier feature representations are integrated to form a comprehensive representation of the intensity field. This integration is achieved by summing the intermediate outputs from each scale:

$$\mathcal{F}(\mathbf{p}) = \sum_{i=1}^{3} \mathbf{g}_i, \tag{10}$$

where $\mathcal{F}(\mathbf{p})$ represents the integrated multi-scale feature at position $\mathbf{p}$.

To further refine this multi-scale representation, we employ a three-layer CNN with ReLU activations. This CNN structure is designed to efficiently process and integrate the information captured in $\mathcal{F}(\mathbf{p})$:

$$E_\theta(\mathbf{p}) = \mathrm{Conv}_3\left(\mathrm{ReLU}\left(\mathrm{Conv}_2\left(\mathrm{ReLU}\left(\mathrm{Conv}_1(\mathcal{F}(\mathbf{p}))\right)\right)\right)\right), \tag{11}$$

where $E_\theta(\mathbf{p})$ is the final output of our FMNF at position $\mathbf{p}$.

Each convolutional layer uses a 3×3 kernel and maintains the spatial dimensions of the input, with 64 channels in the intermediate layers (Conv_1 and Conv_2) and a single-channel output in the final layer (Conv_3). This architecture facilitates adaptive refinement of the multi-scale features, enhancing the overall reconstruction quality by leveraging complementary information present at different resolutions.

3) Network Training: Leveraging the pre-trained iterative unfolding network (Section II-B), we train the FMNF network E_θ to learn a neural representation of the residual image given sparse-view projections. This process incorporates the initial estimate $\mathbf{x}_{\text{init}}$ derived from the sparse projections $\mathbf{y}$. The FMNF network is optimized to refine the residual image $\mathbf{x}_{\text{res}}$, effectively shifting the optimization from the image domain to the parameter space of the FMNF.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/TCI.2025.3536078, IEEE Transactions on Computational Imaging

Training of E_θ is performed by minimizing a composite loss function:

$$\begin{aligned}\mathcal{L}_\theta =&\mathcal{L}_{\text{mse}}(\mathbf{A}(E_\theta(\mathbf{p})+x^n_{\text{init}}), y^n)\\ &+\lambda\left(\alpha\mathcal{L}_{\text{wave}}(E_\theta(\mathbf{p}))+(1-\alpha)\mathcal{L}_{\text{TV}}(E_\theta(\mathbf{p}))\right),\end{aligned} \tag{12}$$

where x^n_{init} represents the initial reconstruction image slice obtained from the iterative unfolding network, y^n denotes the corresponding sparse-view projection, λ is a regularization parameter, and $\alpha \in [0, 1]$ balances the contributions of wavelet and TV regularization.

This loss function comprises three components: a data fidelity loss ($\mathcal{L}_{\text{mse}}$), a wavelet domain sparsity regularization loss ($\mathcal{L}_{\text{wave}}$), and a total variation (TV) regularization loss ($\mathcal{L}_{\text{TV}}$). The TV regularization, based on the anisotropic TV norm, encourages piecewise smoothness in the reconstructed image while preserving important edge details [38].

We define the wavelet loss ($\mathcal{L}_{\text{wave}}$) using a two-level wavelet transform, implemented with the "pytorch_wavelets" library [39]:

$$\mathcal{L}_{\text{wave}}(E_\theta(\mathbf{p})) = \|\text{DWT}_h(E_\theta(\mathbf{p}))\|_1, \tag{13}$$

where DWT_h represents the high-frequency components of the discrete wavelet transform. This formulation promotes sparsity specifically in the high-frequency coefficients of the wavelet domain.

Once the network is trained, the reconstructed image is generated by performing inference over all spatial coordinates. The final reconstructed image is obtained as:

$$x^{*,n} = E_{w^*}(y^n, x^n_0) + E_{\theta^*}(\mathbf{p}), \tag{14}$$

where $\mathbf{x}^{*,n}$ represents the final reconstructed image slice corresponding to the sparse-view projection y^n, E_{w^*} and E_{θ^*} represent the optimized mappings learned by the iterative unfolding network and the FMNF network, respectively. x^n_0 refers to the reconstruction image slice obtained via FBP from y^n.

III. Experiments and Results

This section presents a comprehensive evaluation of the proposed Ca-FMNF method using both simulated and real clinical datasets. Our assessment includes qualitative visual analysis and quantitative measurements. For quantitative evaluation, we employ two widely recognized metrics: Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM). These metrics are calculated for all slices in the test sets, with the average scores reported as the final results.

To benchmark our method, we compare it against five state-of-the-art approaches: FBP, FBPConvNet [39], LEARN [23], SCOPE [31], and NeRF [24]. FBP serves as a baseline, illustrating the extent of artifacts in sparse-view reconstruction. FBPConvNet represents a supervised deep learning approach that applies convolutional neural networks to enhance FBP reconstructions. LEARN exemplifies an iterative unfolding model that integrates deep learning into traditional iterative reconstruction frameworks. SCOPE utilizes implicit neural fields to generate full-projection sinograms through a re-projection strategy. NeRF, another neural field-based method, incorporates prior image information into MLP weights for image reconstruction.

We conducted our experiments using Python and PyTorch libraries on two NVIDIA Tesla V100 GPUs, each equipped with 32GB of memory. Our evaluation encompasses various sparse-view scenarios to thoroughly assess the performance of each method.

A. Simulated Numerical Experiments

1) AAPM Simulated Dataset: This study utilized the AAPM Low-Dose CT Grand Challenge dataset from 2016, comprising CT scans from 10 patients with a total of 2,378 slices. We simulated a fan-beam X-ray setup with a 120kV scan voltage, 600mm source-to-rotation center distance, and 590mm detector-to-rotation center distance. The detector array consisted of 624 elements. We employed the ODL to project original CT slices, generating sparse sinograms at 20 and 60 angles. ODL's FBP algorithm subsequently reconstructed these sinograms into initial CT images. The original slices served as ground truth, while the sparse sinograms and their reconstructed images constituted the network's input data. The dataset was divided into training and testing sets. CT slices from 8 patients (1,943 slices) were used to train supervised baseline methods, while 435 slices from the remaining 2 patients formed the test dataset. All images were standardized to a 256×256 pixel matrix.

2) Parameter Settings: For the proposed Ca-FMNF method, we pre-trained the SSM-based iterative unfolding network on 2,880 slices from the DeepLesion dataset. Training parameters included a batch size of 1, 50 iterations, and an initial learning rate of 1×10^{-4} with cosine annealing to 1×10^{-5}. We employed the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The FMNF network training used a fixed learning rate of 1×10^{-4} for 2000 iterations, with regularization parameters $\lambda = 0.01$ and $\alpha = 0.1$.

For the compared methods, FBPConvNet and LEARN were trained on the AAPM training dataset (1,943 slices) for 50 epochs. Their learning rate was initially set to 1×10^{-4} and gradually decayed to 1×10^{-5} using a cosine annealing strategy. We employed the Adam optimizer with default settings for both methods. SCOPE was trained on sparse projection sinograms from the test set using the Adam optimizer. The training process involved 1000 iterations with a fixed learning rate of 1×10^{-4}. After training, the model estimated full-projection sinograms through a re-projection strategy. These estimated full-projection sinograms were then used to reconstruct the final CT images via the FBP algorithm.

For NeRF, we first pre-trained the network using adjacent slices as prior for 2000 iterations with a learning rate of 5×10^{-4}. Subsequently, the neural field training was conducted for 2000 iterations with a learning rate of 5×10^{-5}.

3) Simulated Reconstruction Results: Fig. 3 presents the reconstruction results of lungs and abdomen under 60 views. The supervised methods, FBPConvNet and LEARN, demonstrated satisfactory performance in suppressing noise and artifacts. However, examination of the ROI reveals a degree of over-smoothing in both methods, particularly in FBPConvNet.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI: 10.1109/TCI.2025.3536078, IEEE Transactions on Computational Imaging

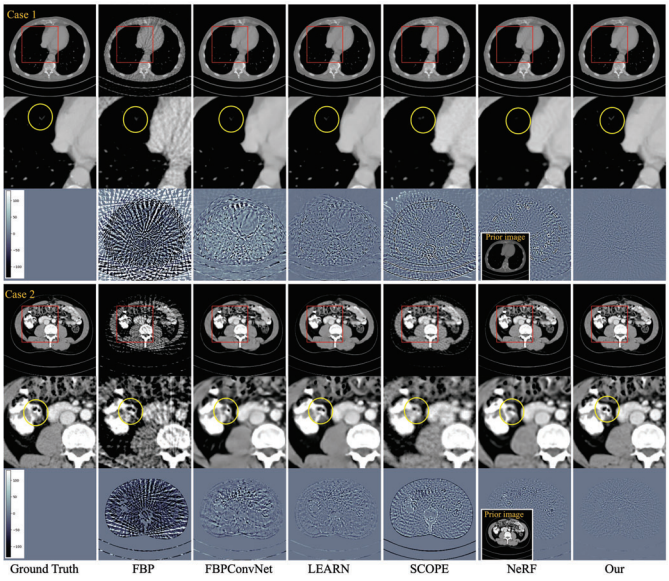

Fig. 3. Compares reconstruction algorithms using 60 views for two cases on AAPM simulated dataset. Each case includes reconstructed images, ROIs, and difference images. The display windows are set to [-1000, 1000] for Case 1 and [-160, 240]for Case 2. For NeRF, the prior image selected for pre-training is presented.

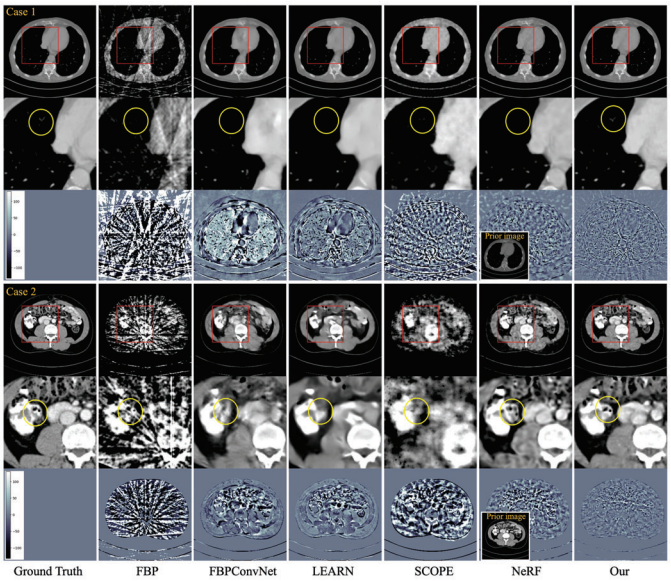

Fig. 4. Compares reconstruction algorithms using 20 views for two cases on AAPM simulated dataset. The display windows are set to [-1000, 1000] for Case 1 and [-160, 240]for Case 2. For NeRF, the prior image selected for pre-training is presented.

This over-smoothing can be attributed to the lack of data-driven constraints in FBPConvNet's reconstruction process, which essentially performs denoising on the initial FBP image, resulting in suboptimal structure preservation. The neural field-based SCOPE method exhibited significant noise and blurred edges in the reconstructed images. In contrast, NeRF, another neural field-based approach, achieved better clarity compared to SCOPE. However, NeRF's performance relies heavily on prior images, and its noise suppression capabilities remain inferior to the supervised LEARN method. Our proposed Ca-FMNF method, which integrates data-driven and neural representation approaches, achieved the most visually appealing results. By capturing data priors from the data-driven model and cascading neural field representation to optimize the residual images, our method successfully reconstructed artifact-free CT images while preserving fine anatomical structures. This superior performance is further corroborated by the difference images, which demonstrate the smallest deviation between our method's results and the ground truth.

Fig. 4 displays the reconstruction outcomes from a mere 20 views projection series. The FBPConvNet method exhibits pronounced streaking artifacts, while LEARN demonstrates superior performance in mitigating noise and artifacts compared to FBPConvNet. However, LEARN still suffers from considerable edge erosion, as evidenced in the difference images. SCOPE's ability to attenuate noise and artifacts remains limited in this highly sparse scenario. In contrast, NeRF achieves notably superior results, preserving finer details. This performance can be attributed to the utilization of priors that closely align with the reconstruction targets. However, it is important to note that obtaining such closely matched priors may not always be feasible in practical applications, potentially limiting NeRF's efficacy when there is significant divergence between the prior and the target. Under the constraints of highly sparse 20-view projections, our proposed Ca-FMNF method consistently demonstrates reduced noise and artifacts. It also captures anatomical details more effectively than the LEARN model. Nevertheless, it is not entirely immune to mild over-smoothing effects.

Table I presents the quantitative evaluation results on the AAPM dataset. Our proposed method, Ca-FMNF, demonstrates superior performance in terms of PSNR and SSIM for both 20 and 60 projections. Notably, Ca-FMNF, when pre-trained on an external dataset (DeepLesion) and further refined with cascaded neural fields, outperforms the supervised training on an internal dataset (AAPM) against the deep iterative unfolding baseline model, LEARN, achieving higher SSIM values. This result underscores the effectiveness of our approach in leveraging external data and advanced neural architectures. In comparison with NeRF, our method shows comparable results with 20 projections but exhibits superior performance with 60 projections. It is important to note that the quality of NeRF's reconstructions is largely dependent on the variance between the prior and target image data distributions, with reconstructions being less constrained by projection data. In our approach to pre-training NeRF, we carefully selected prior images to align closely with the target data distribution, as evidenced by the prior images in the third row of Fig. 3 and 4. Conversely, the SCOPE method did not surpass the performance of the supervised FBPConvNet. This outcome may be attributed to our experimental approach of using FBP to reconstruct the full-projection sinogram without additional denoising steps. Overall, this quantitative assessment highlights the superior performance of our Ca-FMNF method on the AAPM simulated dataset. The results demonstrate the potential of combining external pre-training, cascaded neural fields, and careful consideration of data dis-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/TCI.2025.3536078, IEEE Transactions on Computational Imaging

tributions in improving sparse-view CT reconstruction quality.

TABLE I
QUANTITATIVE EVALUATIONS OF ALL THE COMPARED METHODS ON DIFFERENT VIEWS

Methods	12 views		20 views		60 views	
	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR
FBP	0.196	15.955	0.367	20.170	0.531	25.590
FBPConvNet	0.834	27.950	0.856	30.267	0.919	34.922
LEARN	0.833	27.682	0.903	32.752	0.959	38.152
SCOPE	0.401	19.089	0.786	29.325	0.878	31.872
NeRF	0.827	30.052	0.919	35.825	0.967	40.020
Our	0.882	33.084	0.953	37.265	0.983	43.347

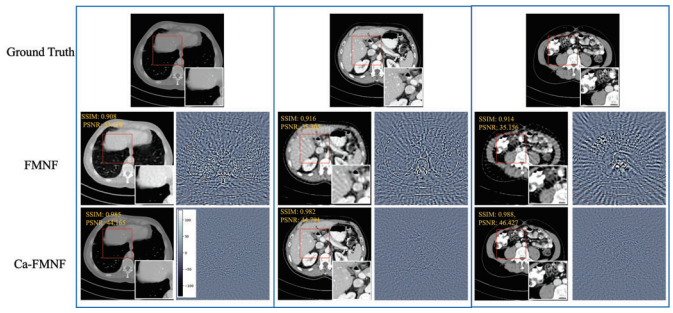

Fig. 5. Display of CT reconstruction results from 60 views on AAPM simulated dataset. The first row presents the GT images for reference. The second row displays the results reconstructed using only the FMNF algorithm, where the FMNF network was trained without the iterative unfolding network. The third row depicts the enhanced reconstructions achieved by employing the complete Ca-FMNF framework, which includes the iterative unfolding network followed by FMNF refinement. Each image has its corresponding SSIM and PSNR values below. The display window for the lung set at [-1000, 1000], and for the abdomen at [-160, 240].

B. Ablation Study

To rigorously assess the efficacy of our proposed Ca-FMNF method, we conducted a comprehensive series of ablation studies. These experiments were designed to validate the effectiveness of key components within our framework, specifically:

1) The cascaded SSM-based iterative model
2) The frequency-encoded multi-scale neural field representation
3) The hybrid loss function

Through these ablation studies, we aim to provide a detailed understanding of how each component contributes to the overall performance of our method in sparse-view CT reconstruction.

1) Effectiveness of the Cascaded SSM-Based Iterative Model: In our framework, the initial CT image prediction is obtained through an iterative unfolding network based on SSBs. The subsequent stage involves the FMNF network, which specifically targets the optimization of the residual image, refining the initial prediction by integrating the Radon transform into its loss function for enhanced joint optimization in both image and projection domains. To validate the efficacy of this cascaded approach, we conducted experiments using the 60-view data, comparing the reconstruction results of using FMNF (without the iterative unfolding network) and those obtained with the cascaded approach. The results, shown in Fig.5, demonstrate the reconstructed images for three samples using the proposed Ca-FMNF algorithm. For comparison, reconstructions without the cascaded approach, labeled as FMNF, were obtained by training the FMNF network without the prior image from the pre-trained iterative unfolding network. The Ca-FMNF reconstructions exhibit high-quality images with clear anatomical structures and well-defined organ boundaries. The comparison between the two indicates that embedding the initial prior image from iterative reconstruction into the subsequent FMNF optimization significantly enhances the reconstruction quality. The difference images further illustrate that the Ca-FMNF, when compared to FMNF alone, yields a smaller discrepancy with the ground truth. Thus, we conclude that the cascaded iterative unfolding network effectively provides valuable prior knowledge, and the targeted optimization of the residual image by FMNF is effective for accurately reconstructing sparse-view CT images, enhancing the representation of high-frequency features and capturing fine details of patient anatomy.

2) Effectiveness of the Frequency-Encoded Multi-Scale Neural Field Representation: To ascertain the effectiveness of our FMNF representation, we conducted a comparative analysis using the 60-view data, employing two different neural field approaches: our FMNF and a neural field using a MLP with position encoding [25]. Fig. 6 presents a comparison of the two methods, with Ca-NeRF depicting the MLP-based neural field reconstruction. It is evident that the Ca-NeRF approach yields images with more noise compared to our Ca-FMNF. Moreover, the enlarged ROIs demonstrate that, despite some smoothing by our method, it maintains superior clarity in edge and structural detail than the Ca-NeRF approach, which is further corroborated by the difference images indicating a smaller deviation from the ground truth with our method. Fig. 6(B) tracks the progression of evaluation metrics through iterations, showing our FMNF's quicker convergence and higher performance across all metrics compared to the Ca-NeRF approach. This evidences that our Ca-FMNF, by integrating multi-scale spatial decomposition and frequency encoding, effectively models high-frequency details and improves reconstruction efficiency.

3) Validating the Efficacy of the Hybrid Loss Function in FMNF Training: To assess the efficacy of the hybrid loss function (Eq. 12) in training the Ca-FMNF model, we conducted a comparative analysis using three distinct training loss configurations. The first configuration, denoted as Ca-FMNF-D, was trained using only the data fidelity loss, focusing on minimizing the mean squared error between the reconstructed and actual CT images. The second variant, Ca-FMNF-T, incorporated both data fidelity loss and TV regularization loss, aiming to constrain the data with the smoothness of the residual image. The third and most comprehensive model, Ca-FMNF (Hybrid), employed a combination of data fidelity loss, TV regularization, and wavelet sparsity loss.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI: 10.1109/TCI.2025.3536078, IEEE Transactions on Computational Imaging

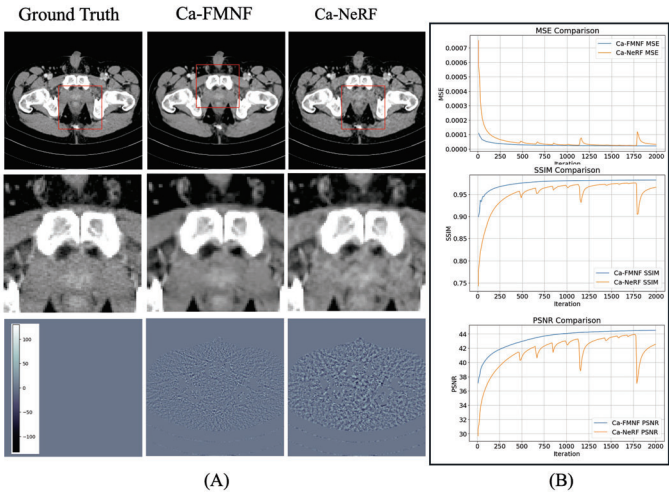

Fig. 6. Performance comparison between the proposed Ca-FMNF method and Ca-NeRF. The first column displays the ground truth CT images, the second column shows reconstructions by Ca-FMNF, and the third column shows those by Ca-NeRF. The insets within the red boxes magnify regions of interest. Difference images at the bottom row emphasize the disparity in reconstruction fidelity. The display window at [-160, 240]. The graphs on the right plot the MSE, SSIM, and PSNR across iterations.

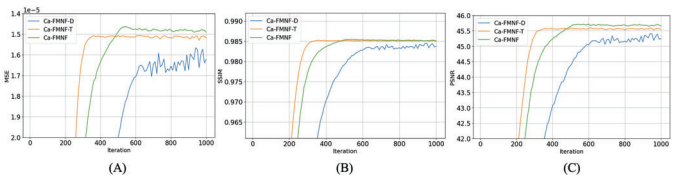

Fig. 7. Quantitative evaluation of the different Ca-FMNF model variations. The graphs depict the change in reconstruction accuracy over iterations as measured by three different metrics. A show the MSE, B illustrates the SSIM, and C represents the PSNR.

For the experimental setup, the parameters of the hybrid loss function, as detailed in Eq. 12, were set with $\lambda = 0.01$ and $\alpha = 0.2$. Fig.7 demonstrates the performance metrics—MSE, SSIM, and PSNR—over 1000 iterations for each of the three models, trained on a single slice from the 60-view data. The results revealed that the Ca-FMNF-T model outperformed the other models in terms of higher metric values and demonstrated a more stable convergence behavior, avoiding the pitfall of overfitting. The inclusion of the TV loss in Ca-FMNF-T contributed to its superior convergence properties compared to the Ca-FMNF-D model, which only utilized the data fidelity loss. The superior performance of Ca-FMNF-T suggests that the regularization, particularly TV regularization, effectively enhances the smoothness of the residual image, thereby increasing the overall stability and robustness of the reconstruction process. Moreover, the hybrid Ca-FMNF model, incorporating both TV and wavelet losses, showed a nuanced balance between data fidelity and regularization, leading to high-quality reconstructions with fewer artifacts and improved stability across iterations.

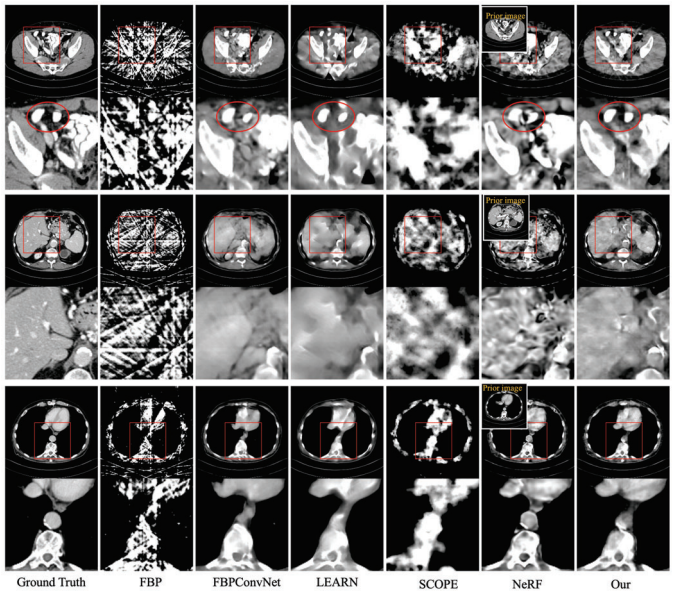

Fig. 8. Compares reconstruction algorithms using 12 views for three cases on AAPM simulated dataset. Each case includes reconstructed images, and ROIs. The display windows are set to [-160, 240]. For NeRF, the prior image selected for pre-training is presented.

C. Reconstruction Using Extremely Sparse Views

In this section, we examine the performance of various methods under extreme sparse-view configurations. To isolate the effects of reconstruction algorithms, we utilized the AAPM simulated dataset, specifically evaluating image reconstruction from 12-view projection data. To ensure fair comparison, all methods were retrained to achieve optimal performance under these conditions.

Fig.8 presents the visual results of different reconstruction methods. As evident from Fig.8, images reconstructed using FBP exhibit severe streak artifacts under such extreme sparse-view conditions. The ill-posed nature of ultra-sparse-view reconstruction poses significant challenges for supervised methods like FBPConvNet and LEARN, which fail to provide reliable image reconstructions. This limitation is further confirmed in the extracted ROIs. SCOPE shows markedly reduced effectiveness under these conditions, failing to produce discernible images.

In contrast, NeRF demonstrates improved performance in artifact suppression and structural preservation, as indicated by the ROIs highlighted. However, NeRF's reconstruction quality is notably influenced by the prior images used in training. Case 3 yields better results compared to Case 2, attributable to the closer alignment between pre-training data and target reconstruction distributions. The red-circled area in Case 1 reveals erroneous structural details, underscoring the impact of prior image selection.

Under severe under-sampling (12 projection angles), our proposed method achieves relatively better reconstruction quality. While the results exhibit some noise, artifacts, and loss of fine details due to over-smoothing, they demonstrate superior preservation of textural details compared to other examined approaches. The difference images provide clear

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/TCI.2025.3536078, IEEE Transactions on Computational Imaging

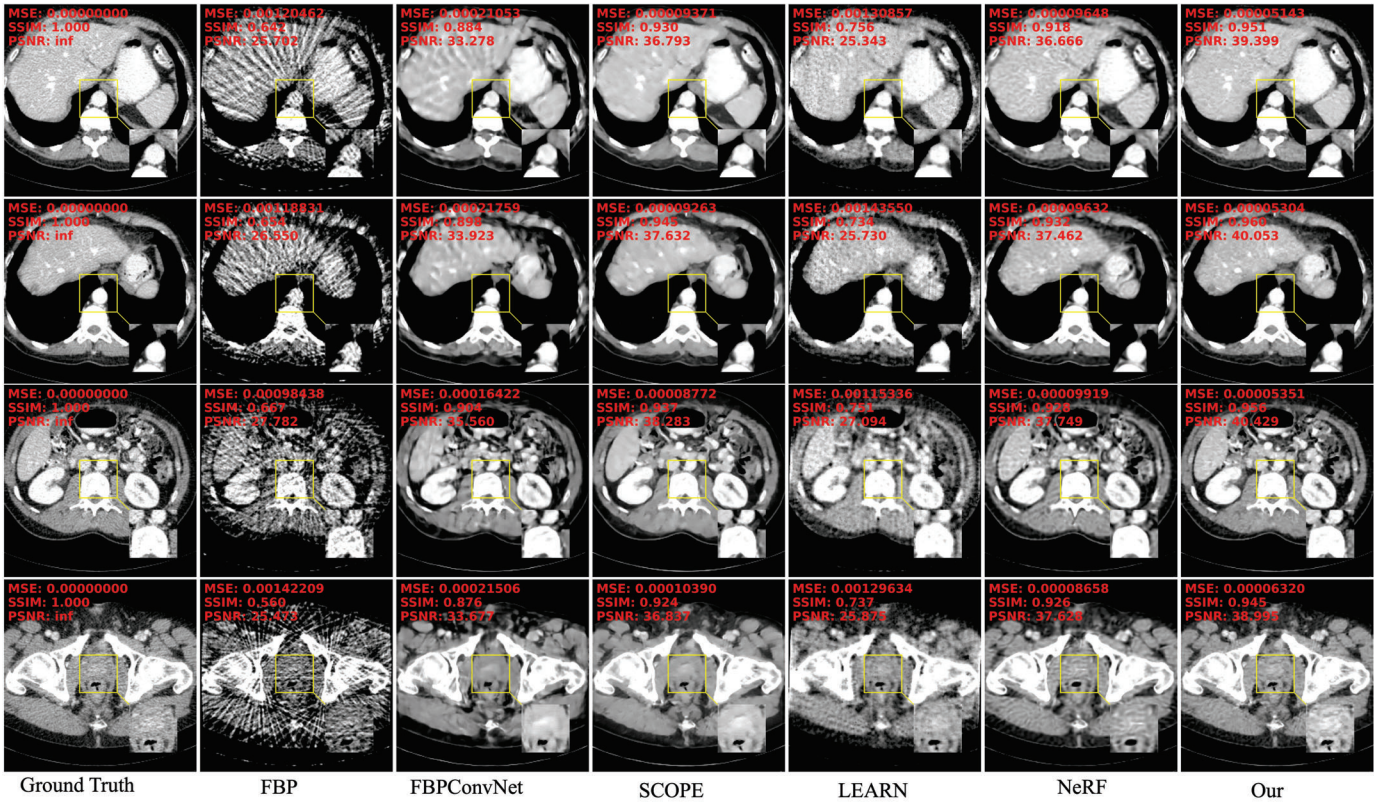

Fig. 9. Compares reconstruction algorithms using 60 views for four cases on AAPM rebinned dataset. Each case includes reconstructed images, and ROIs. The display windows are set to [-160, 240].

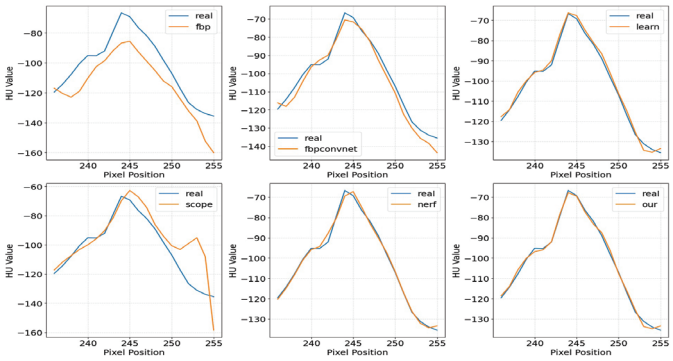

Fig. 10. Intensity profiles along the designated red lines in the ROIs of AAPM rebinned dataset.

evidence that our method's reconstructions show the least deviation from the ground truth.

To provide a comprehensive evaluation, quantitative analysis was conducted, with results presented in Table I.

D. Real Numerical Experiments

1) AAPM Rebinned Dataset: To evaluate our method's performance under realistic clinical conditions, we utilized raw data from the AAPM Low Dose CT Grand Challenge. The dataset was acquired using a Siemens Somatom Definition CT scanner. The scanning parameters included a tube voltage of 120 kV and tube current ranging from 200 mA to 500 mA. A helical scanning mode with flying focal spot was employed, with a rotation time of 500 ms and 2304 projections per rotation.

For computational efficiency, we converted the original helical scan geometry to fan-beam geometry [40]. Full-angle projection data were then rearranged and reconstructed using the FBP algorithm to generate reference images. In our evaluation setup, we selected 64 out of the 2304 original angular projections. Slices of 256×256 pixels were reconstructed using the FBP algorithm, with a voxel size of 1.7 mm. The dataset was divided into training and testing sets to ensure a robust evaluation of our method. We selected six patients, comprising 3573 slices, for training purposes. For testing, we used data from a separate patient, consisting of 703 slices.

The reconstruction parameters for the original helical geometry included a detector resolution of 736×64 and a detector pixel size of 1.2858 mm × 1.0947 mm, with a pitch range of 0.6 to 0.8. For the rearranged fan-beam geometry, the detector resolution was 736 per slice, with a detector pixel size of 1.2858 mm × 1.0 mm. The pitch parameter was not applicable in this case. Common to both geometries were the source-to-center distance of 595.0 mm and the source-to-detector distance of 1085.6 mm.

2) Parameter Settings: For the rebinned AAPM dataset, we employed the same comparative methods as in our simulation experiments. All supervised deep learning methods, including FBPConvNet and LEARN, were retrained to align with the imaging geometry of the rearranged data. The parameter settings for these methods remained consistent with those used

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI: 10.1109/TCI.2025.3536078, IEEE Transactions on Computational Imaging

in the simulation experiments. For neural field-based methods, we directly applied them to the test set without retraining, maintaining the same parameter configurations as in the simulation experiments. In the case of NeRF, we maintained our strategy of pre-training the model using adjacent slices with minimal data distribution differences from the target images.

3) Real Reconstruction Results: The reconstruction results with 64 projections are illustrated in Fig. 9. It's evident that FBP reconstructions suffer from excessive diffusion of artifacts, resulting in poor clarity and inability to discern structural details. While FBPConvNet improves image quality over FBP, it exhibits oversmoothing, failing to capture fine anatomical details accurately. LEARN, applying data constraints, mitigates oversmoothing to a degree, more clearly delineating structural details and tissue edges. However, examining enlarged ROIs reveals some inaccuracies.

In contrast, the SCOPE method, utilizing a neural field and re-projection strategy, shows suboptimal results compared to supervised deep learning methods, with evident artifacts and noise, albeit surpassing traditional FBP reconstruction. A closer inspection of specific ROIs indicates a noticeable blurriness in its outcomes. Conversely, NeRF demonstrates superior performance, with effective noise and artifact suppression and introduction of details and features, though the choice of prior images significantly impacts quality. Ultimately, our method achieves the best clarity in reconstructed images, accurately capturing structural details and edges. Additionally, quantitative reconstruction results in terms of PSNR and SSIM, as shown in the figure, affirm our method's superior performance.

Profile analyses conducted on the regions of interest across all reconstruction algorithms, as illustrated in Fig. 10, indicate that our method's reconstructed profiles align closely with the reference. The comparative plots showcase the performance of various algorithms, including FBP, FBPConvNet, LEARN, SCOPE, and NERF, alongside the reference data. Our proposed method demonstrates good agreement with the reference across the examined pixel range. Other methods exhibit varying degrees of deviation from the reference profile. The FBP reconstruction, for instance, shows more pronounced differences, particularly in areas of rapid intensity change.

IV. Conclusion

In this paper, we have presented a novel cascaded framework, Ca-FMNF, for sparse-view CT reconstruction. The proposed method integrates an iterative unfolding network based on SSMs with a FMNF representation. The SSM-based iterative unfolding network provides an effective initial reconstruction, which is further refined by the FMNF network through a continuous optimization process in the image space. The FMNF network employs a multi-scale grid structure for spatial decomposition and associates each scale with specific frequency bands through Fourier feature encoding, enabling efficient and precise modeling of local features.

Extensive experiments on the AAPM and clinical datasets demonstrate that our Ca-FMNF method outperforms state-of-the-art approaches in terms of both quantitative metrics and visual quality. The cascaded framework effectively leverages the strengths of both data-driven and model-based approaches, achieving superior reconstruction results with preserved edges and structural features. Furthermore, the hybrid loss function, incorporating data fidelity, wavelet sparsity, and total variation regularization, enhances the stability and robustness of the reconstruction process.

In the future, we plan to extend our Ca-FMNF framework to 3D CT reconstruction, which will enable more comprehensive and accurate visualization of anatomical structures. This extension will provide essential volumetric information for clinical applications, such as tumor detection and treatment planning.

[1] G. Zang, R. Idoughi, R. Li, P. Wonka, and W. Heidrich, "Intratomo: self-supervised learning-based tomography via sinogram synthesis and prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1960–1970.

[2] P. J. Withers, C. Bouman, S. Carmignato, V. Cnudde, D. Grimaldi, C. K. Hagen, E. Maire, M. Manley, A. Du Plessis, and S. R. Stock, "X-ray computed tomography," *Nature Reviews Methods Primers*, vol. 1, no. 1, p. 18, 2021.

[3] W. Xia, Z. Lu, Y. Huang, Z. Shi, Y. Liu, H. Chen, Y. Chen, J. Zhou, and Y. Zhang, "MAGIC: Manifold and graph integrative convolutional network for low-dose CT reconstruction," *IEEE transactions on medical imaging*, vol. 40, no. 12, pp. 3459–3472, 2021.

[4] J. Liu, H. Ding, S. Molloi, X. Zhang, and H. Gao, "TICMR: Total image constrained material reconstruction via nonlocal total variation regularization for spectral CT," *IEEE transactions on medical imaging*, vol. 35, no. 12, pp. 2578–2586, 2016.

[5] Q. Xu, H. Yu, X. Mou, L. Zhang, J. Hsieh, and G. Wang, "Low-dose X-ray CT reconstruction via dictionary learning," *IEEE transactions on medical imaging*, vol. 31, no. 9, pp. 1682–1697, 2012.

[6] Z. Li, L. Yu, J. D. Trzasko, D. S. Lake, D. J. Blezek, J. G. Fletcher, C. H. McCollough, and A. Manduca, "Adaptive nonlocal means filtering based on local noise level for CT denoising," *Medical physics*, vol. 41, no. 1, p. 011908, 2014.

[7] J. Wang, C. Wang, Y. Guo, W. Yu, and L. Zeng, "Guided image filtering based limited-angle CT reconstruction algorithm using wavelet frame," *IEEE Access*, vol. 7, pp. 99 954–99 963, 2019.

[8] W. Wu, F. Liu, Y. Zhang, Q. Wang, and H. Yu, "Non-local low-rank cube-based tensor factorization for spectral CT reconstruction," *IEEE transactions on medical imaging*, vol. 38, no. 4, pp. 1079–1093, 2018.

[9] X. Zheng, S. Ravishankar, Y. Long, and J. A. Fessler, "PWLS-ULTRA: An efficient clustering and learning-based approach for low-dose 3D CT image reconstruction," *IEEE transactions on medical imaging*, vol. 37, no. 6, pp. 1498–1510, 2018.

[10] P. Bao, W. Xia, K. Yang, W. Chen, M. Chen, Y. Xi, S. Niu, J. Zhou, H. Zhang, H. Sun, and others, "Convolutional sparse coding for compressed sensing CT reconstruction," *IEEE transactions on medical imaging*, vol. 38, no. 11, pp. 2607–2619, 2019.

[11] K. H. Jin, M. T. McCann, E. Froustey, and M. Unser, "Deep convolutional neural network for inverse problems in imaging," *IEEE transactions on image processing*, vol. 26, no. 9, pp. 4509–4522, 2017.

[12] G. Wang, J. C. Ye, and B. De Man, "Deep learning for tomographic image reconstruction," *Nature machine intelligence*, vol. 2, no. 12, pp. 737–748, 2020.

[13] F. Shamshad, S. Khan, S. W. Zamir, M. H. Khan, M. Hayat, F. S. Khan, and H. Fu, "Transformers in medical imaging: A survey," *Medical Image Analysis*, vol. 88, p. 102802, 2023.

[14] G. Xu, B. Zhang, H. Yu, J. Chen, M. Xing, and W. Hong, "Sparse synthetic aperture radar imaging from compressed sensing and machine learning: Theories, applications, and trends," *IEEE Geoscience and Remote Sensing Magazine*, vol. 10, no. 4, pp. 32–69, 2022.

[15] U. S. Kamilov, C. A. Bouman, G. T. Buzzard, and B. Wohlberg, "Plug-and-play methods for integrating physical and learned models in computational imaging: Theory, algorithms, and applications," *IEEE Signal Processing Magazine*, vol. 40, no. 1, pp. 85–97, 2023.

[16] F. Zhang, M. Zhang, B. Qin, Y. Zhang, Z. Xu, D. Liang, and Q. Liu, "REDAEP: Robust and enhanced denoising autoencoding prior for sparse-view CT reconstruction," *IEEE Transactions on Radiation and Plasma Medical Sciences*, vol. 5, no. 1, pp. 108–119, 2020.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI: 10.1109/TCI.2025.3536078, IEEE Transactions on Computational Imaging

[17] S. Goudarzi, A. Asif, and H. Rivaz, "Fast multi-focus ultrasound image recovery using generative adversarial networks," *IEEE Transactions on Computational Imaging*, vol. 6, pp. 1272–1284, 2020.

[18] Y. Song, L. Shen, L. Xing, and S. Ermon, "Solving Inverse Problems in Medical Imaging with Score-Based Generative Models," in *International Conference on Learning Representations*, 2022.

[19] A. Kazerouni, E. K. Aghdam, M. Heidari, R. Azad, M. Fayyaz, I. Hacihaliloglu, and D. Merhof, "Diffusion models in medical imaging: A comprehensive survey," *Medical Image Analysis*, vol. 88, p. 102846, 2023.

[20] H. Cao, C. Tan, Z. Gao, Y. Xu, G. Chen, P.-A. Heng, and S. Z. Li, "A Survey on Generative Diffusion Models," *IEEE Transactions on Knowledge & Data Engineering*, vol. 36, no. 7, pp. 2814–2830, 2024.

[21] J. He, S. Chen, H. Zhang, X. Tao, W. Lin, S. Zhang, D. Zeng, and J. Ma, "Downsampled imaging geometric modeling for accurate CT reconstruction via deep learning," *IEEE Transactions on Medical Imaging*, vol. 40, no. 11, pp. 2976–2985, 2021.

[22] Y. Zhang, H. Chen, W. Xia, Y. Chen, B. Liu, Y. Liu, H. Sun, and J. Zhou, "LEARN++: Recurrent Dual-Domain Reconstruction Network for Compressed Sensing CT," *IEEE Transactions on Radiation and Plasma Medical Sciences*, vol. 7, pp. 132–142, Feb. 2023.

[23] H. Chen, Y. Zhang, Y. Chen, J. Zhang, W. Zhang, H. Sun, Y. Lv, P. Liao, J. Zhou, and G. Wang, "LEARN: Learned experts' assessment-based reconstruction network for sparse-data CT," *IEEE transactions on medical imaging*, vol. 37, no. 6, pp. 1333–1347, 2018.

[24] L. Shen, J. Pauly, and L. Xing, "NeRP: implicit neural representation learning with prior embedding for sparsely sampled image reconstruction," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 1, pp. 770–782, 2022.

[25] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.

[26] J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove, "Deepsdf: Learning continuous signed distance functions for shape representation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 165–174.

[27] D. Rebain, W. Jiang, S. Yazdani, K. Li, K. M. Yi, and A. Tagliasacchi, "Derf: Decomposed radiance fields," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14 153–14 161.

[28] Y. Chen, S. Liu, and X. Wang, "Learning continuous image representation with local implicit image function," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 8628–8638.

[29] M. Tancik, P. Srinivasan, B. Mildenhall, S. Fridovich-Keil, N. Raghavan, U. Singhal, R. Ramamoorthi, J. Barron, and R. Ng, "Fourier features let networks learn high frequency functions in low dimensional domains," *Advances in neural information processing systems*, vol. 33, pp. 7537–7547, 2020.

[30] Y. Sun, J. Liu, M. Xie, B. Wohlberg, and U. S. Kamilov, "CoIL: Coordinate-Based Internal Learning for Tomographic Imaging," *IEEE Transactions on Computational Imaging*, vol. 7, pp. 1400–1412, 2021.

[31] Q. Wu, R. Feng, H. Wei, J. Yu, and Y. Zhang, "Self-supervised coordinate projection network for sparse-view computed tomography," *IEEE Transactions on Computational Imaging*, vol. 9, pp. 517–529, 2023.

[32] A. Gu and T. Dao, "Mamba: Linear-Time Sequence Modeling with Selective State Spaces," *arXiv preprint arXiv:2312.00752*, 2023.

[33] H. Guo, J. Li, T. Dai, Z. Ouyang, X. Ren, and S.-T. Xia, "Mambair: A simple baseline for image restoration with state-space model," *arXiv preprint arXiv:2402.15648*, 2024.

[34] Y. Liu, Y. Tian, Y. Zhao, H. Yu, L. Xie, Y. Wang, Q. Ye, and Y. Liu, "VMamba: Visual State Space Model," *arXiv preprint arXiv:2401.10166*, 2024.

[35] Z. Wu, Y. Jin, and K. M. Yi, "Neural fourier filter bank," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 14 153–14 163.

[36] K. Yan, X. Wang, L. Lu, and R. M. Summers, "DeepLesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning," *Journal of medical imaging*, vol. 5, no. 3, pp. 036 501–036 501, 2018.

[37] J. Adler, H. Kohr, and O. Öktem, "Operator discretization library (ODL)," *Zenodo*, 2017.

[38] X. Jin, L. Li, Z. Chen, L. Zhang, and Y. Xing, "Anisotropic total variation for limited-angle CT reconstruction," in *IEEE nuclear science symposuim & medical imaging conference*. IEEE, 2010, pp. 2232–2238.

[39] F. Cotter, "Uses of complex wavelets in deep convolutional neural networks," PhD Thesis, 2020.

[40] F. Wagner, M. Thies, L. Pfaff, O. Aust, S. Pechmann, D. Weidner, N. Maul, M. Rohleder, M. Gu, J. Utz *et al.*, "On the benefit of dual-domain denoising in a self-supervised low-dose ct setting," in *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, 2023, pp. 1–5.