DenoMamba: A fused state-space model for low-dose CT denoising

Şaban Öztürk, Oğuz Can Duran, and Tolga Çukur*, *Senior Member*

***Abstract*— Low-dose computed tomography (LDCT) lowers risks linked to radiation exposure, but relies on advanced denoising algorithms to maintain diagnostic image quality. Reigning learning-based models aim to separate noise from tissue signals by projecting LDCT images through multiple network stages that extract latent feature maps. Naturally, separation fidelity depends on the model's ability to capture short- to long-range contextual dependencies across spatial and channel dimensions of these maps. Existing convolutional and transformer models either lack sensitivity to long-range context or suffer from efficiency-related trade-offs, limiting their effectiveness. To achieve high-fidelity LDCT denoising, here we introduce a novel denoising method, DenoMamba, that performs state-space modeling (SSM) to efficiently capture both short- and long-range context in CT images. DenoMamba leverages a novel cascaded architecture equipped with spatial SSM modules to encode spatial context and channel SSM modules comprising a gated convolution network to encode content-aware features of channel context. Contextual feature maps are then consolidated with low-level spatial features via a convolution fusion module (CFM). Comprehensive experiments at 25% and 10% dose reduction demonstrate that DenoMamba outperforms state-of-the-art convolutional, transformer and SSM denoisers with average improvements of 1.6dB PSNR and 1.7% SSIM in image quality.**

***Index Terms*— low-dose computed tomography, denoising, restoration, state space, sequence models**

I. Introduction

A cornerstone in modern medical imaging, CT irradiates the body with a beam of X-rays to furnish detailed cross-sectional views of anatomy [1]. Unlike conventional radiography, CT relies on acquisition of multiple snapshots as the X-ray beam is rotated around the body, causing substantially elevated exposure to ionizing radiation with potential risks including cancer [2]. A mainstream approach to alleviate these health risks involves CT protocols that cap the tube current or exposure time to lower the number of incident photons and thereby the radiation dose [3]. However, as the signal-to-noise ratio (SNR) scales with the number of incident photons, dose reduction inevitably increases the noise component in CT images, significantly degrading image quality and potentially obscuring diagnostic features. Consequently, development of effective denoising methods is imperative to maintaining the diagnostic utility of LDCT images acquired under high levels of dose reduction [4].

CT denoising requires disentanglement of noise components from tissue signals, both of which manifest not only short-range but also long-range contextual dependencies. On one hand, a given tissue type is often distributed broadly across separate regions in anatomical cross-sections [5], so tissue signals can show significant correlations across spatially distant regions [6]. One the other hand, when measurement noise in the sinogram domain is propagated to the image domain, it can generate structured artifacts that span long distances, yielding an inherent degree of long-range contextual dependency among noise components [7]. Moreover, since the noise level in CT measurements scales with the underlying tissue signal intensity, noise components arising from distant tissue regions can show further dependencies, particularly when respective tissue signals are of relatively high intensity [8]. Consequently, both tissue signals and noise components in CT images carry substantial short- to long-range context, which can serve as important cues for effective denoising.

In recent years, deep learning models have superseded traditional approaches in LCDT denoising [9], [10], given their improved adaptability to the distribution of imaging data [11], [12]. These models hierarchically process input images across many network stages, wherein multiple sets of latent feature maps are extracted at each stage that encapsulate different image attributes (e.g., edges, textures) in separate feature channels. Given the diverse contextual dependencies of tissue signals and noise components in CT images, latent feature maps also exhibit significant spatial dependencies over short- to long-range distances [13], [14]. Furthermore, as the network depth increases, higher-levels of latent features are extracted that also manifest strong dependencies across the channel dimension due to overlapping or complementary information. In turn, the success of a denoising model in separating noise from tissue signals depends on its ability to discern idiosyncratic patterns of spatial and channel context in latent feature maps of CT images [15].

Convolutional models for LDCT denoising: Earlier studies in learning-based denoising have predominantly employed convolutional neural network (CNN) models to process LDCT images [16]–[22]. CNN models use compact convolution operators for local filtering driven by spatial distance between image pixels. This locality bias yields linear model complexity with respect to image dimensions and offers high expressiveness for local contextual features that are critical in delineating detailed tissue structure [23]–[25]. A number of advancements over vanilla CNNs have been proposed over

This study was supported in part by TUBA GEBIP 2015 and BAGEP 2017 fellowships awarded to T. Çukur and in part by TUBITAK Grant no. 124E850 awarded to Ş Öztürk. (Corresponding authors: Şaban Öztürk, saban.ozturk@hbv.edu.tr; Tolga Çukur, cukur@ee.bilkent.edu.tr).

Authors are with the Dept. of Electrical-Electronics Engineering and National Magnetic Resonance Research Center (UMRAM), Bilkent University, Ankara, Turkey, 06800. Ş. Öztürk is also with the Ankara Haci Bayram Veli University, Ankara, Turkey. T. Çukur is also with the Neuroscience Graduate Program Bilkent University, Ankara, Turkey, 06800.

the years to improve preservation of tissue boundaries [16], [19]–[21], performance near rare pathology [26], and realism [27]–[30]. However, despite these advances, CNN models characteristically struggle to capture long-range contextual dependencies, which can lead to suboptimal denoising, especially near regions of heterogeneous tissue composition where modeling spatial and channel dependencies is crucial for distinguishing signal from noise [31]–[33].

Transformer models for LDCT denoising: Later studies have instead adopted transformer models that employ self-attention operators to process images as a sequence of tokens (i.e., image pixels or patches) and perform non-local filtering driven by inter-token similarities to improve sensitivity to long-range contextual dependencies [14], [15], [34]–[36]. Note that evaluating similarity between all token pairs induces quadratic complexity with respect to sequence length, compromising computation efficiency [37], [38]. Since deploying vanilla transformers on high-resolution feature maps with pixel-level tokens can be computationally prohibitive, efficient adaptations are typically employed in practice. In these adaptations, feature maps can be downsampled or split into large-sized patches to reduce sequence length, or hybrid architectures can be adopted that reserve transformer branches for low-resolution processing [1], [15], [34], [36], [39], which limit sensitivity to short-range context [40]. Alternatively, efficient transformer adaptations use local windowing or low-rank strategies to build approximate attention operators [14], [41]. However, these strategies either lower the spatial resolution or range of attention operators, which limit sensitivity to long-range context [42]. Thus, it remains a significant challenge in transformer-based methods to maintain a favorable balance between short- and long-range sensitivity in high-resolution medical images, without introducing heavy model complexity that can elevate computational burden and compromise learning efficacy.

SSM models in imaging: An emerging framework in machine learning that promises to capture long-range contextual features while maintaining high computational efficiency is based on SSM [42], [43]. SSMs process images as a sequence of pixels whose relationships are modeled recurrently under linear complexity with respect to sequence length, so they can in principle be an ideal candidate to process LDCT images while avoiding the efficiency-related compromises in transformer models [44], [45]. In medical imaging, SSMs have shown promise for high-level tasks such as segmentation [46]–[48] and classification [49], as well as low-level tasks such as image synthesis [50] and reconstruction [51], [52]. Existing models in this domain either adopt a hybrid CNN-SSM approach with individual encoder-decoder stages constructed using a cascade of CNN and spatial SSM modules, which are fused via either residual connections or concatenation [46], [47], [49]–[53], or an SSM-focused approach with individual stages constructed using spatial SSM modules alone [48], [54]. Reliance on conventional spatial SSM modules can cause poor use of interdependencies across feature channels, degrading quality of feature extraction and downstream task performance. Furthermore, being fundamentally different than abovementioned imaging tasks, denoising typically operates in the original image domain with fine-grained pixel-level targets and requires sensitivity to subtle stochastic noise patterns, which can place distinct demands on architectural design. Thus, the utility of existing SSM models might be limited for LDCT denoising, where sensitive capture of diverse contextual features in CT images is key to model performance.

Proposed method: Here we introduce a novel SSM-based model, DenoMamba, to improve fidelity in LDCT image denoising by effectively capturing short- to long-range context across spatial and channel dimensions under high computational efficiency. To do this, DenoMamba leverages a novel architecture that cascades multiple FuseSSM blocks per network stage (Fig. 1). The proposed FuseSSM blocks convolutionally fuse the spatial context captured by a spatial SSM module with the channel context captured by a novel channel SSM module (Fig. 2). The proposed channel SSM module employs a secondary gated convolution network following the SSM layer in order to refine channel contextual features via content-aware modulation. Meanwhile, to improve preservation of low-level spatial representations in LDCT images, FuseSSM blocks are equipped with an identity propagation path. These building blocks empower DenoMamba to capture diverse contextual information in LDCT images, without necessitating downsampling or patching procedures that restrict spatial precision in transformers. Comprehensive evaluations on LDCT datasets acquired at 25% and 10% of nominal radiation doses demonstrate the superior performance of DenoMamba compared to state-of-the-art baselines. Code to implement DenoMamba is publicly available at https://github.com/icon-lab/DenoMamba.

Distinctions from recent SSM models: While DenoMamba adopts an hourglass design with encoder-decoder stages —a style ubiquitous in modern vision models due to its effectiveness in capturing hierarchical features across scales [55], [56]— it incorporates unique architectural features that distinguish it from recent SSM models for imaging tasks. Existing SSM-based models follow either hybrid CNN–SSM or SSM-focused design that rely solely on spatial SSM modules [46]–[54]. While a number of these models have included channel attention layers in their design, they have not considered use of SSM to capture contextual features in the channel dimension. In contrast, DenoMamba omits a separate CNN module for feature extraction in encoder-decoder stages, and employs learnable fusion of spatial and channel contextual features extracted by independent spatial SSM and channel SSM modules at each stage—an architectural strategy not observed in prior SSM-based imaging models. Furthermore, instead of a trivial adoption of SSM across the channel dimension that would result in spatially decoupled processing, DenoMamba leverages an explicit gating mechanism within the channel SSM modules for content-aware modulation of channel contextual representations.

After the release of our preprint [57], we observed several independent studies also investigating SSMs for LDCT denoising [44], [45], [58], highlighting the growing interest in this direction. However, DenoMamba incorporates key architectural innovations that set it apart from these approaches. Specifically, [44] combines spatial features extracted by parallel CNN, SSM, and max-pooling branches at each

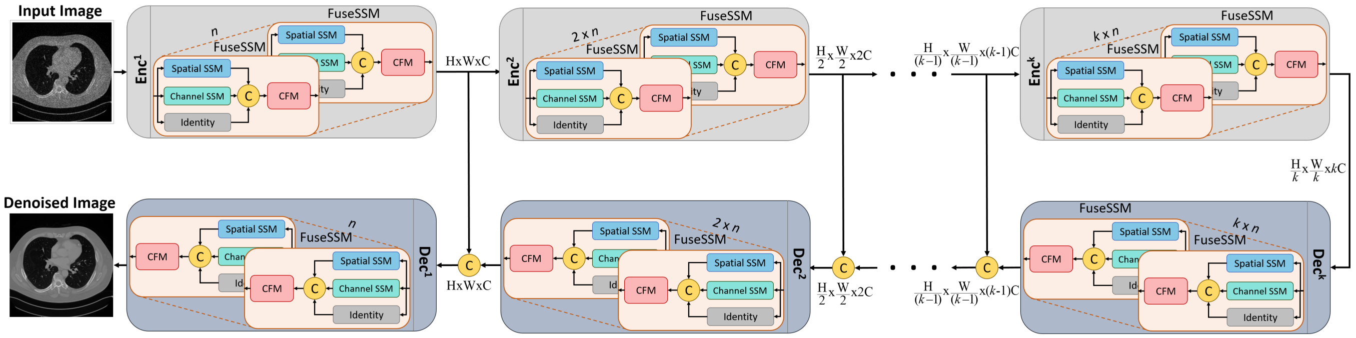
Input Image
Enc¹
FuseSSM
n
Spatial SSM
Channel SSM
Identity
C
CFM
H x W x C
Enc²
2 x n
$\frac{H}{2}$ x $\frac{W}{2}$ x2C
$\frac{H}{(k-1)}$ x $\frac{W}{(k-1)}$ x(k-1)C
Encᵏ
k x n
$\frac{H}{k}$ x $\frac{W}{k}$ xkC
Denoised Image
Dec¹
Dec²
Decᵏ

Fig. 1: The proposed DenoMamba architecture comprises residually-connected encoder-decoder stages, wherein each stage projects input feature maps through a cascade of FuseSSM blocks. The novel FuseSSM blocks use a spatial SSM module to extract spatial context, a channel SSM module to extract channel context, and an identity path to propagate low-level spatial features. Afterwards, low-level spatial features and their spatial- and channel-wise contextualized representations are consolidated via a convolutional fusion module (CFM).

encoder–decoder stage, but omits the multiplicative gating mechanism in the SSM branch that is often considered to enhance stability and expressivity while processing sequences. [45] integrates conventional SSM modules in the encoder but relies exclusively on CNN modules in the decoder, potentially compromising sensitivity to long-range contextual features during recovery of high-resolution denoised images. Meanwhile, [58] replaces the hourglass structure with a wavelet-domain bottleneck between CNN encoder and decoder stages, and serially cascades CNN and SSM branches in its bottleneck—a design based on fixed wavelet filters that may limit aggregation of diverse contextual representations. A shared characteristic of these models is their tight intertwining of CNN and SSM branches for contextual feature extraction that might constrain the propagation of global spatial information due to the local nature of convolutional filters, while depending solely on spatial SSM modules that can neglect channel context. In contrast, DenoMamba is built primarily on SSM layers and introduces dedicated spatial and channel SSM modules to jointly capture spatial and channel interdependencies within a unified framework. To our knowledge, DenoMamba is the first LDCT denoising method to leverage state-space modeling to capture both spatial and channel context in latent feature maps of CT images. Moreover, its novel channel SSM modules apply a gating mechanism after the SSM layers to extract content-aware channel features. Together, these innovations allow DenoMamba to deliver high spatial precision while preserving sensitivity to diverse contextual cues in LDCT images.

Contributions:

- To our knowledge, DenoMamba is the first LDCT denoising method that leverages state-space modeling across spatial and channel dimensions of latent feature maps.
- DenoMamba employs a novel architecture based on convolutional fusion of feature maps extracted via spatial and channel SSM modules along with an identity propagation path, enabling it to effectively consolidate a comprehensive set of CT image features.
- A novel channel SSM module is introduced that extracts content-aware features of channel context by cascading a transposed SSM layer operating over the channel dimension with a subsequent gated convolution network.

II. Theory

A. Problem Definition

LCDT image denoising involves suppression of elevated noise in low-dose CT scans due to reduced number of incident photons from the X-ray beam. Learning-based methods aim to solve this problem by training a neural network model to map noisy LDCT images onto denoised images that would be consistent with a normal dose CT (NDCT) scan. Let $x \in \mathbb{R}^{H\times W}$ denote the noisy LDCT image, and $y \in \mathbb{R}^{H\times W}$ denote the corresponding NDCT image, where H, W are the image height and width, respectively. Given a training set of T image pairs $(x_{tr}[i], y_{tr}[i])$ with $i \in [1\ T]$, a network model $f_\theta(\cdot)$ with parameters θ can be trained as follows:

$$\theta^* = \text{argmin}_\theta \sum_{i=1}^{T} \|f_\theta\left(x_{tr}[i]\right) - y_{tr}[i]\|_2^2. \tag{1}$$

Upon successful training, the optimal parameters θ^* that minimize the loss function should yield a model capable of effectively attenuating noise in LDCT images. The trained model can then be deployed to process novel LDCT images, generating denoised outputs as $\hat{y}_{test}[i] = f_{\theta^*}\left(x_{test}[i]\right)$.

B. DenoMamba

DenoMamba is the first LDCT image denoising method in the literature that uses SSMs to model spatial and channel context, to our knowledge. It employs a novel architecture based on FuseSSM blocks that aggregate low-level spatial features along with a comprehensive set of contextual features across spatial and channel dimensions, hence maintaining a favorable balance between short- and long-range sensitivity. In the following subsections, we describe the overall architecture of DenoMamba and the inner structure of FuseSSM blocks.

1) Overall Model Architecture: As depicted in Fig. 1, DenoMamba employs K encoder and K decoder stages. Each stage is implemented as a cascade of multiple FuseSSM blocks. Starting from the noisy LDCT image x taken as model input, encoder stages serve to extract latent contextualized representations via FuseSSM blocks and to resample the feature map dimensions. Let x^k_{enc} denote the feature map at the output of the kth encoder stage, with $k \in [1, 2, ..., K]$ and $x^0_{\text{enc}} = x$. The mapping through the kth encoder stage can be described

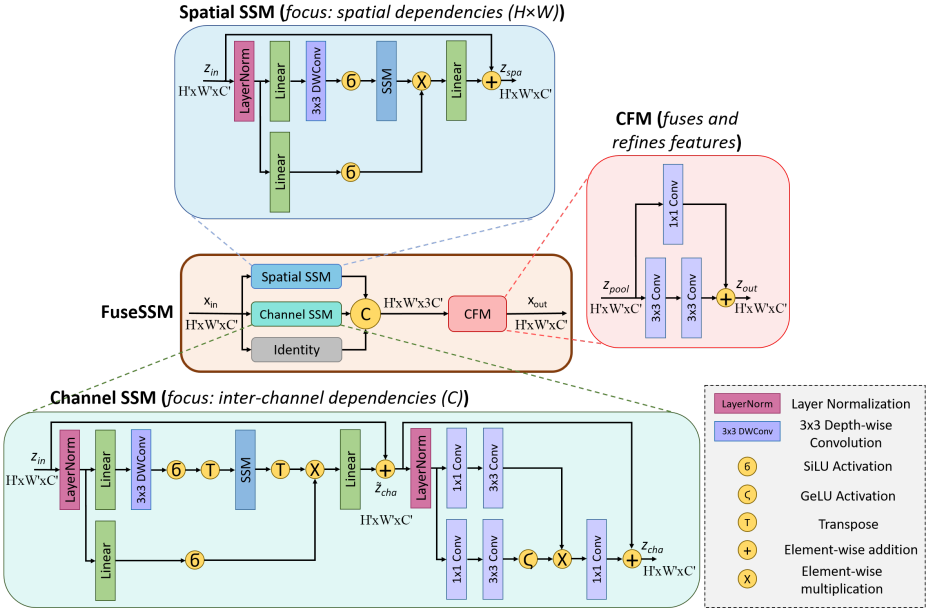

Fig. 2: Each FuseSSM block comprises parallel combination of a spatial SSM module, a channel SSM module and an identity propagation path, followed by a CFM module. The spatial SSM module performs convolutional encoding of image tokens after layer normalization, and processes the feature map via an SSM layer to capture contextual features across the spatial dimension. The channel SSM module performs convolutional encoding of image pixels after layer normalization, and processes the transposed feature map via an SSM layer to derive an initial set of contextual features across the channel dimension, which are projected through a gated convolutional network to obtain higher-level features. The CFM module nonlinearly fuses low-level features from the identity path with contextual features.

as follows:

$$x^k_{\text{enc}} = \begin{cases} \text{Down}(\text{Enc}^k \left(x^{k-1}_{\text{enc}}; \theta^k_{enc}\right)), & \text{if } k \neq K \\ \text{Enc}^k \left(x^{k-1}_{\text{enc}}; \theta^k_{enc}\right), & \text{if } k = K \end{cases} \tag{2}$$

where $\text{Enc}^k(\cdot) := \bigoplus_{r=1}^{E(k)} \text{FuseSSM}(\cdot)$ denotes composition of the kth stage via recursive application of $E(k)$ FuseSSM blocks, θ^k_{enc} denotes the parameters of these FuseSSM blocks, $\text{Down}(\cdot)$ denotes a learnable downsampling operator, and $x^k_{\text{enc}} \in \mathbb{R}^{\frac{H}{2^k} \times \frac{W}{2^k} \times 2^k C}$ for $0 < k < K$, $x^k_{\text{enc}} \in \mathbb{R}^{\frac{H}{2^{K-1}} \times \frac{W}{2^{K-1}} \times 2^{K-1} C}$ for $k = K$. Note that downsampling is performed in all but the final stage (i.e., $k = K$).

Starting from the encoded feature map x^K_{enc}, decoder stages then serve to recover a denoised image $\hat{y}$ from the latent representations via a cascade of FuseSSM blocks and resampling of feature map dimensions. The decoder stages follow a reversed order, such that x^k_{dec} denotes the feature map at the output of the kth decoder stage, with $k \in [1, 2, ..., K]$ and $x^0_{\text{dec}} = x^K_{\text{enc}}$. Thus, the mapping through the kth decoder stage can be described as follows:

$$x^k_{\text{dec}} = \begin{cases} \text{Dec}^k \left(\text{Up}(x^{k-1}_{\text{dec}}) + x^{K-k}_{\text{enc}}; \theta^k_{dec}\right), & \text{if } k \neq K \\ \text{Dec}^k \left(x^{k-1}_{\text{dec}} + x^{K-k}_{\text{enc}}; \theta^k_{dec}\right), & \text{if } k = K \end{cases} \tag{3}$$

where $\text{Dec}^k(\cdot) := \bigoplus_{r=1}^{D(k)} \text{FuseSSM}(\cdot)$ denotes composition of the kth stage via recursive application of $D(k)$ FuseSSM blocks, θ^k_{dec} denotes the parameters of FuseSSM blocks in the kth decoder stage, $\text{Up}(\cdot)$ denotes a learnable upsampling operator, and $x^k_{\text{dec}} \in \mathbb{R}^{\frac{H}{2^{K-k-1}} \times \frac{W}{2^{K-k-1}} \times 2^{K-k-1} C}$ for $0 < k < K$, $x^k_{\text{dec}} \in \mathbb{R}^{H \times W \times C}$ for $k = K$. Note that upsampling is performed on x^k_{dec} in the beginning of all but the final stage (i.e., $k = K$), and encoder feature maps from the respective encoder stage x^{K-k}_{enc} are residually added onto the input decoder maps to improve preservation of low-level structural representations in LDCT images. The final output of DenoMamba is taken as $\hat{y} = x^K_{\text{dec}}$.

2) FuseSSM blocks: DenoMamba is constructed with novel FuseSSM blocks that comprise a spatial SSM module to capture contextual representations in the spatial domain and a channel SSM module to capture contextual representations in the channel domain [59]. We propose to project input feature maps across three parallel pathways that propagate the contextualized representations from spatial and channel SSM modules, along with original input features. Afterwards, these representations are merged using a CFM. For a given FuseSSM block, a schematic of the individual components are depicted in Fig. 2.

The design of FuseSSM blocks in encoder and decoder stages are identical apart from variability in feature map dimensions. Thus, here we will describe the projections through a FuseSSM block without distinguishing between encoder/decoder stages. Assuming that the input feature map at the kth stage is $z_{in} = x^k \in \mathbb{R}^{H' \times W' \times C'}$, the respective FuseSSM block first projects the input through three parallel pathways to compute contextualized representations:

$$\{z_{\text{spa}}, z_{\text{cha}}, z_{\text{in}}\} = \{\text{SSM}_{\text{spa}}(z_{\text{in}}), \text{SSM}_{\text{cha}}(z_{\text{in}}), \text{I}(z_{\text{in}})\}, \tag{4}$$

where SSM_{spa} denotes the spatial SSM, SSM_{cha} denotes the channel SSM, and I denotes the identity propagation path. The extracted contextual representations are then pooled and convolutionally fused within the CFM module:

$$z_{\text{pool}} = \text{Concat}(z_{\text{spa}}, z_{\text{cha}}, z_{\text{in}}), \tag{5}$$

$$z_{\text{out}} = \text{Conv}^{1\times1}\left(z_{\text{pool}}\right) \oplus \text{Conv}^{3\times3}\left(\text{Conv}^{3\times3}\left(z_{\text{pool}}\right)\right), \tag{6}$$

where Concat denotes a concatenation operator that pools feature maps across the channel dimension, $\text{Conv}^{1\times1}$ and $\text{Conv}^{3\times3}$ respectively denote 1×1 and 3×3 convolutional layers, and $\oplus$ is the element-wise addition operator. The feature map $z_{\text{out}} \in \mathbb{R}^{H' \times W' \times C'}$ is taken as the output of the FuseSSM block.

Spatial SSM: Within the spatial SSM module, a first branch linearly embeds the input map and uses a nonlinearity to produce a gating variable $GP_{\text{spa}} \in \mathbb{R}^{H' \times W' \times \alpha C'}$:

$$GP_{\text{spa}} = \sigma(f_{\text{lin}}(f_{\text{LN}}(z_{\text{in}}))), \tag{7}$$

where σ is a SiLU activation function, f_{lin} denotes a learnable linear mapping that expands the feature map across the channel dimension by a factor α and f_{LN} denotes layer normalization over a batch of samples as described in [60]. A second branch performs linear embedding and convolutional encoding, followed by an SSM layer to derive $M_{\text{spa}} \in \mathbb{R}^{H' \times W' \times \alpha C'}$:

$$M_{\text{spa}} = \text{SSM}\left(\sigma(\text{DWConv}^{3\times3}(f_{\text{lin}}(f_{\text{LN}}(z_{\text{in}}))))\right), \quad (8)$$

where SSM denotes a state-space layer, $\text{DWConv}^{3\times3}$ refers to depth-wise convolution of kernel size 3×3.

Here, the state-space layer is implemented based on the Mamba variant in [59]. Accordingly, scanning is performed across two spatial dimensions of the input feature map to the SSM layer in order to expand it onto a sequence $s \in \mathbb{R}^{H'W' \times \alpha C'}$. The sequence is then processed via a discrete state-space model independently for each channel:

$$h[n] = \mathbf{A}h[n-1] + \mathbf{B}s[n], \quad (9)$$

$$\bar{s}[n] = \mathbf{C}h[n], \quad (10)$$

where $n \in [1\ H'W']$ is an integer denoting sequence index, h denotes the hidden state, $s[n]$ is the nth element of the input sequence. $\mathbf{A} \in \mathbb{R}^{N,N}$, $\mathbf{B} \in \mathbb{R}^{N,1}$, $\mathbf{C} \in \mathbb{R}^{1,N}$ are learnable parameters of the state-space model, with N indicating the hidden dimensionality. Note that $\mathbf{B}$ and $\mathbf{C}$ are taken to be functions of the input sequence in Mamba to enable input-adaptive processing [59]. The output sequence $\bar{s} \in \mathbb{R}^{H'W' \times \alpha C'}$ is remapped back onto the feature map $M_{\text{spa}} \in \mathbb{R}^{H' \times W' \times \alpha C'}$.

To compute the module output, M_{spa} is gated with GP_{spa}, and the result is linearly projected and combined with the input through a residual connection:

$$z_{\text{spa}} = z_{\text{in}} + f_{\text{lin}}(GP_{\text{spa}} \odot M_{\text{spa}}), \quad (11)$$

where $\odot$ denotes the Hadamard product operator, and f_{lin} is devised to use an expansion factor of $1/\alpha$ such that $z_{\text{spa}} \in \mathbb{R}^{H' \times W' \times C'}$ has matching dimensionality to z_{in}.

Channel SSM: Within the novel channel SSM module, a first branch produces a gating variable and a second branch performs state-space modeling on the sequentialized input feature map to capture contextual interactions in the channel dimension:

$$GP_{\text{cha}} = \sigma(f_{\text{lin}}(f_{LN}(z_{\text{in}}))), \quad (12)$$

$$M_{\text{cha}} = \text{SSM}\left((\sigma(\text{DWConv}^{3\times3}(f_{\text{lin}}(f_{LN}(z_{\text{in}})))))^{\top}\right)^{\top}, \quad (13)$$

where $\top$ denotes the transpose operator. Differing from the spatial SSM module, the channel SSM module captures channel context by transposing the input sequence prior to and after the SSM layer. This results in an intermediate set of contextual representations $\tilde{z}_{\text{cha}} \in \mathbb{R}^{H' \times W' \times C'}$ derived as:

$$\tilde{z}_{\text{cha}} = z_{\text{in}} + f_{\text{lin}}(GP_{\text{cha}} \odot M_{\text{cha}}). \quad (14)$$

Several components in DenoMamba—such as the depth-wise convolutional layers in FuseSSM blocks and the downsampling/upsampling operations in the encoder–decoder pathway—contribute to learning hierarchical spatial features. In contrast, channel-wise dependencies are primarily modeled by the SSM layers in the channel SSM module, which operate independently at each spatial location. This spatially decoupled processing limits the expressiveness of the resulting channel representations, as it lacks a mechanism to adaptively emphasize spatially relevant channel features. To address this limitation, we enhance the channel SSM module with a content-aware gating network. This gating mechanism adaptively modulates the channel-wise output based on spatial context, enabling the network to extract higher-order, task-relevant features and improving the overall representational power. For this purpose, a second gating variable $GP^2_{\text{cha}} \in \mathbb{R}^{H' \times W' \times C'}$ is first computed:

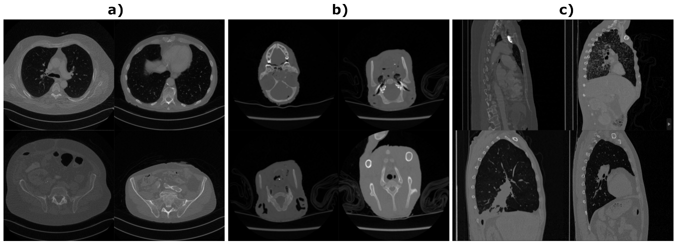

Fig. 3: Sample images from the analyzed datasets: (a) AAPM, (b) Piglet CT, (c) BIMCV-R. Four cross-sections are depicted from each dataset to portray representative samples.

$$GP^2_{\text{cha}} = \zeta(\text{DWConv}^{3\times3}(\text{Conv}^{1\times1}(\tilde{z}_{\text{cha}})), \quad (15)$$

where ζ is an ReLU activation function. GP^2_{cha} is then used to modulate latent features of $\tilde{z}_{\text{cha}}$:

$$z_{\text{cha}} = \text{Conv}^{1\times1}(GP^2_{\text{cha}} \odot \text{DWConv}^{3\times3}(\text{Conv}^{1\times1}(\tilde{z}_{\text{cha}}))) + \tilde{z}_{\text{cha}}. \quad (16)$$

As such, the module output $z_{\text{cha}} \in \mathbb{R}^{H' \times W' \times C'}$ has matching dimensionality to z_{in}.

3) Learning Procedures: Given a training set of image pairs $(x_{tr}[i], y_{tr}[i])$ with $i \in [1\ T]$, DenoMamba with parameters $\theta_{\text{enc}}, \theta_{\text{dec}}$ is trained via a pixel-wise ℓ_1-loss term:

$$\{\theta^*_{\text{enc}}, \theta^*_{\text{dec}}\} = \text{argmin}_{\theta_{\text{enc}}, \theta_{\text{dec}}} \sum_{i=1}^{T} \Big\| \text{Dec}^{(1:K)}\Big(\text{Enc}^{(1:K)}\left(x_{tr}[i]; \theta^{(1:K)}_{\text{enc}}\right); \theta^{(1:K)}_{\text{dec}} \Big) - y_{tr}[i] \Big\|_1. \quad (17)$$

Using the trained parameters $\{\theta^*_{\text{enc}}, \theta^*_{\text{dec}}\}$, the model can be deployed to process a novel LDCT image from the test set $x_{test}[i]$ to estimate a denoised output $\hat{y}_{test}[i]$ as:

$$\hat{y}_{test}[i] = \text{Dec}^{(1:K)}\Big(\text{Enc}^{(1:K)}\left(x_{test}[i]; \theta^{*\,(1:K)}_{\text{enc}}\right); \theta^{*\,(1:K)}_{\text{dec}}\Big) \quad (18)$$

III. Experimental Setup

A. Datasets

AAPM Dataset: Demonstrations of denoising performance were conducted on contrast-enhanced abdominal CT scans from the 2016 AAPM-NIBIB-MayoClinic Low Dose CT Grand Challenge [61]. Two different dose reduction levels were considered, resulting in 25%- and 10%-dose datasets. NDCT scans were acquired at 120 kV reference tube potential with 200 effective mAs as quality reference. LDCT at 25%-dose with 50 effective mAs and LDCT at 10%-dose with 20 effective mAs were simulated from NDCT images assuming a Poisson-Gaussian noise distribution [8], [37]. The training set comprised 760 NDCT-LDCT image pairs, the validation set had 35 pairs, and the test set had 200 pairs. There was no subject overlap among the three sets, and each set contained a mixture of CT images reconstructed at either 1 mm or 3 mm

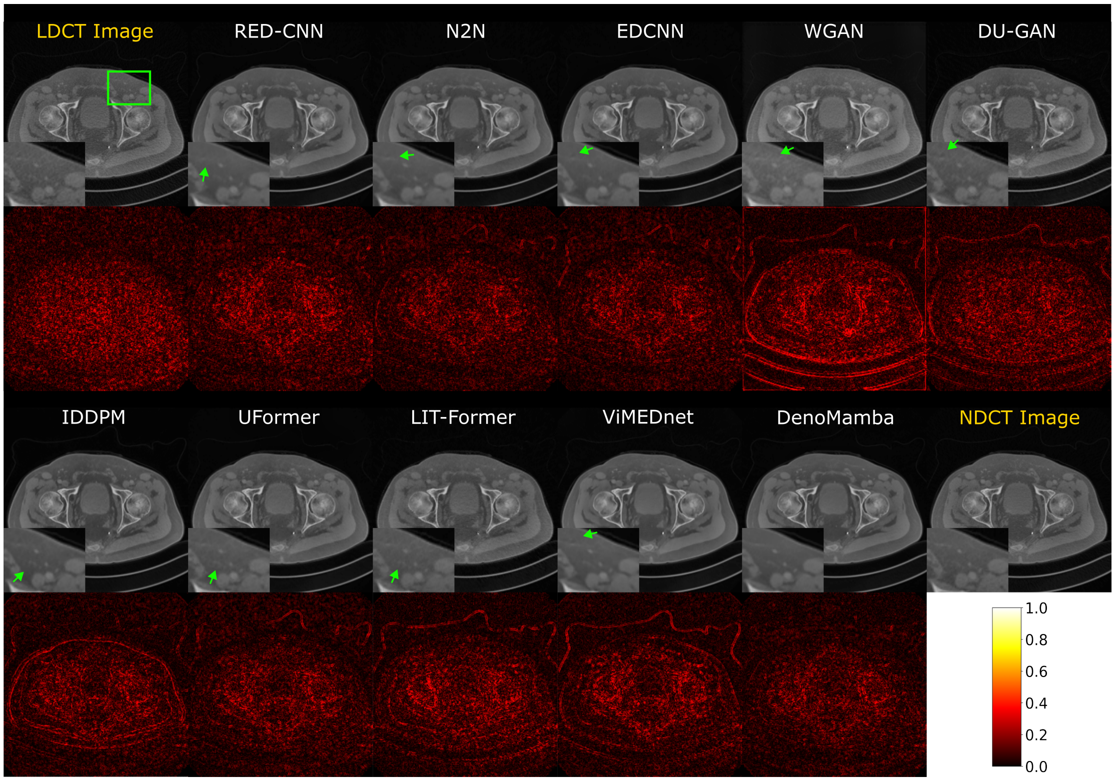

Fig. 4: Denoising results from the 25%-dose AAPM dataset are depicted for representative cross-sections. Images recovered by competing methods are shown along with the LDCT image (i.e., model input), and the NDCT image (i.e., ground truth). A display window of [-150 350] HU is used. To facilitate method comparisons, regions with visible differences are highlighted with zoom-in displays and arrows on denoised images, and absolute error maps with respect to the NDCT image are included (see colorbar).

slice thickness. All images were resized to 256×256 in-plane resolution.

Piglet CT Dataset: This dataset contained CT scans of a deceased piglet acquired at varying radiation doses attained by adjusting the tube current [62]. NDCT scans were acquired at 100 kV reference tube potential with 300 effective mAs as quality reference radiation dose. LDCT scans were acquired at 10%-dose by prescribing 30 effective mAs. As this dataset was primarily used for evaluating the generalization performance of models trained on the AAPM dataset, we only curated a test set comprising 350 NDCT-LDCT image pairs. All images had 0.625 mm slice thickness, and they were resized to 256×256 in-plane resolution.

BIMCV-R Dataset: This dataset included high resolution CT volumes from the BIMCV-R collection that did not have any missing pixel values [63]. NDCT images were acquired at 120 kV reference tube potential with 300 effective mAs as quality reference radiation dose and LDCT at 25%-dose were simulated. A test set comprising 500 NDCT-LDCT volume pairs were curated, where all images were sized to 256×256×8 volumetric resolution.

Sample NDCT images from the datasets are depicted in Fig. 3 to illustrate the general characteristics of CT scans.

B. Architectural Details

In DenoMamba, a $K = 4$ stage encoder-decoder architecture was used, where the number of FuseSSM blocks cascaded within a given stage varied as E = [1, 2, 2, 3] across encoder stages and as D = [2, 2, 1, 2] across decoder stages, respectively. Spatial resolution was lowered by a factor of 2 in each encoder stage except for the final one, while the channel dimensionality was set as [16, 32, 64, 128] across stages. Conversely, spatial resolution was increased by a factor of 2 in each decoder stage except for the final one, with the channel dimensionality set as [64, 32, 16, 16] across stages. Both spatial and channel SSM modules used a state expansion factor of N=16, a local convolution width of 4, and a block expansion factor of α=2.

TABLE I: Denoising performance of competing methods on the 25%-dose AAPM dataset. PSNR (dB), SSIM (%), and BC (%) metrics are listed as mean±std across the test set. Superscripts indicate significance levels relative to DenoMamba ($^{\circ}p<0.05$, $^{\diamond}p<0.005$, $^{\triangle}p<0.0005$). Boldface marks the method that offers the best performance for each metric.

	↑ **PSNR (dB)**	↑ **SSIM (%)**	↑ **BC (%)**
RED-CNN	$41.02 \pm 3.03^{\diamond}$	$96.25 \pm 1.65^{\triangle}$	$94.90 \pm 3.19^{\diamond}$
N2N	$40.72 \pm 2.98^{\diamond}$	$96.37 \pm 1.67^{\circ}$	$93.73 \pm 3.73^{\circ}$
EDCNN	$40.86 \pm 3.06^{\triangle}$	$96.07 \pm 1.72^{\diamond}$	$94.49 \pm 3.69^{\circ}$
WGAN	$39.79 \pm 2.54^{\triangle}$	$94.80 \pm 2.29^{\triangle}$	$91.71 \pm 4.46^{\circ}$
DU-GAN	$40.01 \pm 3.11^{\diamond}$	$94.48 \pm 3.13^{\diamond}$	$92.20 \pm 4.20^{\diamond}$
IDDPM	$41.04 \pm 2.22^{\triangle}$	$96.55 \pm 1.66^{\triangle}$	$93.71 \pm 3.84^{\diamond}$
UFormer	$41.05 \pm 2.79^{\circ}$	$96.76 \pm 1.64^{\triangle}$	$93.92 \pm 3.44^{\diamond}$
LIT-Former	$40.93 \pm 2.82^{\triangle}$	$96.05 \pm 1.87^{\triangle}$	$93.97 \pm 3.71^{\diamond}$
ViMEDNet	$41.73 \pm 3.12^{\triangle}$	$96.24 \pm 1.68^{\triangle}$	$94.62 \pm 3.73^{\diamond}$
DenoMamba	**42.69 ± 2.85**	**97.07 ± 1.74**	**95.24 ± 3.26**

C. Competing Methods

We demonstrated DenoMamba against several state-of-the-art methods for LDCT denoising. For fair comparisons, all competing methods were implemented with a pixel-wise ℓ_1-loss similar to DenoMamba. The only exceptions to this were generative models that were implemented with their original loss terms required to enable adversarial or diffusive learning.

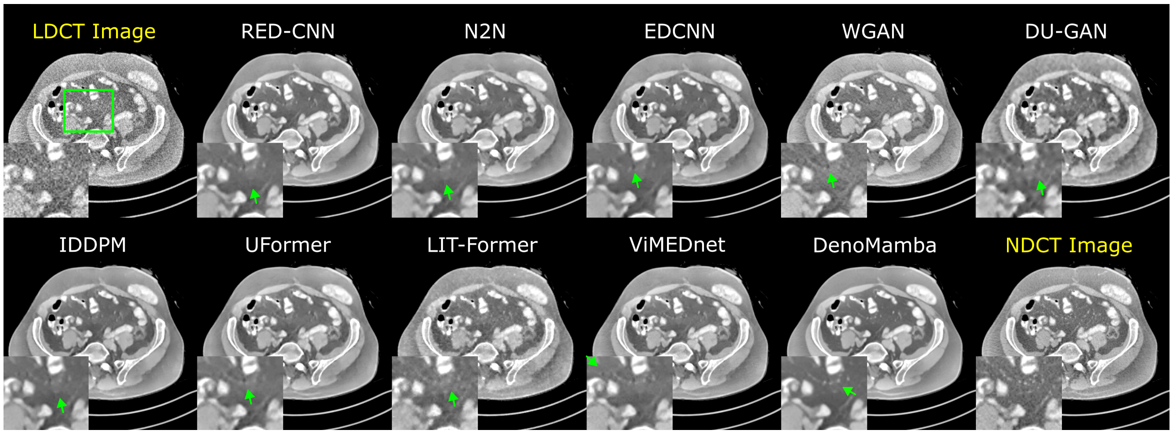

Fig. 5: Denoising results from the 10%-dose AAPM dataset are depicted for a representative cross-section. A display window of [-350 350] HU is used. Please see Supp. Fig. 2 for a visualization of error maps corresponding to the denoised images from competing methods.

TABLE II: Denoising performance of competing methods on the 10%-dose AAPM dataset. PSNR (dB), SSIM (%), and BC (%) metrics are listed as mean±std across the test set. Superscripts indicate significance levels relative to DenoMamba ($^{\circ}$p<0.05, $^{\diamond}$p<0.005, $^{\triangle}$p<0.0005). Boldface marks the method that offers the best performance for each metric.

	↑ **PSNR (dB)**	↑ **SSIM (%)**	↑ **BC (%)**
RED-CNN	$38.27 \pm 2.39^{\circ}$	$95.18 \pm 1.65^{\triangle}$	$92.52 \pm 4.62^{\circ}$
N2N	$37.52 \pm 2.41^{\circ}$	$94.74 \pm 1.74^{\diamond}$	$92.05 \pm 4.64^{\diamond}$
EDCNN	$37.80 \pm 2.49^{\triangle}$	$94.10 \pm 1.78^{\triangle}$	$92.38 \pm 4.78^{\diamond}$
WGAN	$37.37 \pm 2.19^{\circ}$	$94.22 \pm 1.97^{\diamond}$	$89.03 \pm 6.19^{\circ}$
DU-GAN	$37.57 \pm 2.46^{\diamond}$	$94.24 \pm 2.88^{\diamond}$	$90.22 \pm 5.87^{\circ}$
IDDPM	$38.16 \pm 2.60^{\triangle}$	$94.88 \pm 1.73^{\triangle}$	$92.17 \pm 4.91^{\diamond}$
UFormer	$38.77 \pm 2.62^{\diamond}$	$95.82 \pm 1.62^{\circ}$	$92.31 \pm 4.70^{\circ}$
LIT-Former	$37.33 \pm 1.97^{\triangle}$	$92.47 \pm 1.50^{\diamond}$	$91.81 \pm 4.81^{\circ}$
ViMEDNet	$38.88 \pm 2.44^{\triangle}$	$95.90 \pm 1.72^{\triangle}$	$92.40 \pm 4.42^{\diamond}$
DenoMamba	$\mathbf{39.72 \pm 2.43}$	$\mathbf{96.24 \pm 1.73}$	$\mathbf{92.93 \pm 4.52}$

RED-CNN: A convolutional model was considered that uses a hierarchical encoder-decoder architecture equipped with shortcut connections [18].

N2N: A convolutional model was considered that was originally proposed for self-supervised learning on noisy CT images [64]. For fair comparison, the architecture of N2N was adopted to perform supervised learning.

EDCNN: A convolutional model was considered that employs a trainable Sobel convolution kernel for edge detection and dense connections [19].

WGAN: An adversarial model that uses convolutional generator and discriminator subnetworks was considered [27]. Loss term weights were set as λ = 10, λ_1 = 0.1, λ_2 = 0.1.

DU-GAN: An adversarial model that uses convolutional generator and discriminator subnetworks was considered [65]. Loss terms weights were set as λ_{adv} = 0.1, λ_{img} = 1, and λ_{grd} = 20.

IDDPM: A diffusion model with a convolutional backbone augmented with attention mechanism was considered that generated NDCT images starting from Gaussian noise images, with additional guidance from the LCDT image provided as input [29]. The number of diffusion steps was taken as 1000.

UFormer: An efficient transformer model was considered that uses a hierarchical encoder-decoder architecture and local window-based self-attention [66].

TABLE III: Spatial acuity of denoised LDCT images was evaluated using the GMSD metric on the AAPM dataset. Lower scores indicate better preservation of fine structural details. Superscripts indicate significance levels relative to DenoMamba ($^{\circ}$p<0.05, $^{\diamond}$p<0.005, $^{\triangle}$p<0.0005). Boldface marks the method that offers the best performance.

	↓ **GMSD**
RED-CNN	$0.127 \pm 0.025^{\triangle}$
N2N	$0.130 \pm 0.028^{\triangle}$
EDCNN	$0.126 \pm 0.025^{\triangle}$
WGAN	$0.139 \pm 0.025^{\diamond}$
DU-GAN	$0.119 \pm 0.016^{\diamond}$
IDDPM	$0.124 \pm 0.027^{\triangle}$
UFormer	$0.124 \pm 0.024^{\triangle}$
LIT-Former	$0.119 \pm 0.017^{\diamond}$
ViMEDNet	$0.123 \pm 0.023^{\triangle}$
DenoMamba	$\mathbf{0.114 \pm 0.023}$

LIT-Former: An efficient transformer model was considered that was originally proposed for processing 3D images with separate transformer modules for in-plane and through-plane dimensions [1]. LIT-former was adopted for 2D images by removing the through-plane modules.

ViMEDNet: A state-space model was considered that uses a hybrid CNN-SSM architecture equipped with spatial SSM modules [44].

D. Modeling Procedures

Models were implemented using the PyTorch framework and trained on an NVidia RTX 3090 GPU. Training was performed via the Adam optimizer with parameters $\beta_1 = 0.5$ and $\beta_2 = 0.999$ [67]. For all competing methods, the learning rate was set to 1×10^{-4}, and the number of epochs was set to 120. The initial learning rate was halved after every 30 epochs to promote gradual model refinement. Data were split into training, validation and test sets with no subject-level overlap between the three sets (see Supp. Fig. 1 for training/validation curves of DenoMamba). Key model hyperparameters were selected via cross-validation for each competing method. Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) metrics were measured to assess overall similarity to reference images, while Bhattacharyya Coefficient (BC) metric was measured between normalized intensity histograms of denoised and reference images to assess distributional consistency [68]. Gradient-Magnitude Similarity Deviation (GMSD) metric was measured to assess spatial acuity. Note that higher values of PSNR and SSIM, albeit lower values of GMSD are preferable. Significance of differences between competing methods were evaluated via non-parametric Wilcoxon signed-rank tests (p<0.05).

IV. Results

A. Comparison Studies

We demonstrated DenoMamba on abdominal CT scans from the 2016 AAPM Low Dose CT Grand Challenge via comparisons against several state-of-the-art methods from the LDCT denoising literature. Specifically, convolutional models (RED-CNN, N2N, EDCNN), generative models based on adversarial or diffusion learning (WGAN, DU-GAN, IDDPM), and contextually-sensitive models with efficient transformer or SSM backbones (UFormer, LIT-Former, ViMEDNet) were considered. While this study primarily focuses on the utility of network architectures for LDCT denoising, generative models were included in comparisons for a more comprehensive assessment (see Sec. III-C for further details on competing methods). Experiments were first conducted on the 25%-dose dataset to recover NDCT images from LDCT measurements. Table I lists performance metrics for competing methods on the test set. We find that DenoMamba significantly outperforms each competing method (p<0.05). On average, DenoMamba achieves performance improvements of 1.8dB PSNR, 0.8% SSIM, 0.9% BC over convolutional baselines; 2.4dB PSNR, 1.8% SSIM, 2.7% BC over generative baselines, and 1.5dB PSNR, 0.7% SSIM, 1.1% BC over contextually-sensitive baselines.

Representative denoised images recovered by competing methods are displayed in Fig. 4. Among competing methods, convolutional baselines can alleviate local noise patterns in regions of homogeneous tissue signal, but they yield suboptimal depiction of detailed tissue structure that extend over longer distances, particularly near regions of heterogeneous tissue composition. Generative baselines typically yield a higher degree of visual sharpness in denoised images, albeit at the expense of elevated noise in recovered images that is particularly evident for adversarial models. Although contextually-sensitive baselines including ViMEDNet offer improved preservation of tissue structure across heterogeneous regions, they suffer from residual local noise patterns that can manifest as signal intensity fluctuations in homogeneous regions. In comparison, DenoMamba recovers high-quality CT images with more effective suppression of noise patterns, and accurate depiction of tissue structure and contrast.

We also conducted experiments on the 10%-dose dataset to assess competing methods in a relatively more challenging denoising task. Table II lists performance metrics for competing methods on the test set. Corroborating the findings on the 25%-dose dataset, we find that DenoMamba significantly outperforms all competing methods (p<0.05). On average, DenoMamba achieves performance improvements of 1.9dB PSNR, 1.6% SSIM, 0.6% BC over convolutional baselines; 2.0dB PSNR, 1.8% SSIM, 2.5% BC over generative baselines, and 1.4dB PSNR, 1.5% SSIM, 0.8% BC over contextually-sensitive baselines.

Representative denoised images recovered by competing methods are displayed in Fig. 5. Note that prominent noise is apparent in LDCT images given the more aggressive dose reduction in 10%-dose scans. Naturally, this elevates the difficulty of the LDCT denoising task as it becomes challenging to distinguish noise patterns from native variations in tissue signals. We observe that convolutional baselines can still offer reasonable suppression of local noise patterns in homogeneous regions, albeit this suppression comes at the expense of structural artifacts evident in regions of heterogeneous tissue composition. Meanwhile, generative baselines suffer from varying levels of noise amplification that can compromise structural accuracy particularly near tissue boundaries. Although contextually-sensitive baselines including ViMEDNet tend to improve depiction of tissue contrast over heterogeneous regions, they suffer from a degree of spatial blurring that can cause suboptimal depiction of fine tissue structures. Contrarily, DenoMamba offers high-fidelity depiction of detailed tissue structure in CT images and visibly improved suppression of noise. These results suggest that DenoMamba attains a more favorable balance between contextual sensitivity and local precision than competing methods, including ViMEDNet as a conventional SSM baseline.

As denoising methods often leverage frequency-related characteristics to separate tissue signals from noise, there is an inherent risk that they may attenuate fine structural details. To evaluate spatial acuity in denoised LDCT images, we examined GMSD that reflects edge structure consistency across the frequency spectrum [69]. As shown in Table III, convolutional and generative baselines generally exhibit higher GMSD values indicative of lower spatial acuity. In contrast, contextually-sensitive baselines demonstrate relatively improved metrics, suggesting enhanced preservation of fine structures. Notably, DenoMamba achieves the lowest GMSD, significantly outperforming all competing methods (p<0.05). These results indicate that DenoMamba offers enhanced denoising performance, as reflected in the PSNR/SSIM values reported earlier, while also maintaining a level of spatial acuity that is superior to competing methods based on quantitative metrics.

B. Generalization Across Dose Levels

We also assessed denoising performance under shifts in the level of dose reduction. To this end, models trained on 25%-dose scans were tested on 10%-dose scans, and models trained on 10%-dose scans were tested on 25%-dose scans in the AAPM dataset. Testing was performed in zero-shot settings without further model training or adaptation. Table IV lists performance metrics for competing methods. For learning-based models, notable differences in image noise encountered between training and test sets can naturally induce performance losses. Yet, we find that DenoMamba significantly outperforms all competing methods in denoising performance (p<0.05), consistently in both shift direc-

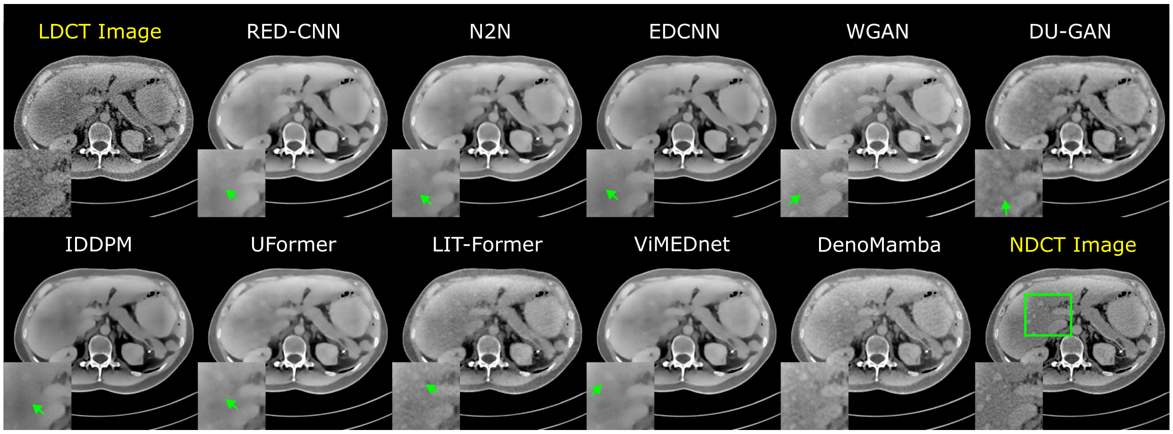

Fig. 6: Generalization across dose levels for the AAPM dataset. Models trained on 25%-dose scans were evaluated on 10%-dose scans. A display window of [-250 450] HU is used. Please see Supp. Fig. 3 for a visualization of error maps corresponding to the denoised images from competing methods.

TABLE IV: Generalization performance across dose levels on the AAPM dataset. Models trained at 25%-dose were tested on 10%-dose scans (left panel), and models trained at 10%-dose were tested on 25%-dose scans (right panel). Superscripts indicate significance levels relative to DenoMamba ($^\circ p<0.05$, $^\diamond p<0.005$, $^\triangle p<0.0005$). Boldface marks the method that offers the best performance for each metric.

	25%-dose → 10%-dose			10%-dose → 25%-dose		
	↑ **PSNR (dB)**	↑ **SSIM (%)**	↑ **BC (%)**	↑ **PSNR (dB)**	↑ **SSIM (%)**	↑ **BC (%)**
RED-CNN	$37.09 \pm 2.40^\triangle$	$93.01 \pm 1.41^\triangle$	$92.03 \pm 4.90^\circ$	$38.03 \pm 2.41^\triangle$	$95.50 \pm 1.81^\diamond$	$92.17 \pm 4.63^\diamond$
N2N	$37.15 \pm 2.52^\triangle$	$93.05 \pm 1.72^\diamond$	$91.98 \pm 4.37^\diamond$	$39.47 \pm 2.46^\triangle$	$96.26 \pm 1.18^\triangle$	$91.96 \pm 4.47^\diamond$
EDCNN	$36.50 \pm 2.10^\triangle$	$92.56 \pm 1.50^\diamond$	$91.84 \pm 4.28^\diamond$	$37.79 \pm 2.42^\triangle$	$94.19 \pm 2.66^\diamond$	$92.16 \pm 4.90^\diamond$
WGAN	$35.91 \pm 2.39^\diamond$	$91.84 \pm 2.28^\diamond$	$89.62 \pm 5.11^\circ$	$37.09 \pm 1.86^\diamond$	$94.32 \pm 2.18^\diamond$	$89.74 \pm 5.09^\circ$
DU-GAN	$35.35 \pm 2.14^\diamond$	$90.65 \pm 2.33^\diamond$	$90.69 \pm 5.06^\circ$	$37.21 \pm 2.33^\diamond$	$94.29 \pm 3.68^\diamond$	$90.88 \pm 4.82^\circ$
IDDPM	$37.75 \pm 2.60^\triangle$	$94.51 \pm 1.73^\diamond$	$91.93 \pm 4.47^\diamond$	$39.34 \pm 2.56^\triangle$	$96.12 \pm 1.55^\diamond$	$92.05 \pm 4.64^\circ$
UFormer	$37.57 \pm 2.62^\triangle$	$94.29 \pm 1.50^\triangle$	$91.26 \pm 4.74^\diamond$	$39.38 \pm 2.42^\triangle$	$96.18 \pm 1.74^\triangle$	$91.81 \pm 4.71^\diamond$
LIT-Former	$36.50 \pm 2.25^\triangle$	$92.71 \pm 1.75^\diamond$	$91.23 \pm 5.03^\diamond$	$38.57 \pm 2.36^\triangle$	$96.15 \pm 1.57^\triangle$	$91.58 \pm 4.72^\diamond$
ViMEDNet	$37.50 \pm 2.35^\triangle$	$93.52 \pm 1.52^\triangle$	$91.98 \pm 4.34^\diamond$	$39.09 \pm 2.43^\triangle$	$96.18 \pm 1.70^\diamond$	$92.19 \pm 4.42^\diamond$
DenoMamba	**38.04 ± 2.20**	**94.88 ± 1.57**	**92.10 ± 4.33**	**39.72 ± 2.42**	**96.33 ± 1.70**	**92.45 ± 4.30**

tions (25%→10%, 10%→25%). On average across directions, DenoMamba achieves performance improvements of 1.2dB PSNR, 1.5% SSIM, 0.3% BC over convolutional baselines; 1.8dB PSNR, 2.0% SSIM, 1.5% BC over generative baselines, and 0.8dB PSNR, 0.8% SSIM, 0.6% BC over contextually-sensitive baselines. We also find that DenoMamba generally offers relatively higher levels of performance benefits over baselines in the shift direction of 25%-dose→10%-dose versus 10%-dose→25%-dose. This result implies that DenoMamba shows improved reliability against elevated task difficulty in the test set compared to baselines. Representative images recovered by competing methods are depicted in Fig. 6. High degrees of spatial blurring are apparent in convolutional baselines, DU-GAN, IDDPM, Uformer and ViMEDNet, which can be attributed to an overestimation of the noise level in LDCT images by the respective domain-transferred models. This spatial blurring yields suboptimal depiction of prominent vessel structures in abdominal images. Meanwhile, remaining methods including WGAN, LIT-Former and DenoMamba that are less amenable to spatial blurring show higher levels of residual noise. Among these methods, DenoMamba offers improved accuracy in depiction of important vascular structures evident in reference NDCT images, despite elevated levels of residual noise. Taken together, these results demonstrate that DenoMamba shows a degree of robustness against shifts in noise levels of CT scans to maintain its superior performance over baselines.

C. Validation on External Dataset

Next, we conducted experiments to validate competing methods on an independent dataset to assess reliability against shifts in the underlying data distribution for CT scans. For this purpose, models separately trained on the 25%-dose and 10%-dose AAPM datasets were tested on the independent 10%-dose Piglet CT dataset. Testing was performed in zero-shot settings without further model training or adaptation. Table V lists performance metrics for competing methods. For both dose levels on which the models were trained, we find that DenoMamba significantly outperforms all competing methods in generalization across datasets ($p<0.05$). On average, DenoMamba achieves performance improvements of 1.3dB PSNR, 1.8% SSIM, 0.3% BC over convolutional baselines; 2.5dB PSNR, 4.5% SSIM, 1.6% BC over generative baselines, and 1.0dB PSNR, 1.5% SSIM, 0.6% BC over contextually-sensitive baselines. We also find that DenoMamba offers comparable levels of performance benefits over baselines in both examined settings, i.e., training on the 25%-dose and training on the 10%-dose AAPM scans. Yet, the absolute denoising performance of several competing methods including DenoMamba are moderately higher when trained on the 25%-dose

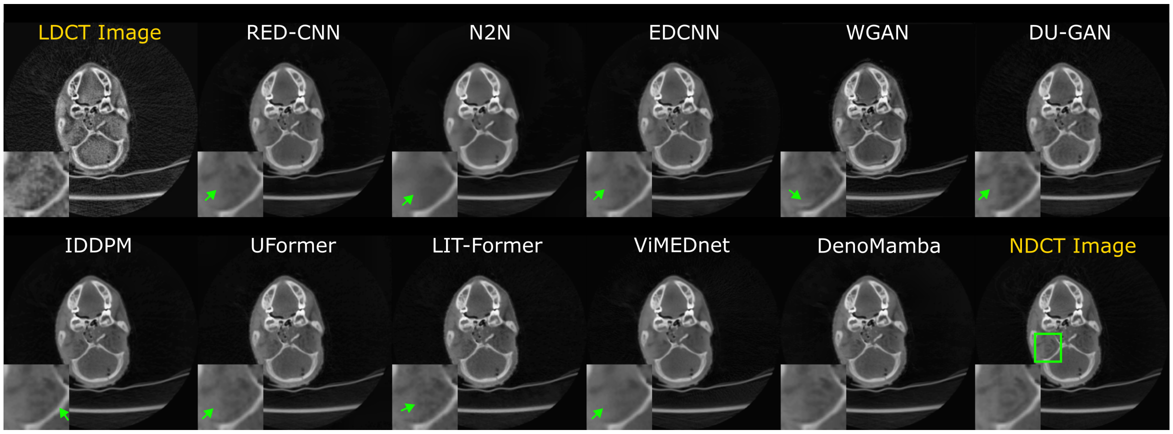

Fig. 7: External validation on the Piglet CT dataset. Models trained on the 25%-dose AAPM dataset were evaluated on the 10%-dose Piglet CT dataset. A display window of [-400 1000] HU is used. Please see Supp. Fig. 4 for a visualization of error maps corresponding to the denoised images.

TABLE V: External validation of competing methods on the 10%-dose Piglet CT dataset. Models trained on either the 25%-dose (left panel) or 10%-dose (right panel) AAPM scans were evaluated on Piglet CT scans. Superscripts indicate significance levels relative to DenoMamba ($^{\circ}p<0.05$, $^{\diamond}p<0.005$, $^{\triangle}p<0.0005$). Boldface marks the method that offers the best performance for each metric.

	25% AAPM → Piglet			10% AAPM → Piglet		
	↑ PSNR (dB)	**↑ SSIM (%)**	**↑ BC (%)**	**↑ PSNR (dB)**	**↑ SSIM (%)**	**↑ BC (%)**
RED-CNN	$38.37 \pm 3.96^{\triangle}$	$96.65 \pm 3.26^{\triangle}$	$92.15 \pm 4.60^{\triangle}$	$38.66 \pm 3.91^{\triangle}$	$96.42 \pm 3.44^{\triangle}$	$92.22 \pm 4.50^{\circ}$
N2N	$37.69 \pm 3.40^{\diamond}$	$95.86 \pm 3.01^{\triangle}$	$92.04 \pm 4.34^{\diamond}$	$38.19 \pm 3.46^{\diamond}$	$96.03 \pm 3.13^{\triangle}$	$92.19 \pm 4.48^{\diamond}$
EDCNN	$38.93 \pm 4.07^{\triangle}$	$96.91 \pm 3.07^{\triangle}$	$92.18 \pm 4.83^{\triangle}$	$38.78 \pm 3.84^{\triangle}$	$96.88 \pm 3.26^{\diamond}$	$92.25 \pm 4.88^{\circ}$
WGAN	$34.36 \pm 3.07^{\circ}$	$88.25 \pm 3.69^{\circ}$	$89.96 \pm 5.71^{\circ}$	$35.11 \pm 2.91^{\circ}$	$88.40 \pm 3.38^{\diamond}$	$89.81 \pm 5.17^{\diamond}$
DU-GAN	$38.42 \pm 3.49^{\diamond}$	$96.03 \pm 5.22^{\triangle}$	$90.76 \pm 5.22^{\circ}$	$38.34 \pm 3.43^{\triangle}$	$96.29 \pm 4.87^{\diamond}$	$91.19 \pm 4.98^{\diamond}$
IDDPM	$38.58 \pm 2.95^{\triangle}$	$96.52 \pm 2.68^{\triangle}$	$91.74 \pm 4.65^{\triangle}$	$38.73 \pm 2.73^{\diamond}$	$97.38 \pm 2.59^{\diamond}$	$91.93 \pm 4.71^{\triangle}$
UFormer	$38.44 \pm 3.72^{\diamond}$	$96.70 \pm 2.71^{\triangle}$	$91.50 \pm 4.70^{\circ}$	$38.45 \pm 3.45^{\diamond}$	$95.99 \pm 2.74^{\triangle}$	$91.58 \pm 4.72^{\circ}$
LIT-Former	$38.69 \pm 3.11^{\diamond}$	$96.63 \pm 3.10^{\triangle}$	$92.09 \pm 4.33^{\diamond}$	$38.53 \pm 3.24^{\diamond}$	$96.50 \pm 2.90^{\diamond}$	$91.76 \pm 4.50^{\circ}$
ViMEDNet	$39.05 \pm 3.86^{\triangle}$	$97.31 \pm 2.88^{\triangle}$	$92.33 \pm 4.78^{\triangle}$	$39.08 \pm 3.83^{\triangle}$	$97.51 \pm 2.76^{\triangle}$	$92.32 \pm 4.78^{\triangle}$
DenoMamba	**39.88 ± 3.73**	**98.40 ± 2.92**	**92.53 ± 4.68**	**39.51 ± 3.53**	**98.19 ± 2.81**	**92.43 ± 4.76**

scans, even though the evaluations are conducted on the 10%-dose Piglet CT scans. Through visual inspection, we confirmed that the 10%-dose Piglet CT scans have more similar levels of noise perturbation to the 25%-dose as opposed to 10%-dose AAPM scans. Therefore, our findings are best attributed to the closer alignment of noise levels between training and test datasets, achieved when models are trained on the 25%-dose AAPM scans. Representative images recovered by competing methods are depicted in Fig. 7. We observe that baseline models either suffer from over-smoothing manifested as spatial blurring (e.g., convolutional baselines, ViMEDNet) or from residual noise patterns manifested as structural artifacts (e.g., generative baselines, transformers) that can both compromise visibility of moderate variations in tissue contrast in denoised CT images. In comparison, DenoMamba recovers high-fidelity images with a closer appearance to reference NDCT images in terms of tissue structure and contrast. Collectively, these results indicate that DenoMamba shows a notable degree of robustness against shifts in the data distribution driven by native variations in anatomy and/or scanner hardware.

D. Computational Efficiency

Computational efficiency is an important consideration for practical implementation of deep learning models. Table VI compares run time, memory usage, and parameter count for competing methods during training and testing phases. As expected, convolutional baselines demonstrate the highest efficiency with minimal computational demands, whereas generative approaches using heavier architectures exhibit substantially lower efficiency. Meanwhile, contextually-sensitive methods, including DenoMamba, occupy a middle ground between these two ends. Considering the range of computational requirements across competing methods, Deno-

TABLE VI: Computational complexity of competing methods. Training and test run time (milliseconds), memory usage (MB), and parameter count (millions) are reported for a single cross-section. Lower values of training time, memory, and parameter count indicate higher computational efficiency.

	Time (ms)		**Memory (MB)**		**Params (M)**
	Training	Test	Training	Test	
RED-CNN	2.4	0.6	47.4	27.6	0.05
N2N	4.9	1.3	345.5	199.4	1.20
EDCNN	3.3	0.9	443.1	139.5	0.08
WGAN	28.3	1.8	1257.4	231.6	3.51
DU-GAN	202.2	6.1	4109.4	2895.5	57.2
IDDPM	151.4	13.4	5764.0	2157.7	88.96
UFormer	38.0	10.4	1405.6	280.4	5.05
LIT-Former	6.42	1.8	317.4	252.6	1.21
ViMEDNet	25.0	7.0	1325.1	347.6	8.91
DenoMamba	92.4	16.2	2434.3	327.8	4.98

TABLE VII: Performance of DenoMamba variants built by replacing SSM modules with vanilla transformers and image downsampling to 128×128 (w ViT+down), with vanilla transformers and split processing of 128×128 image patches (w ViT+patch), and with efficient transformers of linear complexity (w eff. ViT). Inference time and validation PSNR, SSIM are listed for the 25%-dose AAPM dataset. Significance levels of variants relative to DenoMamba are $p<0.05$ for all metrics.

	Time (ms)	**↑ PSNR (dB)**	**↑ SSIM (%)**
w ViT+down	30.9	38.79	92.26
w ViT+patch	37.7	40.45	95.04
w eff. ViT	18.1	41.50	96.11
DenoMamba	16.2	42.70	97.15

Mamba achieves computational metrics generally comparable to other contextually-sensitive approaches, albeit with moderately higher run times and training memory load. These results demonstrate that DenoMamba delivers superior denoising performance while maintaining computational efficiency generally on par with contextually-sensitive alternatives, making it well-suited for practical implementation.

E. Ablation Studies

We conducted a systematic set of ablation studies to examine the importance of key building elements and design parameters in DenoMamba. First, we assessed the efficacy of SSM modules in DenoMamba for capturing contextual representations in comparison to transformer modules. Note that vanilla transformers (ViT) suffer from quadratic complexity with respect to sequence length [70], which prohibits their use for pixel-level processing at the original image resolution given memory limitations on GPUs employed here. Thus, several efficient transformer adaptations were formed based on different strategies to mitigate complexity. A 'w ViT+down' variant was formed by replacing the SSM modules with ViT modules, and spatially downsampling images to a 128×128 size [71]. A 'w ViT+patch' variant was formed by replacing the SSM modules with ViT modules, splitting each image into a set of four 128×128 patches, and processing separate patches individually [72]. A 'w eff. ViT' variant was formed by adopting an efficient module of linear complexity based on transposed attention [73]. Table VII lists performance metrics for DenoMamba and transformer-based variants on the 25%-dose dataset, along with inference times per slice. DenoMamba outperforms all variant models in performance metrics ($p<0.05$), while also offering shorter inference times. These results suggest that transformer adaptations that restrict image resolution or field-of-view, or that employ approximate attention operators to maintain efficiency suffer from notable losses in denoising performance.

TABLE VIII: Performance of DenoMamba variants built by ablating the channel SSM module (w/o cha. SSM), the spatial SSM module (w/o spa. SSM), the CFM module (w/o CFM), the gated convolution network (w/o GCN), and the identity path (w/o Iden.). PSNR, SSIM are listed for the 25%-dose AAPM dataset. Significance levels of variants relative to DenoMamba are $p<0.05$ for all metrics.

	↑ PSNR (dB)	**↑ SSIM (%)**
w/o spa. SSM	42.24	96.84
w/o cha. SSM	42.09	96.99
w/o CFM	42.39	97.01
w/o GCN	42.44	97.08
w/o Iden.	41.64	96.42
DenoMamba	42.70	97.15

TABLE IX: Performance of 3D DenoMamba variants built by ablating the channel SSM module (w/o cha. SSM), the spatial SSM module (w/o spa. SSM), the CFM module (w/o CFM), the gated convolution network (w/o GCN), and the identity path (w/o Iden.). PSNR, SSIM are listed for the BIMCV-R dataset. Significance levels of variants relative to DenoMamba are $p<0.05$ for all metrics.

	↑ PSNR (dB)	**↑ SSIM (%)**
w/o spa. SSM	41.02	97.15
w/o cha. SSM	41.72	97.33
w/o CFM	41.28	97.15
w/o GCN	41.70	97.35
w/o Iden.	41.31	96.89
DenoMamba	42.01	97.49

We then assessed the influence of spatial SSM, channel SSM, and CFM modules in DenoMamba on denoising performance. A 'w/o spa. SSM' variant ablated the spatial SSM modules, a 'w/o cha. SSM' variant ablated channel SSM modules, a 'w/o CFM' variant ablated CFM modules, a 'w/o GCN' variant ablated GCN layers that perform content-aware gating to module channel-wise features, and a 'w/o Iden.' variant ablated identity propagation paths. Table VIII lists performance metrics for DenoMamba variants on the 25%-dose dataset. We find that DenoMamba outperforms all variants ($p<0.05$). Higher performance of DenoMamba over the 'w/o spa. SSM', 'w/o cha. SSM', and 'w/o Iden.' variants indicate that contextual features in spatial and channel dimensions along with lower-level spatial features effectively contribute to LDCT denoising performance. Higher performance of DenoMamba over the 'w/o GCN' variant suggests that the content-aware gating mechanism in GCN layers helps refine channel-wise features based on spatial context. Note that low-level input features can be propagated across FuseSSM blocks in multiple ways, including the identity propagation path feeding into the CFM module where input and contextual features are subjected to nonlinear convolutional fusion, as well as the residual connections in channel and spatial SSM modules that additively fuse the input and contextual features. Taken together, higher performance of DenoMamba against the 'w/o Iden.' and 'w/o CFM' variants indicate that nonlinear convolutional fusion effectively preserves essential low-level representations in CT images.

We then evaluated DenoMamba's applicability to volumetric denoising by adapting its core components—including SSM, CFM modules and GCN layers—to process three-dimensional (3D) CT data. To confirm that DenoMamba's architectural design remains effective in volumetric settings, we conducted ablation analyses on 3D CT scans from the BIMCV-R dataset. As listed in Table IX, the performance metrics for various 3D DenoMamba variants reveal patterns consistent with our 2D findings, with the complete DenoMamba architecture significantly outperforming all variants ($p<0.05$). These results demonstrate that DenoMamba's architectural components generalize effectively to 3D low-dose CT denoising settings.

TABLE X: Performance of DenoMamba while varying the number of encoder-decoder stages K, the number of feature channels C, and the configuration for the number of FuseSSM blocks across stages $E-D$. PSNR, SSIM are listed for the 25%-dose AAPM dataset.

		↑ PSNR (dB)	↑ SSIM (%)
K	3	42.64	97.02
	4	42.70	97.15
	5	42.61	97.11
C	16	42.70	97.15
	32	42.67	97.13
	48	42.53	96.98
$E-D$	$\underline{1}$	42.70	97.15
	$\underline{2}$	42.70	97.13
	$\underline{3}$	42.68	97.15

Finally, we assessed the influence of the number of encoder-decoder stages K, the number of initial feature channels at the first encoder stage C (note that the number of feature channels in remaining stages scale proportionately with C), and the numbers of FuseSSM blocks cascaded across individual encoder-decoder stages $E-D$ (i.e., the number of FuseSSM blocks across K encoder and K decoder stages). In general, prescribing higher values for these design parameters increases model complexity. As learning-based models are subject to an intrinsic trade-off between allowed degrees of freedom versus learning efficacy, we wanted to examine whether the selected design parameters for DenoMamba offer a favorable compromise. For this purpose, models were built by separately varying the values of K, C, and R while remaining parameters were kept fixed. Specifically, we varied K in {3, 4, 5}; C in {16, 32, 48}; and $E-D$ in {$\underline{1}$: [1, 2, 2, 3] - [2, 2, 1, 2], $\underline{2}$: [2, 3, 3, 4] - [3, 3, 2, 2], $\underline{3}$: [4, 6, 6, 8] - [6, 6, 4, 2]}. Table X lists performance metrics of DenoMamba on the 25%-dose dataset. We find that $K=4$, $C=16$, and $E-D=\underline{1}$ yield near-optimal performance, validating the proposed selection of design parameters.

V. Discussion

A. Scope and Practical Implications

Focusing on the image denoising task for CT scans in the current study, we aimed to recover high-quality images from noisy LDCT acquisitions, where existing solutions suffer from several limitations. Previous CNN models offer a high degree of local precision, albeit they are relatively insensitive to long-range relationships between distant anatomical regions in medical images [25]. While transformer models benefit from the long-range contextual sensitivity of self-attention operators, they inherently suffer from quadratic model complexity with respect to sequence length [32]. Meanwhile, common approaches to mitigate this complexity result in inevitable losses in spatial precision [74]. To address these limitations, here we introduced DenoMamba that leverages the emergent framework of state-space modeling to enhance performance and efficiency in LDCT image denoising.

DenoMamba employs a novel SSM-based architecture to improve capture of complex contextual information, and thereby enable more nuanced separation of signal from noise components in CT images. Our demonstrations indicate that it achieves superior performance in LDCT denoising against state-of-the-art CNN, transformer and SSM methods, with apparent quantitative and qualitative benefits in image quality. By effectively denoising images at significantly reduced radiation doses, DenoMamba offers a compelling solution to lower patient radiation exposure without compromising diagnostic utility. This can facilitate more broadspread adoption of CT in clinical scenarios requiring screening of at risk populations (e.g., pediatric or cancer patients) or repeated scans on the same patient (e.g., continual disease monitoring, chronic disease management). Furthermore, the intrinsic efficiency benefits of DenoMamba can be critical for real-world clinical deployment, where processing speed and computational resources are often constrained.

B. Limitations and Future Work

Several technical limitations can be addressed in order to further boost the performance and practicality of DenoMamba. A first line of improvements concerns the nature of denoising tasks targeted during model training. Here, a separate model was trained for LDCT denoising at each reduction level for radiation dose to maintain high performance. Note that this may lower practicality if highly variable reduction levels are expected to be administered in practice. In those cases, DenoMamba can be trained on LDCT images at varying reduction levels, and model specialization to specific radiation doses could be enhanced by adaptive normalization approaches on feature maps [75], [76]. This could improve practicality by building a unified model that can be deployed at various dose reduction levels.

A second line of improvements concerns the datasets on which DenoMamba is trained to perform LDCT denoising. Here, we performed supervised learning relying on the availability of paired LDCT-NDCT images from the same set of subjects [33]. Note that, in practice, the curation of such paired datasets can be challenging as it would require repeated CT scans on a given subject at separate radiation doses. In cases where the amount of paired training data that can be collected is limited, a large training set can be curated by instead adopting cycle-consistent learning procedures on unpaired sets of LDCT and NDCT images [77], or self-supervised learning procedures to train models directly on LDCT measurements [38], [64]. In this study, we primarily conducted experiments on 2D cross-sectional CT images, consistent with many LDCT denoising studies in the literature. While processing slices independently may introduce intensity discontinuities across the anatomical volume, our visual assessments did not reveal notable inconsistencies across consecutive 2D cross-sections. Furthermore, the 2D and 3D variants of DenoMamba demonstrate generally comparable performance metrics, as shown in Tables VIII and IX, suggesting that our 2D results are not significantly affected by such potential artifacts. Nevertheless, future work is warranted to more thoroughly evaluate the relative advantages of 3D versus 2D processing for LDCT denoising tasks.

A third line of improvements concerns the loss terms employed to train DenoMamba. Tissue signal components in CT images characteristically carry diminishing energy towards higher spatial frequencies, whereas noise components often have a relatively balanced distribution across frequencies. This promotes denoising methods to regularize frequency-related attributes for signal-noise separation, resulting in an intrinsic trade-off between the level of noise suppression and preservation of high-frequency details. Here, we performed model training via a simple pixel-wise loss term based on mean absolute error between denoised and reference images. While a degree of blurring was apparent across competing methods, we observed that DenoMamba generally offers similar or superior spatial acuity to baselines in the examined datasets. That said, it might be possible to attain further improvements in recovered image quality by using more sophisticated loss terms including adversarial, score-based or cross-entropy losses [29], [78], which may offer enhanced sensitivity to structural details. Particularly within the context of score-based methods that involve iterative sampling procedures, the long-range contextual sensitivity of DenoMamba combined with task-driven bridge formulations might offer benefits over conventional denoising diffusion models based on CNN backbones [30], [79], [80]. Further work is warranted for a systematic evaluation of the utility of various loss functions on the performance and reliability of DenoMamba.

VI. Conclusion

Here we introduced a novel fused SSM for recovery of high-quality images from noisy LDCT scans. The proposed DenoMamba model leverages an hourglass architecture implemented with novel FuseSSM blocks. Each FuseSSM block extracts contextual features across spatial and channel dimensions via spatial and channel SSM modules, respectively, and performs fusion of contextual and low-level input features via a CFM module. This design enables DenoMamba to leverage contextual relationships in LDCT images without compromising local precision, and thereby to offer superior performance against state-of-the-art LDCT denoising methods. Therefore, DenoMamba holds great promise for performant LDCT image denoising.

References

[1] Z. Chen *et al.*, "Lit-former: Linking in-plane and through-plane transformers for simultaneous ct image denoising and deblurring," *IEEE Trans Med Imaging*, vol. 43, no. 5, pp. 1880–1894, 2024.

[2] S. Li *et al.*, "Dd-dcsr: Image denoising for low-dose ct via dual-dictionary deep convolutional sparse representation," *IEEE Trans Comput Imaging*, vol. 10, pp. 899–914, 2024.

[3] Y. Lei *et al.*, "Shape and margin-aware lung nodule classification in low-dose ct images via soft activation mapping," *Med Image Anal*, vol. 60, p. 101628, 2020.

[4] S.-Y. Jeon *et al.*, "Mm-net: Multiframe and multimask-based unsupervised deep denoising for low-dose computed tomography," *IEEE Trans Rad Plas Med Sci*, vol. 7, no. 3, pp. 296–306, 2023.

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

[6] C. Suo *et al.*, "Cross-level collaborative context-aware framework for medical image segmentation," *Exp Syst Appl*, vol. 236, p. 121319, 2024.

[7] X. Hui *et al.*, "PACT image reconstruction: from sinograms to images using neural networks," in *SPIE BiOS*, vol. 13319. SPIE, 2025, p. 133191U.

[8] M. Meng *et al.*, "Ddt-net: Dose-agnostic dual-task transfer network for simultaneous low-dose ct denoising and simulation," *IEEE J Biomed Health Inf*, vol. 28, no. 6, pp. 3613–3625, 2024.

[9] A. Manduca *et al.*, "Projection space denoising with bilateral filtering and ct noise modeling for dose reduction in ct," *Med Phys*, vol. 36, no. 11, pp. 4911–4919, 2009.

[10] S. Gu *et al.*, "Weighted nuclear norm minimization with application to image denoising," in *IEEE CVPR*, 2014, pp. 2862–2869.

[11] N. Saidulu *et al.*, "Rhlnet: Robust hybrid loss-based network for low-dose ct image denoising," *IEEE Trans Instru Meas*, pp. 1–1, 2024.

[12] J. Huang *et al.*, "Cross-domain low-dose ct image denoising with semantic preservation and noise alignment," *IEEE Trans Multimed*, pp. 1–11, 2024.

[13] D. Ellison *et al.*, *Neuropathology: A Reference Text of CNS Pathology*. Elsevier, 2012.

[14] D. Wang *et al.*, "Ctformer: convolution-free token2token dilated vision transformer for low-dose ct denoising," *Phys Med Biol*, vol. 68, no. 6, p. 065012, 2023.

[15] Z. Zhang *et al.*, "TransCT: Dual-path transformer for low dose computed tomography," in *MICCAI*, 2021, pp. 55–64.

[16] E. Kang *et al.*, "A deep convolutional neural network using directional wavelets for low-dose x-ray ct reconstruction," *Med Phys*, vol. 44, no. 10, pp. e360–e375, 2017.

[17] F. Fan *et al.*, "Quadratic autoencoder (q-ae) for low-dose ct denoising," *IEEE Trans Med Imaging*, vol. 39, no. 6, pp. 2035–2050, 2020.

[18] H. Chen *et al.*, "Low-dose ct with a residual encoder-decoder convolutional neural network," *IEEE Trans Med Imaging*, vol. 36, no. 12, pp. 2524–2535, 2017.

[19] T. Liang *et al.*, "Edcnn: Edge enhancement-based densely connected network with compound loss for low-dose ct denoising," in *IEEE ICSP*, vol. 1, 2020, pp. 193–198.

[20] X. Jiang *et al.*, "Learning a frequency separation network with hybrid convolution and adaptive aggregation for low-dose ct denoising," in *IEEE ICBB*, 2021, pp. 919–925.

[21] Z. Li *et al.*, "Multi-scale feature fusion network for low-dose ct denoising," *J Digit Imaging*, vol. 36, no. 4, pp. 1808–1825, 2023.

[22] J. Wang *et al.*, "Penalized weighted least-squares approach to sinogram noise reduction and image reconstruction for low-dose x-ray computed tomography," *IEEE Trans Med Imaging*, vol. 25, no. 10, pp. 1272–1283, 2006.

[23] X. Yin *et al.*, "Domain progressive 3d residual convolution network to improve low-dose ct imaging," *IEEE Trans Med Imaging*, vol. 38, no. 12, pp. 2903–2913, 2019.

[24] Z. Li *et al.*, "Adaptive weighted total variation expansion and gaussian curvature guided low-dose ct image denoising network," *Biomed Signal Process Cont*, vol. 94, p. 106329, 2024.

[25] Y. Korkmaz *et al.*, "Unsupervised MRI reconstruction via zero-shot learned adversarial transformers," *IEEE Trans Med Imaging*, vol. 41, no. 7, pp. 1747–1763, 2022.

[26] M. Li *et al.*, "SACNN: Self-attention convolutional neural network for low-dose CT denoising with self-supervised perceptual loss network," *IEEE Trans Med Imaging*, vol. 39, no. 7, pp. 2289–2301, 2020.

[27] Q. Yang *et al.*, "Low-dose ct image denoising using a generative adversarial network with wasserstein distance and perceptual loss," *IEEE Trans Med Imaging*, vol. 37, no. 6, pp. 1348–1357, 2018.

[28] J. M. Wolterink *et al.*, "Generative adversarial networks for noise reduction in low-dose ct," *IEEE Trans Med Imaging*, vol. 36, no. 12, pp. 2536–2545, 2017.

[29] W. Xia *et al.*, "Low-dose ct using denoising diffusion probabilistic model for 20x speedup," *arXiv:2209.15136*, 2024.

[30] Q. Gao *et al.*, "Corediff: Contextual error-modulated generalized diffusion model for low-dose ct denoising and generalization," *IEEE Trans Med Imaging*, 2023.

[31] O. Dalmaz *et al.*, "ResViT: Residual vision transformers for multi-modal medical image synthesis," *IEEE Trans Med Imaging*, vol. 44, no. 10, pp. 2598–2614, 2022.

[32] N. Kodali *et al.*, "On convergence and stability of GANs," *arXiv:1705.07215*, 2017.

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

[34] J. Yuan *et al.*, "Hcformer: hybrid cnn-transformer for ldct image denoising," *J Digit Imaging*, vol. 36, no. 5, pp. 2290–2305, 2023.

[35] H. Li *et al.*, "Transformer with double enhancement for low-dose ct denoising," *IEEE J Biomed Health Inf*, vol. 27, no. 10, pp. 4660–4671, 2023.

[36] G. Jiang *et al.*, “Gdaformer: Gradient-guided dual attention transformer for low-dose ct image denoising,” *Biomed Signal Process Cont*, vol. 94, p. 106260, 2024.

[37] L. Yang *et al.*, “Low-dose ct denoising via sinogram inner-structure transformer,” *IEEE Trans Med Imaging*, vol. 42, no. 4, pp. 910–921, 2023.

[38] Y. Korkmaz *et al.*, “Self-supervised mri reconstruction with unrolled diffusion models,” in *MICCAI*, 2023, pp. 491–501.

[39] Q. Yiyu *et al.*, “Low-dose ct image reconstruction method based on cnn and transformer coupling network,” *CT Theory Appl*, vol. 31, no. 6, pp. 697–707, 2022.

[40] J. Liang *et al.*, “Swinir: Image restoration using swin transformer,” in *IEEE Conf Comput Vis*, 2021, pp. 1833–1844.

[41] M. Jian *et al.*, “Swinct: feature enhancement based low-dose ct images denoising with swin transformer,” *Multimed Syst*, vol. 30, no. 1, p. 1, 2024.

[42] M. Heidari *et al.*, “Computation-efficient era: A comprehensive survey of state space models in medical image analysis,” *arXiv:2406.03430*, 2024.

[43] L. Zhu *et al.*, “Vision mamba: Efficient visual representation learning with bidirectional state space model,” *arXiv:2401.09417*, 2024.

[44] J. Huang *et al.*, “A new visual state space model for low-dose ct denoising,” *Med Phys*, 2024.

[45] K. Peng *et al.*, “A low-dose CT image denoising method based on state space model,” *J Phys Conf Ser*, vol. 2858, no. 1, p. 012038, oct 2024.

[46] J. Ma *et al.*, “U-mamba: Enhancing long-range dependency for biomedical image segmentation,” *arXiv:2401.04722*, 2024.

[47] Z. Xing *et al.*, “Segmamba: Long-range sequential modeling mamba for 3d medical image segmentation,” *arXiv:2401.13560*, 2024.

[48] J. Liu *et al.*, “Swin-umamba: Mamba-based unet with imagenet-based pretraining,” *arXiv:2402.03302*, 2024.

[49] Y. Yue *et al.*, “Medmamba: Vision mamba for medical image classification,” *arXiv:2403.03849*, 2024.

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

[51] J. Huang *et al.*, “MambaMIR: An Arbitrary-Masked Mamba for Joint Medical Image Reconstruction and Uncertainty Estimation,” *arXiv:2402.18451*, 2024.

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

[53] H. Gong *et al.*, “nnMamba: 3D Biomedical Image Segmentation, Classification and Landmark Detection with State Space Model,” *arXiv:2402.03526*, 2024.

[54] J. Ruan *et al.*, “VM-UNet: Vision Mamba UNet for Medical Image Segmentation,” *arXiv:2402.02491*, 2024.

[55] O. Ronneberger *et al.*, “U-Net: Convolutional networks for biomedical image segmentation,” *MICCAI*, 2015.

[56] K. He *et al.*, “Deep residual learning for image recognition,” in *IEEE CVPR*, 2016, pp. 770–778.

[57] S. Ozturk *et al.*, “DenoMamba: A fused state-space model for low-dose CT denoising,” *arXiv:2409.13094*, Sep 2024.

[58] L. Li *et al.*, “CT-Mamba: A hybrid convolutional State Space Model for low-dose CT denoising,” *Comput Med Imaging Graph*, vol. 124, p. 102595, 2025.

[59] Y. Liu *et al.*, “Vmamba: Visual state space model,” *arXiv:2401.10166*, 2024.

[60] J. Lei Ba *et al.*, “Layer Normalization,” *arXiv:1607.06450*, 2016.

[61] C. McCollough, “Tu-fg-207a-04: overview of the low dose ct grand challenge,” *Med Phys*, vol. 43, no. 6, pp. 3759–3760, 2016.

[62] X. Yi *et al.*, “Sharpness-aware low-dose CT denoising using conditional generative adversarial network,” *J Digit Imag*, vol. 31, no. 5, pp. 655–669, 2018.

[63] Y. Chen *et al.*, “BIMCV-R: A Landmark Dataset for 3D CT Text-Image Retrieval,” *arXiv:2403.15992*, 2024.

[64] T. Huang *et al.*, “Neighbor2neighbor: Self-supervised denoising from single noisy images,” in *IEEE CVPR*, 2021, pp. 14 776–14 785.

[65] Z. Huang *et al.*, “Du-gan: Generative adversarial networks with dual-domain u-net-based discriminators for low-dose ct denoising,” *IEEE Trans Instru Meas*, vol. 71, pp. 1–12, 2022.

[66] Z. Wang *et al.*, “Uformer: A general u-shaped transformer for image restoration,” in *IEEE CVPR*, 2022, pp. 17 662–17 672.

[67] D. P. Kingma *et al.*, “Adam: A method for stochastic optimization,” in *ICLR*, 2015.

[68] A. Bhattacharyya, “On a measure of divergence between two statistical populations defined by their probability distributions,” *Calcutta Math Soc Bull*, vol. 35, pp. 99–109, 1943.

[69] W. Xue *et al.*, “Gradient magnitude similarity deviation: A highly efficient perceptual image quality index,” *IEEE Trans Image Process*, vol. 23, no. 2, pp. 684–695, 2014.

[70] A. Vaswani *et al.*, “Attention is all you need,” *Adv. Neural Inf. Process. Syst.*, pp. 1–11, 2017.

[71] J. Chen *et al.*, “TransUNet: Transformers make strong encoders for medical image segmentation,” *arXiv:2102.04306*, 2021.

[72] J. Li *et al.*, “Transforming medical imaging with Transformers? A comparative review of key properties, current progresses, and future perspectives,” *Med Image Anal*, vol. 85, p. 102762, 2023.

[73] S. W. Zamir *et al.*, “Restormer: Efficient transformer for high-resolution image restoration,” *arXiv:2111.09881*, 2022.

[74] K. He *et al.*, “Transformers in medical image analysis,” *Intelli Med*, vol. 3, no. 1, pp. 59–78, 2023.

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

[76] X. Zeng *et al.*, “Continual medical image denoising based on triplet neural networks collaboration,” *Comput Biol Med*, vol. 179, p. 108914, 2024.

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

[78] M. U. Mirza *et al.*, “Learning Fourier-Constrained Diffusion Bridges for MRI Reconstruction,” *arXiv:2308.01096*, 2023.

[79] W. Du *et al.*, “Structure-aware diffusion for low-dose ct imaging,” *Phys Med Biol*, vol. 69, no. 15, p. 155008, 2024.

[80] A. Güngör *et al.*, “Adaptive diffusion priors for accelerated MRI reconstruction,” *Med Image Anal*, vol. 88, p. 102872, 2023.