Detail Matters: Mamba-Inspired Joint Unfolding Network for Snapshot Spectral Compressive Imaging

Mengjie Qin[1,2], **Yuchao Feng**[1,2], **Zongliang Wu**[1], **Yulun Zhang**[3], **Xin Yuan**[1*]

[1]School of Engineering, Westlake University, Hangzhou, China.
[2]Westlake Institute for Optoelectronics, Fuyang, Hangzhou, Zhejiang 311421, China.
[3]Shanghai Jiao Tong University, Shanghai, China.
xyuan@westlake.edu.cn

Abstract

In the coded aperture snapshot spectral imaging system, Deep Unfolding Networks (DUNs) have made impressive progress in recovering 3D hyperspectral images (HSIs) from a single 2D measurement. However, the inherent nonlinear and ill-posed characteristics of HSI reconstruction still pose challenges to existing methods in terms of accuracy and stability. To address this issue, we propose a Mamba-inspired Joint Unfolding Network (MiJUN), which integrates physics-embedded DUNs with learning-based HSI imaging. Firstly, leveraging the concept of trapezoid discretization to expand the representation space of unfolding networks, we introduce an accelerated unfolding network scheme. This approach can be interpreted as a generalized accelerated half-quadratic splitting with a second-order differential equation, which reduces the reliance on initial optimization stages and addresses challenges related to long-range interactions. Crucially, within the Mamba framework, we restructure the Mamba-inspired global-to-local attention mechanism by incorporating a selective state space model and an attention mechanism. This effectively **reinterprets Mamba as a variant of the Transformer** architecture, improving its adaptability and efficiency. Furthermore, we refine the scanning strategy with Mamba by **integrating the tensor mode-k unfolding into the Mamba** network. This approach emphasizes the low-rank properties of tensors along various modes, while conveniently facilitating 12 scanning directions. Numerical and visual comparisons on both simulation and real datasets demonstrate the superiority of our proposed MiJUN, and achieving overwhelming detail representation.

Introduction

Coded Aperture Snapshot Spectral Imaging (CASSI) has emerged as a widely developed and utilized method for hyperspectral imaging. This method is characterized by low bandwidth, rapid acquisition, and high throughput. Technically, the CASSI process can be divided into two distinct phases. Initially, the 3D hyperspectral image (HSI) is encoded into a single 2D compressed measurement. Subsequently, the computational reconstruction phase employs reconstruction algorithms to estimate the original HSI from the snapshot measurement. This phase is the critical component of the entire CASSI system; therefore, developing high-quality reconstruction algorithms is imperative for the practical implementation of CASSI systems.

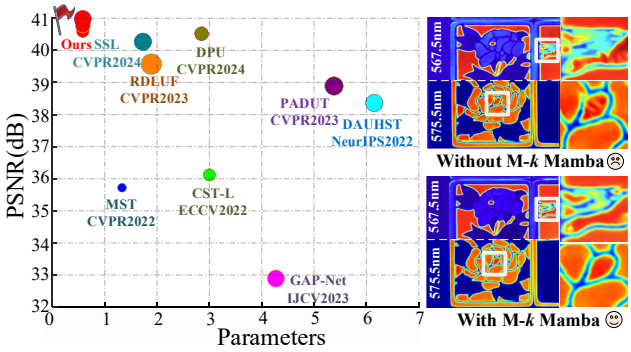

Figure 1: Comparison of reconstruction quality vs. Parameters(M), and FLOPs(G). Our proposed method outperforms comparisons, while utilizing less computational costs. Notably, the images on the right show the feature maps of RDULF and our method, where our features exhibit reduced noise and sharper edges.

To address this challenge, traditional model-based methods (Bioucas-Dias and Figueiredo 2007; Liu et al. 2018; Chen et al. 2023; Luo et al. 2022) often utilize regularization based on image priors to facilitate reconstruction. Although these methods are highly interpretable, they are limited by their reliance on hand-crafted priors, which may lead to sub-optimal results. Recently, numerous deep learning-based approaches for CASSI have been developed. Based on differences in network structure, these approaches are generally divided into three categories: end-to-end methods (Huang et al. 2021; Meng, Ma, and Yuan 2020; Cheng et al. 2022; Meng, Ma, and Yuan 2020), plug-and-play methods (Chan, Wang, and Elgendy 2016; Yuan et al. 2020, 2021; Ebner and Haltmeier 2024), and deep unfolding methods (Wang et al. 2020; Wu et al. 2025; Ma et al. 2019; Zhang et al. 2022). The end-to-end (E2E) method typically constructs a direct mapping from the compressed measurement space to the original image domain. This approach significantly reduces computational complexity and often outperforms traditional model-based methods in terms of efficiency and effectiveness. The plug-and-play (PnP) method incorporates a fixed

*Corresponding author.

pre-trained denoiser into traditional model-based frameworks without additional training. This integration employs pre-trained denoisers, which may not effectively adapt to the specific mappings required by different datasets. Deep unfolding networks (DUNs) reconfigure specific optimization techniques into deep neural architectures. Specifically, DUNs endeavor to construct interpretable deep neural networks by integrating the framework of conventional iterative algorithms. In this work, we focus on DUNs, which have been empirically proven to be successful in resolving optimization challenges.

Typically, DUNs integrate advanced network modules as denoisers to achieve robust interpretability and superior reconstruction capabilities. However, their performance remains uncertain due to reliance on approximated prior settings or insufficient feature learning. Current unfolding algorithms often capture extensive dependencies by leveraging the Transformer framework. Despite these algorithms achieving good results in existing HSI reconstruction tasks, they are still limited by the following issues: (i) These models are developed based on Transformer networks, which have a very high computational cost, as the complexity of the attention being $\mathcal{O}(N^2)$. (ii) There exists a trade-off between computational complexity and effective receptive field, which hinders these methods from exploring long-range dependencies, especially in HSIs. Naturally, this prompts a compelling research question: How can we design an HSI image reconstruction module to achieve a good balance between high performance and low model complexity?

Recently, the state space model (SSM) is a promising backbone for addressing the limitations of Transformers and CNNs. The visualization Mamba model introduces a cross-scanning module, which applies the structured state space sequence (S4) model to visual tasks by unfolding 2D features into 1D arrays along four directions. This allows it to capture long-range context using a global receptive field with $\mathcal{O}(N)$ complexity. However, as the Mamba model unfolds 2D features into 1D sequences, spatially adjacent pixels can become distant in the flattened sequence. This increased separation between neighboring pixels leads to a neglect of local context, resulting in a significant loss of essential local textures, thereby degrading HSI reconstruction performance. To address the aforementioned issues, we propose a Mamba-inspired Joint Unfolding Network (MiJUN) for HSI reconstruction. Specifically, inspired by Mamba, we reformulate the SSM and the attention mechanism in a unified framework, describing Mamba as a variant of the Transformer, thereby leveraging the strengths of both Mamba and Transformer. Furthermore, to address the issue of insufficient spatial and spectral feature representation in HSIs, we are the first to integrate the tensor mode-k unfolding strategy into Mamba. Finally, we introduce an acceleration strategy-based HQS (A-HQS), which can be regarded as a second-order differential equation, featuring improved convex approximation and $\mathcal{O}(1/k^2)$ convergence rates, while the first-order convergence rate is $\mathcal{O}(1/k)$. As shown in Fig. 1, our MiJUN-5stg outperforms the previous SOTA RDLUF-MixS^2-9stg (Dong et al. 2023) by 1.01 dB in PSNR value, with 3$\times$ fewer parameters and 3$\times$ less computational cost.

In summary, we present a **joint unfolding network** for spectral SCI reconstruction, which integrates *mode-k tensor unfolding* into the Mamba framework and then feeds into the accelerated *deep unfolding* network. The principal contributions are as follows:

- We propose a Mamba-inspired accelerated unfolding network for compressive spectral snapshot imaging, which formulates Mamba and Transformer in a unified framework. It retains the inherent advantages of the Mamba, while achieving global-to-local information complementation through the attention module.
- Mode-k tensor unfolding is first incorporated into the Mamba module, which reduces complex tensor operations to relatively easy-to-handle matrix operations by unfolding 3D tensors along each mode. This bridges the high-dimensional input form and the vector form required by Mamba, while conveniently emphasizing low-rankness and achieving 12-direction scanning.
- We introduce an interpretable A-HQS for the solution of the DUN model. Based on this iterative solution framework, redundant elements can be effectively discarded, thereby accelerating the convergence of iterations.
- The comprehensive evaluation conducted on both simulated and real datasets confirms that our proposed method exhibits superior quantitative performance, enhances visual quality, and reduces computational demands. Moreover, it excels at recovering fine details in the image.

Related Work

Vision Transformer for CASSI

Previous studies have employed end-to-end neural networks to develop data-driven priors, which have been extensively applied in SCI applications. Recent related researches (Hu et al. 2022; Zhang and Wu 2021) have also confirmed that CNN-based methods exhibit strong capabilities to model local similarities. However, despite their strengths, CNN-based techniques are constrained by their inductive biases, limiting their ability to identify non-local similarities.

To address the aforementioned issues, Transformer-based approaches (Luo et al. 2024; Wang et al. 2023; Cao et al. 2024) have gained significant popularity in computer vision due to their exceptional ability to model long-range interactions across spatial regions. However, these algorithms exhibit deficiencies in capturing the local features of HSI, failing to adequately represent the detailed and textured information of the images. (Cai et al. 2022b) employ multi-head self-attention (MSA) mechanisms to capture long-range spatial and spectral dependencies in HSI. Using MSA, it computes the spectral dependencies, resulting in an attention map that implicitly encodes the global context. Moreover, (Cai et al. 2022c) introduce a half-shuffle MSA mechanism, which divides attention heads into a local branch and a non-local branch. This method models non-local similarity by shuffling pixels, which brings distant pixels into a local window. However, this technique can only capture non-local similarities of specific pixels, potentially overlooking highly

correlated non-local pixels. Additionally, focusing on pixel-level non-local similarities may miss some object-level non-local similarities. Therefore, designing a network that effectively leverages both local and patch-level non-local priors in HSIs is of great importance.

State Space Model

SSMs were originally developed as a mathematical framework to describe system dynamics in motion. (Gu et al. 2021) introduced the linear State-Space Layer, combining the strengths of recurrent neural networks, temporal convolutions, and neural differential equations to improve model capacity. Building on this, some models (Gu, Goel, and Ré 2021; Xie et al. 2024) leverage the optimization of SSM to address the issue of long-range dependencies, significantly improving computational efficiency.

Recently, SSM has gained increasing attention, being widely applied in natural language processing and gradually extending to visual tasks. Intuitively, Mamba (Gu and Dao 2023) is a state-space model that varies over time based on a gating mechanism, which effectively captures long-sequence dependencies. (Liu et al. 2024) introduce a general visual backbone, Vim, which integrates bidirectional Mamba modules. This approach leverages positional embeddings to encode image sequences and employs a bidirectional state-space model (SSM) to compress visual representations. In (Pei, Huang, and Xu 2024) improves efficiency with a redesigned selective scanning method. However, directly applying Mamba to HSI reconstruction faces challenges, including loss of local context and key textures.

Methodology

Degradation model of CASSI

In the CASSI system, the 3D HSI cube is modulated by a physical mask in the aperture and incorporated different wavelengths through 2D monochrome sensors along the width dimension, ultimately compressed into a single 2D measurement. Fig. 2 illustrates the forward imaging process of the single-disperser CASSI (SD-CASSI). Mathematically, the original HSI data is denoted as $\boldsymbol{X} \in \mathbb{R}^{W\times H\times N_\lambda}$, W and H are the spatial dimensions, and N_λ is the number of spectral channels. Similarly, the physical mask is denoted $\boldsymbol{M} \in \mathbb{R}^{W\times H}$. The coded HSI data cube at n_λ-th wavelength is represented as $\boldsymbol{X}^{'}_{n_\lambda} = \boldsymbol{X}_{n_\lambda} \odot \boldsymbol{M}$, where $\odot$ is the element-wise multiplication.

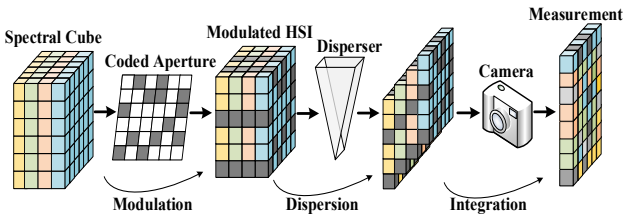

Figure 2: A schematic diagram of CASSI.

Subsequently, passing the disperser, the spatially modulated HSI $\boldsymbol{X}^{'}$ is tilted along the H-axis, which can be formulated as $\boldsymbol{X}^{''} \in \mathbb{R}^{W\times \tilde{H}\times N_\lambda}$ and $\tilde{H} = H + d_{N_\lambda}$. In this context, d_{N_λ} denotes the displacement magnitude experienced by the wavelength corresponding to the N_λ-the order. This operation can be formally described as the modulation of the shifted spectral component, denoted as $\tilde{\boldsymbol{X}} \in \mathbb{R}^{W\times \tilde{H}\times N_\lambda}$, by employing a correspondingly shifted mask $\tilde{\boldsymbol{M}} \in \mathbb{R}^{W\times \tilde{H}\times N_\lambda}$. The relation for $\tilde{\boldsymbol{M}}$ at any given position is articulated as $\tilde{\boldsymbol{M}}(i,j,n_\lambda) = \boldsymbol{M}(w,h+d_\lambda)$. Finally, the imaging sensor acquires the dispersed as a 2D measurement $\boldsymbol{Y}$ can be formulated as follows:

$$\boldsymbol{Y} = \textstyle\sum_{n=1}^{N_\lambda} \tilde{\boldsymbol{X}}(:,:,n_\lambda) \odot \tilde{\boldsymbol{M}}(:,:,n_\lambda) + \boldsymbol{B}, \tag{1}$$

where $\boldsymbol{B}$ denotes the additive noise. Mathematically, by vectorizing $\boldsymbol{X}$ and $\boldsymbol{Y}$, the above equation can be formulated as:

$$\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{x} + \boldsymbol{b}, \tag{2}$$

where $\boldsymbol{x} \in \mathbb{R}^{W\tilde{H}N_\lambda}$, $\boldsymbol{y} \in \mathbb{R}^{W\tilde{H}}$. Here, $\boldsymbol{\Phi} \in \mathbb{R}^{W\tilde{H}\times W\tilde{H}N_\lambda}$ denotes the sensing matrix, which is generally construed as the spatially shifted mask within the imaging apparatus.

Accelerated deep unfolding framework

Recall that directly inferring the HSI $\boldsymbol{x}$ from the degradation model Eq. (2) is intractable. Therefore, it is necessary to utilize the regularizer to constrain the solution space, and the inversion of the Eq. (2) can be construed as an optimization effort aimed at minimizing the cost function:

$$\text{argmin}_{\boldsymbol{x}} \tfrac{1}{2}\|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{x}\|^2 + \tau\mathcal{R}(\boldsymbol{x}), \tag{3}$$

where $\mathcal{R}(\boldsymbol{x})$ is the regularization term, characterizing the prior knowledge of the desired $\boldsymbol{x}$, and τ denotes the noise-balancing factor. Within the research of SCI, previous deep unfolding models disregarded the implications of *accelerated* optimization algorithms, thereby failing to exploit *second-order gradient* information adequately. We augment the HQS algorithm, previously used in DUN, to bridge these two aspects with an improved accelerated variant. For clarity, we hereby briefly describe the iterative frameworks of A-HQS to solve the Eq. (3), which proceed as follows:

$$\boldsymbol{x}_{k+1} = \text{argmin}_{\boldsymbol{x}} \tfrac{1}{2}\|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{x}\|^2 + \tfrac{\mu}{2}\|\boldsymbol{x} - \hat{\boldsymbol{z}}_{k+1}\|^2, \tag{4a}$$

$$\boldsymbol{z}_{k+1} = \text{argmin}_{\boldsymbol{z}} \tfrac{\mu}{2}\|\boldsymbol{x}_{k+1} - \boldsymbol{z}_k\|^2 + \tau\mathcal{R}(\boldsymbol{z}_k), \tag{4b}$$

$$\hat{\boldsymbol{z}}_{k+1} = \boldsymbol{z}_{k+1} + \beta_{k+1}(\boldsymbol{z}_{k+1} - \boldsymbol{z}_k), \tag{4c}$$

where β is the balancing parameter. For the subproblem $\boldsymbol{x}_{k+1}$, it should be noted that Eq. (4a) is differentiable and the gradient descent scheme can be borrowed as a solver:

$$\boldsymbol{x}_{k+1} = (\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi} + \mu\boldsymbol{I})^{-1}(\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{y} + \mu\hat{\boldsymbol{z}}_k), \tag{5}$$

where $\boldsymbol{I}$ denotes the identity matrix with desired dimensions. The matrix can be regarded as a regularized version (by adding $\mu\boldsymbol{I}$) of the Hessian of $\frac{1}{2}\|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{x}\|^2$. Therefore, the optimization process also manifests the application of second-order information in this aspect. In CASSI systems, $\boldsymbol{\Phi}$ is a fat matrix, and $(\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi} + \mu\boldsymbol{I})$ forms a large matrix.

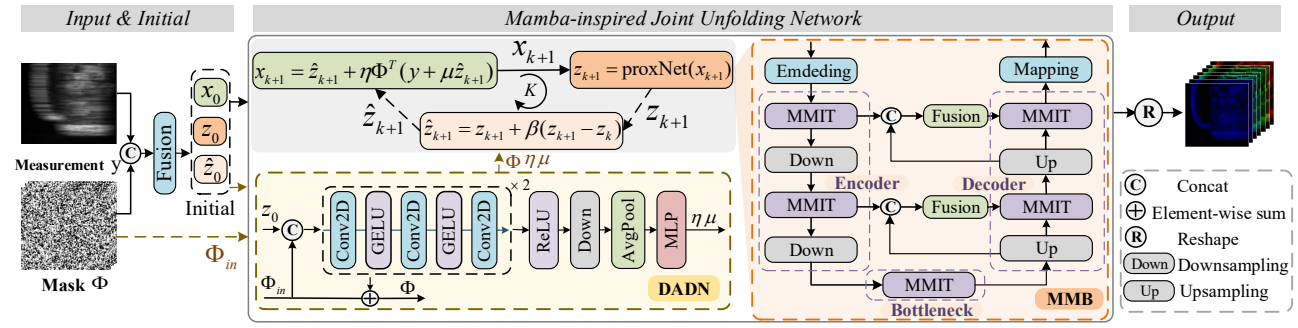

Figure 3: An overview of our proposed MiJUN for HSI reconstruction task, including input & Initial, MiJUN model, and Output. The model includes three iterative operators: $\boldsymbol{x}$, $\boldsymbol{z}$, $\hat{\boldsymbol{z}}$. During the iterative process, the parameters are estimated by the DADN, with $\hat{\boldsymbol{z}}$ learned through the MMB block.

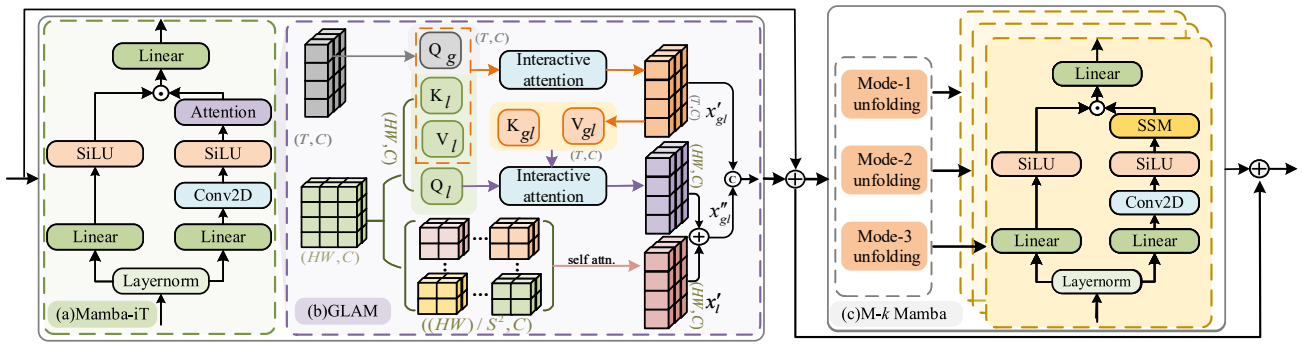

Figure 4: The diagram of the proposed MMIT. Features are first sufficiently modeled with local and global information through the Mamba-i T module, followed by the M-k Mamba to further enhance the low-rank attributes.

Therefore, based on the Sherman-Morrison-Woodbury formula, the equation can be simplified to the following:

$$\begin{aligned}\boldsymbol{x}_{k+1} = &[\mu^{-1}\boldsymbol{I} - \mu^{-1}\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{I} + \boldsymbol{\Phi}\mu^{-1}\boldsymbol{\Phi}^{\mathsf{T}})^{-1}\boldsymbol{\Phi}\mu^{-1}] \\ &\times [\Phi^{\mathsf{T}}\boldsymbol{y} + \mu\hat{\boldsymbol{z}}_k].\end{aligned} \tag{6}$$

For SCI in this paper, $\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathsf{T}}$ corresponds to an identity matrix interspersed with zeros on its diagonal (matching the locations of the undetermined observations) as:

$$\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathsf{T}} = \mathrm{diag}\{r_1, \cdots, r_N\}, \tag{7}$$

where N represents the number of rows in $\boldsymbol{\Phi}$.

Consequently, Eq. (5) simply involves multiplication $(\boldsymbol{\Phi}^{\mathsf{T}}\boldsymbol{y} + \mu\hat{\boldsymbol{z}}_k)$ by $\boldsymbol{\Phi}^{\mathsf{T}}\boldsymbol{\Phi}$, which is an operation with $\mathcal{O}(n)$. Eq. (5) can be simplified as:

$$\boldsymbol{x}_{k+1} = \hat{\boldsymbol{z}}_k + \boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{y} - \boldsymbol{\Phi}\hat{\boldsymbol{z}}_k) \oslash [\mu + \mathrm{diag}(\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathsf{T}})], \tag{8}$$

where $\oslash$ is the element-wise division of Hadamard division.

For the $\boldsymbol{z}$-subproblem, Eq. (4b) is a deterministic approximation operator predicated on a specified prior $\mathcal{R}(\boldsymbol{z})$. Unfortunately, the inherent uncertainty associated with the function $\mathcal{R}(\boldsymbol{z})$ precludes the availability of any closed-form solutions. Thus, we propose a Mamba-inspired and Mamba model to function as the prior extractor, which can generally be formulated as:

$$\boldsymbol{z}_{k+1} = \mathrm{proxNet}_{\tau,\mu}(\boldsymbol{x}_{k+1}). \tag{9}$$

Additionally, $\mathrm{proxNet}_{\tau,\mu}$ can be represented as a denoiser $\mathcal{D}_\eta$ with learnable noise level η. The overall iterative MiJUN framework is shown in Fig. 3. And, we introduce the Mamba and Mamba-inspired Block (MMB) to play the role of proxNet, which features a U-shaped network and mainly includes a submodule of Mamba and Mamba-inspired Transformer (MMIT). The detailed description is as follows.

Prior extractor

Mamba-inspired Transformer. Empirically, Mamba has been shown to perform well in tasks requiring global context understanding. However, despite these strengths, Mamba encounters challenges in adequately representing local texture features. This inadequacy arises because the linear unfolding of 2D features into 1D sequences can lead to the loss of spatially adjacent pixel relationships, crucial for capturing fine-grained local details. The large distance between neighboring pixels in the flattened sequence can result in a neglect of local context, leading to a significant loss of key local textures and reduced performance in tasks that require detailed local feature extraction.

By theoretically and empirically analyzing Mamba from the perspective of linear attention Transformer (Han et al. 2024), and integrating strategies to enhance local feature extraction within the Mamba framework can potentially ad-

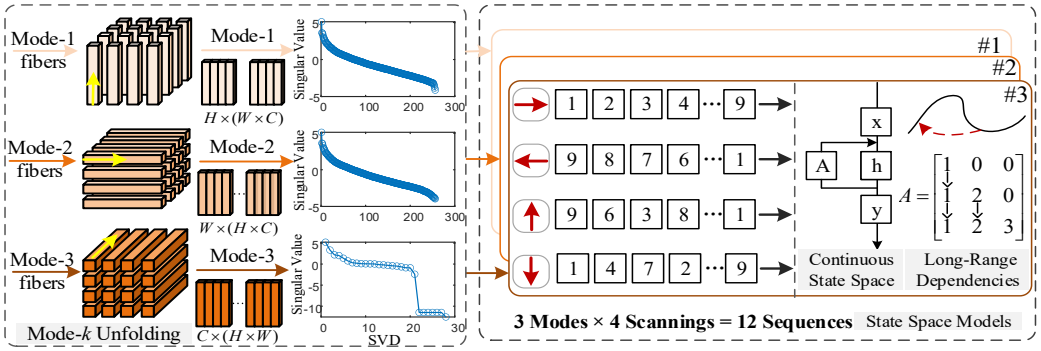

Figure 5: Illustration of Mode-k unfolding along each direction of 3D tensor and linear-overhead SSM with different-direction scanning scheme. The low rank of each matrix after unfolding is demonstrated by singular value decomposition (SVD(log)).

dress this limitation and improve the performance of HSI reconstruction. Specifically, we reformulate selective SSM and attention within a unified framework, describing the Mamba-inspired Transformer (Mamba-i T) as a variant in Fig. 4(a). Following (Liu et al. 2024), the input feature $\boldsymbol{X} \in \mathbb{R}^{W\times H\times C}$ is processed through two parallel branches. One branch consists of channel expansion, a linear layer, and SiLU activation. In the other branch, the channels are first expanded to λC using a linear layer, where λ is a predefined channel expansion factor. Then, features are extracted using a 3×3 convolution followed by SiLU activation. Finally, these features are processed through an attention mechanism. Notably, we adopt a global-local attention mechanism (GLAM) to compensate for Mamba's deficiencies in capturing spatial local features shown in Fig. 4(b). After aggregating the features of both branches, the channels are projected back to C, producing an output $\boldsymbol{X}_{out}$, as follows.

$$\begin{aligned}
&\text{Branch1}: \boldsymbol{X}_1 = \text{SiLU}(\text{Lin}(\text{Ln}(\boldsymbol{X}))),\\
&\text{Branch2}: \boldsymbol{X}_2 = \text{GLAM}(\text{SiLU}(\text{Conv}(\text{Lin}(\text{Ln}(\boldsymbol{X}))))),\\
&\text{Output}: \boldsymbol{X}_{out} = \text{Lin}(\boldsymbol{X}_1 \odot \boldsymbol{X}_2),
\end{aligned}$$

where Ln is layernorm, Lin represents the linear layer, and $\odot$ is Hadamard product.

Mode-k unfolding-based Mamba. Considering the spatial complexity and spectral similarity of HSIs, we adopt a tensor mode-k unfolding strategy to capture both spatial and spectral structures. This approach preserves essential spatial features that might otherwise be lost with traditional channel-slicing methods. As illustrated in Fig. 5 with Scene 1, the schematic shows the image for each mode-k unfolding matrix and their corresponding singular values. Notably, Mode-1 and Mode-2 share similar singular value distributions, while Mode-3 exhibits a distinct pattern.

Therefore, hereby we integrate the tensor mode-k unfolding strategy into the Mamba network, proposing the Mode-k unfolding-based Mamba (M-k Mamba), which is illustrated in the left of Fig. 5. First, the input data undergoes a tensor mode-k unfolding transformation to obtain different tensor unfolded data, *i.e.*, $\boldsymbol{X} \in \mathbb{R}^{B\times WH\times C}$. Then, the data is fed into the model following the previously mentioned Mamba network (Liu et al. 2024). The difference lies in Branch 2, where $\boldsymbol{X}_2$ is computed using the SSM. By structuring the Mamba input to depend on the long-range representation of parameter '$\boldsymbol{A}$', it effectively filters out irrelevant information, allowing more efficient compression of the context into the hidden state. As shown in Fig. 5, in the SSM, the given input is unfolded into four one-dimensional sequences/vectors $\{\boldsymbol{x}_n \in \mathbb{R}^{1\times \hat{H}\hat{W}\hat{C}}\}_{n=1}^{4}$ by scanning pixels along four different traversal paths: from top-left to bottom-right, from top-right to bottom-left, from bottom-right to top-left, and from bottom-left to top-right. Notably, combining 3 modes of mode-k, 12-direction scanning sequences can be conveniently derived. Subsequently, SSM is calculated as follows.

$$\begin{aligned}
&\{\boldsymbol{B}_t, \Delta_t, \boldsymbol{C}_t\} = \mathcal{P}_{\text{proj}}(\boldsymbol{x}_n), \quad \overline{\Delta_t} = \sigma^{+}(\mathcal{P}_{\text{dt}}\Delta_t),\\
&\overline{\boldsymbol{A}_t} = \exp(-\exp(\boldsymbol{A}_{\log}\overline{\Delta_t})), \quad \overline{\boldsymbol{B}_t} = \overline{\Delta_t} \odot \boldsymbol{B}_t,
\end{aligned}$$

where $\mathcal{P}_{\text{proj}}$, $\mathcal{P}_{\text{dt}}$, and $A_{\log}$ are time-invariant weight matrices, σ^{+} is softplus activation function, and $\odot$ is element-wise multiplication. Weight matrices $\boldsymbol{B}$ and $\boldsymbol{C}$ directly depend on input $\boldsymbol{x}_n$, whereas recurrent weight matrix $\boldsymbol{A}$ depends solely on the time-scale parameter Δ. The hidden state $\boldsymbol{h}$ and output $\boldsymbol{y}$ of SSM are calculated as follows.

$$\boldsymbol{h}_t = \overline{\boldsymbol{A}} \odot \boldsymbol{h}_{t-1} + \overline{\boldsymbol{B}} \odot \boldsymbol{x}_n, \quad \boldsymbol{y}_t = \boldsymbol{C}_t\boldsymbol{h}_t + \boldsymbol{D}_t \odot \boldsymbol{x}_n,$$

where $\boldsymbol{D}$ is the scale parameter, $._t$ represents the t-th state.

Experiments

Experimental settings

Datasets. In the simulation experiments, we use two datasets, CAVE and KAIST. The CAVE dataset comprises 32 HSI images with spatial dimensions of 512×512. The KAIST dataset contains 30 HSI images, each with spatial dimensions of 2704×3376. Same as previous researches, we utilize the CAVE dataset as our training set and select 10 scenes from the KAIST dataset for testing. In real experiments, we use five real CASSI datasets (Meng, Ma, and Yuan 2020), with dimensions of $660 \times 714 \times 28$, wavelength range from 450 to 650 nm and a dispersion of 54 pixels.

Implementation Details. The proposed model MiJUN is implemented by Pytorch.During the training process, we utilize the Adam optimizer ($\beta_1 = 0.9$ and $\beta_2 = 0.999$) and

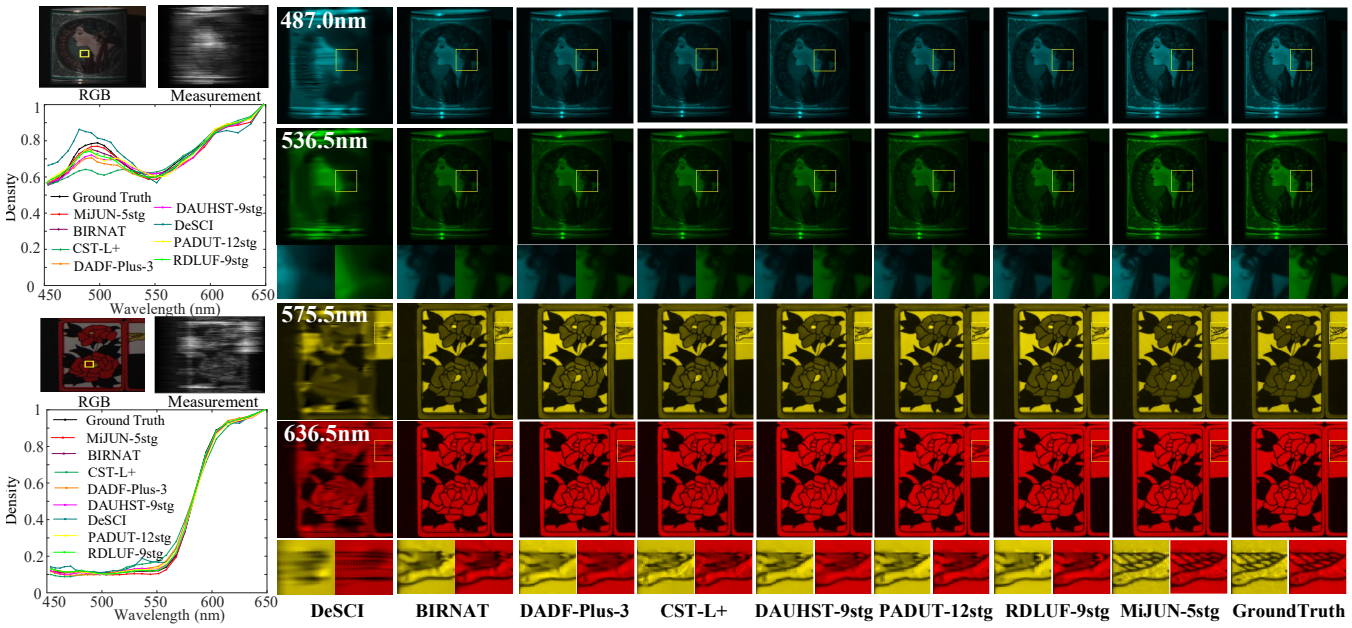

Figure 6: The simulated HSI reconstruction results for Scene 1 (top) & Scene 7 (bottom) with 2 out of 28 spectral channels, including seven state-of-the-art algorithms and our proposed MiJUN-5stg. The left displays the RGB image and measurement. The bottom-left shows the spectral density curves corresponding to the selected yellow box in the RGB image.

a cosine annealing scheduler, running for 200 epochs on a single RTX 4090 GPU. To evaluate the performance, we use the peak signal-to-noise ratio (PSNR) and structure similarity (SSIM) to assess the HSI reconstruction capabilities.

Compare with State-of-the-art

We compare our proposed MiJUN model with several SOTA CASSI algorithms, and the results are analyzed as follows.

Synthetic data. To comprehensively evaluate the quantitative results of all competing methods, we test on ten simulated datasets and presented the corresponding numerical results in Tab. 1. Different colors are used to distinguish the types of algorithms: gray for model-based methods, orange for end-to-end networks, and green for deep unfolding methods. In Tab. 1, it is evident that our MiJUN model achieved the best numerical results in all cases. Fig. 6 shows the visual reconstruction results. It is evident that MiJUN demonstrates a significant advantage over others, particularly in Scene 1, where it excels in detailing hair, and in Scene 7, where it effectively captures the intricacies of bird wings. Furthermore, to evaluate overfitting, we test our pre-trained model on the unseen ICVL dataset [1], as shown in Tab. 2, which demonstrates good generalization performance.

Real data. To further investigate the superiority of this model, we also conduct experiments on real HSI reconstruction tasks. Since the ground truth of real-world scenarios is unattainable, we can only compare qualitative results. Following the experimental setting of (Cai et al. 2022b), we apply MiJUN-5stg to training in the simulated dataset. Fig. 7 presents the visual results of our model compared to other

[1]https://icvl.cs.bgu.ac.il/hyperspectral

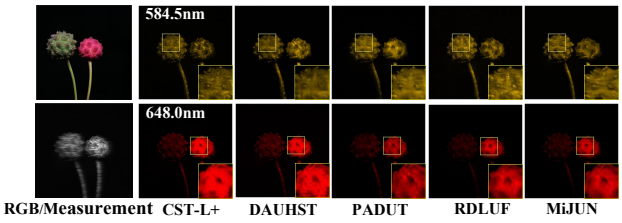

Figure 7: The real data comparisons. 2 out of 28 wavelengths are plotted for visual comparison.

algorithms in Scene 4 (2 out of the 28 spectral channels). In comparison, our model can reconstruct more textures and details, but it still exhibits some blurriness and artifacts. These challenges highlight the difficulties the model faces in handling real-world hyperspectral reconstruction tasks.

Ablation study

Our ablation analysis focuses on three main components of MiJUN, acceleration strategy, Mamba-i T and M-k Mamba. We conduct ablation experiments on public simulated HSI datasets to investigate the effectiveness of each module. Tab. 3 summarizes the performance of different cases compared to our model. We select RDLUF-MixS^2-5stg, which combines the basic module, DADN, and the attention mechanism, as the baseline. As shown in Tab. 3, when we incorporate the acceleration strategy into the baseline and replace the attention mechanism with GLAM, the model performance achieves a certain degree of improvement, with PSNR increasing from 38.59 to 39.64 and SSIM increasing

Algorithms	Params(M)	FLOPs(G)	Scene1	Scene2	Scene3	Scene4	Scene5	Scene6	Scene7	Scene8	Scene9	Scene10	Avg
TwIST (Bioucas-Dias and Figueiredo 2007)	—	—	25.16 0.700	23.02 0.604	21.40 0.711	30.19 0.851	21.41 0.635	20.95 0.644	22.20 0.643	21.82 0.650	22.42 0.690	22.67 0.569	23.12 0.669
GAP-TV (Yuan 2016)	—	—	26.82 0.754	22.89 0.610	26.31 0.802	30.65 0.852	23.64 0.703	21.85 0.663	23.76 0.688	21.98 0.655	22.63 0.682	23.1 0.584	24.36 0.669
DeSCI (Liu et al. 2018)	—	—	27.13 0.748	23.04 0.620	26.62 0.818	34.96 0.897	23.94 0.706	22.38 0.683	24.45 0.743	22.03 0.673	24.56 0.732	23.59 0.587	25.27 0.721
HDNet (Hu et al. 2022)	2.37	154.76	34.96 0.937	35.64 0.943	35.55 0.94	41.64 0.976	32.56 0.948	34.33 0.95	33.27 0.92	32.26 0.945	34.17 0.944	32.22 0.94	34.66 0.946
BIRNAT (Cheng et al. 2022)	4.40	212.55	36.78 0.951	37.89 0.957	40.61 0.971	46.93 0.985	35.42 0.963	35.30 0.959	36.58 0.954	33.95 0.955	39.46 0.969	32.80 0.937	37.57 0.960
DADF-Plus-3 (Xu et al. 2023)	58.13	230.41	37.46 0.965	39.86 0.976	41.03 0.974	45.98 0.989	35.53 0.972	37.02 0.975	36.76 0.958	34.78 0.971	40.07 0.976	34.39 0.962	38.29 0.972
MST++ (Cai et al. 2022b)	1.33	19.42	35.57 0.945	36.22 0.949	37.00 0.959	42.86 0.980	33.27 0.954	35.27 0.960	34.05 0.936	33.50 0.956	36.17 0.956	33.26 0.949	35.72 0.955
CST-L+ (Cai et al. 2022a)	3.00	40.10	35.96 0.949	36.84 0.955	38.16 0.962	42.44 0.975	33.25 0.955	35.72 0.963	34.86 0.944	34.34 0.961	36.51 0.957	33.09 0.945	36.12 0.957
DNU (Wang et al. 2020)	1.19	163.48	31.72 0.863	31.13 0.846	29.99 0.845	35.34 0.908	29.03 0.833	30.87 0.887	28.99 0.839	30.13 0.885	31.03 0.876	29.14 0.849	30.74 0.863
GAP-Net (Meng, Yuan, and Jalali 2023)	4.27	78.58	33.63 0.913	33.19 0.902	33.96 0.931	39.14 0.971	31.44 0.921	32.29 0.927	31.79 0.903	30.25 0.907	33.06 0.916	30.14 0.898	32.89 0.919
DAUHST-9stg (Cai et al. 2022c)	6.15	79.50	37.25 0.958	39.02 0.967	41.05 0.971	46.15 0.983	35.80 0.969	37.08 0.970	37.57 0.963	35.10 0.966	40.02 0.970	34.59 0.956	38.36 0.967
PADUT-5stg (Li et al. 2023)	2.24	37.90	36.68 0.955	38.74 0.969	41.37 0.975	45.79 0.988	35.13 0.967	36.37 0.969	36.52 0.959	34.40 0.967	39.57 0.971	33.78 0.955	37.84 0.967
PADUT-12stg (Li et al. 2023)	5.38	90.46	37.36 0.962	40.43 0.978	42.38 0.979	46.62 0.990	36.26 0.974	37.27 0.974	37.83 0.966	35.33 0.974	40.86 0.978	34.55 0.963	38.89 0.974
RDLUF-MixS^2-9stg (Dong et al. 2023)	1.89	115.34	37.94 0.966	40.95 0.977	43.25 0.979	47.83 0.990	37.11 0.976	37.47 0.975	38.58 0.969	35.50 0.970	41.83 0.978	35.23 0.962	39.57 0.974
MiJUN-5stg	0.56	40.98	38.52 0.969	41.37 0.980	44.29 0.981	**48.84** 0.992	38.58 0.982	38.08 0.978	40.69 0.979	**36.93** 0.977	43.33 0.983	35.41 0.964	40.60 0.978
MiJUN-7stg	0.56	57.32	39.10 0.971	41.42 0.981	44.25 0.981	48.78 0.992	39.04 0.983	37.97 0.978	40.76 0.979	36.46 0.976	43.58 0.984	**35.64** 0.966	40.70 0.979
MiJUN-9stg	0.56	73.67	**39.26** **0.973**	**41.78** **0.983**	**44.31** **0.983**	48.53 **0.994**	**39.30** **0.985**	**38.22** **0.979**	**41.00** **0.983**	36.72 **0.978**	**43.84** **0.985**	35.56 **0.967**	**40.86** **0.982**

Table 1: The results of PSNR in dB (top entry in each cell), SSIM (bottom entry in each cell) on the 10 synthetic spectral scenes.'-5stg' denotes the network with 5 unfolding stages. 'Avg' represents the average of 10 scenes.

Method	eve_0311	BUG-0403	4cam_0411	CC_40D	guCAMP_0514
RDLUF	31.29	30.74	32.90	31.69	37.28
Ours	31.97	31.27	33.82	32.18	37.78

Table 2: Comparison of PSNR on the unseen ICVL dataset.

Methods	PSNR ↑	SSIM ↑
baseline	38.59	0.969
$w/$Acc	38.60	0.971
$w/$ GLAM	39.64	0.974

Methods	PSNR ↑	SSIM ↑
$w/$Mamba-i	39.89	0.976
$w/$Mamba	40.25	0.976
$w/$M-k(ours)	**40.60**	**0.978**

Table 3: Ablation study of key components in our key components. The $w/$ denotes the inclusion of a module.

from 0.969 to 0.974. However, when we further integrate GLAM with the Mamba framework to validate the effectiveness of the Mamba-i T module(GLAB $\rightarrow$ Mamba-i T), we observe a 0.25 dB increase in PSNR. Overall, the Mamba-i T module achieved an improvement of 1.3 dB in PSNR compared to the baseline. Furthermore, we validate the effectiveness of integrating the tensor mode-k unfolding with the Mamba network. In terms of PSNR results, when only the Mamba module is added, the PSNR increased to 40.25 dB. Finally, by effectively integrating the tensor mode-k unfolding with the Mamba module(Mamaba $\rightarrow$ M-k Mamaba), we develop our complete model, MiJUN, which achieved the optimal result with a PSNR of 40.60 dB. This ultimately confirms what is shown in Fig. 1: our model achieved the best results, with sharper image edges and richer details.

Conclusion

In this paper, we introduce a novel joint unfolding network for spectral snapshot compressive imaging, dubbed MiJUN. Firstly, based on the accelerated strategy, we construct an accelerated iteration scheme for DUN, enabling effective elimination of redundant information. Additionally, inspired by Mamba, we incorporate a global-local attention mechanism into the Mamba framework as a variant of the Transformer architecture, effectively enhancing the model's feature extraction capabilities. Furthermore, to fully consider data characteristics, we introduce tensor mode-k unfolding in the Mamba network, which enhances the representation of the intrinsic properties of the data. This approach enables the model to learn features at a fine-grained level, facilitating detailed reconstruction of HSIs. Comprehensive evaluations on both simulated and real datasets confirm the superior quantitative performance of our proposed approach.

Acknowledgments

This work was supported by the National Key R&D Program of China (grant number 2024YFF0505603, 2024YFF0505600), the National Natural Science Foundation of China (grant number 62271414), Zhejiang Provincial Outstanding Youth Science Foundation (grant number LR23F010001), Zhejiang "Pioneer" and "Leading Goose"R&D Program(grant number 2024SDXHDX0006, 2024C03182), the Key Project of Westlake Institute for Optoelectronics (grant number 2023GD007), the 2023 International Sci-tech Cooperation Projects under the purview of the "Innovation Yongjiang 2035" Key R&D Program (grant number 2024Z126), and the Zhejiang Province Postdoctoral Research Excellence Funding Program (grant number ZJ2024086).

References

Bioucas-Dias, J. M.; and Figueiredo, M. A. 2007. A new TwIST: Two-step iterative shrinkage/thresholding algorithms for image restoration. *IEEE Transactions on Image processing*, 16(12): 2992–3004.

Cai, Y.; Lin, J.; Hu, X.; Wang, H.; Yuan, X.; Zhang, Y.; Timofte, R.; and Van Gool, L. 2022a. Coarse-to-fine sparse transformer for hyperspectral image reconstruction. In *European Conference on Computer Vision*, 686–704. Springer.

Cai, Y.; Lin, J.; Hu, X.; Wang, H.; Yuan, X.; Zhang, Y.; Timofte, R.; and Van Gool, L. 2022b. Mask-guided spectral-wise transformer for efficient hyperspectral image reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17502–17511.

Cai, Y.; Lin, J.; Wang, H.; Yuan, X.; Ding, H.; Zhang, Y.; Timofte, R.; and Gool, L. V. 2022c. Degradation-aware unfolding half-shuffle transformer for spectral compressive imaging. *Advances in Neural Information Processing Systems*, 35: 37749–37761.

Cao, M.; Wang, L.; Zhu, M.; and Yuan, X. 2024. Hybrid CNN-Transformer Architecture for Efficient Large-Scale Video Snapshot Compressive Imaging. *International Journal of Computer Vision*, 1–20.

Chan, S. H.; Wang, X.; and Elgendy, O. A. 2016. Plug-and-play ADMM for image restoration: Fixed-point convergence and applications. *IEEE Transactions on Computational Imaging*, 3(1): 84–98.

Chen, Y.; Gui, X.; Zeng, J.; Zhao, X.; and He, W. 2023. Combining low-rank and deep plug-and-play priors for snapshot compressive imaging. *IEEE Transactions on Neural Networks and Learning Systems*, 1–13.

Cheng, Z.; Chen, B.; Lu, R.; Wang, Z.; Zhang, H.; Meng, Z.; and Yuan, X. 2022. Recurrent neural networks for snapshot compressive imaging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2): 2264–2281.

Dong, Y.; Gao, D.; Qiu, T.; Li, Y.; Yang, M.; and Shi, G. 2023. Residual degradation learning unfolding framework with mixing priors across spectral and spatial for compressive spectral imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22262–22271.

Ebner, A.; and Haltmeier, M. 2024. Plug-and-play image reconstruction is a convergent regularization method. *IEEE Transactions on Image Processing*.

Gu, A.; and Dao, T. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.

Gu, A.; Goel, K.; and Ré, C. 2021. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*.

Gu, A.; Johnson, I.; Goel, K.; Saab, K.; Dao, T.; Rudra, A.; and Ré, C. 2021. Combining recurrent, convolutional, and continuous-time models with linear state space layers. *Advances in neural information processing systems*, 34: 572–585.

Han, D.; Wang, Z.; Xia, Z.; Han, Y.; Pu, Y.; Ge, C.; Song, J.; Song, S.; Zheng, B.; and Huang, G. 2024. Demystify Mamba in Vision: A Linear Attention Perspective. *arXiv preprint arXiv:2405.16605*.

Hu, X.; Cai, Y.; Lin, J.; Wang, H.; Yuan, X.; Zhang, Y.; Timofte, R.; and Van Gool, L. 2022. Hdnet: High-resolution dual-domain learning for spectral compressive imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17542–17551.

Huang, T.; Dong, W.; Yuan, X.; Wu, J.; and Shi, G. 2021. Deep gaussian scale mixture prior for spectral compressive imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16216–16225.

Li, M.; Fu, Y.; Liu, J.; and Zhang, Y. 2023. Pixel adaptive deep unfolding transformer for hyperspectral image reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 12959–12968.

Liu, Y.; Tian, Y.; Zhao, Y.; Yu, H.; Wang, Y.; Ye, Q.; and Liu, Y. 2024. VMamba: Visual State Space Model. *arXiv preprint arXiv:2401.10166*.

Liu, Y.; Yuan, X.; Suo, J.; Brady, D. J.; and Dai, Q. 2018. Rank minimization for snapshot compressive imaging. *IEEE transactions on pattern analysis and machine intelligence*, 41(12): 2990–3006.

Luo, F.; Chen, X.; Gong, X.; Wu, W.; and Guo, T. 2024. Dual-Window Multiscale Transformer for Hyperspectral Snapshot Compressive Imaging. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 3972–3980.

Luo, Y.; Zhao, X.; Meng, D.; and Jiang, T. 2022. Hlrtf: Hierarchical low-rank tensor factorization for inverse problems in multi-dimensional imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19303–19312.

Ma, J.; Liu, X.; Shou, Z.; and Yuan, X. 2019. Deep tensor ADMM-net for snapshot compressive imaging. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10223–10232.

Meng, Z.; Ma, J.; and Yuan, X. 2020. End-to-end low cost compressive spectral imaging with spatial-spectral self-attention. In *European conference on computer vision*, 187–204.

Meng, Z.; Yuan, X.; and Jalali, S. 2023. Deep unfolding for snapshot compressive imaging. *International Journal of Computer Vision*, 131(11): 2933–2958.

Pei, X.; Huang, T.; and Xu, C. 2024. EfficientVMamba: Atrous Selective Scan for Light Weight Visual Mamba. *arXiv preprint arXiv:2403.09977*.

Wang, L.; Cao, M.; Zhong, Y.; and Yuan, X. 2023. Spatial-temporal transformer for video snapshot compressive imaging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7): 9072–9089.

Wang, L.; Sun, C.; Zhang, M.; Fu, Y.; and Huang, H. 2020. Dnu: Deep non-local unrolling for computational spectral imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1661–1671.

Wu, Z.; Lu, R.; Fu, Y.; and Yuan, X. 2025. Latent Diffusion Prior Enhanced Deep Unfolding for Snapshot Spectral Compressive Imaging. In *European Conference on Computer Vision*, 164–181. Springer.

Xie, X.; Cui, Y.; Ieong, C.-I.; Tan, T.; Zhang, X.; Zheng, X.; and Yu, Z. 2024. Fusionmamba: Dynamic feature enhancement for multimodal image fusion with mamba. *arXiv preprint arXiv:2404.09498*.

Xu, P.; Liu, L.; Zheng, H.; Yuan, X.; Xu, C.; and Xue, L. 2023. Degradation-aware dynamic fourier-based network for spectral compressive imaging. *IEEE Transactions on Multimedia*.

Yuan, X. 2016. Generalized alternating projection based total variation minimization for compressive sensing. In *2016 IEEE International conference on image processing*, 2539–2543.

Yuan, X.; Liu, Y.; Suo, J.; and Dai, Q. 2020. Plug-and-play algorithms for large-scale snapshot compressive imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1447–1457.

Yuan, X.; Liu, Y.; Suo, J.; Durand, F.; and Dai, Q. 2021. Plug-and-play algorithms for video snapshot compressive imaging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10): 7093–7111.

Zhang, X.; and Wu, X. 2021. Attention-guided image compression by deep reconstruction of compressive sensed saliency skeleton. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13354–13364.

Zhang, X.; Zhang, Y.; Xiong, R.; Sun, Q.; and Zhang, J. 2022. Herosnet: Hyperspectral explicable reconstruction and optimal sampling deep network for snapshot compressive imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17532–17541.