Sp3ctralMamba: Physics-Driven Joint State Space Model for Hyperspectral Image Reconstruction

Ge Meng[1, 2], Jingyan Tu[1, 2], Jingjia Huang[1, 2], Yunlong Lin[1, 2], Yingying Wang[1, 3], Xiaotong Tu[1, 2], Yue Huang[1, 2, 3], Xinghao Ding[1, 2, 3*]

[1] Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, Xiamen University, China
[2]School of Informatics, Xiamen University, China
[3]Institute of Artificial Intelligence, Xiamen University, China
{mengg, tujingyan, huangjj, linyl, wangyingying7}@stu.xmu.edu.cn, {xttu, yhuang2010, dxh}@xmu.edu.cn

Abstract

Hyperspectral image (HSI) reconstruction aims to restore the original 3D HSIs from the 2D hyperspectral snapshot compressive images (SCIs). The key to high-fidelity HSI reconstruction lies in designing refined spatial and spectral attention mechanisms, which are crucial for generating fine-grained representations of HSI based on the limited spatial and spectral information available in SCI. Recently, Mamba has demonstrated remarkable performance and efficiency in modeling spatial correlations. Its implicit attention mechanism generates three orders of magnitude more attention matrices than transformers, significantly raising the performance ceiling for HSI reconstruction. In this paper, we propose a novel joint SSM network named Sp3ctralMamba for HSI reconstruction. Sp3ctralMamba integrates frequency domain knowledge and physical priors to enhance reconstruction quality. Specifically, we first perform hierarchical decomposition of the 3D HSI embedding to mitigate the negative impact of distant bands on reconstruction. Next, we design a joint SSM block S^3Mamba (S^3MAB) to perform parallel scans of the embeddings from different bands. In addition to the conventional vanilla scan, S^3MAB introduces a local scanning scheme to address the reconstruction challenges posed by the spatial sparsity of spectral information. Furthermore, a spiral scanning scheme in the frequency domain is incorporated to enhance the order correlation between different frequency signals. Finally, we introduce energy priors and structural priors to constrain the generation of spectral and spatial representations during the training process. Extensive experiments on both simulated and real datasets demonstrate that Sp3ctralMamba significantly elevates HSI reconstruction performance to a new level, surpassing SOTA methods in both quantitative and qualitative metrics.

Introduction

Objects exhibit varying reflectance under different spectral wavelength. Leveraging this property, hyperspectral imaging is widely used in various fields, such as agriculture (Ishida et al. 2018), environmental monitoring (Wright, Levermore, and Kelly 2019), object tracking (Kim et al. 2012; Pan et al. 2003), and medical image processing (Lu and Fei 2014; Meng et al. 2020). To collect HSIs, traditional imaging systems require sequential scanning of the scene along spatial and spectral dimensions, which lacks flexibility and is time-consuming (Cai et al. 2022b). To this end, a snapshot compressive imaging (SCI) system called coded aperture snapshot spectral imaging (CASSI) is typically used to compress the information of snapshots along the spectral dimension into one single 2D measurement (Yuan, Brady, and Katsaggelos 2021). Then, the complete HSI data can be obtained from these compressed measurements using HSI reconstruction algorithms.

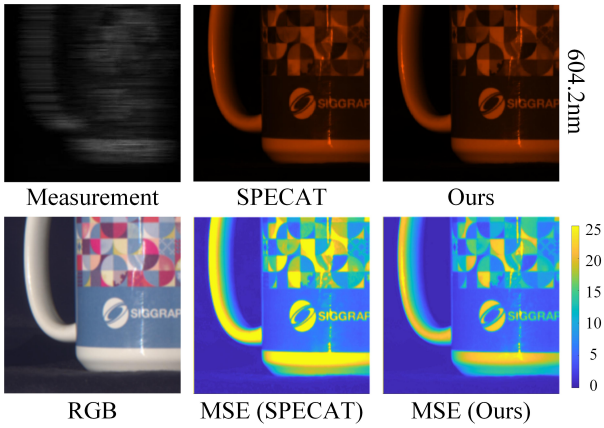


Figure 1: The visual comparisons between our method and ViT-based SOTA method on simulated dataset. The MSE residues between the reconstructed result and the ground truth in the bottom row demonstrate that Sp3ctralMamba has better spectral fidelity.

Due to the powerful ability to model nonlinear mapping functions, Deep Convolutional Neural Networks (CNNs) are applied in reconstructing 3D HSIs from 2D measurements. However, the CNN-based methods (Huang et al. 2021; Meng et al. 2021) show limitations in modeling long-range dependencies in the spatial dimension and spectral self-similarity. The Multi-Head Self-Attention (MSA) module in Vision Transformers (ViTs) excels at capturing non-local similarities and long-range dependencies between pixels (Liu et al. 2021), which addresses the limitations of

*Corresponding Author.

CNN-based methods in HSI reconstruction. Nevertheless, the computational complexity of ViTs is quadratic relative to the image's spatial size. This requires significantly more computational resources when handling HSI data, which has a much higher number of channels compared to RGB images. Although some researchers have attempted to make trade-offs between effective receptive fields and computational efficiency by constraining the size or stride of local windows (Dong et al. 2022; Khan et al. 2022), these approaches may overlook some highly relevant tokens, negatively impacting HSI reconstruction. Figure 1 shows that the proposed method in this paper outperforms the existing ViT-based SOTA method on spectral fidelity.

The selective scan mechanism and efficient hardware design of Vision Mamba (VMamba) make it an effective alternative to ViT. VMamba has demonstrated strong performance and efficiency in various visual tasks, such as image classification (Zhu et al. 2024), object detection (Chen et al. 2024), and image restoration (Guo et al. 2024; Shi et al. 2024). The selective scan mechanism in VMamba requires each element in the array to obtain contextual knowledge only through a compressed hidden state, thereby reducing the computational complexity from quadratic to linear. VMamba addresses the issue of directional dependencies in 2D image data by introducing Cross-Scan Modules (CSM) from different directions. This strategy ensures that each element in the feature map integrates information from all other positions across different directions, achieving a global receptive field without increasing computational complexity. However, there is still room for improvement in existing scanning schemes. For example, pixel-wise scanning increases the burden on computational resources, and the scanning scheme in other transformed domains of image data have yet to be explored.

Without sufficient guidance, the network may easily focus on low-fidelity and less informative image regions (Xia et al. 2024), hindering the reconstruction of spectral and spatial information. The spectral information in HSI exhibits spatial sparsity. Conventional hand-crafted priors, such as sparsity (Kittle et al. 2010; Lin et al. 2014), total variation (Wang et al. 2015; Yuan 2016), and non-local similarity (Liu et al. 2018; Zhang et al. 2019), require manual parameter tuning and typically exhibit poor generalization ability. Additionally, colourants have different sensitivities to texture at different wavelengths (Cheng et al. 2019). This affects the global pixel intensity of images captured at different bands. Each frequency domain signal of an image after Fourier transform corresponds to the global information of the spatial domain image. Therefore, frequency domain knowledge can be used to enhance the representation of global reflectance intensity across different bands.

To address the above issues, we propose a novel physics-driven joint state space model named Sp3ctralMamba for HSI reconstruction. Sp3ctralMamba integrates frequency domain knowledge and HSI priors to assist the Mamba block in learning fine-grained representations of spectral and spatial details. Specifically, we first perform a hierarchical split on the embeddings from the 2D measurement to avoid interference between distant bands. Secondly, we design a joint SSM block named S^3Mamba (S^3MAB), which integrates three scanning methods to jointly process each group of band embeddings both in the spatial and frequency domains. In addition to the conventional vanilla scanning scheme in VMamba, the S^3MAB also includes the Local SSM Block (LSSM), which uses a local scanning scheme to overcome the difficulties of spatial sparsity in HSI. We aim to control the global reflectance intensity of different band images from the frequency domain, so a Fourier SSM Block (FSSM) is specifically designed in the S^3MAB. The FSSM performs a spiral scan from low-frequency signals outward to high-frequency signals, in order to establish the order relationships between different frequency signals. Finally, three physical priors are used to constrain the generation of spectral and spatial representations. The physical mask in the CASSI system provides an imaging prior for the modulation process of HSI. The mask is used to modulate the network's reconstruction results to approximate the input 2D measurements. In addition to constraining the imaging process, we further introduce an energy prior to constrain the local energy variations in the encoded features and a structural prior to constrain the texture information in the decoded features. Experiments on different datasets demonstrate the effectiveness of Sp3ctralMamba.

In summary, the contributions of this work are as follows:

- We propose a novel physics-driven joint state space model named Sp3ctralMamba for HSI reconstruction. Sp3ctralMamba integrates frequency domain knowledge and HSI physical priors into the SSM, significantly enhancing the quality of HSI reconstruction.
- We design a joint SSM block named S^3Mamba (S^3MAB). S^3MAB integrates local and non-local scanning schemes in the spatial domain. Additionally, S^3MAB creatively employs a spiral scanning scheme in the frequency domain to establish order relationships between different frequency signals.
- We introduce three different physical priors to constrain the HSI reconstruction from the perspectives of the compressive imaging process and image content. Ablation experiments demonstrate the effectiveness of these constraints.
- Extensive experiments on both real and simulated data demonstrate the effectiveness and efficiency of Sp3ctralMamba. The reconstruction results from Sp3ctralMamba significantly surpass those of existing SOTA methods.

Methodology

Figure 2 shows the overall architecture of Sp3ctralMamba, which mainly consists of two parts, the joint state space model S^3Mamba and the Physical Prior Constraint Module (PPCM). Figure 3 shows the design of three different SSM blocks in S^3Mamba. The details will be illustrated below.

Joint State Space Model S^3Mamba (S^3MAB)

S^3MAB aims to learn the local structure and global reflectance intensity of each band by integrating information

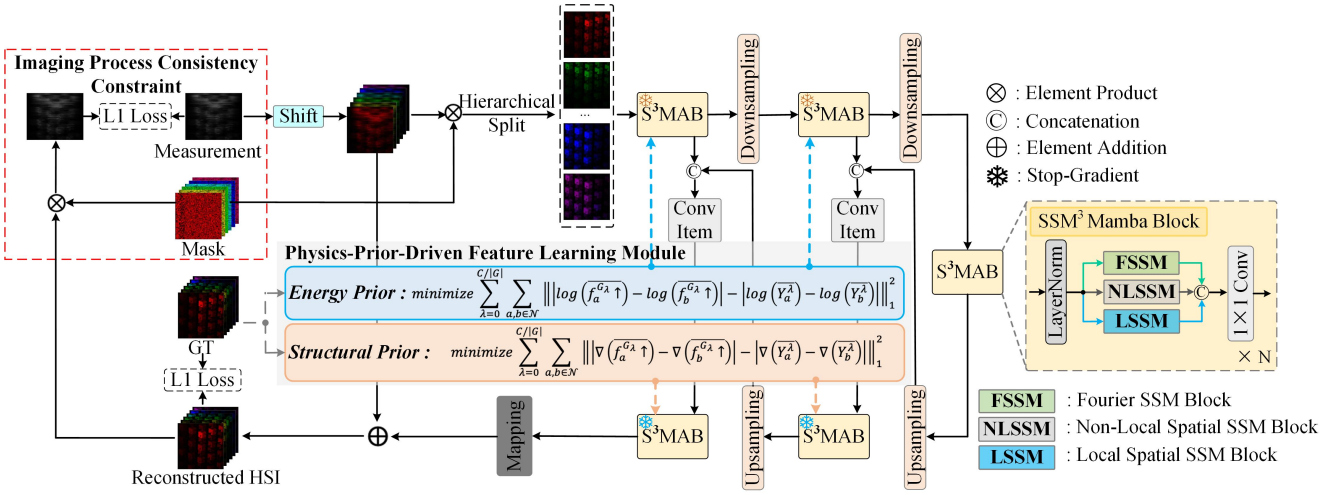
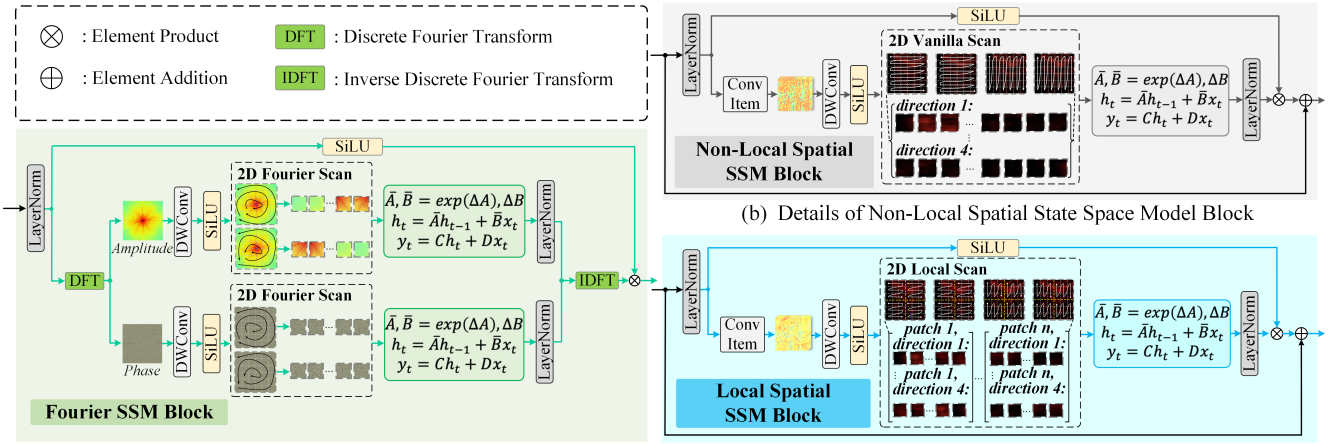


Figure 2: The overall framework of Sp3ctralMamba. It consists of two key components: the joint state space model S^{3}Mamba (S^{3}MAB) and the physics priors constraint module, which includes Physics-Prior-Driven Feature Learning Module (PFLM) and Imaging Process Consistency Constraint (IPCC).



(a) Details of Fourier State Space Model Block

(b) Details of Non-Local Spatial State Space Model Block

(c) Details of Local Spatial State Space Model Block

Figure 3: Architectures of the proposed S^{3}MAB. The sub-figure (a), (b) and (c) depict the details of FSSM, NLSSM and LSSM respectively.

from both the spatial and frequency domains. S^{3}MAB consists of three components: 1) Non-Local Spatial State Space Model Block (NLSSM), 2) Local Spatial State Space Model Block (LSSM), and 3) Fourier State Space Model Block (FSSM). To avoid interference from distant bands, we first perform a hierarchical split on the embedding from the 2D measurement $embed$:

$$\left[feat_0^{G_{\lambda_1}}, feat_0^{G_{\lambda_2}}, \cdots, feat_0^{G_{\lambda_n}}\right] = HS\left(embed\right), \quad (1)$$

where HS $(\cdot)$ represents hierarchical split operation, G_{λ_k} represents a group of features corresponding to the wavelength λ_k, $n = C/|G|$, and C is the number of channels of the HSI $I \in R^{H\times W\times C}$ to be reconstructed.

Non-Local Spatial State Space Model Block (NLSSM). SSMs utilize the framework of linear ordinary differential equations (ODEs) to map the input stimulation $x(t) \in \mathbb{R}^L$ to the output responses $y(t) \in \mathbb{R}^L$ through a hidden state $h(t) \in \mathbb{C}^N$:

$$\begin{aligned} h'(t) &= \mathbf{A}h(t) + \mathbf{B}x(t), \\ y(t) &= \mathbf{C}h(t) + \mathbf{D}x(t), \end{aligned} \quad (2)$$

where N represents the size of the state and $\mathbf{A} \in \mathbb{C}^{N\times N}$, $\mathbf{B}, \mathbf{C} \in \mathbb{C}^N$, and $\mathbf{D} \in \mathbb{C}^1$ are the weighting parameters. Subsequently, Eq.(2) are usually merged using a discretization process:

$$\begin{aligned} \overline{\mathbf{A}} &= \exp(\Delta\mathbf{A}), \\ \overline{\mathbf{B}} &= (\Delta\mathbf{A})^{-1}(\exp(\Delta\mathbf{A}) - I)\cdot\Delta\mathbf{B}, \end{aligned} \quad (3)$$

where Δ is a time scale parameter used to convert the continuous parameters $\mathbf{A}$ and $\mathbf{B}$ into the discrete parameters $\overline{\mathbf{A}}$

and $\overline{\mathbf{B}}$. After discretization, the Eq.(2) can be rewritten as:

$$\begin{aligned} h_t &= \overline{\mathbf{A}} h_{k-1} + \overline{\mathbf{B}} x_k, \\ y_t &= \mathbf{C} h_k + \mathbf{D} x_k, \end{aligned} \tag{4}$$

In NLSSM, as shown in Figure 3 (b), we use the conventional vanilla scanning scheme (Zhu et al. 2024) to obtain the states from four directions and perform cross-merging:

$$nl - feat_i^{G_{\lambda_k}} = \mathrm{CS}\left(feat_{i-1}^{G_{\lambda_k}}\right) \odot \mathrm{SiLU}\left(feat_{i-1}^{G_{\lambda_k}}\right). \tag{5}$$

where $\mathrm{CS}(\cdot)$ represents the cross-scan operation which employs the following operation sequence: $DWConv \rightarrow SiLU \rightarrow SSM \rightarrow LN$. The subscript i denotes the intermediate stage of the encoding and decoding process and $\odot$ is the Hadamard product.

Local Spatial State Space Model Block (LSSM). In addition to NLSSM, we attempt to divide the features into patches to address the spatial sparsity of the spectral information. Different endmembers exhibit varying reflectance at different spectral wavelengths. While the non-local scanning scheme helps build global structural associations, blindly integrating all pixels can interfere with the original reflectance of objects. In LSSM, we perform local scanning on the features:

$$\begin{aligned} feat_{i-1}^{G_{\lambda_k}^j} &= \mathrm{TK}\left(feat_{i-1}^{G_{\lambda_k}}\right), j = 0, 1, \cdots, m \\ l - feat_i^{G_{\lambda_k}^j} &= \mathrm{CS}\left(feat_{i-1}^{G_{\lambda_k}^j}\right) \odot \mathrm{SiLU}\left(feat_{i-1}^{G_{\lambda_k}^j}\right). \end{aligned} \tag{6}$$

where $\mathrm{TK}(\cdot)$ represents the tokenization operation and j is the index of the j-th patch. We perform scanning in four directions and cross-merge for each patch. LSSM enhances the spectral associations within local regions of the HSI.

Fourier State Space Model Block (FSSM). We aim to enhance the learning of reflectance intensity and structure across the entire image by combining frequency domain information. The features $x \in R^{H \times W \times C}$ are first transformed to the frequency domain using discrete Fourier transform (DFT):

$$\mathcal{F}(x)(u,v) = \frac{1}{\sqrt{HW}} \sum_{h=0}^{H-1} \sum_{w=0}^{W-1} x(h,w)\, e^{-j2\pi\left(\frac{h}{H}u + \frac{w}{W}v\right)}, \tag{7}$$

The amplitude component $\mathcal{A}(x)(u,v)$ and the phase component $\mathcal{P}(x)(u,v)$ are expressed as:

$$\mathcal{A}(x)(u,v) = \sqrt{R^2(x)(u,v) + I^2(x)(u,v)}, \tag{8}$$

$$\mathcal{P}(x)(u,v) = \arctan\left[\frac{I(x)(u,v)}{R(x)(u,v)}\right], \tag{9}$$

where $R(x)$ and $I(x)$ represent the real and imaginary part of $\mathcal{F}(x)$ respectively. Unlike spatial scanning, a frequency domain scanning method is used for FSSM:

$$\begin{aligned} \mathcal{A}'(x) &= \mathrm{FCS}(\mathcal{A}(x)), \\ \mathcal{P}'(x) &= \mathrm{FCS}(\mathcal{P}(x)), \end{aligned} \tag{10}$$

where $\mathrm{FCS}(\cdot)$ represents the scanning operation in fourier domain which employs the following operation sequence: $DWConv \rightarrow SiLU \rightarrow SSM \rightarrow LN$. As shown in Figure 3 (a), the SSM in FSSM adopts a spiral scanning scheme, which scans from high frequency to low frequency and vice versa. Unlike convolution operations, this method can establish order associations between signals of different frequencies. Here, x is the input feature. Then, the cross-merged frequency domain features $\mathcal{A}'(x)$ and $\mathcal{P}'(x)$ are transformed back to the spatial domain via inverse discrete Fourier transform (IDFT):

$$feat_{fre} = \mathcal{F}^{-1}(\mathcal{A}'(x), \mathcal{P}'(x)) \odot \mathrm{SiLU}(x). \tag{11}$$

where $\mathcal{F}^{-1}(\cdot)$ is the IDFT operation. FSSM enhances the global reflectance intensity and structural representation of each channel in the HSI.

Physics Priors Constraint Module (PPCM)

The implicit attention mechanism in Mamba models gives rise to three orders of magnitude more attention matrices than transformers (Ali, Zimerman, and Wolf 2024). To ensure the stability of SSM parameter learning, we introduce three physical priors for HSI reconstruction.

Imaging Process Consistency Constraint (IPCC). Inspired by MST (Cai et al. 2022b) , we attempt to explore the modulation effect of the mask $\mathcal{M}$ in the CASSI system. In contrast to MST, we ensure the rationality of the entire reconstruction process by modulating the reconstruction result of Sp3ctralMamba to generate a 2D measurement consistent with the network input:

$$\mathcal{L}_{\mathcal{M}} = \left\| \mathcal{M} \odot \widehat{Y} - X \right\|_1^2. \tag{12}$$

where $\widehat{Y}$ is the reconstructed HSI and X is the input 2D measurement.

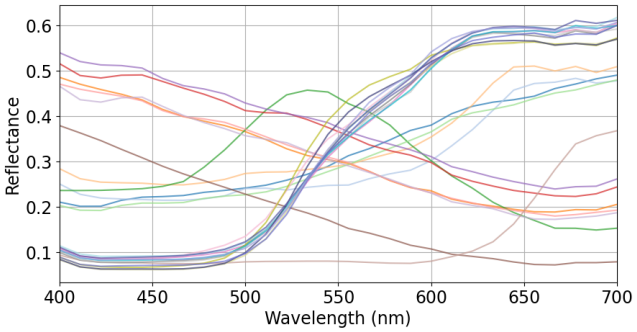


Figure 4: Reflectance versus wavelength curves for different colourants.

Physics-Prior-Driven Feature Learning Module (PFLM). Research (Salamati, Fredembach, and Süsstrunk 2009) showed that near-infrared (NIR) image is transparent to a range of colourants and dyes. The observation results of the NIR image can be extended to HSI data. To further

demonstrate the differences in reflectance variations, multi-spectral reflectance values are sampled over 20 colourants on a colour checker, as illustrated in Figure 4. Note that each colorant's spectral curves exhibit significant fluctuations at different wavelength ranges. It can be observed that some colorants' spectral curves are aggregated, while others exhibit significant differences. Based on the above analysis, we attempt to impose constraints on the reconstruction content with respect to energy and structure priors.

Image Energy Prior Constraint (EPC). Colour Retinex (Grosse et al. 2009) uses the following constraints on the chromaticity and intensity of color images:

$$R(x,y) = \log(I(x,y)) - \log(F(x,y) * I(x,y)), \quad (13)$$

where $R(x,y)$ is the Retinex output at pixel (x,y), $I(x,y)$ is the input image intensity at pixel, and $F(x,y)$ is the surround function that represents the spatial distribution of light around pixel (x,y). We extend Eq.(13) to HSI data and apply the following constraint to the encoder features:

$$\mathcal{L}_E = \sum_{\lambda=0}^{C/|G|} \sum_{a,b\in\mathcal{N}} \left\| \left| log\left(\overline{f_a^{G_\lambda}}\uparrow\right) - log\left(\overline{f_b^{G_\lambda}}\uparrow\right) \right| - \left| log\left(\overline{Y_a^{\lambda}}\right) - log\left(\overline{Y_b^{\lambda}}\right) \right| \right\|_1^2. \quad (14)$$

where $\mathcal{N}$ stands for the set of pixels in a local area, a and b denote the pixels in that area. $\overline{f_a^{G_\lambda}}$ represents the mean value of the feature for each group related to band λ at pixel a, and Y is the pixel value of the corresponding ground truth of the HSI. $\uparrow$ is upsampling operation. When implementing EPC, we stop the gradient propagation of the decoder (as shown in Figure 2).

Image Structural Prior Constraint (SPC). To explore the reasons for the loss of structural information, we further visualize the frequency distribution of features in the encoder and decoder. Figure 5 shows that there is a significant loss of high-frequency signals in the decoder features, which leads to over-smooth reconstruction results. Therefore, we introduce a structural prior to constrain the learning of high-frequency signals in the decoder features:

$$\mathcal{L}_S = \sum_{\lambda=0}^{C/|G|} \sum_{a,b\in\mathcal{N}} \left\| \left| \nabla\left(\overline{f_a^{G_\lambda}}\uparrow\right) - \nabla\left(\overline{f_b^{G_\lambda}}\uparrow\right) \right| - \left| \nabla\left(\overline{Y_a^{\lambda}}\right) - \nabla\left(\overline{Y_b^{\lambda}}\right) \right| \right\|_1^2. \quad (15)$$

where $\nabla(\cdot)$ is gradient operation. In contrast to EPC, when implementing SPC, we enhance the representation of edge in the decoder by stopping the gradient propagation of the encoder.

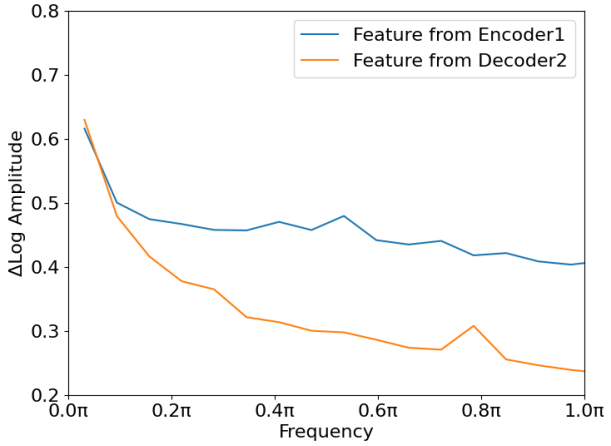


Figure 5: Relative log amplitudes of Fourier in feature maps from Encoder 1 and Decoder 2.

Loss Function

In this paper, we use L1 loss to optimize the reconstructed HSI at the pixel level:

$$\mathcal{L}_1 = \frac{1}{H \times W \times C} \sum_{i=0} \left| \widehat{Y}(i) - Y(i) \right|, \quad (16)$$

where $\widehat{Y}(i)$ represents the predicted value for pixel i, and $Y(i)$ is its corresponding ground truth. In addition to $\mathcal{L}_1$, we also integrate $\mathcal{L}_\mathcal{M}$ throughout the entire training process:

$$\mathcal{L}_{rec} = \mathcal{L}_1 + \mathcal{L}_\mathcal{M}. \quad (17)$$

At different training stages, we introduce $\mathcal{L}_E$ and $\mathcal{L}_S$ to guide feature generation. The details will be described in the Experiments section.

Experiments

Baseline Methods

To demonstrate the effectiveness of Sp3ctralMamba, we compared the performance of our method with several state-of-the-art (SOTA) methods, including TSA-Net (Meng, Ma, and Yuan 2020), GAP-net (Meng, Jalali, and Yuan 2020), MST-L (Cai et al. 2022b), MST++ (Cai et al. 2022c), CST-L (Cai et al. 2022a), DAUHST-9stg (Cai et al. 2022d), PADUT-12stg (Li et al. 2023), SST-LPlus (Cai et al. 2023), and SPECAT (Yao et al. 2024).

Implementation Details

We implemented Sp3ctralMamba on a PC with a single NVIDIA RTX 4090 GPU, and we built our network in the PyTorch framework, training it with the Adam (Kingma and Ba 2014) optimizer for 300 epochs. The learning rate was set to 4×10^{-4} and the batch size was set to 4.

In the initial 200 epochs, we used the reconstruction loss $\mathcal{L}_{rec}$ to optimize the predicted results. In the next 50 epochs, we stopped updating the decoder's gradients and introduced the energy prior $\mathcal{L}_E$ to enhance the encoder's representation of overall pixel intensity. In the final 50 epochs, we did the opposite and introduced the structure prior $\mathcal{L}_S$ to enhance the decoder's representation of edge details.

Methods	Params.	GFLOPs.	Scene1	Scene2	Scene3	Scene4	Scene5	Scene6	Scene7	Scene8	Scene9	Scene10	Avg
TSA-Net	44.2M	135.1	32.30 0.936	31.26 0.906	28.53 0.875	36.36 0.931	30.37 0.937	33.06 0.950	31.04 0.981	30.88 0.924	28.99 0.872	32.62 0.947	31.54 0.917
GAP-net	4.26M	84.5	34.30 0.942	31.59 0.893	28.48 0.831	36.62 0.909	32.74 0.923	34.30 0.921	31.76 0.877	31.52 0.906	30.01 0.867	33.45 0.948	32.48 0.901
MST-L	2.03M	28.5	36.55 0.963	36.29 0.953	33.46 0.904	39.78 0.951	35.40 0.964	36.10 0.963	34.53 0.924	33.08 0.945	34.38 0.926	35.19 0.966	35.48 0.946
MST++	1.34M	19.6	36.65 0.960	37.14 0.963	34.84 0.936	38.94 0.959	36.44 0.968	36.96 0.969	35.37 0.940	34.27 0.953	34.58 0.941	35.08 0.974	36.03 0.956
CST-L	3.00M	40.1	36.98 0.963	38.34 0.963	35.89 0.939	40.98 0.951	36.19 0.967	37.23 0.957	37.75 0.944	34.64 0.946	36.41 0.946	35.93 0.961	36.83 0.954
DAUHST-9stg	6.15M	79.5	38.37 0.972	39.91 0.977	37.71 0.965	42.97 0.967	37.69 0.980	39.05 0.974	37.62 0.964	36.11 0.965	38.45 0.971	37.39 0.976	38.53 0.971
PADUT-12stg	5.38M	90.5	38.42 0.974	40.34 0.983	38.95 0.972	43.50 0.977	38.22 0.983	39.16 0.979	38.21 0.971	36.03 0.969	39.45 0.979	37.30 0.981	38.96 0.977
SST-LPlus	9.71M	162.1	39.49 0.977	40.64 0.978	39.92 0.970	42.79 0.976	38.78 0.980	39.34 0.976	38.21 0.968	36.53 0.966	39.47 0.971	36.17 0.970	39.13 0.973
SPECAT	**0.29M**	**12.4**	40.24 0.982	42.40 0.986	41.43 0.978	44.90 0.982	39.62 0.987	39.90 0.984	39.41 0.977	37.49 0.977	40.45 0.982	37.90 0.983	40.37 0.986
Ours	0.45M	64.65	**40.66** **0.989**	**43.22** **0.992**	**42.17** **0.988**	**45.64** **0.990**	**40.75** **0.993**	**41.70** **0.991**	**39.88** **0.986**	**37.94** **0.988**	**41.43** **0.986**	**38.71** **0.989**	**41.21** **0.989**

Table 1: Comparison of Parameters, GFLOPs, PSNR and SSIM (upper and lower entry in each cell, respectively) of different methods on 10 simulation scenes for optical filter-based HSI system. The best values are bolded.

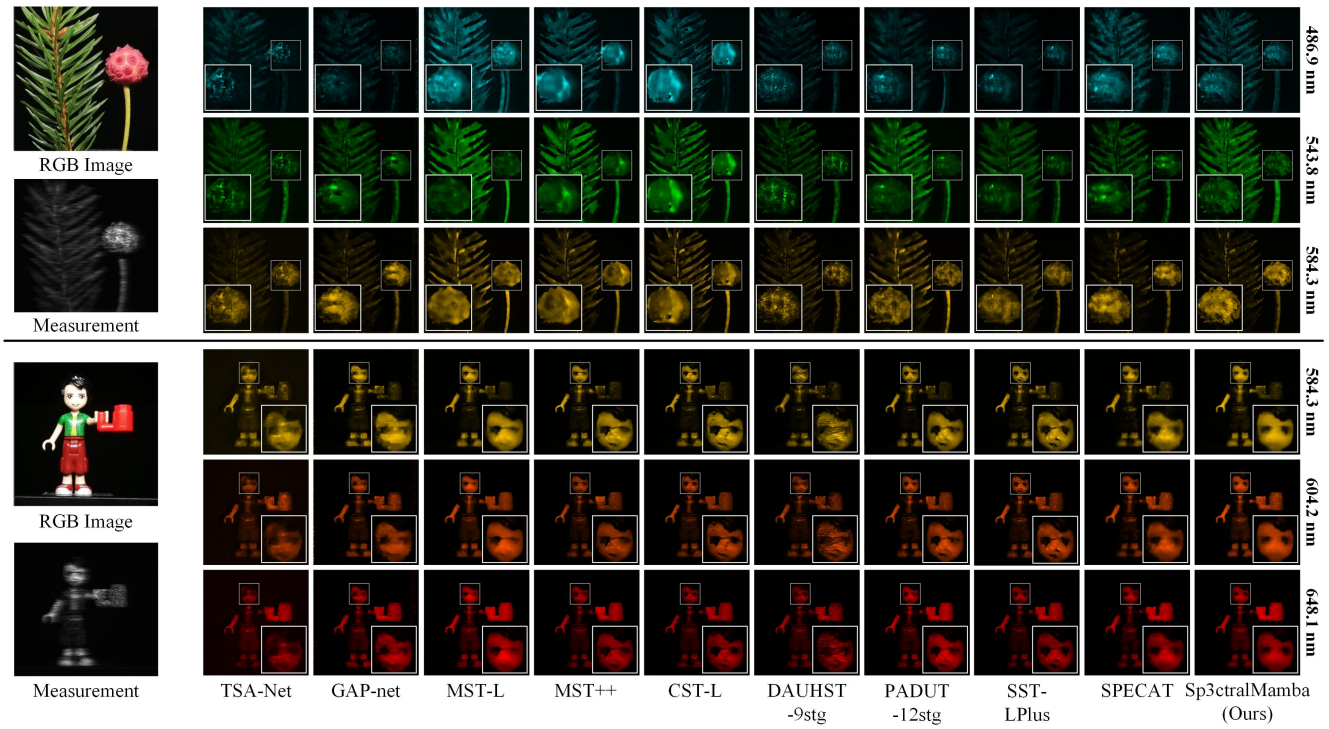


Figure 6: Comparison of reconstruction results on two scenes in the real dataset.

Datasets

We conducted experiments on 28 spectral channels of both simulated and real HSI datasets, with a wavelength range of 450 nm to 650 nm. For simulated data, we used two widely-used hyperspectral datasets: CAVE (Park et al. 2007) and KAIST (Choi et al. 2017). The CAVE dataset consists of 32 hyperspectral images with a spatial resolution of 512×512 pixels. The KAIST dataset includes 30 hyperspectral images with a spatial resolution of 2704×3376 pixels. Consistent with the settings in TSA-Net (Meng, Ma, and Yuan 2020),

we used the CAVE dataset as the training set and selected 10 scenes from KAIST as the testing set. The patch size during training is 256×256. For real data, we used 5 real hyperspectral images obtained from the CASSI system developed in TSA-Net as the testing set.

Metrics

The reconstruction quality of HSI on the simulated dataset was evaluated using peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM). Since the real dataset did not contain ground truth, we used RGB images as a reference for visual comparison.

Performance Comparison

Simulation Results. Table 1 shows the results on 10 simulated scenes on KAIST dataset. It can be observed that our method achieved the best performance across all scenes, proving the effectiveness of Sp3ctralMamba. The average PSNR and average SSIM of our method achieved 41.21 dB and 0.989, outperforming the second-best results by 0.8 dB and 0.003, respectively. Additionally, it can be observed that Sp3ctralMamba achieved optimal reconstruction results without requiring excessive parameters and GFLOPs, indicating that we have made a good tradeoff between performance and computational resources.

Results of CASSI System. To evaluate the effectiveness of our method on real dataset, we further conducted tests on five actual 2D measurements captured by the CASSI system. Figure 6 shows the reconstruction results of all methods on two real scenes. As shown in Figure 6, most methods fail to reconstruct fine structures, but in contrast, Sp3ctralMamba demonstrated a strong advantage. In the first scene, for the spikes on the surface of the plant on the right, our method is able to better preserve the texture details. Similarly, for the face in the second scene, our method also better eliminates the structural shift caused by the compression process of the CASSI system (the black horizontal line between the two eyes).

Ablation Studies

The joint state space model S^3Mamba (S^3MAB), and Physics Priors Constraint Module (PPCM) are two key modules of our network, we conducted a series of ablation experiments on the simulated dataset to demonstrate their effectiveness and necessity.

S^3Mamba (S^3MAB). As shown in Table 2, we explored the impact of different scanning schemes by conducting ablation studies on S^3MAB. For experiment (I), we replaced LSSM and FSSM with NLSSM and found that a single vanilla scan caused a significant decrease in all metrics. For experiments (II) and (III), we sequentially replaced FSSM and LSSM with NLSSM, and compared to the results of experiment (I), the PSNR increased by 1.3dB and 1.7dB, respectively. This is because NLSSM better facilitates the modeling of local spatial correlations, and FSSM enhances the representation of global image reflectance intensity and spatial details by establishing order correlations in frequency domain signals. The introduction of EPC helps improve the PSNR by 0.7dB.

Config	LSSM	FSSM	PSNR↑	SSIM↑
(I)	✗	✗	38.51	0.963
(II)	✓	✗	39.89	0.977
(III)	✗	✓	40.24	0.982
Ours	✓	✓	**41.21**	**0.989**

Table 2: Ablation studies about two scan schemes in S^3MAB. The best values are bolded.

Config	IPCC	EPC	SPC	PSNR↑	SSIM↑
(I)	✗	✗	✗	39.42	0.974
(II)	✓	✗	✗	39.90	0.977
(III)	✗	✓	✗	40.21	0.984
(IV)	✗	✗	✓	40.58	0.986
Ours	✓	✓	✓	**41.21**	**0.989**

Table 3: Ablation studies about three priors on simulated dataset. The best values are bolded.

Physics Priors Constraint Module (PPCM). Table 3 shows the impact on the network of three physics priors. For experiment (I), we did not apply any physical constraints and directly trained Sp3ctraMamba end-to-end for 300 epochs. It is observed that, due to the powerful spatial modeling capability of the Mamba block, the network nearly achieves SOTA reconstruction. For experiment (II), we introduced a mask modulation mechanism to constrain the entire reconstruction process, resulting in a 0.4dB increase in PSNR. For experiment (III), we separately introduced EPC to constrain the pixel intensity of encoded features during the last 100 epochs of training, aiming to prevent excessive information loss in the encoding process. For experiment (IV), we separately introduced SPC to enhance the high-frequency signals in the decoded features during the last 100 epochs of training. This helps increase the network's PSNR by 1.1dB. It is observed that SPC provided more gain to the reconstruction results compared to EPC, indicating that the loss of structural information is a key factor affecting the performance of HSI reconstruction.

Conclusion

In this paper, we propose a novel joint SSM network named Sp3ctralMamba for HSI reconstruction. The implicit attention mechanism in the Mamba block is used to raise the performance ceiling for HSI reconstruction. A novel spiral scanning scheme in the frequency domain is designed in Sp3ctralMamba to establish order correlations between different signals. Furthermore, Sp3ctralMamba constrains the validity of the reconstruction process and enhances the network's representation of spectral and spatial details by introducing three different physical priors. Experimental results reveal that Sp3ctralMamba surpasses the performance of the SOTA HSI reconstruction methods.

Acknowledgments

The work was supported in part by the National Natural Science Foundation of China under Grant 82172033, U19B2031, 61971369, 52105126, 82272071, 62271430, and the Fundamental Research Funds for the Central Universities 20720230104.

References

Ali, A.; Zimerman, I.; and Wolf, L. 2024. The hidden attention of mamba models. *arXiv preprint arXiv:2403.01590*.

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.; Lin, Z.; Wang, H.; Zhang, Y.; Pfister, H.; Timofte, R.; and Van Gool, L. 2022c. Mst++: Multi-stage spectral-wise transformer for efficient spectral reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 745–755.

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

Cai, Z.; Yu, J.; Zhang, Z.; Jin, C.; and Da, F. 2023. SST-ReversibleNet: Reversible-prior-based Spectral-Spatial Transformer for Efficient Hyperspectral Image Reconstruction. *arXiv preprint arXiv:2305.04054*.

Chen, T.; Tan, Z.; Gong, T.; Chu, Q.; Wu, Y.; Liu, B.; Ye, J.; and Yu, N. 2024. Mim-istd: Mamba-in-mamba for efficient infrared small target detection. *arXiv preprint arXiv:2403.02148*.

Cheng, Z.; Zheng, Y.; You, S.; and Sato, I. 2019. Non-local intrinsic decomposition with near-infrared priors. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2521–2530.

Choi, I.; Kim, M.; Gutierrez, D.; Jeon, D.; and Nam, G. 2017. High-quality hyperspectral reconstruction using a spectral prior. Technical report.

Dong, X.; Bao, J.; Chen, D.; Zhang, W.; Yu, N.; Yuan, L.; Chen, D.; and Guo, B. 2022. Cswin transformer: A general vision transformer backbone with cross-shaped windows. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12124–12134.

Grosse, R.; Johnson, M. K.; Adelson, E. H.; and Freeman, W. T. 2009. Ground truth dataset and baseline evaluations for intrinsic image algorithms. In *2009 IEEE 12th International Conference on Computer Vision*, 2335–2342. IEEE.

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

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.

Ishida, T.; Kurihara, J.; Viray, F. A.; Namuco, S. B.; Paringit, E. C.; Perez, G. J.; Takahashi, Y.; and Marciano Jr, J. J. 2018. A novel approach for vegetation classification using UAV-based hyperspectral imaging. *Computers and electronics in agriculture*, 144: 80–85.

Khan, S.; Naseer, M.; Hayat, M.; Zamir, S. W.; Khan, F. S.; and Shah, M. 2022. Transformers in vision: A survey. *ACM computing surveys (CSUR)*, 54(10s): 1–41.

Kim, M. H.; Harvey, T. A.; Kittle, D. S.; Rushmeier, H.; Dorsey, J.; Prum, R. O.; and Brady, D. J. 2012. 3D imaging spectroscopy for measuring hyperspectral patterns on solid objects. *ACM Transactions on Graphics (TOG)*, 31(4): 1–11.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kittle, D.; Choi, K.; Wagadarikar, A.; and Brady, D. J. 2010. Multiframe image estimation for coded aperture snapshot spectral imagers. *Applied optics*, 49(36): 6824–6833.

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.

Lin, X.; Liu, Y.; Wu, J.; and Dai, Q. 2014. Spatial-spectral encoded compressive hyperspectral imaging. *ACM Transactions on Graphics (TOG)*, 33(6): 1–11.

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.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.

Lu, G.; and Fei, B. 2014. Medical hyperspectral imaging: a review. *Journal of biomedical optics*, 19(1): 010901–010901.

Meng, Z.; Jalali, S.; and Yuan, X. 2020. Gap-net for snapshot compressive imaging. *arXiv preprint arXiv:2012.08364*.

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. Springer.

Meng, Z.; Qiao, M.; Ma, J.; Yu, Z.; Xu, K.; and Yuan, X. 2020. Snapshot multispectral endomicroscopy. *Optics Letters*, 45(14): 3897–3900.

Meng, Z.; Yu, Z.; Xu, K.; and Yuan, X. 2021. Self-supervised neural networks for spectral snapshot compressive imaging. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2622–2631.

Pan, Z.; Healey, G.; Prasad, M.; and Tromberg, B. 2003. Face recognition in hyperspectral images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(12): 1552–1560.

Park, J.-I.; Lee, M.-H.; Grossberg, M. D.; and Nayar, S. K. 2007. Multispectral imaging using multiplexed illumination. In *2007 IEEE 11th International Conference on Computer Vision*, 1–8. IEEE.

Salamati, N.; Fredembach, C.; and Süsstrunk, S. 2009. Material classification using color and NIR images. In *Proc. IS&T/SID 17th Color Imaging Conference (CIC)*.

Shi, Y.; Xia, B.; Jin, X.; Wang, X.; Zhao, T.; Xia, X.; Xiao, X.; and Yang, W. 2024. Vmambair: Visual state space model for image restoration. *arXiv preprint arXiv:2403.11423*.

Wang, L.; Xiong, Z.; Gao, D.; Shi, G.; and Wu, F. 2015. Dual-camera design for coded aperture snapshot spectral imaging. *Applied optics*, 54(4): 848–858.

Wright, S. L.; Levermore, J. M.; and Kelly, F. J. 2019. Raman spectral imaging for the detection of inhalable microplastics in ambient particulate matter samples. *Environmental science & technology*, 53(15): 8947–8956.

Xia, J.; Yang, Z.; Li, S.; Zhang, S.; Fu, Y.; Gündüz, D.; and Li, X. 2024. Blind Super-Resolution Via Meta-Learning and Markov Chain Monte Carlo Simulation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Yao, Z.; Liu, S.; Yuan, X.; and Fang, L. 2024. SPECAT: SPatial-spEctral Cumulative-Attention Transformer for High-Resolution Hyperspectral Image Reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 25368–25377.

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

Yuan, X.; Brady, D. J.; and Katsaggelos, A. K. 2021. Snapshot compressive imaging: Theory, algorithms, and applications. *IEEE Signal Processing Magazine*, 38(2): 65–88.

Zhang, S.; Wang, L.; Fu, Y.; Zhong, X.; and Huang, H. 2019. Computational hyperspectral imaging based on dimension-discriminative low-rank tensor recovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10183–10192.

Zhu, L.; Liao, B.; Zhang, Q.; Wang, X.; Liu, W.; and Wang, X. 2024. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*.