Mask-guided Spectral-wise Transformer for Efficient Hyperspectral Image Reconstruction

Yuanhao Cai 1,2,*, Jing Lin 1,2,*, Xiaowan Hu 1,2, Haoqian Wang 1,2,†, Xin Yuan 3, Yulun Zhang 4, Radu Timofte 4, and Luc Van Gool 4
1 Shenzhen International Graduate School, Tsinghua University, 2 Shenzhen Institute of Future Media Technology, 3 Westlake University, 4 ETH Zürich

Abstract

Hyperspectral image (HSI) reconstruction aims to recover the 3D spatial-spectral signal from a 2D measurement in the coded aperture snapshot spectral imaging (CASSI) system. The HSI representations are highly similar and correlated across the spectral dimension. Modeling the inter-spectra interactions is beneficial for HSI reconstruction. However, existing CNN-based methods show limitations in capturing spectral-wise similarity and long-range dependencies. Besides, the HSI information is modulated by a coded aperture (physical mask) in CASSI. Nonetheless, current algorithms have not fully explored the guidance effect of the mask for HSI restoration. In this paper, we propose a novel framework, Mask-guided Spectral-wise Transformer (MST), for HSI reconstruction. Specifically, we present a Spectral-wise Multi-head Self-Attention (S-MSA) that treats each spectral feature as a token and calculates self-attention along the spectral dimension. In addition, we customize a Mask-guided Mechanism (MM) that directs S-MSA to pay attention to spatial regions with high-fidelity spectral representations. Extensive experiments show that our MST significantly outperforms state-of-the-art (SOTA) methods on simulation and real HSI datasets while requiring dramatically cheaper computational and memory costs. https://github.com/caiyuanhao1998/MST/

1. Introduction

Hyperspectral imaging refers to multi-channel imaging where each channel captures the information at a specific spectral wavelength for a real-world scene. Generally, hyperspectral images (HSIs) have more spectral bands than normal RGB images to store richer information and delineate more detailed characteristics of the imaged scene. Relying on this property, HSIs have been widely applied to many computer vision related tasks, e.g., remote sensing [5,34,59], object tracking [21,46], medical image processing [3,33,37], etc. To collect HSIs, conventional imaging systems with spectrometers scan the scenes along the spatial or spectral dimension, usually requiring a long time. Therefore, these traditional imaging systems are unsuitable for capturing and measuring dynamic scenes. Recently, researchers have used snapshot compressive imaging (SCI) systems to capture HSIs. These SCI systems compress information of snapshots along the spectral dimension into one single 2D measurement [58]. Among current existing SCI systems [10,16,32,45,47], the coded aperture snapshot spectral imaging (CASSI) [36,45] stands out and forms one promising mainstream research direction.

Based on CASSI, a large number of reconstruction algorithms have been proposed to recover the 3D HSI cube from the 2D measurement. Conventional model-based methods adopt hand-crafted priors such as sparsity [23,29,46], total variation [24,51,57], and non-local similarity [30,52,61] to regularize the reconstruction procedure. However, these
methods need to tweak parameters manually, resulting in poor generalization ability, unsatisfactory reconstruction quality, and slow restoration speed. With the development of deep learning, HSI reconstruction has witnessed significant progress. Deep convolutional neural network (CNN) applies a powerful model to learn the end-to-end mapping function from the 2D measurement to the 3D HSI cube. Although impressive results have been achieved, CNN-based methods \[20, 35, 36, 39\] show limitations in modeling the inter-spectra similarity and long-range dependencies. Besides, the HSIs are modulated by a physical mask in CASSI. Nonetheless, previous CNN-based methods \[35, 36, 38, 48\] mainly adopt the inner product between the mask and the inter-spectra similarity and long-range dependencies. Hence, we propose a Spectral-wise MSA (S-MSA) to capture the long-range inter-spectra dependencies. Specifically, S-MSA treats each spectral channel feature as a token and calculates the self-attention along the spectral dimension. Secondly, in Fig. 2 (b), a mask is used in the CASSI system to modulate HSIs. The light transmittance of different positions on the mask varies significantly. This indicates that the fidelity of the modulated spectral information is position-sensitive. Therefore, we exploit the mask as a key clue and present a novel Mask-guided Mechanism (MM) that directs the S-MSA module to pay attention to the regions with high-fidelity spectral representations. Meanwhile, MM also alleviates the limitation of S-MSA in modeling the spatial correlations of HSI representations. Finally, with our proposed techniques, we establish a series of extremely efficient MST models that surpass state-of-the-art (SOTA) methods by a large margin, as illustrated in Fig. 1.

Our contributions can be summarized as follows:

- We propose a new method, MST, for HSI reconstruction. To the best of our knowledge, it is the first attempt to explore the potential of Transformer in this task.
- We present a novel self-attention, S-MSA, to capture the inter-spectra similarity and dependencies of HSIs.
- We customize an MM that directs S-MSA to pay attention to regions with high-fidelity HSI representations.
- Our MST dramatically outperforms SOTA methods on all scenes in simulation while requiring much cheaper Params and FLOPS. Besides, MST yields more visually pleasant results in real-world HSI reconstruction.

2. Related Work

2.1. HSI Reconstruction

Traditional HSI reconstruction methods \[18, 23, 29, 30, 43, 46, 52, 57, 61\] are mainly based on hand-crafted priors. For example, GAP-TV \[57\] introduces the total variation prior. DeSCI \[30\] exploits the low-rank property and non-local self-similarity. However, these model-based methods achieve unsatisfactory performance and generality due to the poor representing capacity. Recently, deep CNNs have been applied to learn the end-to-end mapping function of HSI reconstruction \[19, 20, 36, 39, 49\] to achieve promising performance. TSA-Net \[36\] uses three spatial-spectral self-attention modules to capture the dependencies in compressed spatial or spectral dimensions. The additional costs are nontrivial while the improvement is limited. DGSMP \[20\] suggests an interpretable HSI restoration method with learned Gaussian Scale Mixture (GSM) prior. These CNN-based methods yield impressive performance but show limitations in modeling inter-spectra similarity and correlations. Besides, the guidance effect of the mask is understudied.

2.2. Vision Transformer

Transformer is firstly proposed by \[44\] for machine translation. Recently, Transformer has achieved great suc-
cess in many high-level vision tasks, such as image classification [2, 15, 17, 31], object detection [1, 13, 60, 64], segmentation [8, 55, 63], human pose estimation [7, 25, 26, 56], etc. Due to its promising performance, Transformer has also been introduced into low-level vision [6, 9, 11, 14, 27, 28, 54]. SwinIR [27] uses Swin Transformer [27] blocks to build up a residual network and achieve SOTA results in image restoration. However, these Transformers mainly aim to capture long-range dependencies of spatial regions. As for spectrally self-similar and mask-modulated HSIs, directly applying previous Transformers may be less effective in capturing spectral-wise correlations. In addition, the MSA may pay attention to less informative spatial regions.

3. CASSI System

A concise CASSI principle is shown in Fig. 2 (b). Given a 3D HSI cube, denoted by $F \in \mathbb{R}^{H \times W \times N_{\lambda}}$, where $H$, $W$, and $N_{\lambda}$ represent the HSI’s height, width, and number of wavelengths, respectively. $F$ is firstly modulated by the coded aperture (physical mask) $M^* \in \mathbb{R}^{H \times W}$ as

$$F'(i, j, n_{\lambda}) = F(i, j, n_{\lambda}) \odot M^*, \quad (1)$$

where $F'$ denotes the modulated HSIs, $n_{\lambda} \in [1, \ldots, N_{\lambda}]$ indexes the spectral channels, and $\odot$ denotes the element-wise multiplication. After passing through the disperser, $F'$ becomes tilted and is considered to be sheared along the $y$-axis. We use $F'' \in \mathbb{R}^{H \times (W + d(N_{\lambda} - 1)) \times N_{\lambda}}$ to denote the tilted HSI cube, where $d$ represents the shifting step. We assume $\lambda_{c}$ to be the reference wavelength, i.e., $F''(i, j, n_{\lambda_{c}})$ is not sheared along the $y$-axis. Then we have

$$F''(u, v, n_{\lambda}) = F'(x, y + d(\lambda_{n} - \lambda_{c})), \quad (2)$$

where $(u, v)$ represents the coordinate system on the detector plane, $\lambda_{n}$ denotes the wavelength of the $n_{\lambda}$-th channel, and $d(\lambda_{n} - \lambda_{c})$ indicates the spatial shifting for the $n_{\lambda}$-th channel on $F''$. Finally, the captured 2D compressed measurement $Y \in \mathbb{R}^{H \times (W + d(N_{\lambda} - 1))}$ can be obtained by

$$Y = \sum_{n_{\lambda} = 1}^{N_{\lambda}} F''(i, j, n_{\lambda}) + G, \quad (3)$$

where $G \in \mathbb{R}^{H \times (W + d(N_{\lambda} - 1))}$ is the imaging noise on the measurement, generated by the photon sensing detector.

4. Method

4.1. Overall Architecture

The overall architecture of MST is shown in Fig. 3 (a). We adopt a U-shaped structure that consists of an encoder, a bottleneck, and a decoder. MST is built up by Mask-guided Spectral-wise Attention Blocks (MSAB). Firstly, we reverse the dispersion process (Eq. (2)) and shift back the measurement to obtain the initialized signal $H \in \mathbb{R}^{H \times W \times N_{\lambda}}$ as

$$H(x, y, n_{\lambda}) = Y(x, y - d(\lambda_{n} - \lambda_{c})). \quad (4)$$

Then we feed $H$ into the model. Firstly, MST exploits a $conv3 \times 3$ (convolution with kernel size = 3) layer to map $H$ into feature $X_{0} \in \mathbb{R}^{H \times W \times C}$. Secondly, $X_{0}$ undergoes $N_{1}$ MSABs, a downsample module, $N_{2}$ MSABs, and a downsample module to generate hierarchical features. The downsample module is a strided $conv4 \times 4$ layer that downscales the feature maps and doubles the channels. Therefore, the feature of the $i$-th stage of the encoder is denoted as $X_{i} \in \mathbb{R}^{H \times W \times 2^{i}C}$. Thirdly, $X_{2}$ passes through the
bottleneck that consists of $N_3$ MSABs. Subsequently, We follow the spirit of U-Net [42] and design a symmetrical structure as the decoder. In particular, the upsample module is a strided $\text{deconv}2 \times 2$ layer. The skip connections are exploited for feature aggregation between the encoder and decoder to alleviate the information loss caused by the downsample operations. Similarly, the feature of the $i$-th stage of the decoder is denoted as $X_i' \in \mathbb{R}^{H \times W \times 2^i C}$. After passing through the decoder, the feature maps undergo a $\text{conv}3 \times 3$ layer to generate the residual HSIs $R \in \mathbb{R}^{H \times W \times N_3}$. Finally, the reconstructed HSIs $H' \in \mathbb{R}^{H \times W \times N_3}$ can be obtained by the sum of $R$ and $H$, i.e., $H' = H + R$.

In implementation, we set $C$ to 28 and change the combination $(N_1, N_2, N_3)$ to establish a series of MST models with small, medium, and large model sizes and computation costs: MST-S (2,2,2), MST-M (2,4,4), and MST-L (4,7,5).

The basic unit of MST is MSAB. As shown in Fig. 3 (b), MSAB consists of two layer normalization, a Mask-guided Spectral-wise MSA (MS-MSA), and a Feed-Forward Network (FFN). The details of FFN are depicted in Fig. 3 (c).

### 4.2. Spectral-wise Multi-head Self-Attention

The non-local self-similarity is often exploited in HSI reconstruction but is usually not well modeled by CNN-based methods. Due to the effectiveness of Transformer in capturing non-local long-range dependencies and its impressive performance in other vision tasks, we aim to explore the potential of Transformer in HSI reconstruction. However, there are two main issues when directly applying Transformer to HSI restoration. The first problem is that original Transformers model long-range dependencies in spatial dimensions. But the HSI representations are spatially sparse and spectrally correlated, as shown in Fig. 2 (a). Capturing spatial-wise interactions may be less cost-effective than modeling spectral-wise correlations. Hence, we propose S-MSA that treats each spectral feature map as a token and calculates self-attention along the spectral dimension. Fig. 2 (c1) shows the S-MSA used in stage 0 of MST. The input $X_{in} \in \mathbb{R}^{H \times W \times C}$ is reshaped into tokens $X \in \mathbb{R}^{HW \times C}$. Then $X$ is linearly projected into query $Q \in \mathbb{R}^{HW \times C}$, key $K \in \mathbb{R}^{HW \times C}$, and value $V \in \mathbb{R}^{HW \times C}$:

$$Q = XW^Q, K = XW^K, V = XW^V,$$

(5)

where $W^Q$, $W^K$, and $W^V \in \mathbb{R}^{C \times C}$ are learnable parameters; biases are omitted for simplification. Subsequently, we respectively split $Q$, $K$, and $V$ into $N$ heads along the spectral channel dimension: $Q = [Q_1, \ldots, Q_N]$, $K = [K_1, \ldots, K_N]$, and $V = [V_1, \ldots, V_N]$. The dimension of each head is $d_h = \frac{C}{N}$. Please note that Fig. 2 (c1) depicts the situation with $N = 1$ and some details are omitted for simplification. Different from original MSAs, our S-MSA treats each spectral representation as a token and calculates the self-attention for each $head_j$:

$$A_j = \text{softmax} (\sigma_j K_j^T Q_j), \quad \text{head}_j = V_j A_j,$$

(6)

where $K_j^T$ denotes the transposed matrix of $K_j$. Because the spectral density varies significantly with respect to the wavelengths, we use a learnable parameter $\sigma_j \in \mathbb{R}^1$ to adapt the self-attention $A_j$ by re-weighting the matrix multiplication $K_j^T Q_j$ inside $\text{head}_j$. Subsequently, the outputs of $N$ heads are concatenated in spectral wise to undergo a linear projection and then is added with a position embedding:

$$\text{S-MSA}(X) = \left( \text{Concat}(\text{head}_j) \right) W + f_p(V),$$

(7)

where $W \in \mathbb{R}^{C \times C}$ are learnable parameters, $f_p(\cdot)$ is the function to generate position embedding. It consists of two
depth-wise conv3×3 layers, a GELU activation, and re-shape operations. The HSIIs are sorted by the wavelength along the spectral dimension. Therefore, we exploit this embedding to encode the position information of different spectral channels. Finally, we reshape the result of Eq. (7) to obtain the output feature maps $X_{out} \in \mathbb{R}^{H \times W \times C}$.

We analyze the computational complexity of S-MSA and compare it with other MSAs. We only compare the main difference, i.e., the self-attention mechanism in Eq. (6):

$$O(\text{S-MSA}) = \frac{2HWC^2}{N}, \quad O(\text{G-MSA}) = 2(HW)^2C,$$

$$O(\text{W-MSA}) = 2(M^2)(\frac{HW}{M^2})C = 2M^2HWC,$$

where G-MSA denotes the original global MSA [15], W-MSA denotes the local window-based MSA [31], and $M$ represents the window size. The computational complexity of S-MSA and W-MSA is linear to the spatial size $HW$. This cost is much cheaper than that of G-MSA (quadratic to $HW$). Meanwhile, S-MSA treats a whole spectral feature map as a token. Thus, the receptive field of our S-MSA is global and not limited to position-specific windows.

4.3. Mask-guided Mechanism

The second problem of directly using Transformer for HSI restoration is that original Transformers may attend to some less informative spatial regions with low-fidelity HSI representations. In CASSI, a physical mask is used to modulate the HSIIs. Thus, the light transmittance of different positions on the mask varies. As a result, the fidelity of the modulated spectral information is position-sensitive. This observation motivates us that the mask should be used as a clue to direct the model to attend to regions with high-fidelity HSI representations. In this part, we firstly analyze the usage of the mask in previous CNN-based methods, and then introduce our Mask-guided Mechanism (MM).

Previous Mask Usage Scheme. Previous CNN-based methods [35, 36, 38, 48] mainly conduct an inner product between the initialized HSIs $H$ and the mask $M^*$ to generate a modulated input. This scheme introduces spatial fidelity information but suffers from the following limitations: (i) This operation corrupts the input HSI representations, causes the information loss, and leads to spatial dis-continuity. (ii) This scheme only operates at the input. The guidance effect of the mask in directing the network to pay attention to regions with high-fidelity HSI representations is not fully explored. (iii) This scheme does not exploit learnable parameters to model the spatial-wise correlations.

Our MM. Different from previous methods, our MM preserves all the input HSI representations and learns to direct S-MSA to pay attention to the spatial regions with high-fidelity spectral representations. To be specific, given the mask $M^* \in \mathbb{R}^{H \times W}$ shown in Fig. 2 (c2), since the modulated HSIs are shifted by the disperser of the CASSI system, we firstly shift $M^*$ like the dispersion process:

$$M_s(x, y, n) = M^*(x, y + d(\lambda_n - \lambda_c)), \quad (9)$$

where $M_s \in \mathbb{R}^{H \times (W+d(N_s-1)) \times N_r}$ denotes the shifted version of $M^*$. The shifted regions out of the range in $y$-axis on $M^*$ are set to 0. Please note that Fig. 2 (c2) shows the MM used in stage 0 of MST. To match the scale of the feature maps in stage $i$ of MST, $M_s$ needs to pass through the same downsample operations in Fig. 3 (a). Subsequently, $M_s$ undergoes a conv1×1 layer and then is input to two paths. The upper path is an identity mapping to retain the original fidelity information. The lower path undergoes a conv1×1 layer, a depth-wise conv5×5 layer, a sigmoid activation, and an inner product with the upper path.

S-MSA is effective in capturing inter-spectral dependencies but shows limitations in modeling spatial interactions of HSI representations. Thus, the lower path is designed to capture the spatial-wise correlations. Then we have

$$M'_s = (W_1M_s) \odot \left(1 + \delta(f_{dw}(W_2W_1M_s))\right), \quad (10)$$

where $W_1$ and $W_2$ are the learnable parameters of the two conv1×1 layers, $f_{dw}()$ denotes the mapping function of the depth-wise conv5×5 layer, $\delta()$ represents the sigmoid activation, and $M'_s \in \mathbb{R}^{H \times (W+d(N_s-1)) \times C}$ denotes the intermediate feature maps. To spatially align the mask attention map with the modulated HSIs $F'$ in the CASSI system (Fig. 2 (b)) and the initialized input $H$ of MST (Fig. 3 (a)), we reverse the dispersion process and shift back $M'_s$ to obtain the mask attention map $M' \in \mathbb{R}^{H \times W \times C}$ as

$$M'(x, y, n) = M'_s(x, y - d(\lambda_n - \lambda_c), n), \quad (11)$$

where $n_s \in [1, \ldots, C]$ indexes the spectral channels to match the dimensions of $M'_s$. We reshape $M'$ into $M' \in \mathbb{R}^{H \times W \times C}$ to match the dimensions of $V$. Then we split $M$ into $N$ heads in spectral wise: $M = [M_1, \ldots, M_N]$. For each $head_j$, MM conducts its guidance by re-weighting $V_j$ using $M_j \in \mathbb{R}^{HW \times C}$. Hence, when using MM to direct S-MSA, the S-MSA module just needs to make a simple modification by re-formulating $head_j$ in Eq. (6):

$$head_j = (M_j \odot V_j)(A_j). \quad (12)$$

The subsequent steps of S-MSA remain unchanged. By using MM, S-MSA can extract non-corrupted HSI representations, enjoy the guidance of position-sensitive fidelity information, and adaptively model the spatial-wise interactions.

5. Experiments

5.1. Experimental Settings

Following the settings of TSA-Net [36], we adopt 28 wavelengths from 450 nm to 650 nm derived by spectral interpolation manipulation for HSIs. We perform experiments on both simulation and real HSI datasets.

Simulation HSI Data. We use two simulation hyperspectral image datasets, CAVE [41] and KAIST [12]. CAVE
Table 1. Quantitative results on 10 scenes in simulation. PSNR and SSIM are reported. MSTS significantly surpass other competitors.

| Scene | TwIST \cite{4} | GAP-TV \cite{57} | DeSCI \cite{30} | λ-net \cite{39} | HSSP \cite{49} | DNU \cite{50} | DIP-HSI \cite{38} | TSA-Net \cite{36} | DGSMP \cite{20} | MST-S | MST-M | MST-L |
|-------|----------------|------------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|
| PSNR  | SSIM | PSNR  | SSIM | PSNR  | SSIM | PSNR  | SSIM | PSNR  | SSIM | PSNR  | SSIM | PSNR  | SSIM |
| 1     | 34.71 | 0.935 | 35.15 | 0.937 | 35.40 | 0.941 | 34.45 | 0.925 | 35.19 | 0.935 | 35.87 | 0.944 | 35.32 | 0.943 | 36.26 | 0.950 | 36.51 | 0.953 |
| 2     | 34.05 | 0.927 | 35.19 | 0.938 | 35.82 | 0.944 | 34.89 | 0.923 | 35.21 | 0.937 | 35.96 | 0.947 | 35.39 | 0.944 | 36.23 | 0.951 | 36.55 | 0.954 |
| 3     | 34.45 | 0.927 | 35.19 | 0.937 | 35.40 | 0.941 | 34.45 | 0.925 | 35.19 | 0.935 | 35.87 | 0.944 | 35.32 | 0.943 | 36.26 | 0.950 | 36.51 | 0.953 |
| 4     | 34.05 | 0.927 | 35.19 | 0.938 | 35.82 | 0.944 | 34.89 | 0.923 | 35.21 | 0.937 | 35.96 | 0.947 | 35.39 | 0.944 | 36.23 | 0.951 | 36.55 | 0.954 |
| 5     | 34.45 | 0.927 | 35.19 | 0.937 | 35.40 | 0.941 | 34.45 | 0.925 | 35.19 | 0.935 | 35.87 | 0.944 | 35.32 | 0.943 | 36.26 | 0.950 | 36.51 | 0.953 |
| 6     | 34.05 | 0.927 | 35.19 | 0.938 | 35.82 | 0.944 | 34.89 | 0.923 | 35.21 | 0.937 | 35.96 | 0.947 | 35.39 | 0.944 | 36.23 | 0.951 | 36.55 | 0.954 |
| 7     | 34.45 | 0.927 | 35.19 | 0.937 | 35.40 | 0.941 | 34.45 | 0.925 | 35.19 | 0.935 | 35.87 | 0.944 | 35.32 | 0.943 | 36.26 | 0.950 | 36.51 | 0.953 |
| 8     | 34.05 | 0.927 | 35.19 | 0.938 | 35.82 | 0.944 | 34.89 | 0.923 | 35.21 | 0.937 | 35.96 | 0.947 | 35.39 | 0.944 | 36.23 | 0.951 | 36.55 | 0.954 |
| 9     | 34.45 | 0.927 | 35.19 | 0.937 | 35.40 | 0.941 | 34.45 | 0.925 | 35.19 | 0.935 | 35.87 | 0.944 | 35.32 | 0.943 | 36.26 | 0.950 | 36.51 | 0.953 |
| 10    | 34.05 | 0.927 | 35.19 | 0.938 | 35.82 | 0.944 | 34.89 | 0.923 | 35.21 | 0.937 | 35.96 | 0.947 | 35.39 | 0.944 | 36.23 | 0.951 | 36.55 | 0.954 |

Fig. 4. Reconstructed simulation HSI comparisons of Scene 5 with 4 out of 28 spectral channels. 7 SOTA algorithms and our MST-L are included. The spectral curves (bottom-left) are corresponding to the selected green box of the RGB image. Zoom in for a better view.

5.2. Quantitative Results

We compare our MST with several SOTA HSI reconstruction algorithms, including three model-based methods (TwIST \cite{4}, GAP-TV \cite{57}, and DeSCI \cite{30}) and six CNN-based methods (λ-net \cite{39}, HSSP \cite{49}, DNU \cite{50}, PnP-DIP-HSI \cite{38}, TSA-Net \cite{36}, and DGSMP \cite{20}). For fair comparisons, all methods are tested with the same settings as DGSMP \cite{20}. The PSNR and SSIM results of different methods on 10 scenes in the simulation datasets are listed in Tab. 1. The Params and FLOPS (test size = 256 × 256) of open-source CNN-based algorithms are reported in Tab. 2c. It can be observed from these two tables that our MSTs significantly surpass previous methods by a large margin on all the 10 scenes while requiring much cheaper memory and computational costs. More specifically, our best model, MST-L surpasses DGSMP, TSA-Net, and λ-net by 2.55, 3.72, and 6.65 dB while costing 54.0% (2.03 / 646.65), 25.6%, and 23.9% FLOPS. Surprisingly, even our smallest model, MST-S outperforms DGSMP, TSA-Net, PnP-DIP-HSI, DNU, and λ-net by 1.63, 2.80, 3.00, 3.52, and 5.73

RTX 8000 GPU. The training objective is to minimize the Root Mean Square Error (RMSE) and Spectrum Constancy Loss \cite{62} between the reconstructed and ground-truth HSIs.

Figure 4. Reconstructed simulation HSI comparisons of Scene 5 with 4 out of 28 spectral channels. 7 SOTA algorithms and our MST-L are included. The spectral curves (bottom-left) are corresponding to the selected green box of the RGB image. Zoom in for a better view.

Real HSI Data. We use the real HSI dataset collected by the CASSI system developed in TSA-Net \cite{36}.

Evaluation Metrics. We adopt peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) \cite{53} as the metrics to evaluate the HSI reconstruction performance.

Implementation Details. We implement MST in Pytorch. All the models are trained with Adam \cite{22} optimizer (β₁ = 0.9 and β₂ = 0.999) for 300 epochs. The learning rate is set to 4 × 10⁻⁴ in the beginning and is halved every 50 epochs during the training procedure. When conducting experiments on simulation data, patches at a spatial size of 256 × 256 cropped from the 3D cubes are fed into the networks. As for real hyperspectral image reconstruction, the patch size is set to 660 × 660 to match the real-world measurement. The shifting step d in dispersion is set to 2. Thus, the measurement sizes are 256×310 and 660×714 for simulation and real HSI datasets. The shifting step in reversed dispersion is d/4, i = 0, 1, 2 for the i-th stage of MST. The batch size is 5. Random flipping and rotation are used for data augmentation. The models are trained on one

dataset is composed of 32 hyperspectral images at a spatial size of 512 × 512. KAIST dataset consists of 30 hyperspectral images at a spatial size of 2704 × 3736. Following the schedule of TSA-Net \cite{36}, we adopt CAVE as the training set. 10 scenes from KAIST are selected for testing.
Figure 5. Reconstructed real HSI comparisons of Scene 3 with 4 out of 28 spectral channels. Seven SOTA algorithms and our MST-L are included. MST-L reconstructs more detailed contents and suppresses more noise. Please zoom in for better visualization performance.

dB while requiring 24.7%, 2.1%, 2.7%, 78.2%, and 1.5% Params and 2.0%, 11.8%, 20.1%, 7.9%, and 11.0% FLOPS.

To intuitively show the superiority of our MST, we provide PSNR-Params-FLOPS comparisons of different reconstruction algorithms in Fig. 1. The vertical axis is PSNR (performance), the horizontal axis is FLOPS (computational cost), and the circle radius is Params (memory cost). It can be seen that our MSTs take up the upper-left corner, exhibiting the extreme efficiency advantages of our method.

5.3. Qualitative Results

Simulation HSI Reconstruction. Fig. 4 visualizes the reconstructed simulation HSIs of Scene 5 with 4 out of 28 spectral channels using seven SOTA methods and our MST-L. Please zoom in for a better view. As can be seen from the reconstructed HSIs (right) and the zoom-in patches of the selected yellow boxes, previous methods are less favorable to restore HSI details. They either yield over-smooth results sacrificing fine-grained structural contents and textural details, or introduce undesirable chromatic artifacts and blotchy textures. In contrast, our MST-L is more capable of producing perceptually-pleasing and sharp images, and preserving the spatial smoothness of the homogeneous regions. This is mainly because our MST-L enjoys the guidance of modulation information and captures the long-range dependencies of different spectral channels. In addition, we plot the spectral density curves (bottom-left) corresponding to the picked region of the green box in the RGB image (top-left). The highest correlation and coincidence between our curve and the ground truth demonstrate the spectral-wise consistency restoration effectiveness of our MST.

Real HSI Reconstruction. We further apply our proposed approach to real HSI reconstruction. Similar to [20, 36], we re-train our model (MST-L) on all scenes of CAVE [41] and KAIST [12] datasets. To simulate real imaging situations, we inject 11-bit shot noise into the measurements during training. Visual comparisons are shown in Fig. 5. Our MST-L surpasses previous algorithms in terms of high-frequency structural detail reconstruction and real noise suppression.

5.4. Ablation Study

In this part, we adopt the simulation HSI datasets [12, 41] to conduct ablation studies. The baseline model is derived by removing our S-MSA and MM from MST-S.

Break-down Ablation. We firstly conduct a break-down ablation experiment to investigate the effect of each component towards higher performance. The results are listed in Tab. 2a. The baseline model yields 32.29 dB. When we successively apply our S-MSA and MM, the model continuously achieves 0.89 and 1.08 dB improvements. These results suggest the effectiveness of S-MSA and MM.

Self-Attention Scheme Comparison. We compare S-MSA with other self-attentions and report the results in Tab. 2b. For fair comparisons, the Params of models using different self-attention schemes are set to the same value (0.70 M). Please note that we downscale the input feature of global MSA [15] into $\frac{1}{4}$ size to avoid out of memory. The baseline model yields 32.29 dB while costing 0.53 M Params and 7.43 G FLOPS. We respectively apply the global MSA [15], local window-based MSA [31], Swin W-MSA [31], and our S-MSA. The model gains by 0.38, 0.46, 0.57, and 0.89 dB while adding 4.45, 3.64, 3.64, and 2.93 G FLOPS. Our S-MSA yields the most significant improvement but requires the least computational cost. We explain these results by the HSI characteristics that the spectral representations are spatially sparse and spectrally highly self-similar. Hence, capturing spatial interactions may be less cost-effective than modeling inter-spectra dependencies. This evidence clearly
Table 2. Ablations. We train models on CAVE [41] and test on KAIST [12] in simulation. PSNR, SSIM, Params, and FLOPS are reported.

| Method | A-net [39] | DNU [30] | DIP-HSI [38] | TSA-Net [36] | DGSMP [20] |
|--------|-------------|-----------|---------------|---------------|-------------|
| PSNR   | 28.53       | 30.74     | 31.26         | 31.46         | 32.63       |
| SSIM   | 0.841       | 0.863     | 0.894         | 0.894         | 0.917       |
| Params (M) | 62.64     | 11.9     | 33.85         | 44.25         | 3.76        |
| FLOPS (G) | 117.98   | 15.05    | 64.42         | 110.06        | 646.65      |

Table 2c. Performance-Params-FLOPS comparisons with open-source SOTA CNN-based methods.

Method | Baseline Global MSA | Local W-MSA | Swin W-MSA | S-MSA |
|-------|---------------------|-------------|------------|-------|
| PSNR  | 32.29               | 32.67       | 32.75      | 32.86  | 33.18 |
| SSIM  | 0.897               | 0.912       | 0.916      | 0.919  | 0.923 |
| Params (M) | 0.53     | 0.70         | 0.70       | 0.70   |
| FLOPS (G) | 7.43    | 11.88        | 11.07      | 11.07  | 10.36 |

Figure 6. Visualization of the correlation coefficients among spectral channels of HSIs reconstructed by models using different MSAs. The correlation coefficient map of the model equipped with our S-MSA is the most similar one to that of the ground truth. These results demonstrate the promising effectiveness of our S-MSA.

In addition, we further conduct visual analysis about different MSAs in Fig. 6. Specifically, we visualize the correlation coefficients between each spectral pair of HSIs reconstructed by models equipped with different MSAs. It can be observed that the correlation coefficient map of the model using our proposed S-MSA is the most similar one to that of the ground truth. These results demonstrate the promising effectiveness of our S-MSA in modeling the inter-spectra similarity and long-range spectral-wise dependencies.

Mask-guided Mechanism. We conduct ablation studies to investigate the effect of the previous mask usage scheme described in Sec. 4.3, our MM, and their interaction. The adopted network is the baseline model using S-MSA. The results are reported in Tab. 2d. Method A uses our input setting. Method B exploits the previous scheme that adopts \( H \odot M^* \) as the input. B achieves a limited improvement due to the HSI representation corruption and under-utilization of the mask. Method C uses our MM. C yields the most significant improvement by 1.08 dB, showing the guidance advantage of MM for HSI reconstruction. D exploits both the previous scheme and our MM but degrades by 0.19 dB when compared to method C. This degradation may stem from the loss of some input spectral information.

Additionally, to intuitively show the advantages of MM, we visualize the feature maps of the last MSAB in MST-S. As depicted in Fig. 7, the top row shows the original RGB images. The middle and bottom rows respectively exhibit the feature maps without and with MM. The model using MM pays more high-fidelity attention to details.
References

[1] Nicolas arion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In ECCV, 2020.

[2] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. arXiv preprint arXiv:2103.15691, 2021.

[3] V. Backman, M. B. Wallace, L. Perelman, J. Arendt, R. Gurjar, M. Muller, Q. Zhang, G. Zonios, E. Kline, and T. McGillican. Detection of preinvasive cancer cells. Nature, 2000.

[4] J.M. Bioucas-Dias and M.A.T. Figueiredo. A new twist: Two-step iterative shrinkage/thresholding algorithms for image restoration. TIP, 2007.

[5] M. Borengasser, W. S. Hungate, and R. Watkins. Hyperspectral remote sensing: principles and applications. CRC press, 2007.

[6] Yuanhao Cai, Xiaowan Hu, Haoqian Wang, Yulun Zhang, Hanspeter Pfister, and Donglai Wei. Learning to generate realistic noisy images via pixel-level noise-aware adversarial training. In NeurIPS, 2021.

[7] Yuanhao Cai, Zhicheng Wang, Zhengxiong Luo, Binyi Yin, Angang Du, Haoqian Wang, Xinyu Zhou, Erjin Zhou, Xiangyu Zhang, and Jian Sun. Learning delicate local representations for multi-person pose estimation. arXiv preprint arXiv:2003.04030, 2020.

[8] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xioapeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. arXiv preprint arXiv:2105.05537, 2021.

[9] Jiezhang Cao, Yawei Li, Kai Zhang, and Luc Van Gool. Video super-resolution transformer. arXiv preprint arXiv:2106.06847, 2021.

[10] Xun Cao, Tao Yue, Xing Lin, Stephen Lin, Xin Yuan, Qionghai Dai, Lawrence Carin, and David J. Brady. Computational snapshot multispectral cameras: Toward dynamic capture of the spectral world. IEEE Signal Processing Magazine, 2016.

[11] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In CVPR, 2021.

[12] Inchang Choi, MH Kim, D Gutierrez, DS Jeon, and G Nam. High-quality hyperspectral reconstruction using a spectral prior. In Technical report, 2017.

[13] Xiyang Dai, Yinpeng Chen, Jianwei Yang, Pengchuan Zhang, Lu Yuan, and Lei Zhang. Dynamic detr: End-to-end object detection with dynamic attention. In ICCV, 2021.

[14] Zhuo Deng, Yuanhao Cai, Lu Chen, Zheng Gong, Qiqi Bao, Xue Yao, Dong Fang, Shaochong Zhang, and Lan Ma. Rformer: Transformer-based generative adversarial network for real fundus image restoration on a new clinical benchmark. arXiv preprint arXiv:2201.00466, 2022.

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In ICLR, 2021.

[16] Hao Du, Xin Tong, Xun Cao, and Stephen Lin. A prism-based system for multispectral video acquisition. In ICCV, 2009.

[17] Alaaeldin El-Nouby, Hugo Touvron, Mathilde Caron, Piotr Bojanowski, Matthijs Douze, Armand Joulin, Ivan Laptev, Natalia Neverova, Gabriel Synnaeve, Jakob Verbeek, et al. Xcit: Cross-covariance image transformers. In NeurIPS, 2021.

[18] Ying Fu, Yinqiang Zheng, Imari Sato, and Yoichi Sato. Exploiting spectral-spatial correlation for coded hyperspectral image restoration. In CVPR, 2016.

[19] Xiaowan Hu, Yuanhao Cai, Jing Lin, Haoqian Wang, Xin Yuan, Yulun Zhang, Radu Timofte, and Luc Van Gool. Hd-net: High-resolution dual-domain learning for spectral compressive imaging. In CVPR, 2022.

[20] Tao Huang, Weisheng Dong, Xin Yuan, Jinjian Wu, and Guangming Shi. Deep gaussian scale mixture prior for spectral compressive imaging. In CVPR, 2021.

[21] M. H. Kim, T. A. Harvey, D. S. Kittle, H. Rushmeier, R. O. Prum J. Dorsey, and D. J. Brady. 3d imaging spectroscopy for measuring hyperspectral patterns on solid objects. ACM Transactions on on Graphics, 2012.

[22] Diederik P. Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. In ICLR, 2015.

[23] David Kittle, Kerkil Choi, Ashwin Wagadarikar, and David J Brady. Multiframe image estimation for coded aperture snapshot spectral imagers. Applied optics, 2010.

[24] David Kittle, Kerkil Choi, Ashwin Wagadarikar, and David J. Brady. Multiframe image estimation for coded aperture snapshot spectral imagers. OSA Applied Optics, 2010.

[25] Ke Li, Shijie Wang, Xiang Zhang, Yifan Xu, Weijian Xu, and Zhuowen Tu. Pose recognition with cascade transformers. In CVPR, 2021.

[26] Yanjie Li, Shoukui Zhang, Zhicheng Wang, Sen Yang, Wankou Yang, Shu-Tao Xiu, and Erjin Zhou. Tokenpose: Learning keypoint tokens for human pose estimation. In ICCV, 2021.

[27] Jingyu Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In ICCVW, 2021.

[28] Jing Lin, Yuanhao Cai, Xiaowan Hu, Haoqian Wang, Youliang Yan, Xueyi Zou, Henghui Ding, Yulun Zhang, Radu Timofte, and Luc Van Gool. Flow-guided sparse transformer for video deblurring. arXiv preprint arXiv:2201.01893, 2022.

[29] Xing Lin, Yebin Liu, Jiamin Wu, and Qionghai Dai. Spatial-spectral encoded compressive hyperspectral imaging. TOG, 2014.

[30] Yang Liu, Xin Yuan, Jinli Suo, David Brady, and Qionghai Dai. Rank minimization for snapshot compressive imaging. TPAMI, 2019.
[31] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Xiyuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In ICCV, 2021. 2, 3, 5, 7

[32] Patrick Llull, Xuejun Liao, Xin Yuan, Jianbo Yang, David Kittle, Lawrence Carin, Guillermo Sapiro, and David J Brady. Coded aperture compressive temporal imaging. Optics Express, 2013. 1

[33] Guolan Lu and Baowei Fei. Medical hyperspectral imaging: a review. Journal of Biomedical Optics, 2014. 1

[34] Farid Melgani and Lorenzo Bruzzone. Classification of hyperspectral remote sensing images with support vector machines. IEEE Transactions on Geoscience and Remote Sensing, 2004. 1

[35] Ziyi Meng, Shirin Jalali, and Xin Yuan. Gap-net for snapshot compressive imaging. arXiv preprint arXiv:2012.08364, 2020. 2, 5

[36] Ziyi Meng, Jiawei Ma, and Xin Yuan. End-to-end low cost compressive spectral imaging with spatial-spectral self-attention. In ECCV, 2020. 1, 2, 5, 6, 7, 8

[37] Ziyi Meng, Mu Qiao, Jiawei Ma, Zhenming Yu, Kun Xu, and Xin Yuan. Snapshot multispectral endomicroscopy. Optics Letters, 2020. 1

[38] Ziyi Meng, Zhenming Yu, Kun Xu, and Xin Yuan. Self-supervised neural networks for spectral snapshot compressive imaging. In ICCV, 2021. 2, 5, 6, 8

[39] Xin Miao, Xin Yuan, Yunchen Pu, and Vassilis Athitsos. I-net: Reconstruct hyperspectral images from a snapshot measurement. In ICCV, 2019. 2, 6, 8

[40] Z. Pan, G. Healey, M. Prasad, and B. Tromberg. Face recognition in hyperspectral images. TPAMI, 2003. 1

[41] Jong-Il Park, Moon-Hyun Lee, Michael D. Grossberg, and Shree K. Nayar. Multispectral imaging using multiplexed illumination. In ICIP, 2007. 5, 7, 8

[42] Olaf Ronneberger, Philipp Fischer, Thomas Brox, a, and b. U-net: Convolutional networks for biomedical image segmentation. In MICCAI, 2015. 4

[43] Jin Tan, Yanting Ma, Hoover Rueda, Dror Baron, and Gonzalo R. Arce. Compressive hyperspectral imaging via approximate message passing. IEEE Journal of Selected Topics in Signal Processing, 2016. 2

[44] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In NeurIPS, 2017. 2

[45] Ashwin Wagadarikar, Renu John, Rebecca Willett, and David Brady. Single disperser design for coded aperture snapshot spectral imaging. Applied Optics, 2008. 1

[46] Ashwin Wagadarikar, Renu John, Rebecca Willett, and David Brady. Single disperser design for coded aperture snapshot spectral imaging. Applied Optics, 2008. 1, 2

[47] Ashwin A Wagadarikar, Nikos P Pitsianis, Xiaobai Sun, and David J Brady. Video rate spectral imaging using a coded aperture snapshot spectral imager. Optics Express, 2009. 1

[48] Jianjian Wang, Yulin Zhang, Xin Yuan, Yun Fu, and Zhiquiang Tao. A new backbone for hyperspectral image reconstruction. arXiv preprint arXiv:2108.07739, 2021. 2, 5

[49] Lizhi Wang, Chen Sun, Ying Fu, Min H. Kim, and Hua Huang. Hyperspectral image reconstruction using a deep spatial-spectral prior. In CVPR, 2019. 2, 6

[50] Lizhi Wang, Chen Sun, Maqing Zhang, Ying Fu, and Hua Huang. Dnu: Deep non-local unrolling for computational spectral imaging. In CVPR, 2020. 6, 8

[51] Lizhi Wang, Zhiwei Xiong, Dahua Gao, Guangming Shi, and Feng Wu. Dual-camera design for coded aperture snapshot spectral imaging. OSA Applied Optics, 2015. 1

[52] Lizhi Wang, Zhiwei Xiong, Guangming Shi, Feng Wu, and Wenjun Zeng. Adaptive nonlocal sparse representation for dual-camera compressive hyperspectral imaging. TPAMI, 2016. 1, 2

[53] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. TIP, 2004. 6

[54] Zhengdong Wang, Xiaodong Cun, Jianmin Bao, and Jianzhuan Liu. Uformer: A general u-shaped transformer for image restoration. arXiv preprint 2106.03106, 2021. 3

[55] Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Zhicheng Yan, Masayoshi Tomizuka, Joseph Gonzalez, Kurt Keutzer, and Peter Vajda. Visual transformers: Token-based image representation and processing for computer vision. arXiv preprint arXiv:2006.03677, 2020. 3

[56] Sen Yang, Zhibin Quan, Mu Nie, and Wankou Yang. Transpose: Keypoint localization via transformer. In ICCV, 2021. 3

[57] Xin Yuan. Generalized alternating projection based total variation minimization for compressive sensing. In ICIP, 2016. 1, 2, 6

[58] Xin Yuan, David J Brady, and Aggelos K Katsaggelos. Snapshot compressive imaging: Theory, algorithms, and applications. IEEE Signal Processing Magazine, 2021. 1

[59] Yuan Yuan, Xiangtian Zheng, and Xiaohui Li. Hyperspectral image superresolution by transfer learning. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2017. 1

[60] Nicolas ZCarion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. Rend-to-end object detection with transformers. In ECCV, 2020. 3

[61] Shipeng Zhang, Lizhi Wang, Ying Fu, Xiaoming Zhong, and Hua Huang. Computational hyperspectral imaging based on dimension-discriminative low-rank tensor recovery. In ICCV, 2019. 1, 2

[62] Yuanyuan Zhao, Hui Guo, Zhan Ma, Xin Cao, Tao Yue, and Xuemei Hu. Hyperspectral imaging with random printed mask. In CVPR, 2019. 6

[63] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Yuanyuan Zhao, Hui Guo, Zhan Ma, Xun Cao, Tao Yue, and Wankou Yang. Transpose: Keypoint localization via transformer. In ICCV, 2021. 3

[64] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In ICLR, 2021. 3