Improving Spectral Snapshot Reconstruction with Spectral-Spatial Rectification

Jiancheng Zhang [1], Haijin Zeng [2], Yongyong Chen [3], Dengxiu Yu [1], Yin-Ping Zhao [1,*]
[1] Northwestern Polytechnical University, [2] IMEC-UGent, [3] Harbin Institute of Technology (Shenzhen)

Abstract

How to effectively utilize the spectral and spatial characteristics of Hyperspectral Image (HSI) is always a key problem in spectral snapshot reconstruction. Recently, the spectra-wise transformer has shown great potential in capturing inter-spectra similarities of HSI, but the classic design of the transformer, i.e., multi-head division in the spectral (channel) dimension hinders the modeling of global spectral information and results in mean effect. In addition, previous methods adopt the normal spatial priors without taking imaging processes into account and fail to address the unique spatial degradation in snapshot spectral reconstruction. In this paper, we analyze the influence of multi-head division and propose a novel Spectral-Spatial Rectification (SSR) method to enhance the utilization of spectral information and improve spatial degradation. Specifically, SSR includes two core parts: Window-based Spectra-wise Self-Attention (WSSA) and spAtial Rectification Block (ARB). WSSA is proposed to capture global spectral information and account for local differences, whereas ARB aims to mitigate the spatial degradation using a spatial alignment strategy. The experimental results on simulation and real scenes demonstrate the effectiveness of the proposed modules, and we also provide models at multiple scales to demonstrate the superiority of our approach. https://github.com/ZhangJC-2k/SSR

1. Introduction

The advent of compressed sensing introduced the coded aperture snapshot spectral compressive imaging (CASSI) system [13, 29, 36], addressing drawbacks in traditional hyperspectral cameras regarding efficiency and cost-effectiveness, gaining significant attention. This system offers the promise of swift and cost-efficient capture of Hyperspectral Images (HSIs). However, the inherent ill-posedness of decoding 3D HSIs from 2D measurements poses a significant challenge for reconstruction algorithm design. Consequently, various model-based [1, 23, 24, 33, 38, 40] and learning-based [2–6, 17, 18, 28, 37, 42] approaches have emerged to tackle this challenge. Within the realm of reconstruction algorithms, the effective utilization of diverse priors remains a pivotal research challenge.

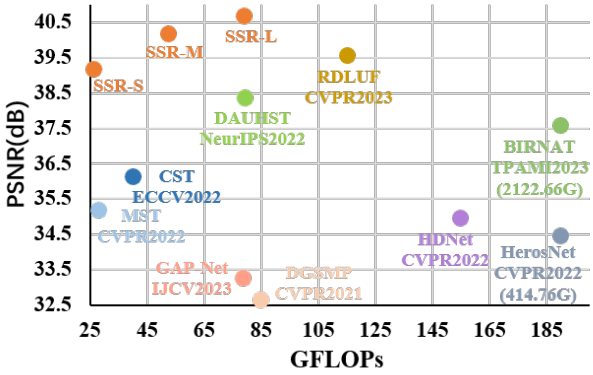

Figure 1. Comparison of PSNR-FLOPs between our SSR and previous reconstruction methods. Our SSR significantly outperforms other methods over 2dB at the same FLOPs when surpassing the state-of-the-art method over 1dB with fewer FLOPs.

Model-based methods initially adopt total variation [20, 40], low-rank [24], sparsity [21, 34] and more traditional hand-craft priors [1, 23, 24, 33, 38] to construct the optimization model, which preliminarily shows the practical feasibility of snapshot spectral reconstruction. With the advancement of deep learning, both convolutional neural networks (CNNs) [6, 17, 18, 28, 37, 42] and transformers [3–5, 11] have been successively introduced for learning the spatial and spectral characteristics of HSIs. They serve as end-to-end denoisers [3, 4, 6, 17, 28] or deep image priors [5, 11, 18, 37, 42] in learning-based methods. These advancements have led to significant progress in both real-world and simulated scenario reconstructions. In addition, the effective utilization of spectral and spatial characteristics has also naturally become a research hotspot.

Previous CNN methods such as TSA-Net [28], HDNet [17], DGSMP [18], and HerosNet [42], along with spatial transformers like Swin Transformer [25] and DAUHST [5], excel in learning spatial representations but do not effectively utilize the distinctive spectral characteristics inherent in HSIs. Conversely, spectra-wise transformers such as MST [3], MST++ [4], and RDLUF [11] show promising

*Corresponding Author

potential in snapshot spectral reconstruction by exploiting inter-spectral similarities. However, their implementations follow the conventional transformer architecture, involving multi-head division in the spectral dimension, which limits the modeling of global spectral similarities and results in mean effect. Furthermore, in spectral snapshot imaging, some spatial degradation occurs in certain bands due to masking, shifting, and compression factors. Unfortunately, existing spatial and spectral networks lack tailored designs to mitigate such degradation effectively.

To optimize the utilization of spectral information and mitigate distinctive spatial degradation, we introduce the Spectral-Spatial Rectification (SSR) method into snapshot spectral reconstruction. Specifically, SSR comprises two key components: Window-based Spectral Self-Attention (WSSA) and spAtial Rectification Blocks (ARB). WSSA divides the features into multiple local windows in the spatial dimension, facilitating global spectral self-attention computations and the local difference consideration. Addressing the interaction gap between windows of WSSA, ARB leverages a Convolution Modulated Block (CMB). This block employs sliding large-kernel convolutions to interact with features between windows and learn effective spatial representation. Subsequently, to mitigate spatial degradation, ARB integrates a novel spatial alignment strategy. It assists low-quality bands in leveraging information from high-quality bands through learning a unified spatial representation and spectral weights, which further optimizes the utilization of spatial information. Overall, our contributions are summarized as follows:

- We analyze the influence of multi-head division on previous spectra-wise transformers and propose Window-based Spectra-wise Self-Attention to model global spectral information while accounting for local differences.
- A Spatial Rectification Block is specially designed to enhance spatial representation and mitigate the spatial degradation in low-quality bands through large-kernel convolutions and a novel spatial alignment strategy.
- The qualitative and quantitative results of real and simulation experiments demonstrate that our SSR method significantly improves spectral snapshot reconstruction.

2. Related Work

The model-based methods [1, 20, 21, 23, 24, 33, 34, 38, 40] construct convex optimization models with some specific priors such as total variation [20, 40], low-rankness [24], and sparsity [21, 34] according to the physical schema of CASSI then solve the problem in an iterative manner to obtain reconstructed images. They have good interpretability and low cost, yet lack performance guarantee. Subsequently, the rise of deep learning [14, 15, 22] has facilitated the rapid development of spectral snapshot reconstruction. To improve reconstruction quality, Plug-and-Play methods [41, 44] plug pre-trained deep networks as priors, which make great progress in performance but are still limited by slow reconstruction speed. Along with faster inference and better results, the CNN-based end-to-end denoisers [17, 28, 31] quickly achieved better results by establishing a mapping between measurements and HSIs, but also brought huge memory and computation costs.

Recently, the transformer-based end-to-end denoisers [2–4] which capture data similarity and long-range dependence have shown impressive performance in both simulation and real scenes. These methods adopt self-attention to learn effective spatial and spectral representation, achieving very high parameter and computational efficiency. Following this, the unfolding methods [5, 11] regard transformers as deep priors for optimization formulas and improve iterative frameworks through end-to-end training. In these approaches, CNNs and transformers play a huge role either as end-to-end denoisers or as deep priors. While previous CNNs and spatial transformers fail to effectively utilize the spectral characteristics inherent in HSIs, spectra-wise transformer [3, 4, 11] has shown great potential in spectral tasks and attracted wide attention. However, current spectra-wise transformers follow classic multi-head design [12, 25, 35] and suffer from the loss of global spectral information and the limitation of ignoring local differences.

3. Spectral Snapshot Imaging Model

Mathematically, we assume a spectral image patch with Λ bands $\{F_\lambda\}_{\lambda=1}^{\Lambda} \in \mathbb{R}^{H\times W}$, where H and W represents the HSI's height and weight. Image frame F_λ is modulated by a physical mask with pattern $M \in \mathbb{R}^{H\times W}$ to get modulated image frame $F_\lambda^{'}$:

$$F_\lambda^{'} = M \circ F_\lambda, \tag{1}$$

where $\circ$ denotes Hadamard's (element-wise) product. Then the modulated image frames of different wavelengths $\left\{F_\lambda^{'}\right\}_{\lambda=1}^{\Lambda}$ are separated to different positions by the disperser. After that, $\left\{F_\lambda^{'}\right\}_{\lambda=1}^{\Lambda}$ distributed across different bands are shifted spatially and summed element-wise. Thus, $\left\{F_\lambda^{'}\right\}_{\lambda=1}^{\Lambda}$ are compressed to a coded measurement:

$$G(m,n) = \sum_{i=1}^{\Lambda} F_\lambda^{'}(m, n + D(\lambda)), \tag{2}$$

where m and n index the spatial coordinates, $D(\lambda) = d(\lambda - 1)$, d represents pixels shift between adjacent bands. Note Eq. (2) assumes a dispersion along the vertical dimension, and the derivation is also applicable for horizontal dispersion. The imaging model in Eq. (2) can be rewritten in the matrix-vector form as follows:

$$g = \Phi f, \tag{3}$$

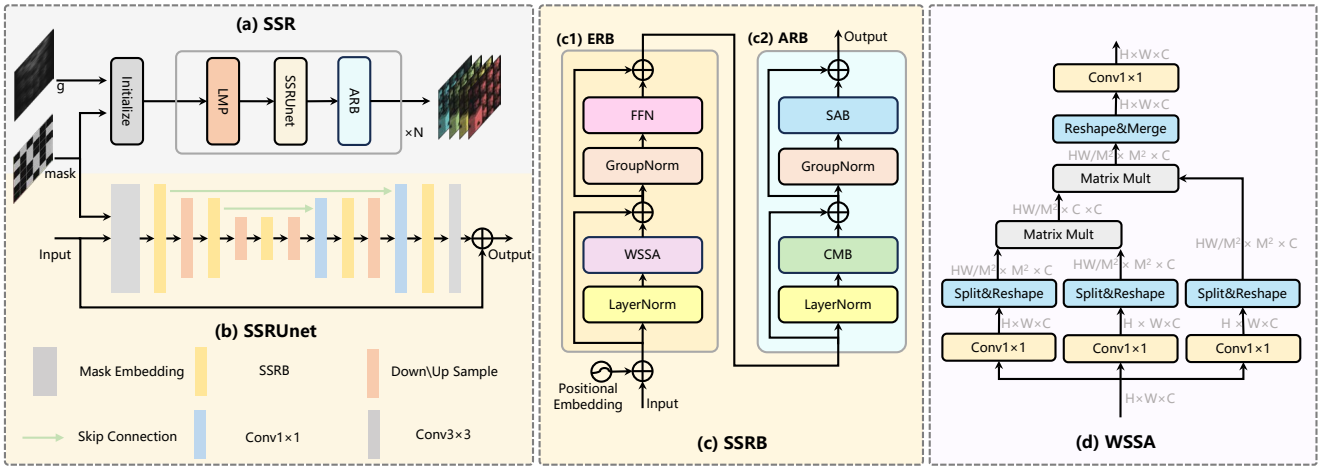

Figure 2. (a)-(c) The overall architecture of SSR, SSRUnet, and SSRB. LMP is achieved via a formula. (d) Details of WSSA.

where $g \in \mathbb{R}^{HW}$, $f \in \mathbb{R}^{HW\Lambda}$ are the vectorized representation of the compressed image G and the original spectral image F respectively, and $\Phi \in \mathbb{R}^{HW \times HW\Lambda}$ is the sensing matrix that describes the system imaging model. In spectral snapshot reconstruction, what we need to do is recover 3D HSI f from the compressed 2D measurement g.

4. Method

Existing spectral and spatial network designs [2–5, 11] are limited by the direct application of traditional multi-head division and unique spatial degradation of snapshot spectral imaging. To address the limitations, we propose the Spectral-Spatial Rectification (SSR) method to better model the spectral information, enhance spatial representation, and improve spatial degradation.

4.1. Overall Architecture

An overview of SSR is presented in Fig. 2 (a), we adopt a multistage network including initialization and N cascaded stages, and each stage consists of Linear Manifold Projection (LMP), Spectral-Spatial Rectification Unet (SSRUnet), and spAtial Rectification Block (ARB). First, the compressed measurement g is reversed to the initial shape [3] and the 2D mask is repeated in channel dimension to obtain the 3D mask. Then the two are inputted to a 1×1 convolution kernel ($conv1 \times 1$) to get the initialization. In each stage, LMP is adopted to assist the reconstruction based on the imaging model. Then the SSRUnet is employed to refine the input through novel spatial and spectral design. Finally, ARB is used to further improve spatial degradation in reconstruction. For the core module SSRUnet, as illustrated in Fig. 2 (b), the $conv3 \times 3$s with residual are utilized to embed mask information into input X as follows:

$$X_{ME} = Conv_{3\times3}(X) \circ (T_I + Conv_{3\times3}(Mask)), \tag{4}$$

where X_{ME} is the features embedded with the mask, T_I is an all-one tensor and $Mask$ is the 3D mask obtained in initialization. Then SSRUnet with three layers and $conv3 \times 3$ are employed to extract the deep feature which is combined with the input as a residual to produce a refined output. The SSRUnet consists of downsampling modules ($conv4 \times 4$), Spectral-Spatial Rectification Block (SSRB), upsampling modules ($deconv2 \times 2$), and Fusion modules ($conv1 \times 1$).

Spectral-Spatial Rectification Block (SSRB). SSRB is the basic module of SSRUnet. As shown in Fig. 2 (c), the implicit positional information [10] is firstly embedded in the input for applying self-attention later in SSRB. Each SSRB is the sequential combination of the spEctral Rectification Block (ERB) and ARB. Fig. 2 (c) illustrates the components of ERB and ARB, i.e., a Feed Forward Network(FFN) and a Window-based Spectral Self-Attention (WSSA) for ERB, a Convolution Moudulated Block (CMB) and a Spatial Alignment Block (SAB) for ARB, and same Layer Normalization (LN) and Group Normalization (GN) for ERB and ARB. WSSA is detailed in Fig. 2 (d) and FFN consists of two linear layers sandwiched with depth-wise $conv3 \times 3$.

4.2. Spectral Rectification Block

Spectral self-attention is a promising approach to utilize spectral characteristics for HSI tasks. However, we find that the previous spectra-wise transformer [3, 4, 11] directly considers a band as a token and follows traditional spatial self-attention design, i.e., multi-head division in spectral (channel) dimension, which ignores the local differences in spectral distribution and results in mean effect and non-global spectral utilization. To solve the problem, we propose the ERB based on WSSA to consider local differences and model global spectral information.

The Influence of Multi-Head Division in Spectral (Channel) Dimension. In the classic transformer design [12, 25, 35], the use of multi-head division in the channel dimension is common and brings some gains. Possibly affected by this, the previous spectra-wise transformer in spectral snapshot reconstruction also divides multi-head in spectral dimension as shown in Fig. 3 (a), which results in the sep-

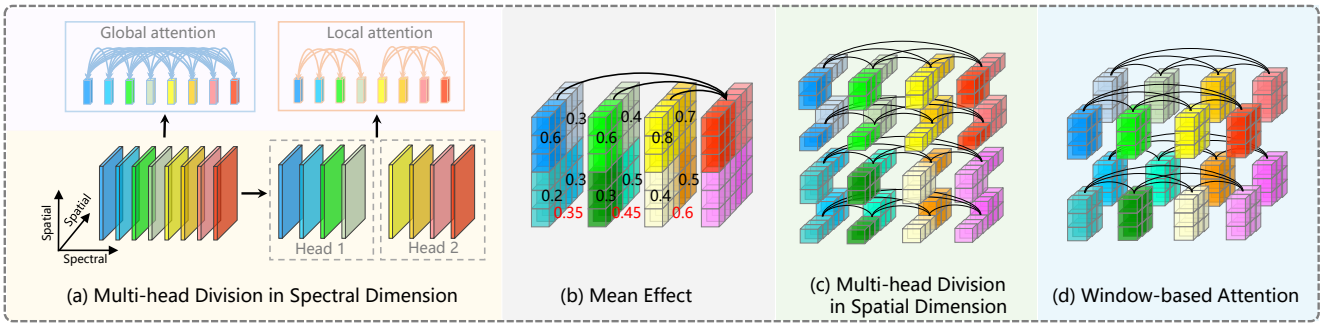

Figure 3. (a)-(c) The influence of multi-head attention and mean effect, (d) The illustration of WSSA.

aration of spectral information and failure to model global spectral information. It's very easy to get confused because the spectral transformer is in contrast to the spatial transformer, where the characteristic dimension of the former is the spatial dimension, and the characteristic dimension of the latter is the spectral (channel) dimension.

The Influence of Mean Effect. In addition, considering that HSIs depict the spectral distribution of scenes, the similarities of different regions should be different but previous multi-head division results in all regions sharing the same similarity. When treating the whole band as a token to calculate spectral self-attention, we obtain the mean of all region similarities in the band and the local differences between regions are lost, which we call the 'Mean Effect'. Define Λ features $\{f_\lambda\}_{\lambda=1}^{\Lambda}$ where feature dimension consists of p different patterns (regions), i.e., $f_\lambda = [f_\lambda^1, f_\lambda^2, \ldots, f_\lambda^p]$, the influence of mean effect on the similarity calculation can be simply described as follows:

$$sim(f_i^k, f_j^k) \rightarrow \frac{1}{p}\sum_{m=1}^{p} sim(f_i^m, f_j^m), \tag{5}$$

where the left part denotes the expected similarity of the kth pattern between f_i^k and f_j^k, and the right part represents the obtained similarity in previous approaches. We also present an example in Fig. 3 (b), the average similarity and local region similarity between the red band and other bands are marked with red numbers and black numbers respectively. To demonstrate the validity of our theory, we carried out ablation experiments of different region sizes, i.e., window size under the same conditions later. In addition, the mean effect may explain why multi-head attention is more effective than single-head attention.

Multi-Head Division in Spatial Dimension. When realizing appropriate feature dimension and mean effect, it was natural to treat the entire band frame as a token and then directly divide multi-head [35] in the merged spatial dimension to calculate self-attention, which we call MSSA later for short. However, as shown in Fig. 3 (c), this would split the entire image into (incomplete) strip shapes on each head, which mitigates the mean effect to some extent but breaks the spatial correlation of features. That is, adjacent areas are divided into different heads, while areas that are far apart in space are divided into the same head.

Window-based Spectral Self-Attention (WSSA). To capture the global spectral similarity, solve the mean effect, and maintain spatial correlation, WSSA divides features into a number of windows spatial-wisely, and then the spectra-wise self-attention is calculated within windows to avoid the interference of each other, which is shown in Fig. 3 (d). Specifically, define input $X \in \mathbb{R}^{H\times W\times C}$, as shown in Fig. 2 (d), X is first linearly projected to $3C$ channels tensor via a 1×1 convolution then uniformly divides the tensor channel-wisely into the query (Q), key (K), value(V). Subsequently, Q, K, V are split spatial-wise into HW/M^2 windows with size of $\mathbb{R}^{M\times M\times C}$ separately:

$$\{Q_i, K_i, V_i\}_{i=1}^{HW/M^2} = Q, K, V, \tag{6}$$

where $Q_i, K_i, V_i \in \mathbb{R}^{M\times M\times C}$ and are then reshaped into shape $\mathbb{R}^{M^2\times C}$ as tokens to calculate attention as follows:

$$Attention_i = SoftMax(Q_i^T K_i/M)V_i, \tag{7}$$

where M is used as the scale factor before applying the $softmax$ function. Then the outputs of HW/M^2 attention are reshape in $\mathbb{R}^{M\times M\times C}$ and merge together in the original arrangement to undergo a $conv1\times 1$ projection:

$$X' = \{Attention_i\}_{i=1}^{HW/M^2} W_1, \tag{8}$$

where $W_1 \in \mathbb{R}^{C\times C}$ are weight matrices of a $conv1\times 1$ and X' is the final output $\in \mathbb{R}^{H\times W\times C}$. Different from multi-head self-attention which implements token embedding and then divides heads, WSSA first implements windows split then reshaping into tokens, which preserves the spatial correlation of each token. Regarding WSSA as a special multi-head self-attention, the number of WSSA heads is more, often hundreds of thousands, far more than the number of multi-head attention.

Computational Complexity. The computational complexity of WSSA is displayed as follows:

$$O(WSSA) = 4HWC^2 + 2M^2C^2\frac{HW}{M^2} = 6HWC^2. \tag{9}$$

The computational complexity of WSSA is linear to the spatial size HW and window size-independent, meaning that there is no additional computation cost when we improve spectral utilization.

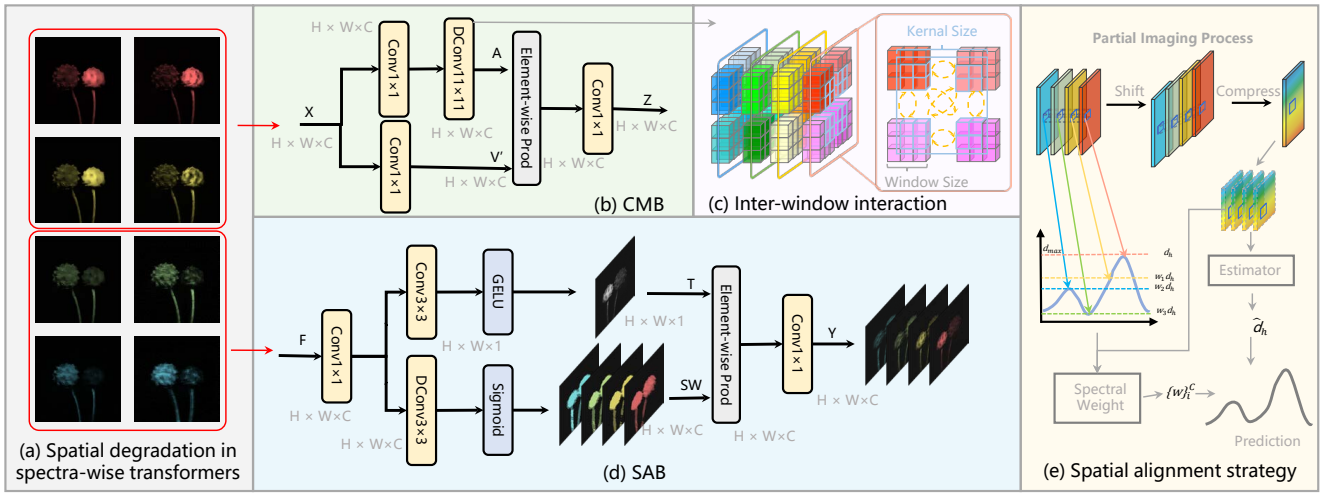

Figure 4. (a) Display of spatial degradation in spectra-wise transformers. (b) The details of CMB. (c) The effect of large kernel convolution in CMB. (d) The details of SAB. (e) Spatial alignment strategy of SAB.

4.3. Spatial Rectification Block

The utilization of WSSA enhances the capability of spectral information capture. However, a limitation arises from the absence of interaction between adjacent windows within WSSA, potentially leading to discontinuous spatial representation and the occurrence of blocking artifacts, as highlighted by [7]. Additionally, unique spatial degradation occurs due to mask, shift, and compression in spectral snapshot imaging. To address these concerns, we propose a spAtial Rectification Block (ARB) comprising a Convolution Modulated Block (CMB) and a Spatial Alignment Block (SAB). This ARB aims to foster interaction between adjacent windows and alleviate spatial degradation in the HSIs.

Convolution Modulated Block (CMB). Here, we try to solve two problems: (1) When different windows cannot interfere with each other, WSSA also fails to exchange information, which leads to discrete spatial expression and may lead to blocking artifacts [7]. (2) As shown in the top two rows of Fig .4 (a), over-smoothing demonstrates that the spatial representation ability of a single spectra-wise transformer is not sufficient. To solve these two problems and inspired by the development of large kernel convolution in high-level tasks, we adopted a Convolution Modulated Block (CMB) proposed in [16] to modulate spatial information and make adjacent windows interact with each other through large kernel convolution sliding as shown in Fig. 4 (b). CMB modulates spatial information in a similar manner to self-attention but much more efficiently. Specifically, given the input tokens $X \in \mathbb{R}^{H\times W\times C}$, we use a simple depth-wise convolution with kernel size 11×11 to calculate the modulated attention A as follows:

$$A = DConv_{11\times 11}(XW_2), \tag{10}$$

where $W_2 \in \mathbb{R}^{H\times W\times C}$ are weight matrices of $conv1 \times 1$s, and $DConv_{11\times 11}$ denotes a depth-wise convolution with kernal size 11×11. Then we adopt the element-wise product and attention A to modulate the project feature V' to get the output Z:

$$Z = (A \circ V')W_4, V' = XW_3, \tag{11}$$

where W_3 and $W_4 \in \mathbb{R}^{C\times C}$ are weight matrices of $conv1\times 1$s. On the one hand, CMB could effectively improve the spatial representation of features through modulation operation. On the other hand, CMB enables the pixels between adjacent windows of WSSA to interact spatial-wisely when depth-wise convolution kernel size 11×11 is greater than patch size 8×8, as demonstrated in Fig. 4 (c).

Spatial Alignment Block (SAB). As shown in the bottom two rows of Fig. 4 (a), we observe that the reconstruction results of some bands have more spatial degradation such as distortion and deformation, which is not common in other HSI tasks. Thus, we think this may be related to the unique imaging process. As shown in Fig. 4 (e), shift and compression in the imaging process mix information from different spatial locations, and spatial texture recovery in bands with low spectral density may be difficult, which is neglected by previous methods. Considering that the low-density bands can be obtained by multiplying the high-density bands by weights, we can utilize the high-quality bands to improve the spatial texture of the low-quality bands through a novel spatial alignment strategy, as shown in Fig. 4 (e). To achieve this, a Spatial Alignment Block (SAB) is proposed to improve spatial degradation. As shown in 4 (d), we adopt different convolution kernels to estimate spatial texture distribution which corresponds to the high-density part, and learn spectral weights in SAB when two linear layers are used to mix channel information. Specifically, given the input feature $F \in \mathbb{R}^{H\times W\times C}$, we use a convolution with kernel size 3×3 and $GELU$ activation function to learn the spatial texture T as follows:

$$T = \psi(Conv_{3\times 3}(FW_5)), \tag{12}$$

Table 1. The PSNR (upper entry in each cell) in dB and SSIM (lower entry in each cell) results of the test methods on 10 scenes.

Algorithms	Reference	Params	GFLOPs	Scene1	Scene2	Scene3	Scene4	Scene5	Scene6	Scene7	Scene8	Scene9	Scene10	Avg
DIP-HSI [30]	ICCV 2021	33.85	64.42	32.68 0.892	27.26 0.858	31.30 0.915	40.54 0.953	29.79 0.884	30.39 0.908	28.18 0.878	29.44 0.888	34.51 0.890	28.51 0.874	31.26 0.894
TSA-Net [28]	ECCV 2020	44.25	110.06	32.03 0.892	31.00 0.858	32.25 0.915	39.19 0.953	29.39 0.884	31.44 0.908	30.32 0.878	29.35 0.888	30.01 0.890	29.59 0.874	31.46 0.894
DGSMP [18]	CVPR 2021	3.58	84.77	33.26 0.915	32.09 0.898	33.06 0.925	40.54 0.964	28.86 0.882	33.08 0.937	30.74 0.886	31.55 0.923	31.66 0.911	31.44 0.925	32.63 0.917
GAP-Net [27]	IJCV 2023	4.27	78.58	33.74 0.911	33.26 0.900	34.28 0.929	41.03 0.967	31.44 0.919	32.40 0.925	32.27 0.902	30.46 0.905	33.51 0.915	30.24 0.895	33.26 0.917
HerosNet [42]	CVPR 2022	11.27	414.76	35.69 0.973	35.01 0.968	34.82 0.967	38.07 0.985	33.18 0.969	34.94 0.976	33.58 0.962	33.19 0.968	33.04 0.964	33.01 0.965	34.45 0.970
HDNet [17]	CVPR 2022	2.37	154.76	35.14 0.935	35.67 0.940	36.03 0.943	42.30 0.969	32.69 0.946	34.46 0.952	33.67 0.926	32.48 0.941	34.89 0.942	32.38 0.937	34.97 0.943
MST [3]	CVPR 2022	2.03	28.15	35.40 0.941	35.87 0.944	36.51 0.953	42.27 0.973	32.77 0.947	34.80 0.955	33.66 0.925	32.67 0.948	35.39 0.949	32.50 0.941	35.18 0.948
CST [2]	ECCV 2022	3.0	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
BIRNAT [8]	TPAMI 2023	4.40	2122.66	36.79 0.951	37.89 0.957	40.61 0.971	46.94 0.985	35.42 0.964	35.30 0.959	36.58 0.955	33.96 0.956	39.47 0.970	32.80 0.938	37.58 0.960
DAUHST [5]	NeurIPS 2022	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
RDLUF [11]	CVPR 2023	1.81	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
SSR-S	Ours	**1.73**	**26.37**	38.22 0.963	40.05 0.972	42.57 0.975	47.46 0.986	36.50 0.972	37.33 0.971	37.60 0.964	35.70 0.968	41.37 0.975	35.06 0.959	39.19 0.971
SSR-M	Ours	3.45	52.65	38.61 0.967	41.35 0.978	43.94 0.979	48.32 0.988	37.80 0.977	38.07 0.975	38.54 0.969	36.82 0.974	42.72 0.980	35.79 0.965	40.20 0.975
SSR-L	Ours	5.18	78.93	**39.07** **0.970**	**42.04** **0.981**	**44.49** **0.980**	**48.80** **0.990**	**38.64** **0.980**	**38.50** **0.978**	**39.16** **0.971**	**36.96** **0.976**	**43.12** **0.982**	**36.08** **0.968**	**40.69** **0.978**
SSR-L*	Ours	**1.73**	78.93	38.81 0.968	41.51 0.979	43.76 0.979	48.62 0.988	38.32 0.979	37.85 0.975	38.50 0.969	36.85 0.974	42.64 0.980	35.82 0.965	40.27 0.976

where $T \in \mathbb{R}^{H\times W\times 1}$, ψ is $GELU$ activation function and $W_5 \in \mathbb{R}^{C\times C}$ are weight matrices of the linear layer. Meanwhile, we use a depth-wise convolution with kernel size 3×3 and $Sigmoid$ activation function to learn spectral element-wise weight as follows:

$$SW = \sigma(DConv_{3\times 3}(FW_5)), \tag{13}$$

where $SW \in \mathbb{R}^{H\times W\times C}$, σ is $Sigmoid$ activation function. Finally, we use multiple element-wise products and a $conv1 \times 1$ projection W_6 to calculate the output Y as follows:

$$Y = \{T \circ SW_i\}_{i=1}^{C} W_6, \tag{14}$$

where $Y \in \mathbb{R}^{H\times W\times C}$ and $SW_i \in \mathbb{R}^{H\times W\times 1}$.

4.4. Linear Manifold Projection (LMP)

Eq. (3) is a key constraint in spectral compression reconstruction. To take advantage of this prior, we introduce a Linear Manifold Projection (LMP) proposed in GAP-Net [27] to assist the reconstruction. We add an additional parameter to control projection intensity and perform the LMP as follows:

$$f' = f + \rho\Phi^T[(g - \Phi f)./(\Phi\Phi^T)], \tag{15}$$

where ρ is the parameter that is estimated through a simple network similar to [5].

4.5. Implementation Details

We change the stage numbers N to establish a series of SSR models with small, medium, and large scales: SSR-S ($N = 3$), SSR-M ($N = 6$), and SSR-L ($N = 9$). The proposed SSR is implemented by PyTorch and the window size in WSSA is empirically set to 8×8. Adam [19] optimizer ($\beta_1 = 0.9$ and $\beta_2 = 0.999$) and Cosine Annealing [26] scheduler are adopted to train SSR on a single RTX 3090 GPU. Training samples are patches with spatial sizes of 256×256 and 384×384 randomly cropped from 3D HSI data cubes for simulation and real experiments separately. The shifting step d of the imaging model is 2. The channels of SSRUnet layers are set to $\Lambda, 2\Lambda, 4\Lambda$ in sequence and basic bands $\Lambda = 28$. The training loss function is the root mean square error (RMSE) between reconstructed and ground-truth HSIs and adopts the multi-stage loss setting proposed in [43].

5. Experiment

5.1. Experimental Settings

Following [5, 11, 17, 28, 42], we select 28 wavelengths from 450nm to 650nm by using spectral interpolation manipulation to derive HSIs. We conduct experiments on simulation and real datasets.

Simulation and Real Data. Two simulation datasets, CAVE [32] and KAIST [9], and five real HSIs captured by the CASSI system developed in [28] are adopted. The CAVE dataset provides 32 HSIs with a spatial size of 512×512. The KAIST dataset includes 30 HSIs with a spatial size of 2704×3376. We use CAVE for simulation training and select 10 scenes from KAIST for simulation testing.

Evaluation Metrics. We adopt two image quality indexes including peak signal-to-noise ratio (PSNR), and structure

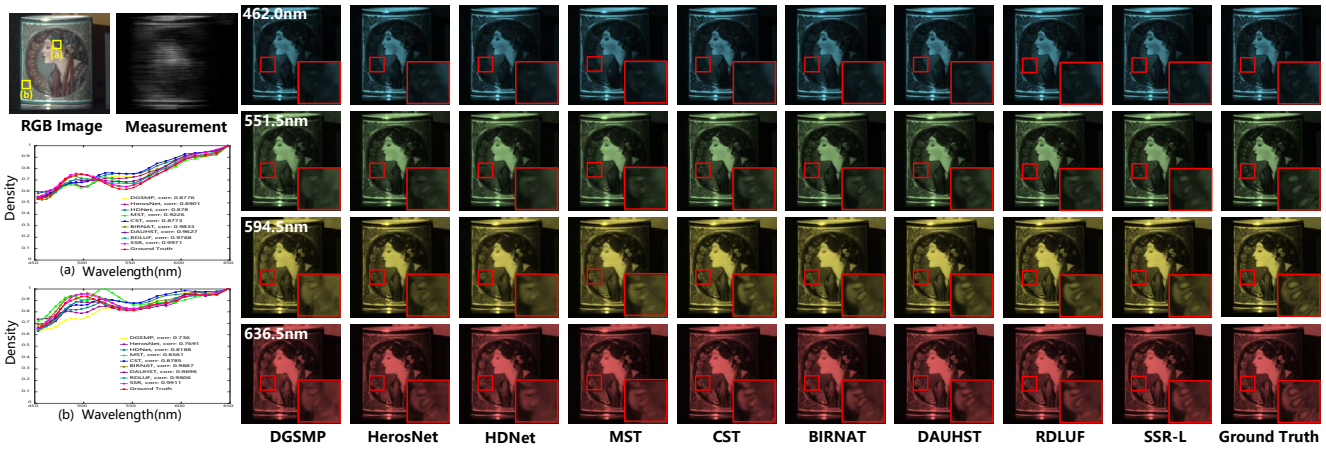

Figure 5. Reconstructed images of simulation scene1 with 4 out of 28 spectral channels by the state-of-the-art methods. Two regions in scene1 are selected for analyzing the spectra of the reconstructed results. The figure is better viewed in a zoomed-in PDF.

similarity (SSIM) [39] for quantitative evaluation. Specifically, PSNR measures the visual quality, while SSIM measures the structure similarity. Generally, higher values of PSNR and SSIM mean better reconstruction results.

5.2. Simulation Scene

Quantitative results. Table. 1 and Fig. 1 show the quantitative comparison of our SSR and the state-of-the-art (SOTA) methods: DIP-HSI [30], TSA-Net [28], DGSMP [18], GAPnet [27], HerosNet [42], HDNet [17], MST [3], CST [2], BIRNAT [8], DAUHST [5], RDLUF [11]. We can see that SSR significantly outperforms the other methods by over 2dB at the same FLOPs and exceeds the leading method RDLUF by more than 1dB with reduced computational demands. Concretely, our best model, SSR-L surpasses SOTA methods RDLUF, DAUHST, and DAUHST by 1.12, 2.33, and 3.11dB while less computational effort is required. Surprisingly, our small model SSR-S surpasses most methods when requiring the least parameters and FLOPs. Moreover, our middle model outperforms the best method RDLUF by 0.63dB when less than $1/2$ FLOPs are required. It is worth noting that RDLUF adopts a stage parameter-sharing strategy and we take the same strategy to establish our SSR-L*, which brings performance degradation but still achieves a clear advantage over other methods with the least parameters. We found that the degradation comes from the stability of the method, that is, as the number of parameters increases, the performance of our SSR steadily increases and RDLUF decreases.

Visual comparison. We provide the visual comparison of simulation scene7 with 4 out of 28 spectral channels in Fig. 5. SSR-L successfully recovers the clear pattern and sharp edge on the cup when the other methods all suffer from blurred or distorted effects. In addition, we plot the spectral curves of two regions in scene1 in the bottom-left of Fig. 5. It's intuitive that SSR-L has a prominent higher correlation with the reference spectra, which demonstrates the effectiveness of our spectra-wise attention.

Table 2. Ablation study of WSSA and ARB.

Baseline-1	WSSA	ARB	PSNR	SSIM	Params (M)	FLOPs (G)
✓			36.60	0.956	1.06	14.43
✓	✓		37.76	0.965	1.30	18.82
✓	✓	✓	39.19	0.971	1.73	26.37

5.3. Real Scene Results

To verify the effect of the proposed method on the real scenes, five measurements captured by the CASSI system are utilized for testing and the ground truths of the scenes are unavailable. For fair comparisons, all methods are trained on the CAVE and KAIST datasets jointly using the fixed real mask with 11-bit shot noise injected. The bottom two rows of Fig. 6 plot the visual comparisons of the proposed SSR-S and the existing SOTA methods while the top two rows plot the visual comparisons of different spectra-wise transformers [3, 4]. It is intuitive to see that our SSR-S restores clearer spatial textures and sharper edges than previous spectra-wise transformers in the top two rows, which shows the effect of spatial modulation. Compared with other methods in the bottom two rows, SSR-S obtains clearer results in these two bands when other methods fail to recover these two bands, demonstrating the effectiveness of our spatial alignment strategy.

5.4. Ablation Experiment

In this part, we adopt the CAVE and KAIST datasets to conduct ablation studies.

Effectiveness of WSSA and ARB. We first conduct a break-down ablation experiment to investigate the effect of each component towards higher performance. The results are listed in Table. 2. The baseline-1 model is derived by removing our WSSA and ARB from SSR-S and yields 36.60dB. When we successively apply our WSSA and ARB, the model continuously achieves 1.16dB and 1.43dB improvements. These results suggest the effectiveness of WSSA and ARB.

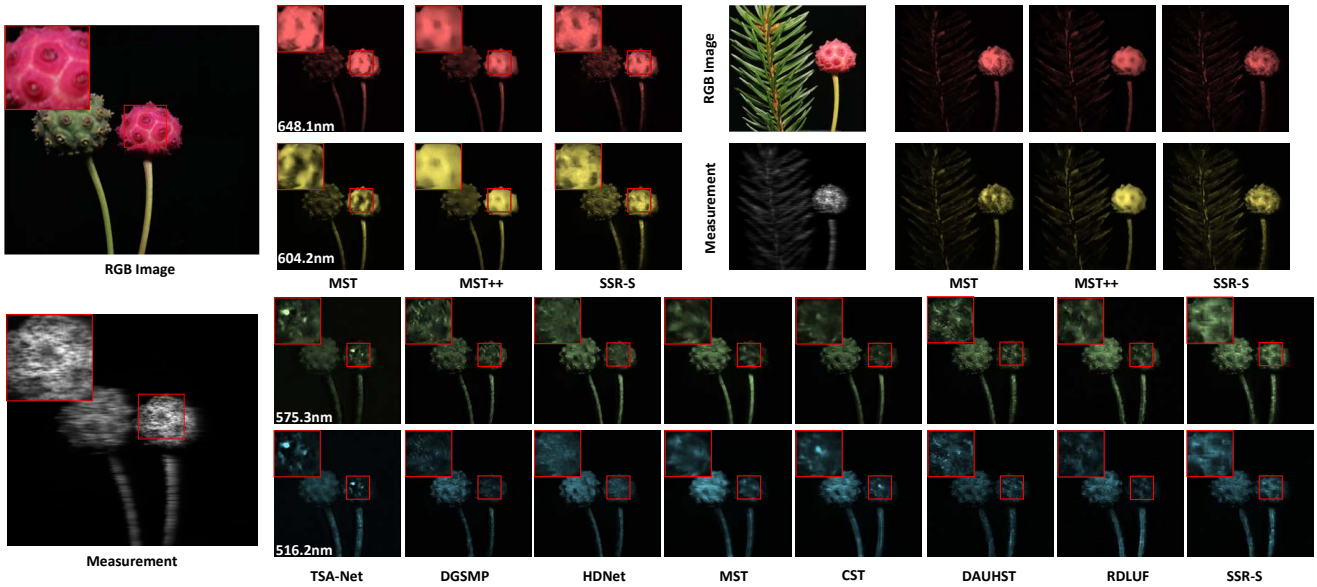

Figure 6. Reconstructed images of real scene2 and scene4 with 2 and 4 out of 28 spectral channels separately by the state-of-the-art methods. Compared with other competing methods, our SSR recovers more details and clear content.

Table 3. Ablation study of different self-attention schemes.

Metric	Baseline-2	S-MSA [3]	Swin-MSA [25]	HS-MSA [5]	WSSA*	WSSA
PSNR	32.01	32.85	33.01	33.09	33.19	33.30
SSIM	0.908	0.916	0.918	0.919	0.922	0.924
Params (M)	1.03	1.29	1.29	1.29	1.29	1.29
FLOPs (G)	13.84	18.15	19.04	19.04	18.15	18.65

Self-Attention Scheme Comparison. We compare WSSA with other self-attentions and report the results in Table. 3. We adopt the half operation [5] in Swin-MSA [25] to keep the computation almost the same and use multi-head attention in the spectral dimension of WSSA* to align with S-MSA[3]. Baseline-2 is SSR-S that retains only the ERB and removes attention and yields 32.01dB. WSSA yields the most significant improvement of 1.29dB, which is 0.45dB, 0.29dB, and 0.21dB higher than S-MSA [3], Swin-MSA [25], and HS-MSA [5], which shows the effectiveness of WSSA. In addition, WSSA* and WSSA achieve 0.34dB and 0.45dB gain than S-MSA respectively, which demonstrates the role of considering local differences and modeling global spectral information.

Influence of Multi-head attention and Mean effect. We demonstrate the influence of multi-head attention and mean effect on the spectra-wise self-attention through further experiments on the relationship between performance and token dim (window size) in WSSA and MSSA, which is shown in Fig. 7. Intuitively, the large window size would suffer from limited performance, which illustrates the influence of the mean effect. Performance degradation also occurs when the window size is too small to retain complete feature information. In our experiments, 8×8 window size optimally balances retaining feature information while minimizing the mean effect, which produces the highest performance of 33.3dB. When the token dim is small, WSSA clearly performs better than MSSA, which shows the role of maintaining spatial correlation. When the token dim is large, MSSA is slightly better than WSSA, which may be related to the shift in the imaging process and the strip pattern can contain some shift information.

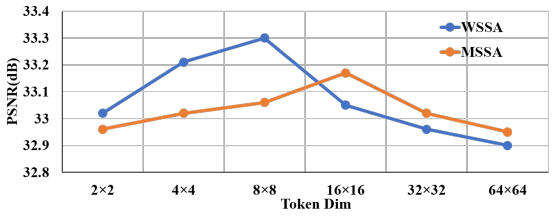

Figure 7. Influence of token dim in spectra-wise attention.

6. Conclusion

In this paper, we analyze the influence of multi-head attention and mean effect on the spectra-wise transformer, and a novel SSR method is proposed to improve spectral snapshot reconstruction. To model the global spectral information, consider the local difference, and maintain spatial correlation, WSSA is proposed to better utilize spectral similarity. ARB leverages CMB to address the interaction between adjacent windows of WSSA and learn spatial representation when SAB is specially designed to mitigate spatial degradation in low-quality bands through a novel spatial alignment strategy. Extensive experiments on simulation and real scenes show the effectiveness of the proposed modules. Our SSR at different scales also significantly outperforms the state-of-the-art methods with less cost.

Acknowledgements: This work was supported by the National Natural Science Foundation of China under Grant 62302394 and 62106063, the China Postdoctoral Science Foundation under Grant 317751, and the Natural Science Foundation of Shaanxi under Grant 2023-JC-QN-0757.

References

[1] José M. Bioucas-Dias and Mário A. T. Figueiredo. A new twist: Two-step iterative shrinkage/thresholding algorithms for image restoration. *IEEE Transactions on Image Processing*, 16:2992–3004, 2007. 1, 2

[2] Yuanhao Cai, Jing Lin, Xiaowan Hu, Haoqian Wang, Xin Yuan, Yulun Zhang, Radu Timofte, and Luc Van Gool. Coarse-to-fine sparse transformer for hyperspectral image reconstruction. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVII*, pages 686–704. Springer, 2022. 1, 2, 3, 6, 7

[3] Yuanhao Cai, Jing Lin, Xiaowan Hu, Haoqian Wang, Xin Yuan, Yulun Zhang, Radu Timofte, and Luc Van Gool. Mask-guided spectral-wise transformer for efficient hyperspectral image reconstruction. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17481–17490, 2022. 1, 2, 3, 6, 7, 8

[4] Yuanhao Cai, Jing Lin, Zudi Lin, Haoqian Wang, Yulun Zhang, Hanspeter Pfister, Radu Timofte, and Luc Van Gool. Mst++: Multi-stage spectral-wise transformer for efficient spectral reconstruction. In *CVPRW*, 2022. 1, 2, 3, 7

[5] Yuanhao Cai, Jing Lin, Haoqian Wang, Xin Yuan, Henghui Ding, Yulun Zhang, Radu Timofte, and Luc Van Gool. Degradation-aware unfolding half-shuffle transformer for spectral compressive imaging. In *Advances in Neural Information Processing Systems*, 2022. 1, 2, 3, 6, 7, 8

[6] Yuanhao Cai, Yuxin Zheng, Jing Lin, Xin Yuan, Yulun Zhang, and Haoqian Wang. Binarized spectral compressive imaging. In *Proc. Conf. Neural Inf. Process. Syst.*, 2023. 1

[7] Xiangyu Chen, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Activating more pixels in image super-resolution transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22367–22377, 2023. 5

[8] Ziheng Cheng, Bo Chen, Ruiying Lu, Zhengjue Wang, Hao Zhang, Ziyi Meng, and Xin Yuan. Recurrent neural networks for snapshot compressive imaging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):2264–2281, 2023. 6, 7

[9] Inchang Choi, Daniel S. Jeon, Giljoo Nam, Diego Gutierrez, and Min H. Kim. High-quality hyperspectral reconstruction using a spectral prior. *ACM Transactions on Graphics*, 36: 1–13, 2017. 6

[10] Xiangxiang Chu, Zhi Tian, Bo Zhang, Xinlong Wang, and Chunhua Shen. Conditional positional encodings for vision transformers. In *The Eleventh International Conference on Learning Representations*, 2022. 3

[11] Yubo Dong, Dahua Gao, Tian Qiu, Yuyan Li, Minxi Yang, and Guangming Shi. 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 (CVPR)*, pages 22262–22271, 2023. 1, 2, 3, 6, 7

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2, 3

[13] Michael E Gehm, Renu John, David J Brady, Rebecca M Willett, and Timothy J Schulz. Single-shot compressive spectral imaging with a dual-disperser architecture. *Optics express*, 15(21):14013–14027, 2007. 1

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 630–645. Springer, 2016. 2

[16] Qibin Hou, Cheng-Ze Lu, Ming-Ming Cheng, and Jiashi Feng. Conv2former: A simple transformer-style convnet for visual recognition. *arXiv preprint arXiv:2211.11943*, 2022. 5

[17] Xiaowan Hu, Yuanhao Cai, Jing Lin, Haoqian Wang, Xin Yuan, Yulun Zhang, Radu Timofte, and Luc Van Gool. Hdnet: High-resolution dual-domain learning for spectral compressive imaging. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17521–17530, 2022. 1, 2, 6, 7

[18] Tao Huang, Weisheng Dong, Xin Yuan, Jinjian Wu, and Guangming Shi. Deep gaussian scale mixture prior for spectral compressive imaging. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16211–16220, 2021. 1, 6, 7

[19] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[20] David Kittle, Kerkil Choi, Ashwin Wagadarikar, and David J Brady. Multiframe image estimation for coded aperture snapshot spectral imagers. *Applied optics*, 49(36):6824–6833, 2010. 1, 2

[21] David Kittle, Kerkil Choi, Ashwin Wagadarikar, and David J Brady. Multiframe image estimation for coded aperture snapshot spectral imagers. *Applied optics*, 49(36):6824–6833, 2010. 1, 2

[22] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. 2

[23] Xing Lin, Yebin Liu, Jiamin Wu, and Qionghai Dai. Spatial-spectral encoded compressive hyperspectral imaging. *ACM Transactions on Graphics (TOG)*, 33:1 – 11, 2014. 1, 2

[24] Yang Liu, Xin Yuan, Jinli Suo, David J. Brady, and Qionghai Dai. Rank minimization for snapshot compressive imaging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41:2990–3006, 2019. 1, 2

[25] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In

Proceedings of the IEEE/CVF international conference on computer vision, pages 10012–10022, 2021. 1, 2, 3, 8

[26] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 6

[27] Ziyi Meng, Shirin Jalali, and Xin Yuan. Gap-net for snapshot compressive imaging. *arXiv: Image and Video Processing*, 2020. 6, 7

[28] 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, 6, 7

[29] Ziyi Meng, Mu Qiao, Jiawei Ma, Zhenming Yu, Kun Xu, and Xin Yuan. Snapshot multispectral endomicroscopy. *Optics Letters*, 45(14):3897–3900, 2020. 1

[30] Ziyi Meng, Zhenming Yu, Kun Xu, and Xin Yuan. Self-supervised neural networks for spectral snapshot compressive imaging. In *2021 IEEE/CVF International Conference on Computer Vision*, pages 2602–2611, 2021. 6, 7

[31] X. Miao, X. Yuan, Y. Pu, and V. Athitsos. lambda-net: Reconstruct hyperspectral images from a snapshot measurement. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 2

[32] J. Park, M. Lee, M. D. Grossberg, and S. K. Nayar. Multispectral Imaging Using Multiplexed Illumination. In *IEEE International Conference on Computer Vision*, 2007. 6

[33] Jin Tan, Yanting Ma, Hoover F. Rueda, Dror Baron, and Gonzalo R. Arce. Compressive hyperspectral imaging via approximate message passing. *IEEE Journal of Selected Topics in Signal Processing*, 10:389–401, 2016. 1, 2

[34] Jin Tan, Yanting Ma, Hoover F. Rueda, Dror Baron, and Gonzalo R. Arce. Compressive hyperspectral imaging via approximate message passing. *IEEE Journal of Selected Topics in Signal Processing*, 10:389–401, 2016. 1, 2

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2, 3, 4

[36] 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*, 17(8): 6368–6388, 2009. 1

[37] Lizhi Wang, Chen Sun, Maoqing Zhang, Ying Fu, and Hua Huang. Dnu: Deep non-local unrolling for computational spectral imaging. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1658–1668, 2020. 1

[38] Minghua Wang, Qiang Wang, and Jocelyn Chanussot. Tensor low-rank constraint and l_0 total variation for hyperspectral image mixed noise removal. *IEEE Journal of Selected Topics in Signal Processing*, 15:718–733, 2021. 1, 2

[39] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. 7

[40] Xin Yuan. Generalized alternating projection based total variation minimization for compressive sensing. *2016 IEEE International Conference on Image Processing (ICIP)*, pages 2539–2543, 2016. 1, 2

[41] Xin Yuan, Yang Liu, Jinli Suo, and Qionghai Dai. Plug-and-play algorithms for large-scale snapshot compressive imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1447 – 1457, 2020. 2

[42] Xuanyu Zhang, Yongbing Zhang, Ruiqin Xiong, Qilin Sun, and Jian Zhang. Herosnet: Hyperspectral explicable reconstruction and optimal sampling deep network for snapshot compressive imaging. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17511–17520, 2022. 1, 6, 7

[43] Yin-Ping Zhao, Jiancheng Zhang, Yongyong Chen, Zhen Wang, and Xuelong Li. Rcump: Residual completion unrolling with mixed priors for snapshot compressive imaging. *IEEE Transactions on Image Processing*, 33:2347–2360, 2024. 6

[44] Siming Zheng, Yang Liu, Ziyi Meng, Mu Qiao, Zhishen Tong, Xiaoyu Yang, Shensheng Han, and Xin Yuan. Deep plug-and-play priors for spectral snapshot compressive imaging. *Photonics Research*, 9(2):B18–B29, 2021. 2