SSTHyper: Sparse Spectral Transformer for Hyperspectral Image Reconstruction

Meng Xu, Mingying Lin, Qi Ren, and Sen Jia *

College of Computer Science and Software Engineering, Shenzhen University, China
senjia@szu.edu.cn

Abstract. Transformer-based methods have improved the quality of hyperspectral images (HSIs) reconstructed from RGB by effectively capturing their remote relationships. The self-attention mechanisms in existing Transformer models have not fully considered the spatial sparsity and spectral continuity characteristics of HSIs and fail to effectively filter out significant features, resulting in lower-quality reconstruction results. This paper proposes a sparse spectral transformer model for HSI reconstruction (SSTHyper) to address this limitation, adaptively preserving crucial features. The network consists of an encoder-decoder structure capable of learning shallow and deep spatial-spectral priors, primarily composed of sparse spectral self-attention groups. Introducing a sparse spectral self-attention mechanism allows adaptive masking of non-significant details, enhancing reconstruction accuracy. Meanwhile, a lightweight cross-level fusion network is proposed to reduce model parameters and computational costs to enhance spatial-spectral feature extraction. Experimental results on two benchmark datasets demonstrate the outstanding performance of the proposed method. The code will be released at SSTHyper.

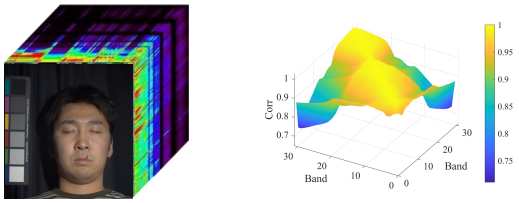
Keywords: Spectral super-resolution · Transformer · Sparse spectral self-attention

1 Introduction

Compared to natural images, hyperspectral images (HSIs) contain rich spectral information and are applied in various domains such as scene classification [41,36], object detection [59,47], semantic segmentation [39,9], and object tracking [23,45]. However, HSIs face challenges such as low spatial resolution, difficulty in acquisition, and high costs. To address this challenge, researchers have developed a new research direction, namely spectral reconstruction. The reconstruction process aims to generate HSIs from either multispectral or high-resolution RGB images.

HSI reconstruction is an ill-posed inverse problem that aims to recover tens or hundreds of HSI bands from a limited number of multispectral or RGB bands. Currently, most convolutional neural network (CNN)-based methods adopt the

* Corresponding author



(a) (b)

Fig. 1: Visualization of the inter-band similarity in an HSI image. (a) An HSI image from the CAVE dataset. (b) A three-dimensional plot illustrating the inter-band similarity of (a).

end-to-end learning framework, where they learn the intrinsic mappings of spectral and spatial signals from many samples. Different from methods that rely on manually designed features [55,33,54,48,1], CNN-based methods have demonstrated powerful implicit learning capabilities in HSI reconstruction tasks [30,24,29]. However, CNNs are constrained by their local receptive fields, limiting their ability to capture long-range dependencies, which in turn affects their overall representational capacity.

To address these limitations, Transformer [21,13] have been applied in reconstruction tasks. However, directly applying Transformers faces several important challenges. First, HSI datasets are much smaller than natural image datasets, which challenges Transformer that typically require large amounts of data for optimal performance. Second, while Transformer excel at capturing spatial relationships in images, they struggle to model important spectral dependencies in HSIs. This is because although the spectral features in HSIs are discontinuous in coordinate space, they exhibit high similarity in the spectral domain. Third, the computational complexity of the self-attention mechanism in global Transformer [18] is quadratic for each attention head, posing a bottleneck for model scalability and efficiency.

To address the issues, this paper proposes a sparse spectral transformer hyperspectral reconstruction network (SSTHyper) to enhance the performance of reconstructed HSIs. As shown in Figure. 1, the high similarity and continuity between spectral bands in HSIs reveal significant data redundancy, allowing sparse representation and inverse reconstruction to create a more lightweight model. Therefore, we develop a sparse spectral self-attention (SSSA) mechanism to model spectral similarity in the spectral dimension and construct spectral attention maps to generate sparse vectors, thereby preserving key information and masking non-important features. In addition, a cross-level fusion network (CLFN) is introduced to emphasize specific spectral features and effectively model local spectral characteristics. Simultaneously, the sparse spectral self-attention block (SSSAB) is constructed in conjunction with the SSSA mechanism. Finally, by utilizing the SSSA groups (SSSAGs), deep spatial-spectral features and contextual information of the HSIs are extracted, gradually im-

proving the reconstruction quality. In summary, the main contributions of this paper are as follows:

- A lightweight SSTHyper for spectral restoration from RGB images is developed. The model can maintain vital spectral features throughout self-attention computation and achieve an optimal balance between accuracy and efficiency.
- To enhance the key information from HSIs while filtering out irrelevant details, the high similarity and continuity between HSI adjacent bands are extracted through the SSSA mechanism.
- The proposed CLFN aims to improve the joint spatial-spectral feature extraction capabilities while simultaneously minimizing model parameters.
- Extensive experiments demonstrate that SSTHyper outperforms state-of-the-art (SOTA) methods on two benchmark datasets with the least computational costs.

2 Related Work

2.1 Spectral Reconstruction

Deep learning-based spectral reconstruction approaches have shown remarkable performance with data-driven strategies. In the early stages, most approaches are based on CNN [30,26,38,25] or GAN architectures [2,32,56]. For example, Xiong *et al.* [46] introduced the HSCNN network, which utilizes RGB images and compressed measurements to reconstruct HSIs. Building upon HSCNN, residual and dense connections were employed to prevent vanishing gradients in the neural network [37,20]. Furthermore, Zhang *et al.* [52] designed a pixel-aware deep function fusion network to learn different receptive fields, modeling from RGB to HSI. TSA-Net [34] utilizes spatial-spectral self-attention to achieve band recovery in the HSI. In the NTIRE 2022 Spectral Reconstruction Challenge [5], dense residual channel re-calibration network (DRCR) [28] achieved third place due to its outstanding performance. GAN-based approaches have the generator-discriminator structure and reconstruct the HSI by the adversarial learning between these two networks. Overall, CNNs struggle to capture global information, particularly with weak modeling capabilities for long-range dependencies. GANs heavily rely on the quality and diversity of the training dataset. Insufficient RGB-HSI paired training data may result in suboptimal performance.

2.2 Transformer-based HSI Reconstruction

Transformer [40] was initially introduced in the field of natural language processing and has been recently applied in computer vision tasks such as image restoration [58,17,14,27], object detection [19,16], and image classification [44,35]. Many researchers have also applied the Transformer to spectral reconstruction tasks. Cai *et al.* pioneered the end-to-end application of the Transformer for HSI reconstruction [7]. Building on this, MST++ [8] utilized the sparse spatial and

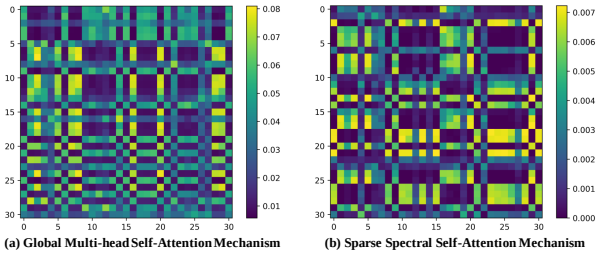


Fig. 2: Visualizations of original and sparse attention maps. Darker colors highlight regions of denser information, indicating more focused attention areas.

spectral self-similarity characteristics of HSIs by designing spectral multi-head self-attention and employing a U-shaped structure to extract multi-resolution contextual information for spectral reconstruction. Meanwhile, further embedding HSI sparsity into deep learning for HSI reconstruction [6]. Wang *et al.* [42] designed a local window attention mechanism and incorporated it into a U-Net architecture to enhance reconstruction in local regions of HSI. In contrast, the cross aggregation transformer (CAT) model [15] extended the local window to a rectangular shape. As illustrated in Figure. 2, we present the attention maps for the global multi-head attention mechanism (G-MSA) and our proposed SSSA approach. We have conducted a thorough analysis of the computational complexity of the SSSA and compared it with the G-MSA [18]. The computational complexities are expressed as:

$$O(\text{SSSA}) = \frac{HWC^2}{k}, \quad O(\text{G-MSA}) = (HW)^2C, \tag{1}$$

where H, W and C denotes the height, the weight, channels, respectively, and k denotes the top k most informative elements. The computational complexity of SSSA is linear with the spatial dimensions HW, making it more efficient than G-MSA, which has high computational demands. In response, this paper introduces a sparse attention mechanism tailored for spectral reconstruction, which effectively addresses the limitations of G-MSA in exploring spectral features.

3 Method

3.1 Network Architecture

Our goal is to develop a model to reconstruct HSIs from RGB images. This task can be formulated as follows:

$$\mathbf{Y} \cdot \mathbf{A} = \mathbf{X}, \tag{2}$$

where $\mathbf{X} \in \mathbb{R}^{H\times W\times 3}$ represent the RGB image, and $\mathbf{Y} \in \mathbb{R}^{H\times W\times C}$ denote the HSI reconstructed from $\mathbf{X}$. Here, C signifies the number of channels in the

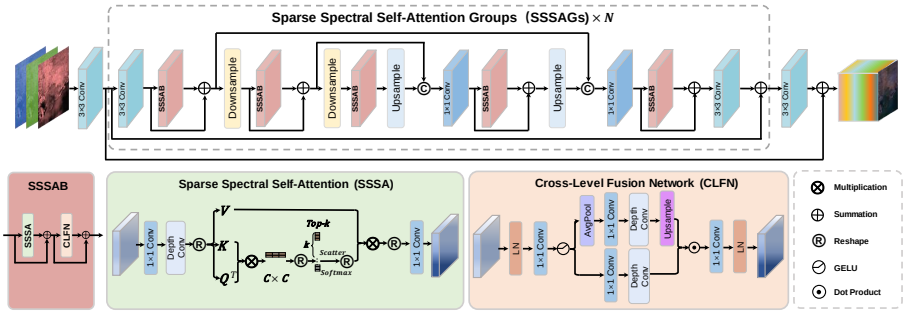


Fig. 3: Illustration of sparse spectral transformer hyperspectral image reconstruction network (SSTHyper).

HSI, while H and W correspond to the height and width, respectively. $\mathbf{A} \in \mathbb{R}^{C\times 3}$ denotes the transformation matrix from $\mathbf{Y}$ to $\mathbf{X}$. We assume that through extensive data-driven learning, it is possible to obtain $\mathbf{A}$ and address the issue of the ill-posed inverse problem.

As illustrated in Figure. 3, the SSTHyper adopts a single encoder-decoder architecture for spectral reconstruction, featuring N SSSAGs as its main components. N is set to 3 according to the analysis in the ablation experiment (Section 4.3). To achieve spectral reconstruction, the RGB image $\mathbf{X}$ is first processed through a $3 \times 3 Conv$ layer to generate shallow features consistent with the number of channels in the HSIs. Before entering the SSSAGs, shallow features are preserved and feature loss during processing is reduced by integrating them with the output through skip connections. The SSSAGs consist of multiple sparse spectral self-attention blocks (SSSABs) and convolutional layers forming long-short connections, used for encoding and decoding deep features. Further, extract deep spatial and spectral contextual features through downsampling and upsampling layers. The SSSAB serves as an intermediate transition layer for extracting spatial and spectral features. To prevent information loss during deep feature extraction, the residual connection features are fused with the upsampling output features. The output of the SSSAGs is obtained by summing the feature map and the output of the jump-joining connection. Finally, the feature maps from three SSSAGs, along with those from the $3\times3 Conv$ layer, are summed with the original features preserved by the long jump connections to obtain the reconstructed HSIs.

3.2 Sparse Spectral Self-Attention

In the standard Transformer self-attention mechanism, the process entails computing an attention map for query $\mathbf{Q}$, key $\mathbf{K}$ and value $\mathbf{V}$ with the size of $R \times L \times d$. Here, R denotes the number of heads, L represents the sequence length, and d is the embedding dimension. The attention map computation is

commonly denoted as:

$$\text{Att}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}}\right)\mathbf{V}. \tag{3}$$

By employing multiple heads for $\mathbf{Q}$, $\mathbf{K}$ and $\mathbf{V}$, the outputs of dimension C/R are obtained where C is the number of channels, and these outputs are concatenated to yield the final output. The standard self-attention mechanism relies on dense fully-connectedness, maintaining attention weights for all query-key pairs without additional processing of these weights.

Due to the high similarity among adjacent bands in HSIs, as shown in Figure 1. Meanwhile, if only all channel features are calculated in the attention mechanism [8], it will increase the computational cost and be vulnerable to the interference and influence of other subtle features. Therefore, inspired by DRSformer [12], we designed the SSSA module as shown in Figure 3. This mechanism handles channel attention weights, retaining significant features in attention while ignoring irrelevant features, thereby enhancing the computational efficiency of the model.

We first contextually encode the channels by applying a $1 \times 1 Conv$ layer and a $3 \times 3 DepthConv$ layer, expanding the feature channels to three times their original size [49]. Assuming the input is denoted as $\mathbf{X}_{in}$, the inputs $\mathbf{Q} = \mathbf{X}_{in}^q$, $\mathbf{K} = \mathbf{X}_{in}^k$, $\mathbf{V} = \mathbf{X}_{in}^v$ of the attention function can be generalized as:

$$\mathbf{X}_{in}^{q,k,v} = W_p W_d(\mathbf{X}_{in}), \tag{4}$$

where $W_p W_d$ respectively represents $1 \times 1 Conv$ and $3 \times 3 DepthConv$. In the subsequent stages, the $L2$ norms of the matrices $\mathbf{Q}$ and $\mathbf{K}$ are computed to optimize the model's stability during training. Subsequently, matrix multiplication is applied in the spatial-spectral dimension between $\mathbf{K}$ and $\mathbf{Q}^T$, resulting in the derivation of an attention matrix $\mathbf{M}$ with dimensions $C \times C$.

Considering that attention maps contain redundant features and interfere with the reconstruction process, the top-k elements are selected by prioritizing the parts with the highest weights in the attention maps to ensure focused attention on effective key features, thereby enhancing computational efficiency. In the attention maps, we select to retain the top-k elements based on a threshold, implicitly capturing the interrelations between channels, thereby transforming them into sparse attention maps. The top-k strategy minimizes redundancy while efficiently retaining essential information (Section 4.3). SSSA treats the entire spectral feature map as a single marker, allowing for a global receptive field. Our proposed mechanism selectively focuses on the most informative first k elements, lowering the overall computational complexity.

Following this, $\mathbf{M}$ is reshaped into a vector with C^2 elements, where the first k elements are selectively retained. Values smaller than the minimum value below the top-k elements are set to 0 using the scatter function, resulting in the creation of a new sparse vector. Subsequently, the vector is reverted to a matrix with dimensions $C \times C$ through the softmax function. The obtained attention matrix effectively retains essential global information from the original

attention map while masking information with lower attention weights. Subsequently, the resulting matrix is subjected to multiplication with the matrix $\mathbf{V}$. This computational process can be succinctly summarized as

$$\mathrm{SSSAtt}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathrm{softmax}\left(T_k(v)\right)\mathbf{V}, \tag{5}$$

where v represents the one-dimensional vector reshaped from $\mathbf{M}$, and T_k denotes the operation of selecting the values of the first k largest elements, along with the operation of setting the values of elements smaller than the first k largest elements to 0. The details of T_k can be summarized as follows:

$$T_k(v_i) = \begin{cases} v_i, & \text{if } v_i \geq \min(\text{top-}k), \\ 0, & \text{otherwise}, \end{cases} \tag{6}$$

where v_i denotes the value of the i-th element of the vector v. Finally, the resulting feature map is fed into the $1 \times 1 Conv$ for output mapping, yielding the final output.

3.3 Cross-Level Fusion Network

The features obtained through SSSA are inputted into the CLFN module for processing. First, layer normalization (LN), $1 \times 1 Conv$ layer, and gaussian error linear unit (GELU) function processing are applied to obtain the feature maps. The $1 \times 1 Conv$ layer serves to double the size of the feature channels. For the input feature denoted as $\mathbf{X}^{''} \in \mathbb{R}^{H \times W \times D}$, the process is succinctly expressed as:

$$\hat{\mathbf{X}} = \phi(W_p(\mathrm{LN}(\mathbf{X}^{''}))), \tag{7}$$

where ϕ and W_p symbolize GELU and the $1 \times 1 Conv$ layer, respectively. In the CLFN mechanism, for further feature enhancement, the input features are divided along the channel into two features of the same dimension, each of which is input into two different branches. Both branches utilize a depth-separable convolution, encompassing point-by-point convolution and depth convolution, to capture information from neighboring pixel positions in the spatial domain. The distinction lies in the utilization of spatial average pooling on one branch, resulting in a reduction of the spatial dimensions of the feature map to half its original size. Conventional channel attention mechanisms utilizing global pooling may pose the risk of interferences among spatially distant pixels, potentially compromising the model's performance. To overcome these challenges, our study incorporates spatial average pooling to systematically reduce the spatial dimensions of the feature map, mitigating such interferences and effectively decreasing the size of the local features to half of the original aggregated feature map. To enhance local prominent spectral features, the features of the upsampling branch are upsampled to restore the initial feature size. Subsequently, the features of the two branches are multiplied element-wise to highlight important features, achieving cross-level feature fusion.

Table 1: Comparisons with SOTA methods on CAVE and NTIRE2022 HSI datasets

Model	CAVE				NTIRE2022				Params (M)	FLOPs (G)	FPS (s)
	RMSE↓	PSNR↑	SSIM↑	SAM ↓	MRAE↓	RMSE ↓	PSNR↑	SSIM↑			
EDSR [31]	5.85	33.78	0.9761	7.87	0.3637	0.052	27.01	0.8676	2.45	36.86	0.1924
HSCNN+ [37]	14.75	26.46	0.9381	7.47	0.3849	0.059	26.29	0.8281	4.65	70.88	0.2123
HRNet [57]	5.25	34.86	0.9808	6.92	0.3567	0.053	27.28	0.8520	31.70	38.14	0.2505
FMNet [52]	7.63	31.68	0.9728	8.05	0.3377	0.048	27.69	0.8743	11.78	179.67	0.2313
MIRNet [50]	5.29	35.11	0.9817	6.81	0.2190	0.030	32.28	0.9448	3.57	6.44	0.1917
MPRNet [51]	5.39	34.85	0.9836	**5.97**	0.1818	0.025	34.25	0.9524	3.62	61.75	0.1997
HINet [11]	5.08	35.22	0.9755	7.51	0.2392	0.032	31.89	0.9391	4.98	7.29	0.1892
HDNet [22]	5.65	34.41	0.9801	6.71	0.2785	0.038	29.81	0.9058	2.66	40.46	0.2109
Restormer [49]	5.15	35.36	0.9834	6.53	0.1764	0.026	33.66	0.9494	15.09	19.93	0.1987
MST [7]	5.11	35.24	0.9821	6.72	0.2655	0.032	31.26	0.9537	2.45	7.10	0.2556
MST++ [8]	5.17	35.01	0.9820	7.68	0.1646	0.025	34.32	0.9546	1.62	5.43	0.1950
CST [6]	5.59	34.01	0.9797	7.10	0.2625	0.039	29.49	0.9076	3.67	7.95	0.1648
ESSAformer [53]	6.54	33.57	0.9796	6.91	0.2847	0.031	31.29	0.9492	6.11	231.27	0.1788
RepCPSI [43]	5.51	34.70	0.9805	6.71	0.3009	0.052	28.65	0.8815	2.07	31.63	0.2097
SSTHyper (Ours)	**4.83**	**35.62**	**0.9841**	6.51	**0.1584**	**0.023**	**34.64**	**0.9576**	**1.39**	**4.55**	**0.2598**

After the feature fusion of the dual branches, the features are processed through $1 \times 1 Conv$ and LN to further extract channel features, speeding up model training. In the CLFN mechanism, the two branches can be represented as:

$$\tilde{\mathbf{X}} = \mathrm{CLFN}(\hat{\mathbf{X}}), \quad \mathrm{CLFN}(\hat{\mathbf{X}}) = \mathrm{LN}(W_p(\mathrm{CLFN}_1(\hat{\mathbf{X}}) \odot \mathrm{CLFN}_2(\hat{\mathbf{X}}))), \tag{8}$$

$$\mathrm{CLFN}_1(\hat{\mathbf{X}}) = \mathrm{Up}(W_p W_d(\mathrm{Avg}(\hat{\mathbf{X}}))), \quad \mathrm{CLFN}_2(\hat{\mathbf{X}}) = W_p W_d(\hat{\mathbf{X}}), \tag{9}$$

where Avg denotes average pooling operation and Up denotes nearest neighbor interpolation. The proposed CLFN in this paper effectively enhances the model's spatial and spectral feature extraction capabilities, thereby improving the overall performance of the model. Compared to traditional feedforward layer networks, which expand feature channels to four times their original size, the CLFN proposed in this paper only requires expanding feature channels to twice their original size, reducing the model's parameters and computational burden. Additionally, it enhances local prominent spectral features. The SSSAB module, delineated in Figure. 3, first extracts spatial-spectral contextual information of the HSI using SSSA, preserving representative key features. Then, it enters the CLFN to enhance the extraction of spatial-spectral features. Skip connections are added to reduce information loss during this process.

4 Experiment

4.1 Experimental Settings

Datasets. To evaluate our work, we conducted experiments using two HSI datasets: CAVE [10] and NTIRE2022 [5]. The CAVE dataset consists of 32 pairs

of RGB-HSI images, each with a resolution of 512 × 512 pixels and covering 31 spectral bands from 400 nm to 700 nm in 10 nm intervals. The NTIRE2022 dataset, designated for the Spectral Reconstruction Competition of 2022, contains 1000 pairs of images, with 900 for training, 50 for validation, and 50 for testing. The spatial size of each image in the NTIRE2022 dataset is 482 × 512 pixels, covering 204 spectral bands from 400 nm to 1000 nm. To ensure consistency, all images have been resampled to contain 31 spectral bands from 400 nm to 700 nm, with a uniform interval of 10 nm.

Implementation details. For training, paired RGB and HSI are cropped to 128 × 128. The batch size is set to 20. The SSTHyper model is optimized using the ADAM optimizer with β_1=0.9 and β_2=0.99, and epsilon set to 1e-8. The initial learning rate is 0.0004, and training is performed for 300 epochs on the NTIRE2022 dataset, and 100 epochs on CAVE.

In the NTIRE2022 dataset, the mean relative absolute error (MRAE) is used as the loss function. However, in the CAVE dataset, the presence of zero values precipitates a phenomenon of gradient explosion when employing MRAE as the loss function. Consequently, the mean absolute error (MAE) is selected as the loss function during the training phase.

Evaluation metrics. We employ various metrics to assess the performance of SSTHyper, including root mean square error (RMSE), peak signal-to-noise ratio (PSNR), structural similarity index measure (SSIM), spectral angle mapper (SAM), and MRAE.

Comparison methods. SSTHyper is benchmarked against fourteen SOTA methods to assess its effectiveness across CAVE and NTIRE2022 datasets, including EDSR [31], HSCNN+ [37], HRNet [57], FMNet [52], MIRNet [50], MPRNet [51], HINet [11], HDNet [22], Restormer [49], MST [7], MST++ [8], CST [6], ESSAformer [53] and RepCPSI [43]. It is worth noting that HSCNN+, HRNet, and MST++ respectively achieved first place in the NTIRE 2018 [3], NTIRE 2020 [4], and NTIRE 2022 [5] Spectral Reconstruction Challenges. Restormer, CST, MST, MST++, and ESSAformer are spectral reconstruction methods based on the Transformer architecture.

4.2 Reconstruction Results

Quantitative comparisons. The quantitative assessment results for CAVE and NTIRE2022 datasets are presented in Table 1. The results demonstrate that our method obtains consistent improvements over the SOTA methods. In the CAVE dataset, SSTHyper surpasses MST with a 0.28 RMSE reduction, and 0.21 SAM decrease. Compared to MPRNet, SSTHyper achieves a 0.0005 SSIM improvement and a 0.77 dB PSNR increase. On the NTIRE2022 dataset, SSTHyper achieved SOTA performance, surpassing the previous SOTA method MST++ with notable reductions in MRAE and RMSE by 3.77% and 8.00%, respectively, and improvements in PSNR and SSIM by 0.32 dB and 0.003, respectively. This advancement highlights the effectiveness and superiority of SSTHyper on the given datasets. We provide frames per second (FPS) for each model,

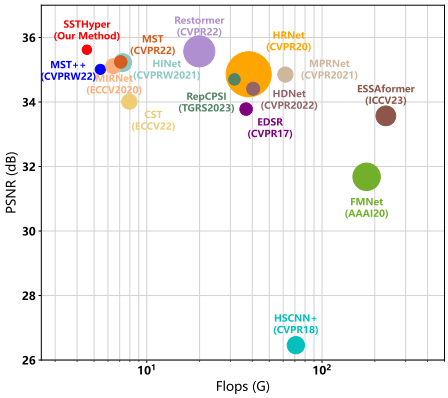


Fig. 4: The proposed SSTHyper surpasses other spectral reconstruction algorithms in terms of PSNR (performance) while significantly reducing FLOP (computational complexity).

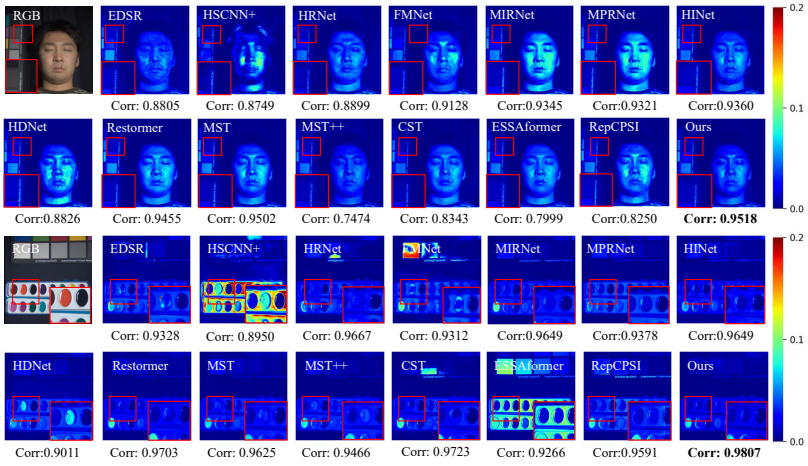
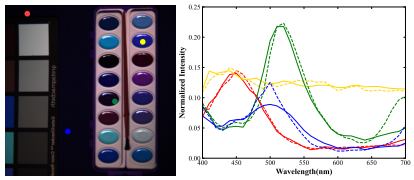


Fig. 5: The reconstructed error maps of two images in the validation set of the CAVE dataset by all methods. The coefficients below the image represent the average correlation coefficients between the reconstruction results and the ground truth spectra within the red region.

where SSTHyper exhibits the highest FPS, underscoring its superior computational efficiency. This represents the most efficient computational performance among all the compared methods in Figure. 4.

Visual results. To further assess the efficacy of the proposed method in terms of visual perception, we selected the 8th and 23rd images from the test set of CAVE. As shown in Figure. 5, the RMSE heatmap (along the spectral direction) between the reconstructed HSIs and the ground truths was analyzed



(a) (b)

Fig. 6: (a) Randomly annotate four points on the RGB images in the CAVE dataset. (b) Visualize the spectral curves of ground truth and HSI reconstructed by the proposed algorithm at four points, represented by solid and dashed lines, respectively.

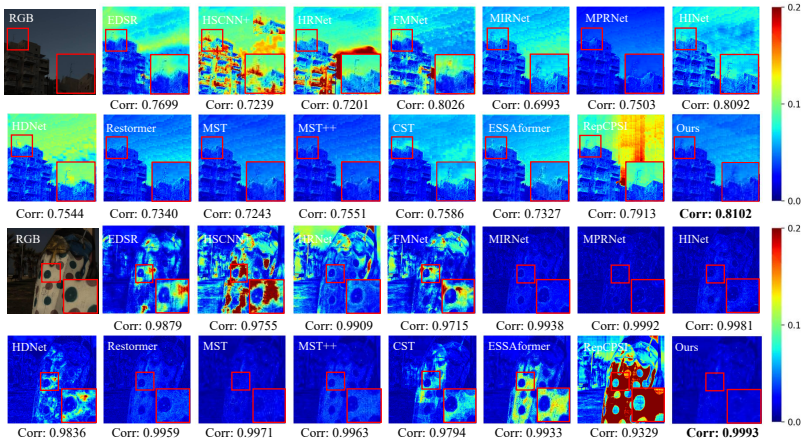


Fig. 7: The reconstructed error maps of two images in the validation set of the NTIRE2022 dataset by all methods. The coefficients below the image represent the average correlation coefficients between the reconstruction results and the ground truth spectra within the red region.

and visualized. Most algorithms exhibit noticeable errors in the color palette and around the center of the face in the reconstructed images. The images of human portraits displayed in the first and second rows reveal evident distortion in the reconstructions produced by HSCNN+, HDNet, MIRNet, and MPRNet. For the pigment board scene, in contrast to MST++, SSTHyper shows more detailed handling of texture in fine details while maintaining smoothness in homogeneous regions. We annotate a fixed region in red and calculate the spectral similarity between this region and the ground truth. It can be observed that our proposed method exhibits higher similarity to the ground truth.

Furthermore, in Figure. 6(a), four random points are selected to visualize the spectral curves. From the bright and dark areas in the images, it can be

observed that the reconstructed spectral response curves exhibit high similarity to the ground truth.

The 1st and 30th images from the validation data in NTIRE2022 are selected in this study (Figure. 7), comparing error maps of reconstruction HSIs are presented. From the visualization results, in the first scenario, most algorithms exhibit evident errors in delineating the outlines of buildings and depicting the sky scene. It can also be observed that the images reconstructed by EDSR, HSCNN+, FMNet, HRNet, HDNet, ESSAformer, and RepCPSI exhibit significant errors on the 30th image. Although there is very little difference in visual appearance between MPRNet, MST, and SSTHyper, the correlation coefficients between the reconstructed HSIs and the ground truth spectra are computed, as shown below each image in Figure. 7. By comparing the correlation coefficients, the proposed method exhibits better spectral reconstruction performance in both dark and bright regions.

4.3 Ablation Study

In this section, we conducted ablation experiments on the CAVE dataset to validate the effectiveness of SSSA and CLFN modules in SSTHyper.

Table 2: The impact of the self-attention module on the performance of the proposed method on the CVAE dataset

Model	RMSE↓	PSNR↑	SSIM↑	Params (M)	FLOPS (G)
Baseline	5.97	33.48	0.9797	1.07	15.25
Baseline+TKSA	5.12	35.07	0.9822	1.39	19.55
Baseline+S-MSA	5.72	34.30	0.9828	1.38	19.39
Baseline+MDTA	5.13	35.08	0.9835	1.39	19.56
Baseline+SSSA	**4.83**	**35.62**	**0.9841**	1.39	19.55

Effectiveness of SSSA. To assess the efficacy of SSSA, we conducted ablation experiments comparing three spectral self-attention mechanisms: top-k sparse attention (TKSA) [12], spectral-wise multi-head self-attention (S-MSA) [8], and multi-dimensional convolutional head transposed self-attention (MDTA) [49]. The corresponding results are documented in Table 2. Here, the baseline represents the backbone model without the self-attention module. Compared to the baseline, SSSA significantly improves performance, with RMSE decreasing by 1.14 and PSNR increasing by 2.14. Furthermore, compared to TKSA, S-MSA, and MDTA, RMSE decreases by 0.29, 0.89, and 0.3, respectively. Moreover, the complexity of SSSA is consistent with other self-attention mechanisms. Additionally, we further provide visual analysis in Figure. 8, indicating that SSSA can better integrate details, especially in prominent areas.

Effectiveness of CLFN. To assess the effectiveness of incorporating CLFN layer, we conducted ablation experiments comparing two feedforward networks:

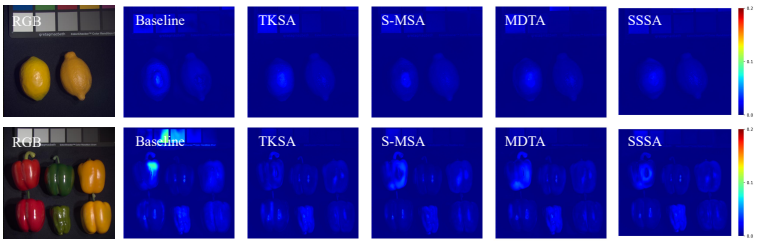


Fig. 8: Ablation qualitative comparison for different self-attention modules of SSTHyper. The models are consistent with the settings in Table 2.

Table 3: The impact of the feedforward network on the performance of the proposed method on the CVAE dataset

Model	RMSE ↓	PSNR ↑	SSIM ↑	Params (M)	FLOPS (G)
FFN	5.09	35.43	0.9827	1.63	23.41
GDFN	5.21	35.08	0.9828	1.63	23.58
CLFN	**4.83**	**35.62**	**0.9841**	**1.39**	**19.55**

the feedforward network (FFN) [8] employed by MST++ and the gated-dconv feedforward network (GDFN) [49]. The results are presented in Table 3. Compared to FFN and GDFN, GFFN reduces the model parameters by 14.72%, decreases the floating-point operations by 16.49% and 17.10%, respectively. Meanwhile, the RMSE also decreases to a certain extent compared to the latter two models, while the PSNR increases by 0.19 dB and 0.54 dB. These experimental results verify that the proposed CLFN model has the advantage of lightweightness, which also promotes performance improvement.

Table 4: The performance of different numbers of SSSAGs on the CAVE

N	RMSE ↓	PSNR ↑	SSIM ↑	SAM ↓	Params (M)	FLOPS (G)
1	5.12	34.93	0.9816	7.11	0.47	1.61
3	**4.83**	**35.62**	**0.9841**	6.51	**1.39**	**4.55**
5	5.2	35.18	0.9831	**6.26**	2.31	7.49
7	4.84	35.28	0.9839	6.55	3.08	9.87

The number of SSSAGs. To evaluate the effect of the number of SSSAGs modules, we use different N values to conduct the experiments. The results are recorded in Table 4. When $N = 1$, both the parameter count and performance of the method are relatively low. When $N > 3$, the performance of the method improves, but the computational complexity increases. Therefore, setting N to 3 maintains stable performance and computational efficiency for the method.

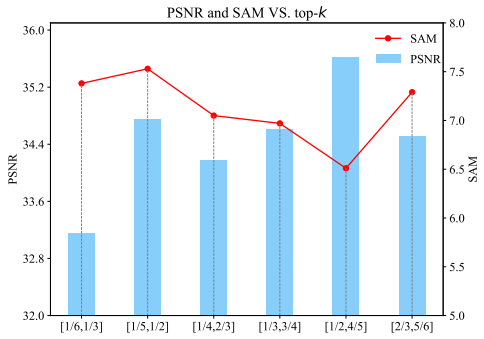


Fig. 9: Ablation qualitative comparison for different number k in the SSSA.

The number of k. The key parameter of our proposed SSSA is k, and its influence is investigated in Figure. 9. The optimal choice of k determines the sparsity rate. To avoid an exhaustive search, we set a controllable interval range of k to dynamically learn the most contribute score. When k is small, we find that the performance will undoubtedly decline sharply due to insufficient global information aggregation. The best result can achieve 32.18 dB and 6.51 for PSNR and SAM when the range in the SSSA is assigned to [1/2, 4/5]. As k continues to increase, the final reconstructed performance gradually decreases due to the introduction of irrelevant and useless features.

5 Conclusion

We introduce a novel Transformer network, namely SSTHyper, specifically designed for HSI reconstruction tasks. By fully considering the similarity of adjacent channels in HSI, we employ the SSSA mechanism to preserve crucial information. Simultaneously, the CLFN is introduced to enhance the computational efficiency of the network. The overall architecture of the network consists of an encoder-decoder, capable of learning deep contextual information in the spatial-spectral domain. Due to its unique design, our model can effectively model remote dependencies and produce superior reconstruction results with reduced computational costs. Extensive experimental results validate the effectiveness of this approach, and its performance surpasses current SOTA algorithms.

Acknowledgments. This work was supported in part by the National Natural Science Foundation of China under Grant 42271336, 62271327, and 41971300; in part by Natural Science Foundation of Guangdong Province under Grant 2024A1515011079 and 2022A1515011290; in part by the Shenzhen Science and Technology Program under Grant RCJC20221008092731042 and JCYJ20220818100206015; and in part by Guangdong Province Key Laboratory of Popular High Performance Computers 2017B030314073 and Guangdong Province Engineering Center of China-made High Performance Data Computing System.

References

1. Aeschbacher, J., Wu, J., Timofte, R.: In defense of shallow learned spectral reconstruction from RGB images. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. pp. 471–479 (2017)
2. Alvarez-Gila, A., Van De Weijer, J., Garrote, E.: Adversarial networks for spatial context-aware spectral image reconstruction from rgb. In: Proceedings of the IEEE international conference on computer vision workshops. pp. 480–490 (2017)
3. Arad, B., Ben-Shahar, O., Timofte, R., Van Gool, L., Zhang, L., Yang, M., et al.: NTIRE 2018 Challenge on Spectral Reconstruction from RGB Images. In: CVF conference on computer vision and pattern recognition workshops (CVPRW). pp. 1042–1042 (2018)
4. Arad, B., Timofte, R., Ben-Shahar, O., Lin, Y.T., Finlayson, G.D.: NTIRE 2020 Challenge on Spectral Reconstruction From an RGB Image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 446–447 (2020)
5. Arad, B., Timofte, R., Yahel, R., Morag, N., Bernat, A., Cai, Y., Lin, J., Lin, Z., Wang, H., Zhang, Y., et al.: NTIRE 2022 Spectral Recovery Challenge and Data Set. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 863–881 (2022)
6. Cai, Y., Lin, J., Hu, X., Wang, H., Yuan, X., Zhang, Y., Timofte, R., Van Gool, L.: Coarse-to-Fine Sparse Transformer for Hyperspectral Image Reconstruction. In: European Conference on Computer Vision. pp. 686–704. Springer (2022)
7. Cai, Y., Lin, J., Hu, X., Wang, H., Yuan, X., Zhang, Y., Timofte, R., Van Gool, L.: Mask-guided Spectral-wise Transformer for Efficient Hyperspectral Image Reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17502–17511 (2022)
8. Cai, Y., Lin, J., Lin, Z., Wang, H., Zhang, Y., Pfister, H., Timofte, R., Van Gool, L.: MST++: Multi-stage Spectral-wise Transformer for Efficient Spectral Reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 745–755 (2022)
9. Cao, X., Zhou, F., Xu, L., Meng, D., Xu, Z., Paisley, J.: Hyperspectral Image Classification With Markov Random Fields and a Convolutional Neural Network. IEEE Trans. Image Processing **27**, 2354–2367 (2017)
10. Chakrabarti, A., Zickler, T.: Statistics of real-world hyperspectral images. In: CVPR 2011. pp. 193–200. IEEE (2011)
11. Chen, L., Lu, X., Zhang, J., Chu, X., Chen, C.: HINet: Half Instance Normalization Network for Image Restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. pp. 182–192 (June 2021)
12. Chen, X., Li, H., Li, M., Pan, J.: Learning A Sparse Transformer Network for Effective Image Deraining. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5896–5905 (2023)
13. Chen, Y., Dai, X., Chen, D., Liu, M., Dong, X., Yuan, L., Liu, Z.: Mobile-Former: Bridging MobileNet and Transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5270–5279 (2022)
14. Chen, Z., Zhang, Y., Gu, J., Kong, L., Yang, X., Yu, F.: Dual Aggregation Transformer for Image Super-Resolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 12312–12321 (October 2023)

15. Chen, Z., Zhang, Y., Gu, J., Kong, L., Yuan, X., et al.: Cross Aggregation Transformer for Image Restoration. Advances in Neural Information Processing Systems **35**, 25478–25490 (2022)
16. Cui, Y., Yang, L., Yu, H.: Learning dynamic query combinations for transformer-based object detection and segmentation. In: International Conference on Machine Learning. pp. 6591–6602 (2023)
17. Cui, Y., Ren, W., Yang, S., Cao, X., Knoll, A.: IRNeXt: Rethinking convolutional network design for image restoration. In: International Conference on Machine Learning (2023)
18. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
19. Fang, Y., Liao, B., Wang, X., Fang, J., Qi, J., Wu, R., Niu, J., Liu, W.: You only look at one sequence: Rethinking transformer in vision through object detection. Advances in Neural Information Processing Systems **34**, 26183–26197 (2021)
20. Galliani, S., Lanaras, C., Marmanis, D., Baltsavias, E., Schindler, K.: Learned Spectral Super-Resolution. arXiv preprint arXiv:1703.09470 (2017)
21. He, J., Yuan, Q., Li, J., Xiao, Y., Liu, X., Zou, Y.: DsTer: A dense spectral transformer for remote sensing spectral super-resolution. International Journal of Applied Earth Observation and Geoinformation **109**, 102773 (2022)
22. Hu, X., Cai, Y., Lin, J., Wang, H., Yuan, X., Zhang, Y., Timofte, R., Van Gool, L.: Hdnet: High-resolution dual-domain learning for spectral compressive imaging. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17542–17551 (2022)
23. Iwaneczko, P., Jedrasiak, K., Daniec, K., Nawrat, A.: A Prototype of Unmanned Aerial Vehicle for Image Acquisition. In: International Conference on Computer Vision and Graphics (2012)
24. Jiang, J., Sun, H., Liu, X., Ma, J.: Learning Spatial-Spectral Prior for Super-Resolution of Hyperspectral Imagery. IEEE Transactions on Computational Imaging **6**, 1082–1096 (2020)
25. Kaya, B., Can, Y.B., Timofte, R.: Towards spectral estimation from a single RGB image in the wild. In: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). pp. 3546–3555. IEEE (2019)
26. Koundinya, S., Sharma, H., Sharma, M., Upadhyay, A., Manekar, R., Mukhopadhyay, R., Karmakar, A., Chaudhury, S.: 2D-3D CNN based architectures for spectral reconstruction from RGB images. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 844–851 (2018)
27. Li, A., Zhang, L., Liu, Y., Zhu, C.: Feature Modulation Transformer: Cross-Refinement of Global Representation via High-Frequency Prior for Image Super-Resolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 12514–12524 (October 2023)
28. Li, J., Du, S., Wu, C., Leng, Y., Song, R., Li, Y.: DRCR Net: Dense Residual Channel Re-calibration Network with Non-local Purification for Spectral Super Resolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1259–1268 (2022)
29. Li, Y., Zhang, L., Ding, C., Wei, W., Zhang, Y.: Single Hyperspectral Image Super-Resolution with Grouped Deep Recursive Residual Network. 2018 IEEE Fourth International Conference on Multimedia Big Data (BigMM) pp. 1–4 (2018)
30. Li, Y., Hu, J., Zhao, X., Xie, W., Li, J.: Hyperspectral image super-resolution using deep convolutional neural network. Neurocomputing **266**, 29–41 (2017)

31. Lim, B., Son, S., Kim, H., Nah, S., Mu Lee, K.: Enhanced Deep Residual Networks for Single Image Super-Resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 136–144 (2017)
32. Liu, P., Zhao, H.: Adversarial networks for scale feature-attention spectral image reconstruction from a single rgb. Sensors **20**(8), 2426 (2020)
33. Liu, Y., Yuan, X., Suo, J., Brady, D.J., Dai, Q.: Rank Minimization for Snapshot Compressive Imaging. IEEE Transactions on Pattern Analysis and Machine Intelligence **41**, 2990–3006 (2018)
34. Meng, Z., Ma, J., Yuan, X.: End-to-End Low Cost Compressive Spectral Imaging with Spatial-Spectral Self-Attention. In: European Conference on Computer Vision (2020)
35. Novack, Z., McAuley, J., Lipton, Z.C., Garg, S.: Chils: Zero-shot image classification with hierarchical label sets. In: International Conference on Machine Learning. pp. 26342–26362. PMLR (2023)
36. Ravì, D., Fabelo, H., Callic, G.M., Yang, G.Z.: Manifold Embedding and Semantic Segmentation for Intraoperative Guidance With Hyperspectral Brain Imaging. IEEE Transactions on Medical Imaging **36**, 1845–1857 (2017)
37. Shi, Z., Chen, C., Xiong, Z., Liu, D., Wu, F.: HSCNN+: Advanced CNN-Based Hyperspectral Recovery from RGB Images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 939–947 (2018)
38. Stiebel, T., Koppers, S., Seltsam, P., Merhof, D.: Reconstructing spectral images from RGB-images using a convolutional neural network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 948–953 (2018)
39. Tu, B., Ren, Q., Li, J., Cao, Z., Chen, Y., Plaza, A.J.: NCGLF2: Network combining global and local features for fusion of multisource remote sensing data. Inf. Fusion (2023)
40. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
41. Wang, Q., Huang, W., Xiong, Z., Li, X.: Looking Closer at the Scene: Multiscale Representation Learning for Remote Sensing Image Scene Classification. IEEE Trans. Neural Netw. Learn. Syst. **33**, 1414–1428 (2020)
42. Wang, Z., Cun, X., Bao, J., Zhou, W., Liu, J., Li, H.: Uformer: A General U-Shaped Transformer for Image Restoration. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 17683–17693 (2022)
43. Wu, C., Li, J., Song, R., Li, Y., Du, Q.: RepCPSI: Coordinate-Preserving Proximity Spectral Interaction Network With Reparameterization for Lightweight Spectral Super-Resolution. IEEE Transactions on Geoscience and Remote Sensing **61**, 1–13 (2023)
44. Wu, K., Peng, H., Chen, M., Fu, J., Chao, H.: Rethinking and improving relative position encoding for vision transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10033–10041 (2021)
45. Xiong, F., Zhou, J., tao Qian, Y.: Material Based Object Tracking in Hyperspectral Videos. IEEE Trans. Image Process. **29**, 3719–3733 (2018)
46. Xiong, Z., Shi, Z., Li, H., Wang, L., Liu, D., Wu, F.: HSCNN: CNN-Based Hyperspectral Image Recovery from Spectrally Undersampled Projections. 2017 IEEE International Conference on Computer Vision Workshops (ICCVW) pp. 518–525 (2017)

47. Yan, L., Yamaguchi, M., Noro, N., Takara, Y., Ando, F.: A novel two-stage deep learning-based small-object detection using hyperspectral images. Optical Review **26**, 597 – 606 (2019)
48. Yuan, X.: Generalized alternating projection based total variation minimization for compressive sensing. 2016 IEEE International Conference on Image Processing (ICIP) pp. 2539–2543 (2015)
49. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H.: Restormer: Efficient Transformer for High-Resolution Image Restoration. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5728–5739 (2022)
50. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H., Shao, L.: Learning Enriched Features for Real Image Restoration and Enhancement. In: ECCV (2020)
51. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H., Shao, L.: Multi-Stage Progressive Image Restoration. In: CVPR (2021)
52. Zhang, L., Lang, Z., Wang, P., Wei, W., Liao, S., Shao, L., Zhang, Y.: Pixel-Aware Deep Function-Mixture Network for Spectral Super-Resolution. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 12821–12828 (2020)
53. Zhang, M., Zhang, C., Zhang, Q., Guo, J., Gao, X., Zhang, J.: ESSAformer: Efficient Transformer for Hyperspectral Image Super-resolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 23073–23084 (2023)
54. Zhang, S., Dong, Y., Fu, H., Huang, S.L., Zhang, L.: A Spectral Reconstruction Algorithm of Miniature Spectrometer Based on Sparse Optimization and Dictionary Learning. Sensors (Basel, Switzerland) **18** (2018)
55. Zhang, S., Wang, L., Fu, Y., Zhong, X., Huang, H.: Computational Hyperspectral Imaging Based on Dimension-Discriminative Low-Rank Tensor Recovery. 2019 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 10182–10191 (2019)
56. Zhao, Y., Po, L.M., Lin, T., Yan, Q., Liu, W., Xian, P.: HSGAN: Hyperspectral Reconstruction From RGB Images With Generative Adversarial Network. IEEE Transactions on Neural Networks and Learning Systems (2023)
57. Zhao, Y., Po, L.M., Yan, Q., Liu, W., Lin, T.: Hierarchical Regression Network for Spectral Reconstruction from RGB Images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 422–423 (2020)
58. Zhou, M., Huang, J., Guo, C.L., Li, C.: Fourmer: An efficient global modeling paradigm for image restoration. In: International Conference on Machine Learning. pp. 42589–42601 (2023)
59. Zou, Z., Shi, Z.: Hierarchical Suppression Method for Hyperspectral Target Detection. IEEE Trans. Geosci. Remote Sens. **54**, 330–342 (2016)