Learning Exhaustive Correlation for Spectral Super-Resolution: Where Spatial-Spectral Attention Meets Linear Dependence

Hongyuan Wang[1], Lizhi Wang[2,⋆], Jiang Xu[1]
Chang Chen[3], Xue Hu[3], Fenglong Song[3], and Youliang Yan[3]

[1] Beijing Institute of Technology,
[2] Beijing Normal University,
[3] Huawei Noah's Ark Lab

Abstract. Spectral super-resolution that aims to recover hyperspectral image (HSI) from easily obtainable RGB image has drawn increasing interest in the field of computational photography. The crucial aspect of spectral super-resolution lies in exploiting the correlation within HSIs. However, two types of bottlenecks in existing Transformers limit performance improvement and practical applications. First, existing Transformers often separately emphasize either spatial-wise or spectral-wise correlation, disrupting the 3D features of HSI and hindering the exploitation of unified spatial-spectral correlation. Second, existing self-attention mechanism always establishes full-rank correlation matrix by learning the correlation between pairs of tokens, leading to its inability to describe linear dependence widely existing in HSI among multiple tokens. To address these issues, we propose an Exhaustive Correlation Transformer (ECT) for spectral super-resolution. First, we propose a Spectral-wise Discontinuous 3D (SD3D) splitting strategy, which models unified local-nonlocal spatial-spectral correlation by integrating spatial-wise continuous splitting strategy and spectral-wise discontinuous splitting strategy. Second, we propose a Dynamic Low-Rank Mapping (DLRM) model, which captures linear dependence among multiple tokens through a dynamically calculated low-rank dependence map. By integrating unified spatial-spectral attention and linear dependence, our ECT can model exhaustive correlation within HSI. The experimental results on both simulated and real data indicate that our method achieves SOTA.

Keywords: Spectral super-resolution · Exhaustive correlation · Spatial-Spectral attention · Linear dependence

1 Introduction

Hyperspectral image (HSI) consists of multiple channels, with each channel representing the response in a specific spectral band. In comparison to the 3-channel RGB image, HSI excels in capturing detailed spectral information from a scene. Owing to this advantage, HSI finds extensive applications in image classification [12, 23, 35], object detection [29], face recognition [55], and more.

⋆ Corresponding Author: Lizhi Wang (wanglizhi@bnu.edu.cn)

However, acquiring 3D HSI with 2D sensors is challenging due to the mismatch of dimensions. Traditional scanning-based methods typically require multiple exposures to capture a full HSI, which is disadvantageous for dynamic and rapidly changing scenes.

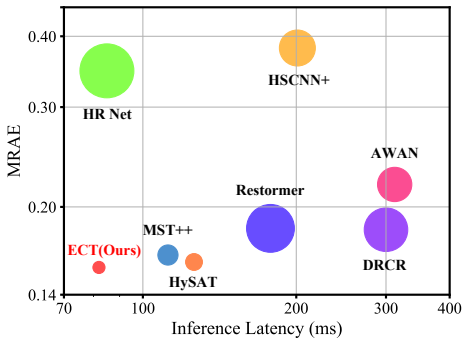

Fig. 1: Comparisons of MRAE, inference latency, and parameters on the NTIRE 2022 dataset are presented. The circle radius represents the number of parameters.

To address this issue, researchers have designed snapshot compressive imaging (SCI) systems with customized optical modulation and reconstruction algorithms, enabling snapshot acquisition of HSI [8, 26, 42, 46, 49]. However, these methods are often expensive and bulky in system implementation. Consequently, the task of HSI reconstruction from the easily obtainable RGB image [22, 52], known as spectral super-resolution, has emerged as a popular solution with the advantages of being cheap and lightweight.

The crucial aspect of spectral super-resolution lies in exploiting correlations within HSI. Early research utilizes sparse coding [3] or low-rank representation [47] for spectral super-resolution. However, these methods often suffer from limited expressive power and generalization ability, thus failing to achieve satisfactory results. With the increasing computing power, learning-based methods have made significant progress in recent years and have become the mainstream solution for spectral super-resolution. Currently, Transformers [9, 39] have attained the state-of-the-art performance for spectral super-resolution by leveraging spectral-wise correlation through a spectral-wise self-attention mechanism. However, two types of bottlenecks exist that limit performance improvement and practical applications. First, existing Transformers predominantly focus on spectral-wise correlation while overlooking spatial-wise correlation in spectral super-resolution. Some works in other tasks [15, 31, 40] attempt to model both spectral-wise and spatial-wise correlations together but often utilize separate network modules. The neglect and separation undermine the 3D nature of HSI and hinder the exploitation of unified spectral-spatial correlation. Second, existing spectral-wise self-attention mechanism always captures the full-rank correlation matrix by learning the correlation between pairs of spectral bands, *i.e.* tokens, in the Transformer. These characteristics result in the inability to establish linear dependence widely existing in HSI among multiple tokens.

In this paper, we propose an Exhaustive Correlation Transformer (ECT) to model the unified spatial-spectral correlation and linear dependence, both of which we believe are crucial for spectral super-resolution. The first motivation behind our method stems from the spatial-spectral similarity in HSI. Thus, we propose a Spectral-wise Discontinuous 3D (SD3D) splitting strategy to simultaneously model unified attention along the spectral and spatial dimensions. The

SD3D splitting strategy contains continuous splitting in the spatial dimension and discontinuous splitting in the spectral dimension, allowing for an effective focus on spectral-wise non-local features without disrupting the continuous structure in the spatial dimension. The second motivation behind our methods arises from the information redundancy in HSI and its low-rank characteristic [16, 53]. Thus, we propose a Dynamic Low-Rank Mapping (DLRM) module to capture the linear dependence among multiple tokens. The DLRM module simultaneously interacts among multiple tokens and maps them into a low-rank space, thereby learning a low-rank dependence map. By integrating unified spatial-spectral attention and linear dependence, our ECT can model the exhaustive correlation within HSI and achieves SOTA in extensive experiments on simulated and real data, with the lowest error achieved under the smallest number of parameters and the lowest inference latency. Codes and pretrained models will be available at `https://github.com/HW-VMCL/ECT_SSR`.

Our contributions are summarized as follows:

- We propose an Exhaustive Correlation Transformer (ECT) to model unified spatial-spectral correlation and linear dependence for spectral super-resolution, which achieves SOTA performance.
- We propose a Spectral-wise Discontinuous 3D (SD3D) splitting strategy to exploit the unified spatial-spectral correlation within HSI by concurrently adopting spatial-wise continuous and spectral-wise discontinuous splitting.
- We propose a Dynamic Low-Rank Mapping (DLRM) module to model the linear dependence within HSI by dynamically calculating a low-rank dependence map among multiple tokens.

2 Related Work

2.1 Spectral Reconstruction

HSI acquisition is typically carried out using push-broom cameras, which is time-consuming and challenging to capture dynamic or rapidly changing scenes. To address this issue, coded aperture snapshot spectral imaging (CASSI) systems have been widely used, generating 2D measurements [33, 49], which are then processed through a series of reconstruction algorithms [7, 10, 11, 41, 42, 46, 50] to obtain HSI.

However, CASSI systems are often expensive. Reconstructing HSIs from RGB images is a cost-effective alternative. Arad *et al.* [3] employ sparse coding for spectral super-resolution, while Aeschbacher *et al.* [1] use shallow learning models and achieve improved results. Due to the presence of substantial redundant information in HSI, a low-rank prior is critical for spectral reconstruction. There are several spectral reconstruction works [16–18, 47, 53, 54] inspired by the low-rank prior. Recently, Three spectral super-resolution challenges [3–5] are held and significantly inspire the research. With the development of deep learning, convolutional neural networks are widely used in the spectral super-resolution task. Shi *et al.* [36] propose a convolutional neural network for spectral super-resolution, which win the NTIRE 2018 Challenge on Spectral Reconstruction

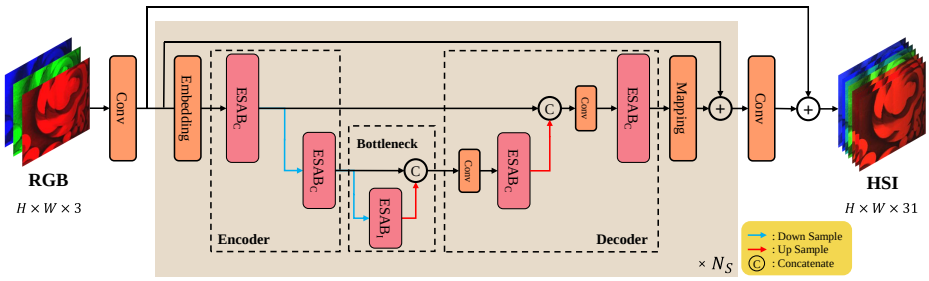

Fig. 2: The macro design of Exhaustive Correlation Transformer (ECT).

from RGB Images. Li *et al.* [24, 25] introduce the channel attention mechanism into the convolutional neural network to improve the performance. Thanks to dynamic weights and long-range correlation modeling, Cai *et al.* [9] are the first to introduce Transformers into the field of spectral super-resolution for modeling spectral-wise correlation and win first place in the NTIRE 2022 challenge [5]. Wang *et al.* [39] further improve the modeling ability of spectral-wise correlation and achieve SOTA performance recently. Though recent methods based on Transformer achieve remarkable performance improvements, they only focus on modeling spectral-wise correlation while ignoring spatial-wise correlation. Moreover, both of them neglect the critical low-rank characteristic of HSI.

2.2 Transformer Model

In the field of NLP, to capture long-range dependencies and enable parallel processing, Vaswani *et al.* [38] introduced the Transformer model based on the self-attention mechanism. Thanks to its capability to capture long-range dependencies, global receptive fields, and dynamic weight computation, Dosovitskiy *et al.* [21] applied the Transformer to image classification, achieving outstanding results. The Transformer architecture has found widespread use in high-level computer vision tasks such as image classification [2, 14, 27, 32, 48], semantic segmentation [34, 37, 45], and object detection [13, 19]. Furthermore, in low-level computer vision tasks, Transformer-based models have also demonstrated remarkable performance in tasks like image super-resolution [15, 28, 30, 58], deraining [43, 44, 51], and denoising [28, 30, 43, 51, 56]. Transformers that leverage the self-attention mechanism can capture long-range correlations between Transformer tokens through dot-product similarity calculations and adaptively fuse tokens based on these correlations, offering strong expressive power.

From the perspective of feature maps, token splitting occurs in the spatial [21, 32] or spectral dimensions [2, 9, 51], allowing for modeling the relationships between pixels or patches or between channels. While there are some efforts to combine these two types of Transformers to model spatial and spectral correlations [15, 31, 40], most of these works directly treat spatial and spectral Transformers as separate modules, which destroys the 3D nature and can not fully exploit the unified correlations.

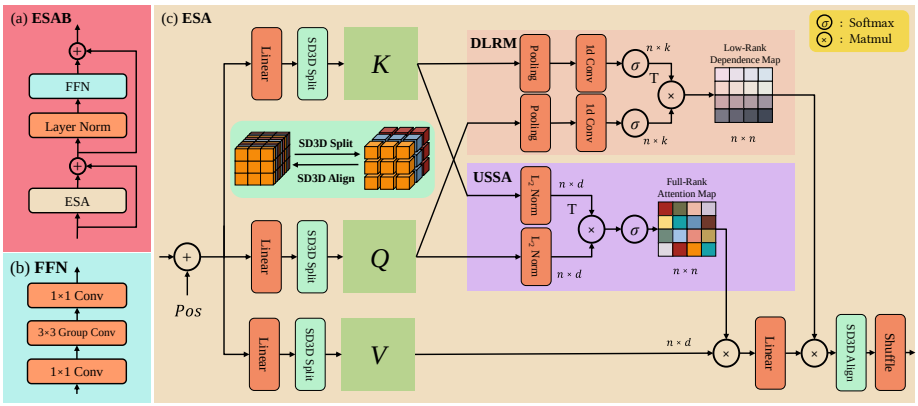

Fig. 3: Micro Design of ECT. (a) Exhaustive Self-Attention Block (ESAB). (b) Feed Forward Network (FFN). (c) Exhaustive Self-Attention (ESA). Key designs in ESA are the Spectral-wise Discontinuous 3D (SD3D) splitting and alignment strategies, the Dynamic Low-Rank Mapping (DLRM) model, and the Unified Spatial-Spectral self-Attention (USSA) model.

3 Method

In this paper, we propose an Exhaustive Correlation Transformer (ECT), which can model unified spatial-spectral attention and linear dependence simultaneously. In this section, we first introduce the macro design of ECT. Then, we delve into the micro design within ECT.

3.1 Macro Design

We propose an Exhaustive Correlation Transformer (ECT) for spectral super-resolution. The overall network employs a multi-stage U-shaped architecture, as shown in Figure 2. For a 3-channel RGB input, it is expanded to 31 channels using a 3×3 convolution and then processed through N_s U-shaped modules. Each U-shaped module consists of Embedding, Encoder, Bottleneck, Decoder, and Mapping components. Embedding and Mapping are implemented with 3×3 convolutions, expanding the channel dimensions to 32 on the input side and reducing them back to 31 on the output side. The main components of the Encoder and Decoder are the Cross Exhaustive Self-Attention Blocks (ESAB_C), utilizing 4×4 convolutions with a stride of 2 for downsampling and 2×2 transpose convolutions with a stride of 2 for upsampling. The Bottleneck includes a layer of Inter Exhaustive Self-Attention Block (ESAB_I). ESAB_C is employed to model the correlations between tokens, while ESAB_I is used to model the correlations within tokens. ESAB_C can model spatial-wise non-local and global-aware spectral-wise local correlation, while ESAB_I can model spatial-wise local and spectral-wise non-local correlation. The spatial resolution of the feature map becomes 1/4 after downsampling, while the channel doubles. The number of attention heads

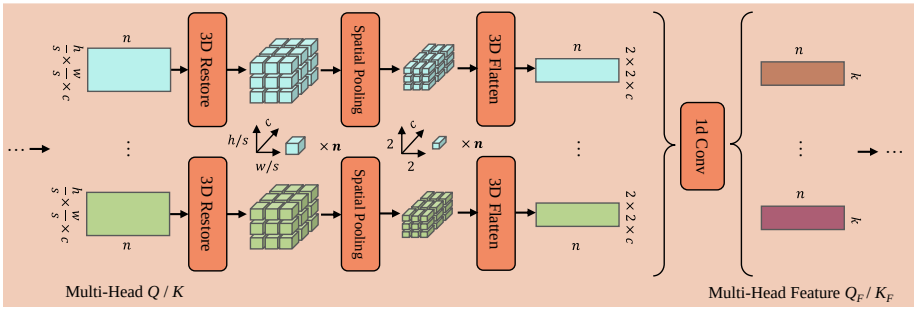

Fig. 4: Detailed design of the Dynamic Low-Rank Mapping (DLRM) module.

scales with the channel changes. Residual connections exist between the Encoder and Decoder, retaining more input information for reconstruction. Furthermore, a long-range residual connection is added to stabilize the training.

3.2 Micro Design

Since the main difference between ESAB_C and ESAB_I lies in the subsequent processing, whether it is for correlations between tokens or within tokens, let us focus on ESAB_C to illustrate the micro design. Furthermore, from this point onward, we will no longer distinguish between ESAB_C and ESAB_I in the mathematical notation below.

As depicted in Figure 3, an Exhaustive Self-Attention Block (ESAB) comprises Exhaustive Self-Attention (ESA), Layer Norm, and a Feed-Forward Network. The key process in ESA is summarized as follows: Firstly, the Spectral-wise Discontinuous 3D (SD3D) splitting strategy is applied to generate tokens, facilitating the exploitation of unified spatial-spectral correlation. Following that, the Low-Rank Dependence Map is generated through the Dynamic Low-Rank Mapping (DLRM) module to model linear dependence among multiple tokens. The Full-Rank Attention Map is generated through the Unified Spatial-Spectral self-Attention (USSA) module to model the independent correlation between pairs of tokens. Next, we introduce the implementation details of ESA.

First, the feature map undergoes two layers of grouped convolutions to learn dynamic positional encoding, which is added to the feature map to model the position of each token. Then, the feature map is linearly transformed into a hidden space. A Spectral-wise Discontinuous 3D (SD3D) splitting operation is performed to generate Q, K, and V. SD3D splitting strategy contains continuous splitting in the spatial dimension and discontinuous splitting in the spectral dimension, which allows for a more effective focus on spectral-wise non-local features without disrupting the continuous structure in the spatial dimension. The original feature map has dimensions $H \times W \times C$, after the SD3D splitting, the number of tokens, denoted as n, becomes $C \times s/c$, and the dimension of each token, denoted as d, becomes $H \times W \times c/s^2$, where s and c are hyperparameters. To simplify the expression, the multi-head attention mechanism is omitted here.

Then, the Unified Spatial-Spectral self-Attention (USSA) is applied to capture independent full-rank correlations between pairs of tokens. The calculation of the Full-Rank Attention Map in USSA is expressed by

$$\mathrm{USSA}(Q,K) = \sigma\left(\tau \frac{K^T \times Q}{||K|| \cdot ||Q||}\right), \tag{1}$$

where σ denotes softmax and τ is a learnable parameter. $Q = W_Q X$, $K = W_K X$, and $V = W_V X$. L_2 normalization is performed within each Token to stabilize the training, and then the expressive power is improved by the learnable parameter τ. It is worth noting that the L_2 normalization and learnable parameter τ are designed by [2] to accommodate variable token sizes, which have been followed by abundant spectral-wise self-attention based methods [8, 9, 39, 51]. We align with these designs in this paper. Since the dimension of tokens d is greater than the number of tokens n in this scenario and the Softmax after the dot product, there is a lower risk of rank reduction in the attention maps, as discussed in [6]. Typically, the learned attention maps have a full rank or nearly full rank. From the optimization point of view, since the linear correlation is different in different HSIs, the loss can only be minimized when the attention mechanism learns full-rank or nearly full-rank attention maps. The experiments confirm this point as well, we found through statistics that the vast majority of attention maps are full rank. Furthermore, the scaled dot-product attention is calculated independently between paired tokens and cannot capture the linear dependence among multiple tokens.

To address the limitation of self-attention in modeling linear dependence within HSI, we propose a Dynamic Low-Rank Mapping (DLRM) module. The details of DLRM are illustrated in Figure 4. Initially, the tokens from the multi-head Q (K) are restored in a 3D manner and subsequently spatially pooled, reducing the spatial dimensions of the tokens from $h/s \times w/s$ to 2×2. Following this step, the tokens are flattened in a 3D fashion to form a two-dimensional matrix. Finally, interactions take place among various heads and tokens through a 1d convolution, yielding a feature Q_F (K_F) with dimensions $n \times k$, where k is a hyperparameter and $k < n$. The matrices Q_F and K_F then undergo a Softmax function, followed by transposition and multiplication to generate a dynamic $n \times n$ matrix, which is a low-rank matrix with a rank no greater than k. The calculation of the Low-Rank Dependence Map in DLRM is expressed by

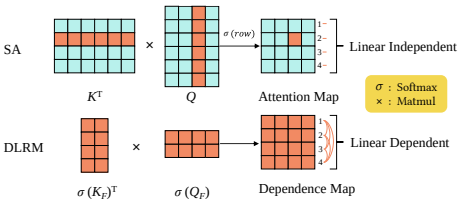

Fig. 5: Comparison of the Attention Map in Self-Attention (SA) and the Dependence Map in DLRM.

$$\mathrm{DLRM}(Q,K) = \sigma(K_F)^T \times \sigma(Q_F). \tag{2}$$

The difference between the Attention Map in self-attention and the Dependence Map in DLRM is illustrated in Figure 5. As shown in the figure, the

Table 1: The experimental results on the NTIRE 2022 [5] dataset, NTIRE 2020 Real World Track [4] dataset, and ICVL [3] dataset are as follows, with **bold** indicating the first place and underlined indicating the second place.

Method	NTIRE 2022		NTIRE 2020		ICVL		Params	FLOPs	Latency
	MRAE	RMSE	MRAE	RMSE	MRAE	RMSE	(M)	(G)	(ms)
HSCNN+ [36]	0.3814	0.0588	0.0684	0.0182	0.2322	0.0424	4.65	304.45	201
HR Net [57]	0.3476	0.0550	0.0682	0.0179	0.1139	0.0313	31.70	163.81	85
AWAN [25]	0.2191	0.0349	0.0668	0.0175	0.1040	0.0252	4.04	270.61	312
Restormer [51]	0.1833	0.0274	0.0645	0.0157	0.0945	0.0230	15.11	93.77	178
DRCR [24]	0.1823	0.0288	0.0664	0.0171	0.0763	0.0164	9.48	586.61	300
MST++ [9]	0.1645	0.0248	0.0624	0.0155	0.0691	0.0144	1.62	22.29	112
HySAT [39]	0.1599	0.0246	0.0589	**0.0142**	0.0654	0.0154	1.40	21.08	126
ECT (Ours)	**0.1564**	**0.0236**	**0.0588**	**0.0142**	**0.0635**	**0.0142**	**1.19**	**16.75**	**82**

calculation of the correlation in self-attention is token-to-token, with each row of the correlation matrix independently calculated. Each element in the matrix is solely associated with two tokens. In contrast, DLRM first facilitates information exchange among various tokens and attention heads. Therefore, each element in the Dependence Map gathers information from multiple tokens. Consequently, each element in the Dependence Map aggregates information from multiple tokens. Moreover, the Dependence Map can implicitly model the linear correlations among multiple tokens due to the low-rank nature. In summary, the Dependence Map can effectively capture the linear dependence among multiple tokens, which is absent in the Attention Map.

Then the correlations learned by USSA and DLRM are used for token fusion. First, V is multiplied by the Full-Rank Attention Map learned by USSA. Following this, V undergoes a linear transformation with the learnable parameter W and is subsequently multiplied by the Low-Rank Dependence Map learned by DLRM. The overall arithmetic process of ESA can be summarized by

$$\mathrm{ESA}(X) = \mathrm{DLRM}(Q, K) \times W \times \mathrm{USSA}(Q, K) \times V. \tag{3}$$

After the token fusion, a Spectral-wise Discontinuous 3D (SD3D) alignment is performed to restore the feature map to its original shape. Finally, channel shuffling is applied to fully explore spectral-wise non-local features.

4 Experiments

4.1 Dataset

For spectral super-resolution experiments on simulated data, we utilized open-source datasets, including NTIRE 2022 [5], NTIRE 2020 Real World Track [4], and ICVL [3]. To further validate the generalization ability of the algorithm, we conduct spectral super-resolution experiments on real RGB data.

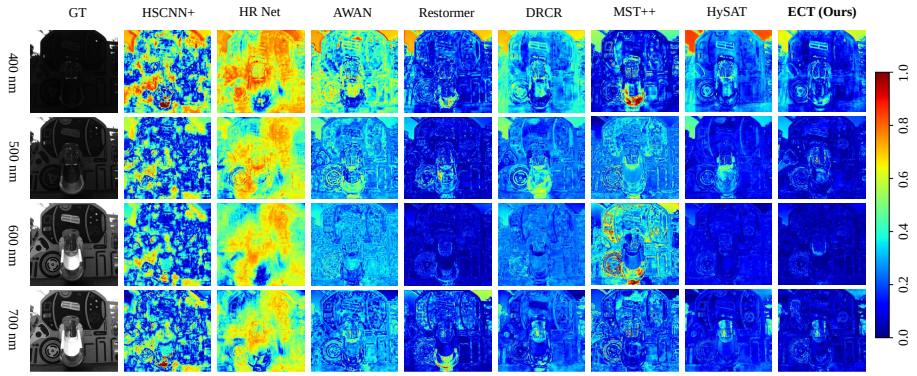

Fig. 6: The MRAE heatmaps, including 400 nm, 500 nm, 600 nm, and 700 nm bands on *ARAD_0944* from the NTIRE 2022 validation data.

The NTIRE 2022 dataset is currently the most complex dataset for spectral super-resolution. It is captured using the Specim IQ camera and includes a wide variety of scenes and colors. The dataset contains a total of 950 available images, including 900 training images and 50 validation images. The spatial resolution of the images is 482×512, and they consist of 31 spectral channels sampled at $10nm$ intervals, covering the wavelength range from $400nm$ to $700nm$.

The NTIRE 2020 dataset comprises 460 available images, including 450 training images and 10 validation images. The spatial and spectral resolutions of these images are consistent with the NTIRE 2022 dataset.

The ICVL dataset comprises 203 available HSI images. The spatial resolution is 1392×1300, and the spectral resolution is consistent with the NTIRE datasets. We randomly select 20 images for the validation and others for the training.

For the real data experiments, there is no available real dataset and most existing spectral super-resolution methods focus on fitting simulated data. Hence, we capture several real RGB images using FLIR Blackfly S BFS-U3-31S4 and obtained the spectral curves of the flattened regions for validation using Specim IQ. We use HSI from the NTIRE 2022 training set and simulate RGB images to create paired training data. We built the data simulation pipeline by referring to the data simulation methods used in NTIRE 2020 and NTIRE 2022.

4.2 Implementation Details

For the hyperparameters in the network structure, we set the SD3D Splitting scale $c = 4, s = 2$ and low-rank factor $k = 12$ for ESAB_C. For ESAB_I, we set the SD3D Splitting scale $c = 16, s = 4$ and low-rank factor $k = 8$. The number of network stages N_s is set to 2.

For the evaluation metrics, following the NTIRE challenges, we use the Mean Relative Absolute Error (MRAE) and Root Mean Squared Error (RMSE) metrics to evaluate the performance of each network. We primarily use MRAE as

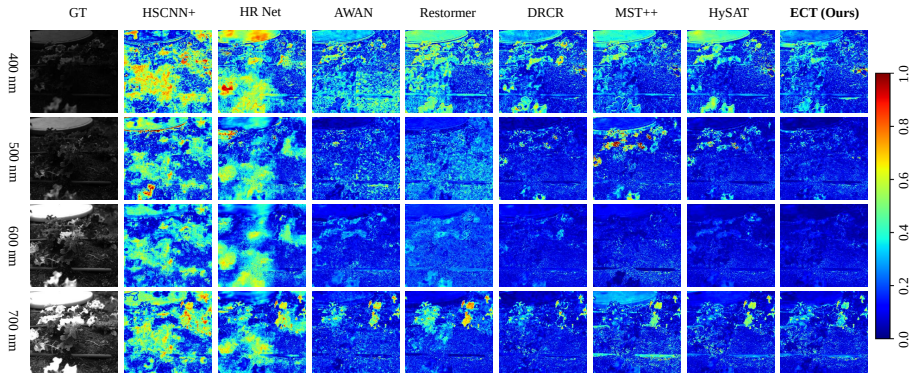

Fig. 7: The MRAE heatmaps, including 400 nm, 500 nm, 600 nm, and 700 nm bands on *ARAD_0940* from the NTIRE 2022 validation data.

the main metric and also adopt it as the training objective. RMSE is used as an auxiliary metric. We further evaluate Spectral Angle Mapper (SAM) on real data and Peak Signal-to-Noise Ratio (PSNR) for CASSI-based spectral reconstruction method.

For the network training details, we utilize a batch size of 40 and employ a learning rate schedule that follows the cosine annealing scheme, decreasing from 4e−4 to 1e−6 over 3e5 iterations. We choose the AdamW optimizer with parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$ and weight decay is set to 1e−4. RGB images are first split into 128×128 patches and undergo random rotations and flips for data augmentation before being input into the network. The codes and pretrained models will be made publicly available.

4.3 Results on Simulated Data

Quantitative Results On NTIRE 2022 [5], NTIRE 2020 [4], and ICVL [3] datasets, we compared ECT with various neural networks, as presented in Table 1. HSCNN+ is the champion in the NTIRE 2018 Clean Track and Real World Track. HR Net and AWAN are the champions in the NTIRE 2020 Clean track and Real World track, respectively. MST++ and DRCR are the first and third place in NTIRE 2022, respectively. Restormer is an advanced algorithm in image reconstruction that has a core design similar to MST++. HySAT is the SOTA spectral super-resolution method published recently. Among all the methods, ECT achieves the lowest MRAE with the lowest computational costs and the smallest number of parameters. We further test the inference latency of all models with the same input size $512 \times 512 \times 3$ on the same 3090 GPU. We select the average inference latency over 30 runs when the inference latency is stabilized. The results demonstrate the superior performance of our method and highlight the significance of modeling unified spatial-spectral correlation and linear dependence.

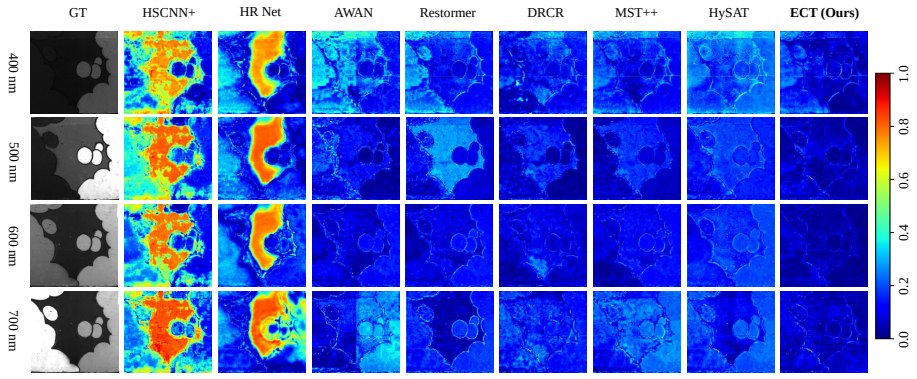

Fig. 8: The MRAE heatmaps, including 400 nm, 500 nm, 600 nm, and 700 nm bands on *ARAD_0903* from the NTIRE 2022 validation data.

Table 2: Experimental results on real data.

Method	Outdoor Scene			Indoor Scene		
	MRAE	RMSE	SAM	MRAE	RMSE	SAM
DRCR	0.2143	0.0070	0.1896	0.2499	0.0104	0.2340
MST++	0.2622	0.0084	0.2245	0.2257	0.0091	0.2055
HySAT	0.2135	0.0070	0.1868	0.2202	0.0088	0.1974
ECT	**0.2012**	**0.0065**	**0.1730**	**0.2114**	**0.0082**	**0.1831**

Table 3: Ablation study of N_s.

N_s	MRAE	RMSE	Params	Latency
1	0.1648	0.0243	0.60 M	41 ms
2	0.1564	0.0236	1.19 M	82 ms
3	0.1542	0.0231	1.78 M	124 ms

Qualitative Results We showcase the visual effects of the MRAE heatmaps in Figure 6, Figure 7, and Figure 8. We also present a comparison of spectral curves in small regions among the Ground Truth and various reconstruction algorithms in Figure 9. The visual results indicate that ECT exhibits the best reconstruction performance across different wavelengths.The reconstruction effect of ECT in the spatial direction is also superior, which further confirms the effectiveness of our approach.

4.4 Results on Real Data

To further investigate the generalization ability of ECT, we conduct some experiments on real RGB data. We choose flattened regions for validation to avoid the influence of misalignment. We capture RGB images of a color chart both indoors under halogen lights and outdoors under sunlight, along with corresponding HSIs. We evaluated the algorithms using the average error of the 18 color patches on the color chart. We further evaluate Spectral Angle Mapper (SAM) on real data. The quantitative and qualitative results of ECT compared with three advanced spectral super-resolution methods MST++ [9], DRCR [24] and HySAT [39] are shown in Table 2 and Figure 10. The experimental results on real data demonstrate the advanced performance and the strong generalization ability of ECT.

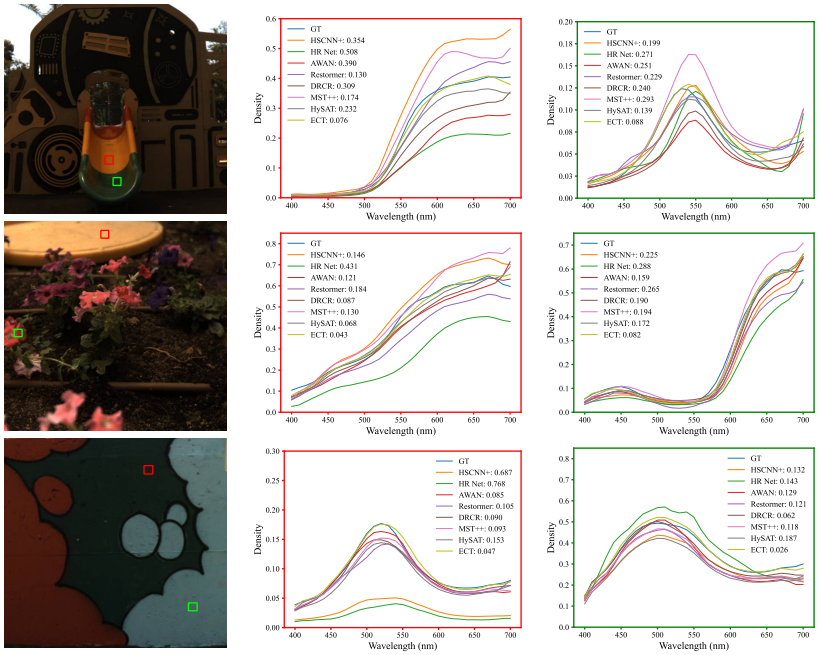

Fig. 9: Comparison of reconstructed spectral curves and MRAE in the small regions. The images in the first line are the official RGB of *ARAD_0944*, *ARAD_0940* and *ARAD_0903* from the NTIRE 2022 validation data. The spectral curves with the red and green axes correspond to the red and green boxes in the corresponding figure.

Table 4: Ablation study of SD3D splitting strategy and the DLRM module.

SD3D	DLRM	MRAE	RMSE	Params	FLOPs
✗	✗	0.1761	0.0266	0.55 M	7.84 G
✓	✗	<u>0.1700</u>	<u>0.0255</u>	0.58 M	8.24 G
✗	✓	0.1733	0.0261	0.56 M	8.27 G
✓	✓	**0.1648**	**0.0243**	0.60 M	8.69 G

Table 5: Ablation study of the token splitting strategy.

Splitting	MRAE	RMSE	Params	FLOPs
Spectral-wise	<u>0.1740</u>	<u>0.0257</u>	0.59 M	8.69 G
Spatial-wise	0.1937	0.0285	0.94 M	14.35 G
SD3D	**0.1648**	**0.0243**	0.60 M	8.69 G

4.5 Ablation Study

To fully explore the effects and the working mechanics of the whole architecture, Spectral-wise Discontinuous 3D (SD3D) splitting strategy, and the Dynamic Low-Rank Mapping (DLRM) module, we introduce some ablation studies here. All ablation studies are conducted on the NTIRE 2022 dataset.

Ablation Study of the Network Stage N_s We conduct an ablation study on the number of stages (N_s) in the network, and the results are shown in Table 3. We primarily set $N_s = 2$ balancing performance and inference latency. All ablation studies below are conducted using a 1-stage structure for efficiency.

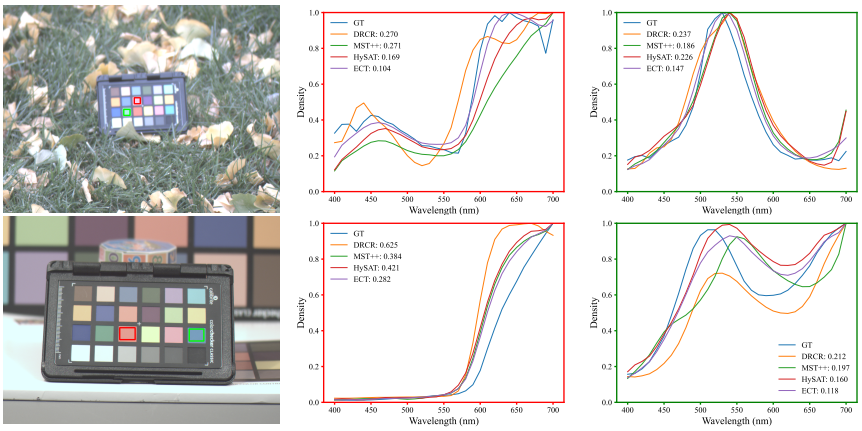

Fig. 10: Comparison of reconstructed spectral curves and MRAE in the small regions. Each region is normalized to remove the influence of brightness. The images in the first line are the real RGBs captured outdoors and indoors.

Table 6: Ablation study of the continuity in SD3D splitting strategy.

Spatial-wise	Spectral-wise	MRAE↓	RMSE↓
Continuous	Continuous	0.1769	0.0263
Continuous	Discontinuous	**0.1648**	**0.0243**
Discontinuous	Discontinuous	0.1739	0.0263

Table 7: Ablation study of the low-rank factor k.

Factor k	MRAE	RMSE	Params	FLOPs
8	0.1689	0.0260	0.60 M	8.69 G
12	**0.1648**	**0.0243**	0.60 M	8.69 G
16	0.1671	0.0249	0.60 M	8.69 G
32	0.1701	0.0259	0.62 M	8.69 G

Ablation Study of SD3D Splitting and DLRM The Spectral-wise Discontinuous 3D (SD3D) splitting strategy and the Dynamic Low-Rank Mapping (DLRM) are our two significant contributions. The SD3D splitting strategy is used to model unified spatial-spectral correlation, while DLRM is employed to capture linear dependence. We test the performance improvement of SD3D and DLRM compared to MST++ [9] without these two structures. The experimental results, as shown in Table 4, demonstrate that both the SD3D splitting strategy and DLRM can lead to improvements. When used together, they achieve even greater performance improvements. The results indicate the effectiveness of our two key designs.

Ablation Study of SD3D Splitting Strategy First, We compare our SD3D splitting strategy with traditional spatial-wise and spectral-wise splitting strategies. The results shown in Table 5 demonstrate the effectiveness of our method. Moreover, the key characteristic of the SD3D splitting strategy lies in its spatial-wise continuity and spectral-wise discontinuity splitting approach. Discontinuous splitting allows for a greater focus on non-local information, while continuous splitting helps preserve the local structure. We conducted an ablation study on the continuity and discontinuity in both spectral and spatial directions, as

Table 8: Comparison with the SOTA method of CASSI-based spectral reconstruction.

Method	NTIRE 2022			NTIRE 2020			ICVL			Params	FLOPs
	MRAE	RMSE	PSNR	MRAE	RMSE	PSNR	MRAE	RMSE	PSNR	(M)	(G)
PADUT	0.1850	0.0271	33.04	0.0624	0.0158	37.07	0.0798	0.0193	36.77	1.71	20.29
ECT	**0.1564**	**0.0236**	**34.81**	**0.0588**	**0.0142**	**37.71**	**0.0635**	**0.0142**	**38.50**	**1.19**	**16.75**

shown in Table 6. Experiments indicate that spectral super-resolution benefits from non-local features in the spectral direction, and adverse effects arise when disrupting spatial continuity.

Ablation Study of Low-Rank Factor k The critical parameter in the DLRM module is the number of feature Q_F (K_F) columns, denoted as k. The rank of the Low-Rank Dependence Map in DLRM is not greater than k. The experimental results for different values of k in ESAB_C are shown in Table 7. When $k = 32$, it means $k = n$, which does not constrain the dependence map to be low-rank. The experimental results highlight the importance of the low-rank characteristic of the dependence map.

5 Discussion

Recently, spectral reconstruction algorithms based on CASSI have received wide interest. We find that performance improvements brought by recent advanced CASSI-based spectral reconstruction algorithms [7, 20, 26] are typically reflected in sharper spatial reconstructions. However, spatial details reconstruction is not a necessary challenge in the spectral super-resolution task. Therefore, recent algorithms used to improve CASSI-based spectral reconstruction have difficulty enhancing the performance of spectral super-resolution. We retrain the SOTA algorithm for CASSI-based spectral reconstruction PADUT [26] for the spectral super-resolution task. The experimental results shown in Table 8 demonstrate that our ECT achieves significantly better performance.

6 Conclusion

In this paper, we analyze the limitations of existing spectral super-resolution Transformers in modeling unified spatial-spectral correlation and linear dependence. We propose an Exhaustive Correlation Transformer (ECT) to model these correlations for spectral super-resolution. Specifically, we propose a Spectral-wise Discontinuous 3D (SD3D) splitting strategy to model unified spatial-spectral correlation and a Dynamic Low-Rank Mapping (DLRM) module to capture linear dependence. Experimental results demonstrate that our approach achieves state-of-the-art performance on both simulated and real data.

Acknowledgements: This work was supported in part by the National Natural Science Foundation of China under Grant 62322204, Grant 62131003, Grant 62072038.

References

1. Aeschbacher, J., Wu, J., Timofte, R.: In defense of shallow learned spectral reconstruction from rgb images. In: ICCVW. pp. 471–479 (2017)
2. Ali, A., Touvron, H., Caron, M., Bojanowski, P., Douze, M., Joulin, A., Laptev, I., Neverova, N., Synnaeve, G., Verbeek, J., et al.: Xcit: Cross-covariance image transformers. NeurIPS **34**, 20014–20027 (2021)
3. Arad, B., Ben-Shahar, O.: Sparse recovery of hyperspectral signal from natural rgb images. In: ECCV. pp. 19–34. Springer (2016)
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: CVPRW. 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: CVPRW. pp. 863–881 (2022)
6. Bhojanapalli, S., Yun, C., Rawat, A.S., Reddi, S., Kumar, S.: Low-rank bottleneck in multi-head attention models. In: ICML. pp. 864–873 (2020)
7. 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: ECCV. pp. 686–704. Springer (2022)
8. 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: CVPR. pp. 17502–17511 (2022)
9. 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: CVPRW. pp. 745–755 (2022)
10. Cai, Y., Lin, J., Wang, H., Yuan, X., Ding, H., Zhang, Y., Timofte, R., Gool, L.V.: Degradation-aware unfolding half-shuffle transformer for spectral compressive imaging. NeurIPS **35**, 37749–37761 (2022)
11. Cai, Y., Zheng, Y., Lin, J., Yuan, X., Zhang, Y., Wang, H.: Binarized spectral compressive imaging. NeurIPS **36** (2023)
12. Camps-Valls, G., Bruzzone, L.: Kernel-based methods for hyperspectral image classification. IEEE TGRS **43**(6), 1351–1362 (2005)
13. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: ECCV. pp. 213–229. Springer (2020)
14. Chen, C.F.R., Fan, Q., Panda, R.: Crossvit: Cross-attention multi-scale vision transformer for image classification. In: ICCV. pp. 357–366 (2021)
15. Chen, X., Wang, X., Zhou, J., Qiao, Y., Dong, C.: Activating more pixels in image super-resolution transformer. In: CVPR. pp. 22367–22377 (2023)
16. Dian, R., Fang, L., Li, S.: Hyperspectral image super-resolution via non-local sparse tensor factorization. In: CVPR. pp. 5344–5353 (2017)
17. Dian, R., Li, S.: Hyperspectral image super-resolution via subspace-based low tensor multi-rank regularization. IEEE TIP **28**(10), 5135–5146 (2019)
18. Dian, R., Li, S., Fang, L.: Learning a low tensor-train rank representation for hyperspectral image super-resolution. IEEE TNNLS **30**(9), 2672–2683 (2019)
19. Ding, J., Xue, N., Long, Y., Xia, G.S., Lu, Q.: Learning roi transformer for oriented object detection in aerial images. In: CVPR. pp. 2849–2858 (2019)
20. Dong, Y., Gao, D., Qiu, T., Li, Y., Yang, M., Shi, G.: Residual degradation learning unfolding framework with mixing priors across spectral and spatial for compressive spectral imaging. In: ICCV. pp. 22262–22271 (2023)

21. 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. In: ICLR (2020)
22. He, J., Yuan, Q., Li, J., Xiao, Y., Liu, D., Shen, H., Zhang, L.: Spectral super-resolution meets deep learning: Achievements and challenges. Information Fusion p. 101812 (2023)
23. Hu, W., Huang, Y., Wei, L., Zhang, F., Li, H.: Deep convolutional neural networks for hyperspectral image classification. Journal of Sensors **2015**, 1–12 (2015)
24. 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: CVPR. pp. 1259–1268 (2022)
25. Li, J., Wu, C., Song, R., Li, Y., Liu, F.: Adaptive weighted attention network with camera spectral sensitivity prior for spectral reconstruction from rgb images. In: CVPRW. pp. 462–463 (2020)
26. Li, M., Fu, Y., Liu, J., Zhang, Y.: Pixel adaptive deep unfolding transformer for hyperspectral image reconstruction. In: ICCV. pp. 12959–12968 (2023)
27. Li, Y., Yuan, G., Wen, Y., Hu, J., Evangelidis, G., Tulyakov, S., Wang, Y., Ren, J.: Efficientformer: Vision transformers at mobilenet speed. NeurIPS **35**, 12934–12949 (2022)
28. Li, Y., Fan, Y., Xiang, X., Demandolx, D., Ranjan, R., Timofte, R., Van Gool, L.: Efficient and explicit modelling of image hierarchies for image restoration. In: CVPR. pp. 18278–18289 (2023)
29. Liang, J., Zhou, J., Bai, X., Qian, Y.: Salient object detection in hyperspectral imagery. In: ICIP. pp. 2393–2397. IEEE (2013)
30. Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R.: Swinir: Image restoration using swin transformer. In: ICCV. pp. 1833–1844 (2021)
31. Liu, Q., Wu, Z., Xu, Y., Wei, Z.: A unified attention paradigm for hyperspectral image classification. IEEE TGRS **61**, 1–16 (2023)
32. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: ICCV. pp. 10012–10022 (2021)
33. Llull, P., Liao, X., Yuan, X., Yang, J., Kittle, D., Carin, L., Sapiro, G., Brady, D.J.: Coded aperture compressive temporal imaging. Optics express **21**(9), 10526–10545 (2013)
34. Lu, Z., He, S., Zhu, X., Zhang, L., Song, Y.Z., Xiang, T.: Simpler is better: Few-shot semantic segmentation with classifier weight transformer. In: ICCV. pp. 8741–8750 (2021)
35. Melgani, F., Bruzzone, L.: Classification of hyperspectral remote sensing images with support vector machines. IEEE TGRS **42**(8), 1778–1790 (2004)
36. Shi, Z., Chen, C., Xiong, Z., Liu, D., Wu, F.: Hscnn+: Advanced cnn-based hyperspectral recovery from rgb images. In: CVPRW. pp. 939–947 (2018)
37. Strudel, R., Garcia, R., Laptev, I., Schmid, C.: Segmenter: Transformer for semantic segmentation. In: ICCV. pp. 7262–7272 (2021)
38. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. NeurIPS **30** (2017)
39. Wang, H., Wang, L., Chen, C., Hu, X., Song, F., Huang, H.: Learning spectral-wise correlation for spectral super-resolution: Where similarity meets particularity. In: ACM MM. p. 7676–7685 (2023)
40. Wang, J., Li, K., Zhang, Y., Yuan, X., Tao, Z.: Sˆ 2-transformer for mask-aware hyperspectral image reconstruction. arXiv preprint arXiv:2209.12075 (2022)

41. Wang, L., Sun, C., Fu, Y., Kim, M.H., Huang, H.: Hyperspectral image reconstruction using a deep spatial-spectral prior. In: CVPR. pp. 8032–8041 (2019)
42. Wang, L., Sun, C., Zhang, M., Fu, Y., Huang, H.: Dnu: Deep non-local unrolling for computational spectral imaging. In: CVPR. pp. 1661–1671 (2020)
43. Wang, Z., Cun, X., Bao, J., Zhou, W., Liu, J., Li, H.: Uformer: A general u-shaped transformer for image restoration. In: CVPR. pp. 17683–17693 (2022)
44. Xiao, J., Fu, X., Liu, A., Wu, F., Zha, Z.J.: Image de-raining transformer. IEEE TPAMI (2022)
45. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. NeurIPS **34**, 12077–12090 (2021)
46. Xiong, Z., Shi, Z., Li, H., Wang, L., Liu, D., Wu, F.: Hscnn: Cnn-based hyperspectral image recovery from spectrally undersampled projections. In: ICCVW. pp. 518–525 (2017)
47. Xue, J., Zhao, Y.Q., Bu, Y., Liao, W., Chan, J.C.W., Philips, W.: Spatial-spectral structured sparse low-rank representation for hyperspectral image super-resolution. IEEE TIP **30**, 3084–3097 (2021)
48. Yu, W., Luo, M., Zhou, P., Si, C., Zhou, Y., Wang, X., Feng, J., Yan, S.: Metaformer is actually what you need for vision. In: CVPR. pp. 10819–10829 (2022)
49. Yuan, X., Brady, D.J., Katsaggelos, A.K.: Snapshot compressive imaging: Theory, algorithms, and applications. IEEE Signal Processing Magazine **38**(2), 65–88 (2021)
50. Yuan, X., Liu, Y., Suo, J., Dai, Q.: Plug-and-play algorithms for large-scale snapshot compressive imaging. In: CVPR. pp. 1447–1457 (2020)
51. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H.: Restormer: Efficient transformer for high-resolution image restoration. In: CVPR. pp. 5728–5739 (2022)
52. Zhang, J., Su, R., Fu, Q., Ren, W., Heide, F., Nie, Y.: A survey on computational spectral reconstruction methods from rgb to hyperspectral imaging. Scientific reports **12**(1), 11905 (2022)
53. Zhang, S., Wang, L., Fu, Y., Zhong, X., Huang, H.: Computational hyperspectral imaging based on dimension-discriminative low-rank tensor recovery. In: ICCV. pp. 10183–10192 (2019)
54. Zhang, S., Wang, L., Zhang, L., Huang, H.: Learning tensor low-rank prior for hyperspectral image reconstruction. In: CVPR. pp. 12006–12015 (2021)
55. Zhang, X., Zhao, H.: Hyperspectral-cube-based mobile face recognition: A comprehensive review. Information Fusion **74**, 132–150 (2021)
56. Zhao, H., Gou, Y., Li, B., Peng, D., Lv, J., Peng, X.: Comprehensive and delicate: An efficient transformer for image restoration. In: CVPR. pp. 14122–14132 (2023)
57. Zhao, Y., Po, L.M., Yan, Q., Liu, W., Lin, T.: Hierarchical regression network for spectral reconstruction from rgb images. In: CVPRW. pp. 422–423 (2020)
58. Zou, W., Ye, T., Zheng, W., Zhang, Y., Chen, L., Wu, Y.: Self-calibrated efficient transformer for lightweight super-resolution. In: CVPR. pp. 930–939 (2022)