Hyperspectral Image Reconstruction via Combinatorial Embedding of Cross-Channel Spatio-Spectral Clues

Xingxing Yang[1], Jie Chen[1]*, Zaifeng Yang[2]

[1]Department of Computer Science, Hong Kong Baptist University
[2]Institute of High Performance Computing, Agency for Science Technology and Research
csxxyang@comp.hkbu.edu.hk, chenjie@comp.hkbu.edu.hk, yang_zaifeng@ihpc.a-star.edu.sg

Abstract

Existing learning-based hyperspectral reconstruction methods show limitations in fully exploiting the information among the hyperspectral bands. As such, we propose to investigate the chromatic inter-dependencies in their respective hyperspectral embedding space. These embedded features can be fully exploited by querying the inter-channel correlations in a combinatorial manner, with the unique and complementary information efficiently fused into the final prediction. We found such independent modeling and combinatorial excavation mechanisms are extremely beneficial to uncover marginal spectral features, especially in the long wavelength bands. In addition, we have proposed a spatio-spectral attention block and a spectrum-fusion attention module, which greatly facilitates the excavation and fusion of information at both semantically long-range levels and fine-grained pixel levels across all dimensions. Extensive quantitative and qualitative experiments show that our method (dubbed CESST) achieves SOTA performance. Code for this project is at: https://github.com/AlexYangxx/CESST.

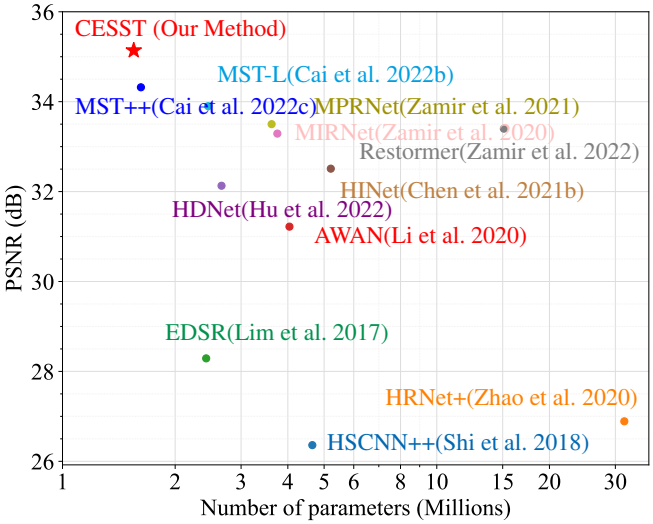


Figure 1: HSI reconstruction on NTIRE2022 HSI dataset (Arad et al. 2022).

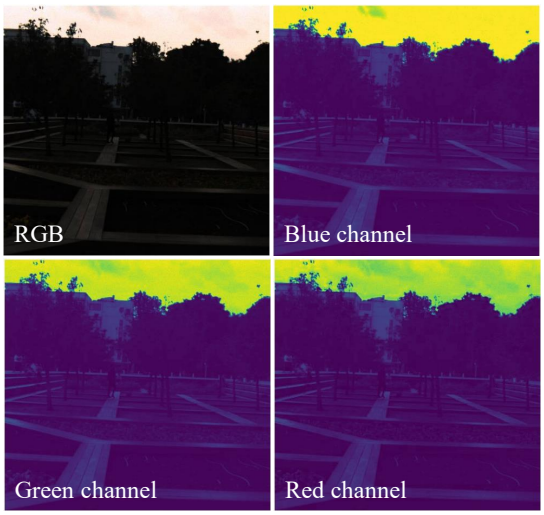
Introduction

Combining spectroscopy and image processing techniques, the hyperspectral imaging system (HIS) records rich spectral information along long-range-distributed spectral bands as well as spatial information. In the past few years, HIS has emerged as a powerful tool in remote sensing (Yuan, Zheng, and Lu 2017), medical image processing (Lu and Fei 2014), agriculture (Adão et al. 2017), etc. Nonetheless, HIS usually requires a long acquisition time and captures images with limited spatial resolution, which constrained its applications, especially in dynamic or real-time scenarios (Arad et al. 2020). To facilitate and promote the applications of HIS, recent studies have explored efficient data captures, *e.g.*, snapshot compressive imaging system that records 3D hyperspectral cube into the 2D measurement (Channing 2022; Cai et al. 2022a). However, these methods require expensive, bulky equipment and complicated reconstruction processing for high-fidelity 3D hyperspectral cubes. To this end, an increased interest in the hyperspectral image (HSI) reconstruction from RGB images using deep learning methods has emerged, which shows great potential due to the handy RGB image capture devices and satisfactory HSI reconstruction performance (Yan et al. 2020; Cai et al. 2022c).

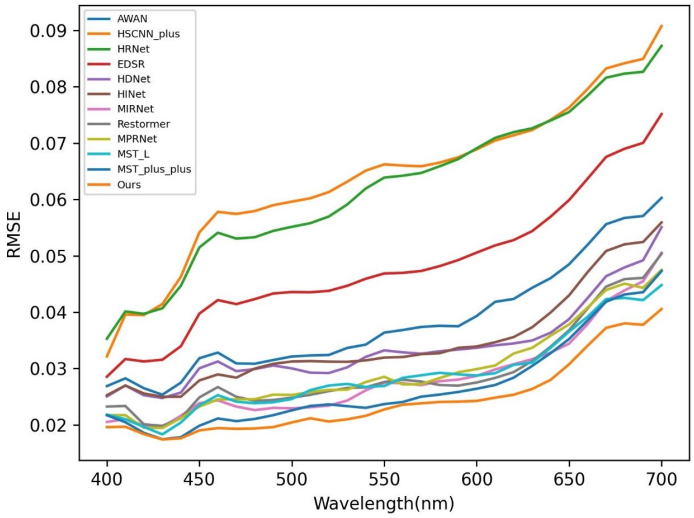
Conventional hyperspectral reconstruction methods are mainly model-based, *e.g.*, sparse coding (Arad and Ben-Shahar 2016), which fails in exploring the intrinsic spectral relations between input RGB images and the corresponding hyperspectral images and suffers from representation capacity. Over the years, an enormous amount of deep-learning-based research has been developed that mainly focuses on deep convolutional neural networks (CNNs) in an end-to-end manner (Wang et al. 2018a, 2019). However, CNNs still fail to capture long-range dependencies and rely on delicately designed modules.

Recently, vision transformer (Liu et al. 2021) (ViT) has been introduced into computer vision from natural language processing (NLP) and shows great potential in learning long-range dependencies and non-local self-similarities. However, existing frameworks for HSI reconstruction have two main limitations: (**i**) the complexity of the standard global transformer (Dosovitskiy et al. 2020) is quadratic to the spatial dimension, which occupies substantial computational re-

*Corresponding Author.



(a) Channel differences.



(b) RMSE loss.

Figure 2: Limitation of existing methods. Fig. 2a shows an example of visual differences between R, G, and B channels of the same RGB image. As can be seen, the red channel contains more texture energy than the other two channels, and most existing methods show compromised performance over the long-wavelength range. Fig. 2b shows the RMSE loss across all bands in the validation set. All existing methods show a dramatic rise of RMSE loss in the long-wavelength bands, while our method slows such deterioration and achieves the lowest RMSE loss across all bands.

sources and thus limits its applications for high-resolution inputs. (**ii**) Although the Swin-transformer (Liu et al. 2021) and spectral-wise transformer (Cai et al. 2022c) achieve linear complexity to the spatial dimension via window-based multi-head self-attention (MSA) or calculating spectral-wise self-attention maps, they all focus on one dimension of the 3D hyperspectral cube, *i.e.*, spatial or spectral dimension. Most importantly, as shown in Fig. 2, ***all these existing reconstruction methods treat the problem trivially by directly studying the correlations in the hyperspectral space***. Specifically, these methods brutally combine and project the features (with different energy and noise characteristics) from RGB channels into the high dimensional spectral space in the early stages, which would inevitably sacrifice some critical information from the R, G, or B channels.

In this study, we propose a novel hyperspectral image reconstruction framework that excavates the unique and complementary information among the RGB input channels in a **C**ombinatorial manner for efficient **E**mbedding of **S**patio-**S**pectral clues based on a **T**ransformer structure (CESST), which achieves the best PSNR-Params performance compared with SOTA methods in Fig. 1. The novelty and technical contributions are generalized as follows:

- We propose a novel framework for hyperspectral image reconstruction, which first fully excavates the intra-channel spatio-spectral features in the projected high-dimensional embedding space before inter-channel fusion. Such channel-wise independent modeling procedure ensures unique local spectral features are well uncovered and preserved;
- We propose a novel Spectrum-Fusion Attention Module (SFAM) that exhaustively queries and explores cross-channel correlations in a combinatorial manner via six parallel transformer branches. SFAM fully excavates complementary information for comprehensive inter-channel fusion;
- An efficient plug-and-play spatio-spectral attention block (SSAB) is designed to simultaneously extract spatio-spectral features at both semantically long-range levels and fine-grained pixel levels across all dimensions, while keeping the complexity linear to the spatial dimension;
- Both quantitative and qualitative experiments demonstrate that our CESST framework significantly outperforms SOTA methods while requiring fewer Parmas.

Related Work

Hyperspectral Image Reconstruction

HSI reconstruction methods can mainly be categorized into two groups: recover 3D HSI cubes from corresponding 2D measurements recorded by SCI systems(Huang et al. 2021; Miao et al. 2019; Wang et al. 2020, 2016; Meng, Ma, and Yuan 2020), or recover from corresponding RGB images(Shi et al. 2018; Yan et al. 2020; Cai et al. 2022c; Akhtar and Mian 2018). In terms of the former methodology, many efforts(Huang et al. 2021; Miao et al. 2019; Wang et al. 2020; Meng, Ma, and Yuan 2020) focus on the coded aperture snapshot spectral imaging (CASSI) system(Meng, Ma, and Yuan 2020; Wagadarikar et al. 2008). However, SCI systems are often expensive, which limits their applications. Compared with the former, given an HSI, the corresponding RGB image can be generated using its camera response function, so recovering HSI from its corresponding RGB image is much cheaper. Thus this topic has significant research

and practical value. In this work, we mainly focus on the latter one.

Two basic approaches are currently being adopted in research into HSI reconstruction from RGB images: model-based and deep learning-based methods. Most model-based methods (Arad and Ben-Shahar 2016; Aeschbacher, Wu, and Timofte 2017; Jia et al. 2017; Robles-Kelly 2015) mainly focus on using hand-crafted priors to conduct spectrum interpolation along the channel dimension. For example, Arad *et al.* (Arad and Ben-Shahar 2016) proposed to address this issue by leveraging hyperspectral prior to creating a sparse dictionary of HSIs and their corresponding RGB projections. Robles *et al.* (Robles-Kelly 2015) further proposed to employ color and texture information to assist the reconstruction process subject to the material properties of the objects in the scene. However, these model-based methods rely heavily on hand-crafted priors and suffer from poor representation capacities. Meanwhile, they do not take the spatial context into consideration.

Recently, inspired by the rapid progress of deep learning in image restoration (Tu et al. 2022; Helminger et al. 2021; Chen et al. 2022; Shao et al. 2020), CNNs have been widely exploited to learn the implicit mapping relation from RGB to HSI (Wang et al. 2018a, 2019; Shi et al. 2018; Fubara, Sedky, and Dyke 2020), which learn the spatial contextual information in a statistic sense. For instance, HSCNN (Xiong et al. 2017) proposed to upsample the input RGB image along the spectral dimension and learn the corresponding enhanced HSI using deep residual convolutional blocks. Alvarez-Gila *et al.* (Alvarez-Gila, Van De Weijer, and Garrote 2017) further proposed a conditional generative adversarial framework to deal with the paired-data insufficiency. Many other methods have produced impressive results by designing delicate architectures, including UNet (Can and Timofte 2018), Resnet (Stiebel et al. 2018), and self-attention mechanisms (Wang et al. 2020). However, these CNN-based methods show limitations in capturing long-range inter-dependencies and non-local self-similarities.

Vision Transformer

Since the vision Transformer (ViT) (Dosovitskiy et al. 2020) was first introduced into vision tasks, there has been a wave of enthusiasm due to its strength in capturing long-range correlations between spatial contexts. Since the complexity of standard global transformers (Chen et al. 2021a; Dosovitskiy et al. 2020) is quadratic to the spatial dimension of input images, many researchers introduce the local-processing idea of CNNs into transformer blocks (Liu et al. 2021; Zhou et al. 2021) to reduce the computational burden. For instance, Liu *et al.* proposed to leverage local window-based MSA, whose computational complexity is linear to the spatial dimension. Cai *et al.* (Cai et al. 2022b) further proposed a spectral-wise MSA to calculate the self-attention map along the channel dimension for HSI reconstruction. Nonetheless, neither spatial window-based MSA nor spectral-wise MSA considers spatial and spectral information, limiting the representation capacity.

The Proposed Method

Motivation. Existing frameworks (Cai et al. 2022c; Wang et al. 2019; Shi et al. 2018; Fubara, Sedky, and Dyke 2020) project the RGB image directly to the high-dimensional hyperspectral space in an early stage. Such brutal transformation sacrifices potentially crucial intra-channel features, as shown in Fig. 2, and it would be more difficult to learn from the inter-channel correlations in subsequent stages. As such, we propose to first fully excavate the intra-channel spatio-spectral features in the projected high-dimensional embedding space before inter-channel fusion, ensuring local spectral features are well uncovered and preserved.

Network Architecture

We propose a multi-scale encoder-decoder architecture for HSI reconstruction, which has three layers of similar structures as shown on the top row of Fig. 3, with each layer focusing on different scales (full, half, and quarter sizes).

At each scale, three encoder-decoder feature extraction blocks (FEBs) are designed to learn the contextual features of each channel independently (*e.g.*, R, G, or B). Unlike the other methods that brutally combine and project the RGB channels into the high-dimensional spectral space in the early stages, our channel-wise independent modeling procedure ensures unique local spectral features are well uncovered and preserved. FEBs adopt UNet (Yang et al. 2021) as the backbone to extract both contextual and spectral features crucial for spectrum reconstruction. Specifically, as shown in Fig. 3 (b), each FEB comprises two encoder blocks, one bottleneck block, and two decoder blocks. Each block consists of a spatio-spectral attention block (SSAB), shown in Fig. 3 (c). Subsequently, a spectrum-fusion attention module (SFAM), illustrated in Fig. 3 (a), is cascaded to the output of the three FEBs. The SFAM exhaustively queries and explores cross-channel correlations in a combinatorial manner via six parallel transformer branches and comprehensively fuses the complementary information from R, G, and B channels. Finally, inspired by (Zamir et al. 2021), a supervised spectrum-consistency module (SCCM) is cascaded to generate spectrum-consistent predictions as well as cross-scale features with the supervision of ground-truth HSI signals (as shown in Fig. 3 (d)), which is only embedded in the middle and bottleneck branches. The cross-scale features provide informative guidance from a lower scale to a larger scale, thus formulating a coarse-to-fine reconstruction (indicated by the light blue lines in Fig. 3).

Spatio-Spectral Attention Block

HSI contains plentiful spatio-spectral clues; however, existing CNN-based feature extraction blocks struggle to model the non-local self-similarities. Meanwhile, transformer-based blocks only take one-dimensional features into account (*i.e.*, spatial (Liu et al. 2021) or spectral (Cai et al. 2022c)). To address these issues, we propose a dual-dimensional transformer-based feature extraction block embedded in the FEB, denoted as spatio-spectral attention block (SSAB), to extract both spatial and spectral features and increase the learning capacity. As shown in Fig.

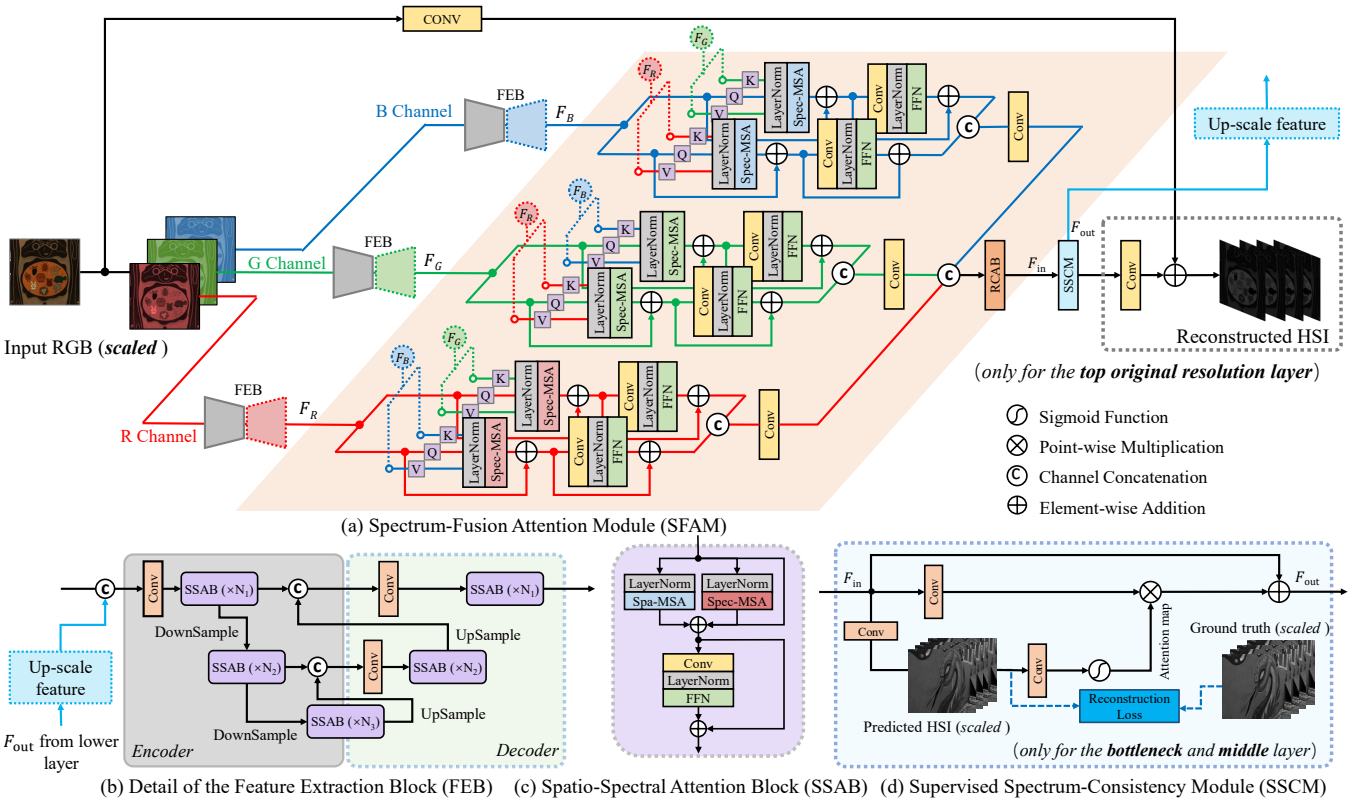


Figure 3: Illustration of the proposed CESST framework, which consists of three layers of similar structures as shown on the top row, which represents the top original resolution layer. A middle layer and a bottleneck layer share similar structures but with slight differences.

3 (c), the SSAB consists of a parallel spatial-MSA and spectral-MSA, which calculates both the spatial multi-head self-attention and the spectral multi-head self-attention in parallel, and then feeds both features to enhance cross-dimensional interaction.

Note that our proposed spatial-MSA is different from conventional window-based MSA (Liu et al. 2021), which suffers from limited receptive fields within non-overlapping position-specific windows. As shown in Fig. 4, our spatial-MSA consists of a normal window-based MSA, followed by a shuffle-window MSA, to build long-range cross-window interactions. The main difference between conventional window-based MSA and shuffle-window MSA is the spatial shuffle mechanism. To be specific, we assume a Window-based MSA with window size M whose input has N tokens; we first reshape the output spatial dimension into $(M, N/M)$, transpose and then flatten it back as the input of the next layer. This operation puts the tokens from distant windows together and helps build long-range cross-window connections. Note that spatial shuffle requires the spatial alignment operation to adjust the spatial tokens into the original positions for spatially aligning features and image content. The spatial alignment operation first reshapes the output spatial dimension into $(N/M, M)$, transposes it, and then flattens it, which is an inverse process of the spatial shuffle. Moreover, considering that the "grid issue" widely exists when using window-based transformers to deal with high-resolution images, we introduce a depth-wise convolution layer between the normal window-based MSA and shuffle-window MSA via a residual connection. The kernel size of the convolution layer is the same as the window size. On the other hand, the spectral-MSA is mostly inspired by (Zamir et al. 2022; Cai et al. 2022c), which treats the spectral feature map as a token and thus focuses on more non-local spectral self-similarities.

MSA scheme	Receptive Field	Complexity-HW	Calculation
ViT	Global	Quadratic	Spatial
Swin-T	local	Linear	Spatial
Restormer	Global	Linear	Spectral
MST++	Global	Linear	Spectral
CESST	Global	Linear	Spatio-spectral

Table 1: Comparison of the properties of different MSAs.

We further summarize the main properties of existing transformer-based blocks, including ViT (Dosovitskiy et al. 2020) (global MSA), Swin-Transformer (Liu et al. 2021) (Window-based MSA), Restormer (Zamir et al. 2022) (spectral MSA), MST++ (Cai et al. 2022c) (spectral MSA), and our CESST (spatio-spectral MSA) in Table 1. Our CESST computes global receptive fields and models both spatial and spectral self-similarities with linear computational costs.

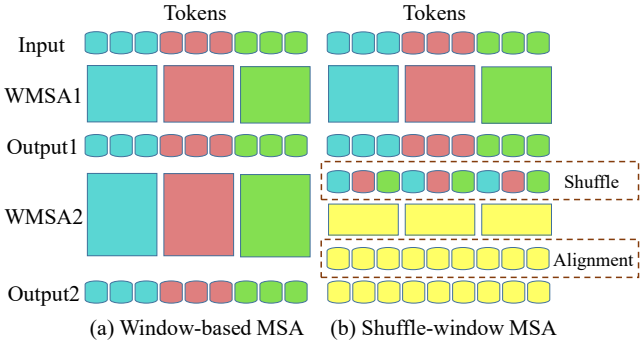


Figure 4: Comparison of traditional window-based MSA and our shuffle-window MSA. WMSA represents window-based multi-head self-attention. (a) two stacked window-based transformer blocks, where each output token only relates to tokens within the same window, without any cross-window interaction; (b) WMSA2 takes data from different windows after WMSA1 by spatial shuffle and alignment, which introduces global cross-window interaction.

Spectrum-Fusion Attention Module

To further improve the feature utilization and interactivity within the three learned hyperspectral representations (*i.e.*, F_R, F_G, F_B), we design an effective feature fusion module, named spectrum-fusion attention module, including two parts: channel learning and spectrum fusion.

In **channel learning**, we propose to extract the correlations between each two of the three learned hyperspectral representations and then fuse them to generate the final fine-grained reconstructed HSI. As shown in Fig. 3 (a), taking F_R as an example, we take the F_G as the *Value* and *Query*, and F_R as the *Key*, and then feed them into the spectral-MSA to learn the correlation between F_R and F_G. The learning of the correlation between F_R and F_B is also carried out in a similar way. Then, the F_G-enriched $\mathcal{F}_R^{RG}$ and the F_B-enriched $\mathcal{F}_R^{RB}$ are concatenated to fed into a convolutional layer to generate F_G-F_B-enriched $\mathcal{F}_R$, which can be formulated as:

$$\mathcal{F}_R^{RG} = F_{\text{S-MSA}}\left(\mathbf{Q}_{F_R}, \mathbf{K}_{F_G}, \mathbf{V}_{F_G}\right) \in \mathbb{R}^{31\times H\times W}, \quad (1)$$

$$\mathcal{F}_R^{RB} = F_{\text{S-MSA}}\left(\mathbf{Q}_{F_R}, \mathbf{K}_{F_B}, \mathbf{V}_{F_B}\right) \in \mathbb{R}^{31\times H\times W}, \quad (2)$$

$$\mathcal{F}_R = F_{\text{conv}}\left(F_{\text{concat}}\left[\mathcal{F}_R^{RG}, \mathcal{F}_R^{RB}\right]\right) \in \mathbb{R}^{31\times H\times W}, \quad (3)$$

where $F_{\text{S-MSA}}(\cdot)$ is spectral-MSA operation (including layernorm, residual connection, and feed-forward operations for simplicity, $F_{\text{conv}}(\cdot)$ is a 3×3 convolutional layer, $F_{\text{concat}}(\cdot)$ is a concatenation process. Similarly, F_R-F_B-enriched $\mathcal{F}_G$ and F_R-F_G-enriched $\mathcal{F}_B$ also have such two channel learning branches. Thus, the channel learning part has six branches. In **spectrum fusion**, the three representative hyperspectral features $\mathcal{F}_R$, $\mathcal{F}_G$, $\mathcal{F}_B$ are concatenated first and then fed into a residual coordinate attention block (Yang et al. 2021) (RCAB) to generate fine-grained pixel-level HSI signals, which can be formulated as:

$$\mathbf{X} = F_{\text{RCAB}}\left(F_{\text{concat}}\left[\mathcal{F}_R, \mathcal{F}_G, \mathcal{F}_B\right]\right) \in \mathbb{R}^{31\times H\times W}. \quad (4)$$

Objective Function

To supervise reconstructed HSIs at any given scale $s = 1, 2, ..., S$, we employ the L_{MIX} loss (Zhao et al. 2016), which combines both SSIM loss and L1 loss, as well as the mean relative absolute error (MRAE), to formulate a supervised consistency constraint on both pixel and feature levels:

$$\mathcal{L}_{\text{MIX}} = \sum_{s=1}^{3}\left[(\mathbf{X}^s, \mathbf{Y}^s)_{mix}\right], \quad (5)$$

$$\mathcal{L}_{\text{MRAE}} = \sum_{s=1}^{3}\left[\frac{|\mathbf{Y}^s - \mathbf{X}^s|}{\mathbf{Y}^s}\right]. \quad (6)$$

Here $\mathbf{Y}^s$ represents the ground-truth image in each scale.

Total Loss. The full objective function is expressed as:

$$\mathcal{L} = \mathcal{L}_{\text{MIX}} + \lambda_1 \mathcal{L}_{\text{MRAE}}, \quad (7)$$

Where λ_1 is the hyperparameter that controls the relative importance of the two loss terms, empirically set to 100.

Experiments and Analysis

Experimental Settings

Datasets. We adopt two datasets: NTIRE2022 HSI dataset (Arad et al. 2022) and ICVL HSI dataset (Arad and Ben-Shahar 2016), to evaluate the performance of our CESST. In NTIRE2022 HSI dataset, there are 950 available RGB-HSI pairs, including 900 for training and 50 for validation. All the HSIs are captured at 482×512 spatial resolution over 31 channels from 400nm to 700nm. Besides, ICVL dataset contains 201 high-resolution HSIs. Considering that it does not provide aligned RGB images, we adopt the method proposed by Magnusson*et al.* (Magnusson et al. 2020) to produce corresponding RGB images. Since it contains 18 images within different resolutions, we only use the left 183 image pairs (147 pairs for training and 36 pairs for testing).

Implementation Details. We implement our CESST with Pytorch. All the models are trained with Adam (Kingma and Ba 2014) optimizer ($\beta_1 = 0.9$ and $\beta_2 = 0.999$) for 300 epochs. The learning rate is initialized as 0.0002, and the Cosine Annealing scheme is adopted. During the training phase, RGB-HSI pairs are first cropped into 128×128 and the input RGB images are linearly rescaled to $[0, 1]$. We employ random rotation and flipping to augment training data. The whole training time of the proposed CESST is about 40 hours with a single NVIDIA Ampere A100-40G. All the RGB images are also rescaled to $[0, 1]$ during the validation procedure. Our CESST takes 0.141s per image (size of $482\times 512\times 3$) for HSI reconstruction.

Evaluation Metrics. Mean relative absolute error(MRAE), Peak Signal-to-Noise Ratio(PSNR), error relative global dimensionless synthesis(ERGAS), and spectral angle mapper(SAM) are employed to evaluate HSI reconstruction methods.

Quantitative Results

We compared our CESST with both HSI reconstruction and image restoration methods, including four RGB-HSI reconstruction methods: HSCNN+ (Shi et al. 2018) (winner of

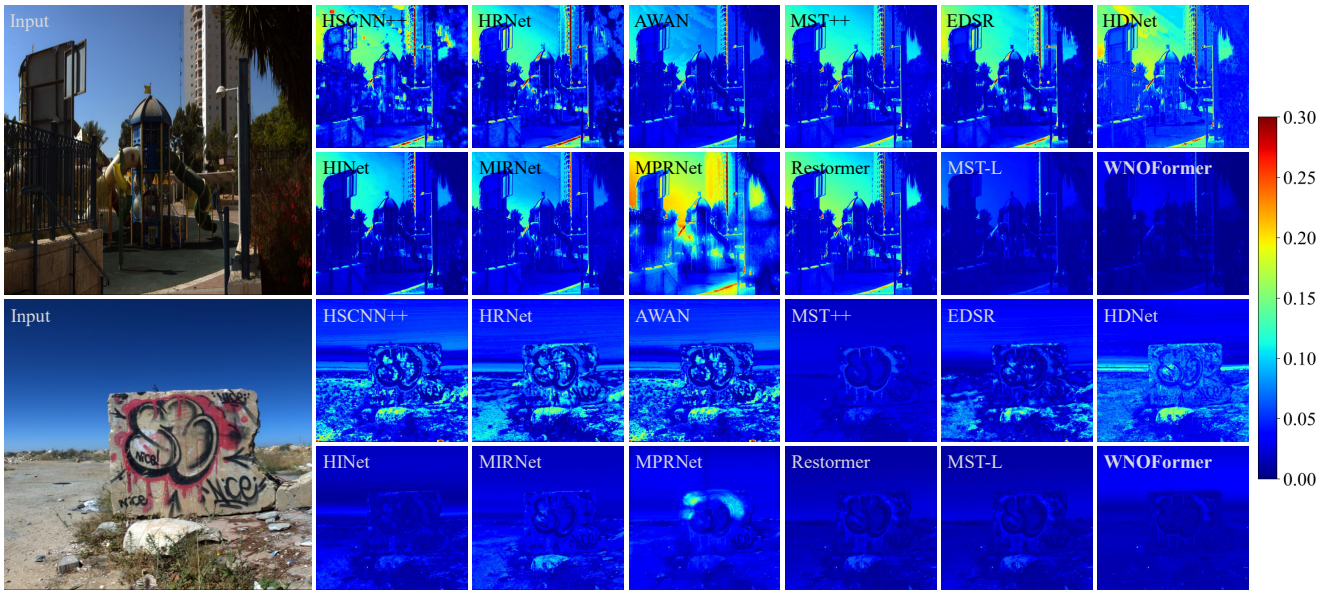


Figure 5: The reconstruction error images of two images chosen from the validation set of ICVL dataset. The error images are the heat maps of the root mean square error (RMSE) (along spectral direction) between ground truths and reconstructed HSIs.

Method	NTIRE2022 HSI Dataset						ICVL Dataset			
	Params(M)	FLOPs(G)	ERGAS	SAM	MRAE	PSNR	ERGAS	SAM	MRAE	PSNR
HSCNN+ (Shi et al. 2018)	4.65	304.45	228.80	0.1093	0.3814	26.36	247.31	0.1108	0.2178	26.28
HRNet (Zhao et al. 2020)	31.70	163.81	107.23	0.0855	0.3476	26.89	**101.02**	**0.0806**	0.2155	26.93
AWAN (Li et al. 2020)	4.04	270.61	147.57	0.0960	0.2500	31.22	169.93	0.1054	0.1887	30.33
MST++ (Cai et al. 2022c)	1.62	**23.05**	107.23	0.0852	0.1645	34.32	118.33	0.0972	0.1776	31.41
EDSR (Lim et al. 2017)	2.42	158.32	212.51	0.0983	0.3277	28.29	235.83	0.1039	0.1972	27.51
HDNet (Hu et al. 2022)	2.66	173.81	133.72	0.1006	0.2048	32.13	177.47	0.1153	0.1942	29.28
HINet (Chen et al. 2021b)	5.21	31.04	140.82	0.0937	0.2032	32.51	152.45	0.0983	0.1663	30.43
MIRNet (Zamir et al. 2020)	3.75	42.95	115.38	0.0944	0.1890	33.29	131.43	0.0998	0.1797	30.76
Restormer (Zamir et al. 2022)	15.11	93.77	112.05	0.0983	0.1833	33.40	130.17	0.1003	0.1689	31.01
MPRNet (Zamir et al. 2021)	3.62	101.59	101.50	0.0901	0.1817	33.50	131.11	0.0979	0.2138	29.09
MST-L (Cai et al. 2022b)	2.45	32.07	112.57	0.0931	0.1772	33.90	122.36	0.0893	0.1845	30.75
CESST (ours)	**1.54**	90.18	**98.74**	**0.0791**	**0.1497**	**35.14**	109.47	0.0917	**0.1230**	**33.25**

Table 2: Comparison with SOTA methods on NTIRE2022 HSI dataset. The best results are highlighted in bold.

NTIRE2018 HSI challenge), HRNet (Zhao et al. 2020), AWAN (Li et al. 2020) (winner of NTIRE2020 HSI challenge), MST++ (Cai et al. 2022c) (winner of NTIRE2022 HSI challenge); two compressive HSI recovery methods: HDNet (Hu et al. 2022) and MST-L (Cai et al. 2022b); five image restoration methods: EDSR (Lim et al. 2017), HINet (Chen et al. 2021b), MIRNet (Zamir et al. 2020), Restormer (Zamir et al. 2022), and MPRNet (Zamir et al. 2021). For fair comparisons, all the methods were retrained and tested with the same settings as MST++ (Cai et al. 2022c). As shown in Table. 2, it can be observed that our method obtains the best results of all five metrics while costing the least Params on NTIRE2022 HSI dataset. To more intuitively illustrate the competitiveness of our method, we provide the PSNR-Params comparisons in Fig. 1, including both HSI reconstruction methods and image restoration methods. The horizontal axis is the Params (memory cost), and the vertical axis is the PSNR (performance). As can be seen, our method takes up the upper-left corner, indicating the best efficiency. Note that although the FLOPs of our model are larger than MST++ (Cai et al. 2022c), benefited from our parallel calculation design, *e.g.*, the three parallel branches of feature extraction, the parallel spatial-MSA and spectral-MSA of SSAB, our model achieves comparable inference time compared with MST++ on the same GPU.

Qualitative Results

To evaluate the visual quality of our method, we provide visual comparisons in Fig. 5. Almost all existing methods fail to generate chromatic-consistent content and artifact-free results, especially for the high-frequency components (*i.e.*, the sky area). In contrast, our method is capable of recovering more precise texture information and better pixel-level quality over other methods. This is because we treat each channel of input RGB images as a unique feature and model them into high dimensional space respectively, rather than

Baseline	Independent	SSAB	SFAM	Params(M)	FLOPs(G)	MRAE	PSNR
✓				0.92	54.24	0.3207	28.14
✓	✓			1.03	74.75	0.2514	31.83
✓	✓	✓		1.37	81.38	0.1917	33.71
✓	✓	✓	✓	1.54	90.18	**0.1497**	**35.14**

(a) Break-down ablation study of different pipelines.

Method	MRAE	PSNR
Baseline	0.1917	33.71
SFT	0.1652	34.22
EFF	0.1611	34.73
SFAM	**0.1497**	**35.14**

(b) Ablations of different fusion.

Table 3: Ablation studies of break-down ablation and fusion scheme comparison.

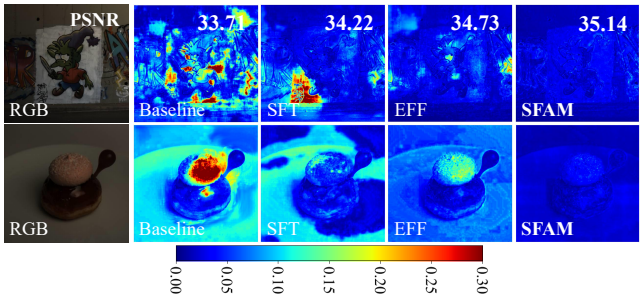


Figure 6: The visual analysis of error images chosen from the validation set of NTIRE2022 HSI dataset. The error images are the heat maps of the root mean square error (RMSE) between the ground truth and the reconstructed HSI.

equally modeling them and studying correlations in the high dimensional space directly (*i.e.*, this is why our method performs better in the long-range spectral bands, as shown in 2b). Besides benefiting from the independent modeling and combinatorial excavation mechanism, as well as our spatio-spectral attention mechanism, both spectral self-similarities and spatial details can be efficiently explored and well fused to generate fine-grained pixel-level predictions. Meanwhile, the intensity of our results is closest to ground truths compared with existing methods.

Ablation Study

In this section, we perform ablation studies to investigate the effectiveness of our proposed structure. The baseline model is derived by removing the independent modeling structure, including spatio-spectral attention block (SSAB), and spectrum-fusion attention module (SFAM) from CESST, and using the widely used ResNet (He et al. 2016) block.

Break-down Ablation. To investigate the effect of each module, we first perform a break-down ablation and provide the quantitative results in Table 3a. From the first and the second rows, we find that the independent modeling significantly improves the performance of the whole model, which yields a 3.69dB improvement in PSNR. When we successively apply both SSAB and SFAM, the reconstruction performance further achieves 1.88dB and 1.43dB improvement. These results demonstrate the effectiveness of our independent modeling, SSAB, and SFAM.

Fusion Scheme Comparison. As the fusion module is one of the main contributions of this work, we further compare our SFAM with the other popular fusion schemes in Table 3b, including concatenation-convolution (acts as the baseline in this scenario, which is the same settings as the third row in Table 3a), SFT layer (Wang et al. 2018b), and EFF layer (Hu et al. 2022). As can be seen, our module gains 1.43dB, 0.92dB, and 0.41dB in PSNR, which verifies the effectiveness of our SFAM. In addition, we further provide visual analysis in Fig. 6, which shows that our SFAM is more capable of fusing fine-grained details, especially in the salient regions.

MSA scheme	MRAE	PSNR
Spatial-MSA (Dosovitskiy et al. 2020)	0.1614	34.75
Shifted-window MSA (Liu et al. 2021)	0.1783	33.92
Spectral-MSA (Zamir et al. 2022)	**0.1483**	34.96
Shuffle-window MSA	0.1639	34.51
***Spatio-spectral MSA** (ours)*	0.1497	**35.14**

Table 4: Ablation study of different MSAs.

MSA Comparison. To further validate the effectiveness of our proposed spatio-spectral attention mechanism, we compare our SSAB with different MSA variations, including spatial-MSA (Dosovitskiy et al. 2020), shifted-window MSA (Liu et al. 2021), spectral-MSA (Cai et al. 2022c) and shuffle-window MSA (a variation of our original spatio-spectral MSA, which disables the spectral-MSA and keeps other settings consistent). We switch these modules directly in our framework. As shown in Table 4, our spatio-spectral MSA performs best in RMSE and PSNR and is comparable with spectral-MSA in MRAE.

Conclusion

In this paper, we have proposed a novel hyperspectral image reconstruction framework that excavates the unique and complementary information among the RGB input channels in a combinatorial manner for efficient embedding of spatio-spectral clues based on a Transformer structure: CESST. We have proposed a spatio-spectral attention module, and a spectrum-fusion attention module, which greatly facilitates the excavation and fusion of information at both semantically long-range levels and fine-grained pixel levels across all dimensions. The effectiveness of each module has been validated by ablation studies. Extensive visual comparisons and quantitative experiments have demonstrated that our proposed method achieves superior HSI reconstruction performance compared with SOTA methods.

Acknowledgments

This research was supported by A*STAR C222812026.

References

Adão, T.; Hruška, J.; Pádua, L.; Bessa, J.; Peres, E.; Morais, R.; and Sousa, J. J. 2017. Hyperspectral imaging: A review on UAV-based sensors, data processing and applications for agriculture and forestry. *Remote sensing*, 9(11): 1110.

Aeschbacher, J.; Wu, J.; and Timofte, R. 2017. In defense of shallow learned spectral reconstruction from RGB images. In *ICCVW*, 471–479.

Akhtar, N.; and Mian, A. 2018. Hyperspectral recovery from RGB images using Gaussian processes. *TPAMI*, 42(1): 100–113.

Alvarez-Gila, A.; Van De Weijer, J.; and Garrote, E. 2017. Adversarial networks for spatial context-aware spectral image reconstruction from rgb. In *ICCVW*, 480–490.

Arad, B.; and Ben-Shahar, O. 2016. Sparse Recovery of Hyperspectral Signal from Natural RGB Images. In *ECCV*.

Arad, B.; Timofte, R.; Ben-Shahar, O.; Lin, Y.-T.; and Finlayson, G. D. 2020. Ntire 2020 challenge on spectral reconstruction from an rgb image. In *CVPRW*, 446–447.

Arad, B.; Timofte, R.; Yahel, R.; Morag, N.; Bernat, A.; Cai, Y.; Lin, J.; Lin, Z.; Wang, H.; Zhang, Y.; et al. 2022. NTIRE 2022 Spectral Recovery Challenge and Data Set. In *CVPRW*, 863–881.

Cai, Y.; Lin, J.; Hu, X.; Wang, H.; Yuan, X.; Zhang, Y.; Timofte, R.; and Van Gool, L. 2022a. Coarse-to-fine sparse transformer for hyperspectral image reconstruction. In *ECCV*, 686–704. Springer.

Cai, Y.; Lin, J.; Hu, X.; Wang, H.; Yuan, X.; Zhang, Y.; Timofte, R.; and Van Gool, L. 2022b. Mask-guided spectral-wise transformer for efficient hyperspectral image reconstruction. In *CVPR*, 17502–17511.

Cai, Y.; Lin, J.; Lin, Z.; Wang, H.; Zhang, Y.; Pfister, H.; Timofte, R.; and Van Gool, L. 2022c. Mst++: Multi-stage spectral-wise transformer for efficient spectral reconstruction. In *CVPRW*, 745–755.

Can, Y. B.; and Timofte, R. 2018. An efficient CNN for spectral reconstruction from RGB images. *arXiv preprint arXiv:1804.04647*.

Channing, G. 2022. Spectral DefocusCam: Compressive Hyperspectral Imaging from Defocus Measurements. In *AAAI*, volume 36, 13128–13129.

Chen, H.; Wang, Y.; Guo, T.; Xu, C.; Deng, Y.; Liu, Z.; Ma, S.; Xu, C.; Xu, C.; and Gao, W. 2021a. Pre-trained image processing transformer. In *CVPR*, 12299–12310.

Chen, J.; Yang, Z.; Chan, T. N.; Li, H.; Hou, J.; and Chau, L.-P. 2022. Attention-Guided Progressive Neural Texture Fusion for High Dynamic Range Image Restoration. *TIP*, 31: 2661–2672.

Chen, L.; Lu, X.; Zhang, J.; Chu, X.; and Chen, C. 2021b. HINet: Half instance normalization network for image restoration. In *CVPR*, 182–192.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Fubara, B. J.; Sedky, M.; and Dyke, D. 2020. Rgb to spectral reconstruction via learned basis functions and weights. In *CVPRW*, 480–481.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

Helminger, L.; Bernasconi, M.; Djelouah, A.; Gross, M.; and Schroers, C. 2021. Generic image restoration with flow based priors. In *CVPR*, 334–343.

Hu, X.; Cai, Y.; Lin, J.; Wang, H.; Yuan, X.; Zhang, Y.; Timofte, R.; and Van Gool, L. 2022. Hdnet: High-resolution dual-domain learning for spectral compressive imaging. In *CVPR*, 17542–17551.

Huang, T.; Dong, W.; Yuan, X.; Wu, J.; and Shi, G. 2021. Deep gaussian scale mixture prior for spectral compressive imaging. In *CVPR*, 16216–16225.

Jia, Y.; Zheng, Y.; Gu, L.; Subpa-Asa, A.; Lam, A.; Sato, Y.; and Sato, I. 2017. From RGB to spectrum for natural scenes via manifold-based mapping. In *ICCV*, 4705–4713.

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

Li, J.; Wu, C.; Song, R.; Li, Y.; and Liu, F. 2020. Adaptive weighted attention network with camera spectral sensitivity prior for spectral reconstruction from RGB images. In *CVPRW*, 462–463.

Lim, B.; Son, S.; Kim, H.; Nah, S.; and Mu Lee, K. 2017. Enhanced deep residual networks for single image super-resolution. In *CVPRW*, 136–144.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 10012–10022.

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

Magnusson, M.; Sigurdsson, J.; Armansson, S. E.; Ulfarsson, M. O.; Deborah, H.; and Sveinsson, J. R. 2020. Creating RGB images from hyperspectral images using a color matching function. In *IGARSS 2020-2020 IEEE International Geoscience and Remote Sensing Symposium*, 2045–2048.

Meng, Z.; Ma, J.; and Yuan, X. 2020. End-to-end low cost compressive spectral imaging with spatial-spectral self-attention. In *ECCV*, 187–204. Springer.

Miao, X.; Yuan, X.; Pu, Y.; and Athitsos, V. 2019. l-net: Reconstruct hyperspectral images from a snapshot measurement. In *ICCV*, 4059–4069.

Robles-Kelly, A. 2015. Single image spectral reconstruction for multimedia applications. In *ACM MM*, 251–260.

Shao, Y.; Li, L.; Ren, W.; Gao, C.; and Sang, N. 2020. Domain adaptation for image dehazing. In *CVPR*, 2808–2817.

Shi, Z.; Chen, C.; Xiong, Z.; Liu, D.; and Wu, F. 2018. Hscnn+: Advanced cnn-based hyperspectral recovery from rgb images. In *CVPRW*, 939–947.

Stiebel, T.; Koppers, S.; Seltsam, P.; and Merhof, D. 2018. Reconstructing spectral images from rgb-images using a convolutional neural network. In *CVPRW*, 948–953.

Tu, Z.; Talebi, H.; Zhang, H.; Yang, F.; Milanfar, P.; Bovik, A.; and Li, Y. 2022. Maxim: Multi-axis mlp for image processing. In *CVPR*, 5769–5780.

Wagadarikar, A.; John, R.; Willett, R.; and Brady, D. 2008. Single disperser design for coded aperture snapshot spectral imaging. *Applied optics*, 47(10): B44–B51.

Wang, L.; Sun, C.; Fu, Y.; Kim, M. H.; and Huang, H. 2019. Hyperspectral image reconstruction using a deep spatial-spectral prior. In *CVPR*, 8032–8041.

Wang, L.; Sun, C.; Zhang, M.; Fu, Y.; and Huang, H. 2020. Dnu: Deep non-local unrolling for computational spectral imaging. In *CVPR*, 1661–1671.

Wang, L.; Xiong, Z.; Shi, G.; Wu, F.; and Zeng, W. 2016. Adaptive nonlocal sparse representation for dual-camera compressive hyperspectral imaging. *TPAMI*, 39(10): 2104–2111.

Wang, L.; Zhang, T.; Fu, Y.; and Huang, H. 2018a. Hyperreconnet: Joint coded aperture optimization and image reconstruction for compressive hyperspectral imaging. *TIP*, 28(5): 2257–2270.

Wang, X.; Yu, K.; Dong, C.; and Loy, C. C. 2018b. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *CVPR*, 606–615.

Xiong, Z.; Shi, Z.; Li, H.; Wang, L.; Liu, D.; and Wu, F. 2017. Hscnn: Cnn-based hyperspectral image recovery from spectrally undersampled projections. In *ICCVW*, 518–525.

Yan, L.; Wang, X.; Zhao, M.; Kaloorazi, M.; Chen, J.; and Rahardja, S. 2020. Reconstruction of hyperspectral data from RGB images with prior category information. *TCI*, 6: 1070–1081.

Yang, X.; Chen, J.; Yang, Z.; and Chen, Z. 2021. Attention-Guided NIR Image Colorization via Adaptive Fusion of Semantic and Texture Clues. *arXiv preprint arXiv:2107.09237*.

Yuan, Y.; Zheng, X.; and Lu, X. 2017. Hyperspectral image superresolution by transfer learning. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 10(5): 1963–1974.

Zamir, S. W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F. S.; and Yang, M.-H. 2022. Restormer: Efficient transformer for high-resolution image restoration. In *CVPR*, 5728–5739.

Zamir, S. W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F. S.; Yang, M.-H.; and Shao, L. 2020. Learning enriched features for real image restoration and enhancement. In *ECCV*, 492–511. Springer.

Zamir, S. W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F. S.; Yang, M.-H.; and Shao, L. 2021. Multi-stage progressive image restoration. In *CVPR*, 14821–14831.

Zhao, H.; Gallo, O.; Frosio, I.; and Kautz, J. 2016. Loss functions for image restoration with neural networks. *TCI*, 3(1): 47–57.

Zhao, Y.; Po, L.-M.; Yan, Q.; Liu, W.; and Lin, T. 2020. Hierarchical regression network for spectral reconstruction from RGB images. In *CVPRW*, 422–423.

Zhou, Z.; Qiu, S.; Wang, Y.; Zhou, M.; Chen, X.; Hu, M.; Li, Q.; and Lu, Y. 2021. Swin-Spectral Transformer for Cholangiocarcinoma Hyperspectral Image Segmentation. In *2021 14th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, 1–6. IEEE.