MST++: Multi-stage Spectral-wise Transformer for Effcient Spectral Reconstruction

Yuanhao Cai[1,*], Jing Lin[1,*], Zudi Lin[2], Haoqian Wang[1,†], Yulun Zhang[3], Hanspeter Pfister[2], Radu Timofte[3,4], Luc Van Gool[3]

1 Shenzhen International Graduate School, Tsinghua University 2 Harvard University 3 ETH Zürich 4 JMU Würzburg

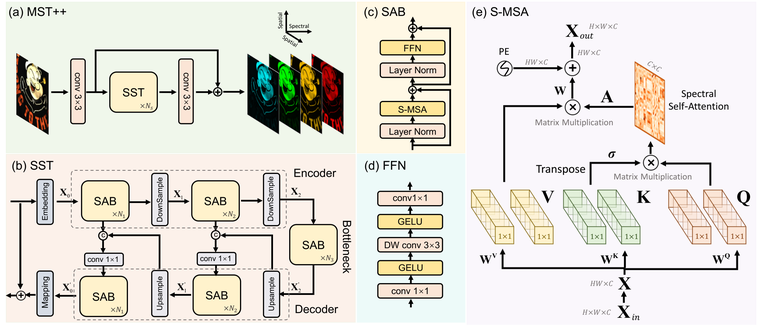
Introduction

Existing leading methods for spectral reconstruction (SR) focus on designing deeper or wider convolutional neural networks (CNNs) to learn the end-to-end mapping from the RGB image to its hyperspectral image (HSI). These CNN-based methods achieve impressive restoration performance while showing limitations in capturing the long-range dependencies and self-similarity prior. To cope with this problem, we propose a novel Transformer-based method, Multi-stage Spectral-wise Transformer (MST++), for efficient spectral reconstruction. In particular, we employ Spectral-wise Multi-head Self-attention (S-MSA) that is based on the HSI spatially sparse while spectrally self-similar nature to compose the basic unit, Spectral-wise Attention Block (SAB). Then SABs build up Single-stage Spectral-wise Transformer (SST) that exploits a U-shaped structure to extract multi-resolution contextual information. Finally, our MST++, cascaded by several SSTs, progressively improves the reconstruction quality from coarse to fine. Comprehensive experiments show that our MST++ significantly outperforms other state-of-the-art methods. Our MST++ is based on our CVPR 2022 work MST and has won the first place in NTIRE 2022 Spectral Recovery Challenge.

Code for MST：https://github.com/caiyuanhao1998/MST
Code for MST++：https://github.com/caiyuanhao1998/MST-plus-plus

Method

Overall Architecture



In figure (a), MST++ is cascaded by N_s Single-stage Spectral-wise Transformers (SSTs). In figure (b), SST adopts a three-level U-shaped architecture. Details of Spectral-wise Attention Block (SAB), Feed-Forward Network (FFN), and Spectral-wise Multi-head Self-Attention (S-MSA) are shown in figure (c), (d), and (e).

Spectral-wise Multi-head Self-Attention

Suppose the input tokens of S-MSA as $\mathbf{X}$ that is projected into $\mathbf{Q}$, $\mathbf{K}$, $\mathbf{V}$ as

$$\mathbf{Q} = \mathbf{X}\mathbf{W}^{\mathbf{Q}}, \mathbf{K} = \mathbf{X}\mathbf{W}^{\mathbf{K}}, \mathbf{V} = \mathbf{X}\mathbf{W}^{\mathbf{V}}$$

Subsequently, $\mathbf{Q}, \mathbf{K}, \mathbf{V}$ are split into N heads along the spectral dimension and the self-attention is calculated inside each $head_j$ as

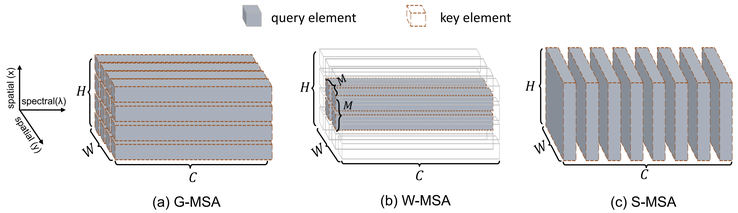
$$\mathbf{A}_j = \mathrm{softmax}(\sigma_j \mathbf{K}^{\mathrm{T}}{}_j \mathbf{Q}_j),\ head_j = \mathbf{V}_j \mathbf{A}_j$$

Then the outputs of N heads are aggregated by a linear projection and is added with a position embedding that is produced by function $f_p(\cdot)$ as

$$\mathbf{S}-\mathbf{MSA}(\mathbf{X}) = \left(\mathrm{Concate}_{j=1}^{N}(head_j)\right)\mathbf{W} + f_p(\mathbf{V})$$

$f_p(\cdot)$ consists of two depth-wise *conv* 3×3, a GELU activation, and reshape operation. Finally, we reshape the above results to get the output feature maps

Discussion with Previous MSA Modules



We compare the computational complexity of our S-MSA, global spatial-wise MSA (G-MSA), and local window-based MSA (W-MSA) as

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

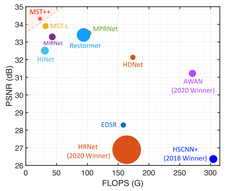
$$O(\mathrm{S-MSA}) = N\left(\left(\frac{C}{N}\right)^2 HW + \left(\frac{C}{N}\right)^2 HW\right)^2 = \frac{2HWC^2}{N}$$

Ensemble Strategies

We adopt self ensemble, multi-scale ensemble, and top-k multi-model ensemble when testing on the test-challenge dataset. The top-k multi-model ensemble averages the outputs of MIRNet, MPRet, Restormer, HINet, MST, and MST++.

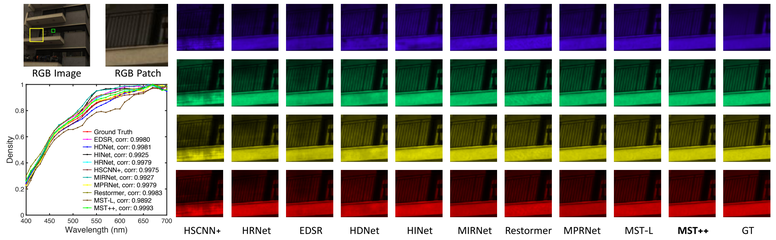
Experiment

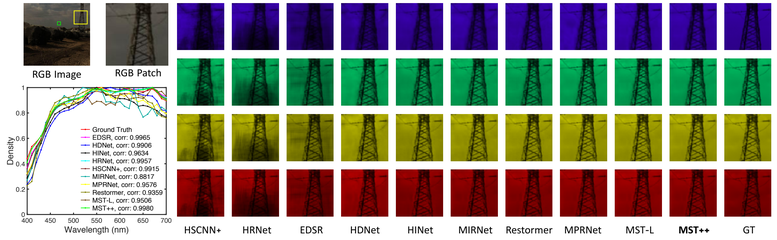
Quantitative Results



Method	Params (M)	FLOPS (G)	MRAE	RMSE	PSNR	Username	MRAE	RMSE
	NTIRE 2022 HSI Dataset - Valid					NTIRE 2022 HSI Dataset - Test		
HSCNN+ [67]	4.65	304.45	0.3814	0.0588	26.36	pipixia	0.2434	0.0411
HRNet [88]	31.70	163.81	0.3476	0.0550	26.89	uslab	0.2377	0.0391
EDSR [45]	2.42	158.32	0.3277	0.0437	28.29	orange_dog	0.2377	0.0376
AWAN [36]	4.04	270.61	0.2500	0.0367	31.22	askldklasfj	0.2345	0.0361
HDNet [29]	2.66	173.81	0.2048	0.0317	32.13	HSHAJii	0.2308	0.0364
HINet [21]	5.21	31.04	0.2032	0.0303	32.51	ptdoge_hot	0.2107	0.0365
MIRNet [84]	3.75	42.95	0.1890	0.0274	33.29	test_pseudo	0.2036	0.0324
Restormer [83]	15.11	93.77	0.1833	0.0274	33.40	gkdgkd	0.1935	0.0322
MPRNet [85]	3.62	101.59	0.1817	0.0270	33.50	deeppf	0.1767	0.0322
MST-L [13]	2.45	32.07	0.1772	0.0256	33.90	mialgo_ls	0.1247	0.0257
MST++	**1.62**	**23.05**	**0.1645**	**0.0248**	**34.32**	**MST++***	**0.1131**	**0.0231**

Qualitative Results





Conclusion

our contributions:

- The first Transformer MST++
- A novel S-MSA
- SOTA results
- A Strong Baseline

Code and models are publicly available at