Supplementary Material: Spatial-Spectral Transformer for Hyperspectral Image Denoising

Miaoyu Li [1], **Ying Fu** [1] *, **Yulun Zhang** [2]

[1] Beijing Institute of Technology, [2] ETH Zürich
{miaoyli, fuying}@bit.edu.cn, yulun100@gmail.com

Overview

Contents of Supplementary Material are structured as:

1) Additional implementation details of proposed SST.
2) Experimental results on Washington DC Mall dataset.
3) Model generalization ability from ICVL to CAVE.
4) Detailed analysis of ablation study. The inference time comparison and other variants are also included.
5) More visual examples of our denoising results.

Implementation details

In this section, we provide additional implementation details on ICVL dataset. The proposed attention block in our Spatial-Spectral Transformer consists of three sub networks 1) Non-Local Spatial-Attention (NLSA), 2) Global Spectral Attention (GSA), and 3) Multilayer Perceptron (MLP). Now we discuss the detailed architecture information of each component in Table 1, Table 2, and Table 3.

Supposing the input HSI from ICVL datasets is of size $H \times W \times 31$. H and W stand for the spatial resolution and spectral resolution. The shallow feature of size $H \times W \times 90$ is extracted by a convolution layer in advance. The NLSA module could be regarded as a collection of attention operations on a group of windows in size $8 \times 8 \times 90$. It firstly projects the input features to $8 \times 8 \times 270$ and splits them to Q, K, and V in channel-wise to conduct the attention operation. The GSA conducts a global attention on image, thus its window size can be regarded as $H \times W$.

Num	Layer type	Inp Dim	Out Dim	Win size
1	Linear	90	270	8×8
2	Spatial Attention	90	90	8×8
3	Linear	90	90	8×8

Table 1: The hyperparameters of our NLSA layer.

Num	Layer type	Inp Dim	Out Dim	Win Size
1	Linear	90	270	$H \times W$
2	Spectral Attention	90	90	$H \times W$
3	Linear	90	90	$H \times W$

Table 2: The hyperparameters of our GSA layer.

*Corresponding author

Num	Layer type	Input Dim	Output Dim
1	Linear	90	180
2	Gelu	180	180
3	Linear	180	90

Table 3: The hyperparameters of our MLP layer.

Experiment results on Washington DC Mall

Setup. We carry out an additional experiment on Washington DC Mall dataset, which consists of 1280×307 pixels with 191 bands. Following (Shi et al. 2021), the whole HSI is split into two parts for training and testing. The part of size $1080 \times 303 \times 191$ is used for training and the part of size $200 \times 200 \times 191$ is used for testing. The training dataset and testing HSI are scaled into the range [0,1] before adding simulated noise. It is worth mentioning that the training samples are quite limited in this experiment.

We follow the noise settings mentioned in the main paper to conduct the simulated experiments. In the first experiment, we add i.i.d Guassian noise with known variance on clean HSI to evaluate the denoising performance. In the second experiment, we conduct noise removal under mixture noise. Each band is firstly contaminated by non-i.i.d Guassian noise with unknown variance. Then all the bands are randomly degraded by complex noises, including stripe noise, deadline noise, and impulse noise.

Quantitative Results. The quantitative results of simulated i.i.d Gaussian noise and mixed noise on Washington DC Mall are shown in Table 4. We can see that our method achieves better quantitative results under most metrics.

Visual Comparison. The visual comparison between different HSI denoising methods is shown in Figure 1. Results are made with bands 60, 27, and 17 for the red, green, and blue colors for better visual effect. The noisy HSI is corrupted by severe mixture noise. We can observe that BM4D and LLRT restore the HSI with obvious stripe noise and incomplete textures. NGMeet and QRNN3D preserve the edge information but fail to perfectly remove the stripe noise. Compared to T3SC, our proposed SST can reconstruct detailed texture with high fidelity in spectral dimension.

Sigma	Metrics	Methods								
		Noisy	BM4D	LLRT	TSLRLN	NGMeet	HSID-CNN	QRNN3D	T3SC	Ours
10	PSNR	28.132	36.859	**39.471**	36.870	37.291	34.455	37.077	38.491	39.360
	SSIM	0.9897	0.9985	0.9990	0.9982	0.9983	0.9973	0.9983	0.9988	**0.9991**
	SAM	9.3403	3.8209	3.0128	3.7575	3.4537	4.1638	3.7708	3.2415	**2.8288**
30	PSNR	18.589	29.946	33.879	33.682	34.525	31.544	33.797	35.421	**35.448**
	SSIM	0.9143	0.9929	0.9965	0.9967	0.9974	0.9949	0.9968	0.9979	**0.9979**
	SAM	22.116	6.7652	4.6160	4.8418	4.3104	5.6865	4.6302	3.9436	**3.7852**
50	PSNR	14.147	26.756	30.934	31.406	31.847	28.939	31.277	**32.999**	32.976
	SSIM	0.7933	0.9853	0.9937	0.9947	0.9952	0.9911	0.9946	0.9964	**0.9964**
	SAM	31.656	8.5073	5.5127	5.6736	5.2037	7.4539	5.5461	4.6118	**4.4457**
70	PSNR	11.229	24.688	28.963	29.715	29.642	26.344	29.071	**31.215**	31.214
	SSIM	0.6624	0.9763	0.9905	0.9923	0.9922	0.9841	0.9913	0.9946	**0.9946**
	SAM	39.062	9.7828	6.1752	6.3233	6.1058	9.7970	6.8078	5.1741	**5.0022**
10-70	PSNR	14.592	27.073	31.225	31.635	32.113	29.280	31.551	**33.259**	33.252
	SSIM	0.8097	0.9864	0.9941	0.9950	0.9955	0.9917	0.9949	0.9966	**0.9966**
	SAM	30.605	8.3394	5.4025	5.5950	5.1163	7.2106	5.4306	4.5261	**4.3932**
mix	PSNR	13.121	20.678	22.719	25.457	23.939	26.901	29.827	26.707	**30.148**
	SSIM	0.6613	0.9138	0.9496	0.9371	0.9218	0.9842	0.9920	0.9851	**0.9934**
	SAM	38.328	14.939	11.221	14.761	15.627	8.6240	7.3468	11.038	**5.4279**

Table 4: Quantitative comparison under various noise levels on Washington DC Mall dataset. Best results are in bold.

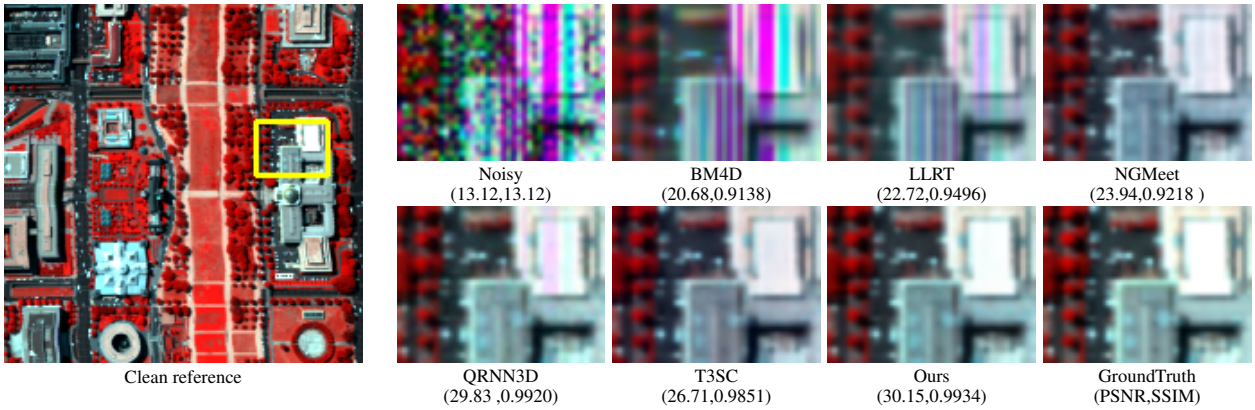

Figure 1: Visual quality comparison under mixture noise on Washington DC Mall dataset.

Metrics	PNSR (dB)↑	SSIM↑	SAM↓	ERGAS↓
T3SC	36.57	0.9916	14.72	22.26
QRNN3D	36.54	0.9910	19.17	22.67
Ours	**37.30**	**0.9916**	**14.64**	**20.34**

Table 5: Results under Guassian noise level [10,70] on CAVE daaset with models trained on ICVL dataset.

Model Generalization Ability.

To show the generalization ability of our proposed SST model, we directly verify the denoising results on CAVE dataset (Yasuma et al. 2010) with models trained on ICVL dataset. CAVE dataset has similar wavelength and band number to ICVL dataset. We include all 32 HSIs in CAVE dataset as testing HSIs. All competing networks are under the same settings. In Table 5, our model achieves better results than T3SC and QRNN3D, showing the superior generalization ability of our proposed model.

Detailed Analysis of Ablation Study

Components of SSMA

More comparisons and detailed analyses about our proposed attention module and other choices of attention module are discussed here. As shown in Figure 2, there are several different strategies to conduct the spatial self-attention with spectral self-attention for HSI denoising. Specifically, NLSA stands for the non-local spatial self-attention layer, and GSA stands for the global spectral self-attention layer. When changing the components of attention block, other settings of the network stay the same.

Single layer of NLSA/GSA. The most basic application of attention module are shown as Figure 2 (a) and Figure 2 (b), which employ a single attention module to extract spatial or spectral information of HSI. From Table 6, it can be observed that the application of single spectral attention achieves the worst performance, showing the importance of preserving non-local spatial information.

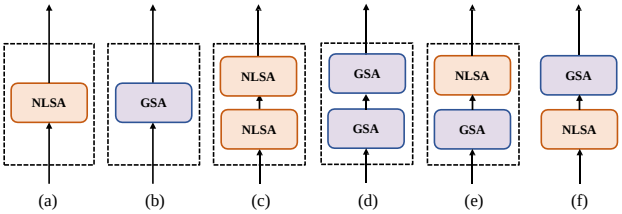

Figure 2: Different types of attention Block. NLSA stands for the non-local spatial self-attention layer, and GSA stands for the global spectral self-attention layer.

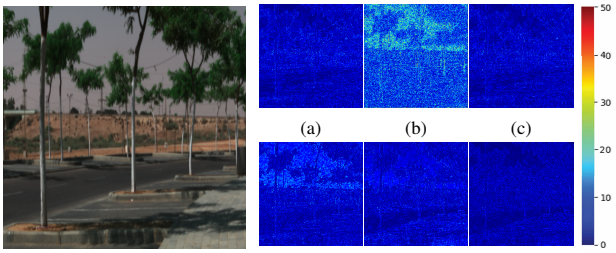

Figure 3: The error maps at the 19th band of our method with different components with $\sigma = 50$ on ICVL dataset. (a)-(f) refer to the results of different structures in Figure 2.

Duplicate Layer of NLSA/GSA. To represent the effectiveness of proposed attention module under fail computation cost. We propose attention blocks that consist of two NLSA modules or two GSA modules. The detailed architectures are shown in Figure 2 (c) and Figure 2 (d). Duplicate layers perform better results than single layer, but are not as good as the NLSA-GSA layer, in line with our expectations.

The position of NLSA and GSA. The exchanged position of NLSA and GSA is under consideration in Figure 2 (e) and Figure 2 (f). With same numbers of params and GFLOPs, our proposed module still obtains better results compared to Spectral-Spatial Attention, indicating the effectiveness of conducting spectral attention after spatial attention.

Figure 3 shows the images of ground truth and error maps at the 19th band on one HSI from ICVL dataset with σ=50. The error maps are the absolution errors between the ground truth and the denoised results. The visual results imply that our proposed attention module produce better denoising result that is closer to the groundtruth.

Model Size

Table 7 reports the params, GFLOPs, and inference time of two comparative deep learning-based methods and our proposed method on ICVL dataset. Specifically, the GLOPs are calculated on $64\times64\times31$ patch while the inference time is calucated on testing HSI with size of $512\times512\times31$. We also include the SST-S and SST-B for comparison.

We observe that with deeper layers, QRNN3D gets few improvement on PSNR but has a large growth in model complexity. With comparable inference time, our proposed SSTs significantly achieve better performance.

Method	Params (M)	GFLOPs↓	PSNR (dB)↑
(a) w/o NLSA	3.00	14.3	34.67
(b) w/o GSA	2.98	13.1	40.44
(c) NLSA-NLSA	4.23	20.1	40.56
(d) GSA-GSA	4.08	21.4	39.82
(e) GSA-NLSA	4.15	20.7	40.69
(f) NLSA-GSA (Ours)	4.15	20.7	41.09

Table 6: Ablation study related to the effectiveness of our proposed spatial-spectral multi-head self-attention.

Method	Params (M)	GFLOPs ↓	Inference time (s) ↓	PSNR (dB) ↑
QRNN3D	0.89	19.6	0.62	39.70
QRNN3D-L	1.34	30.6	1.15	39.82
T3SC	0.83	N/A	1.04	40.39
SST-S (Ours)	1.4	10.6	0.86	40.74
SST-B (Ours)	3.47	17.4	1.01	40.92
SST-L (Ours)	4.14	20.6	1.10	41.09

Table 7: Comparison with state-of-the-art models on model size and inference time on ICVL dataset.

Network Components

We explore two variants of our proposed SST and the results are shown in Table 8. Method A is our proposed SST without the global skip-connection that adds the noisy input to the output. Method B replace the GSA layer with spectral self-attention that calculated in local windows. Though these two variants have same parameters and GFLOPs with the proposed SST, they are less effective to restore the clean HSI from noisy observation, proving the effect of global skip-connection and global spectral self-attention. With global skip-connection, the proposed SST could make direct connection between input and output and preserve critical information. With global spectral self-attention, the comprehensive features across HSI are carefully considered.

Method	global skip connection	sepctral self-attention	PSNR (dB) ↑
A	✗	global	40.60
B	✓	local	40.78
Ours	✓	global	41.09

Table 8: Quantitative comparison of our proposed SST with two variants

More visual examples

We provide more visual results on ICVL dataset in Figure 4 and Figure 5. We also use the pseudo color images to show the denoising results. It can be concluded that our method capture the textures and structures well, and preserve more spatial and sepctral information than competing methods.

References

Shi, Q.; Tang, X.; Yang, T.; Liu, R.; and Zhang, L. 2021. Hyperspectral image denoising using a 3-D attention denoising network. *IEEE Transactions on Geoscience and Remote Sensing*, 59(12): 10348–10363.

Yasuma, F.; Mitsunaga, T.; Iso, D.; and Nayar, S. K. 2010. Generalized assorted pixel camera: postcapture control of resolution, dynamic range, and spectrum. *IEEE transactions on image processing*, 19(9): 2241–2253.

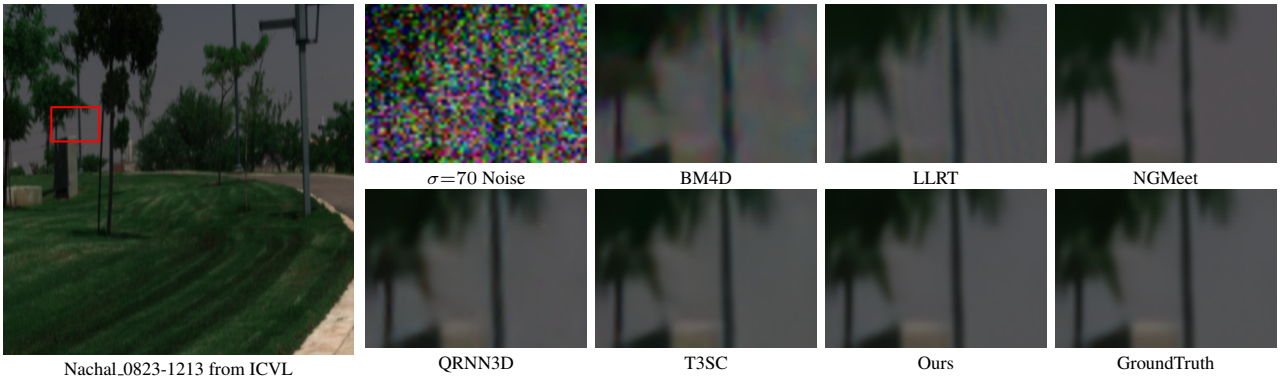

Figure 4: Visual quality comparison under σ=70 on ICVL dataset using pesudo color image.

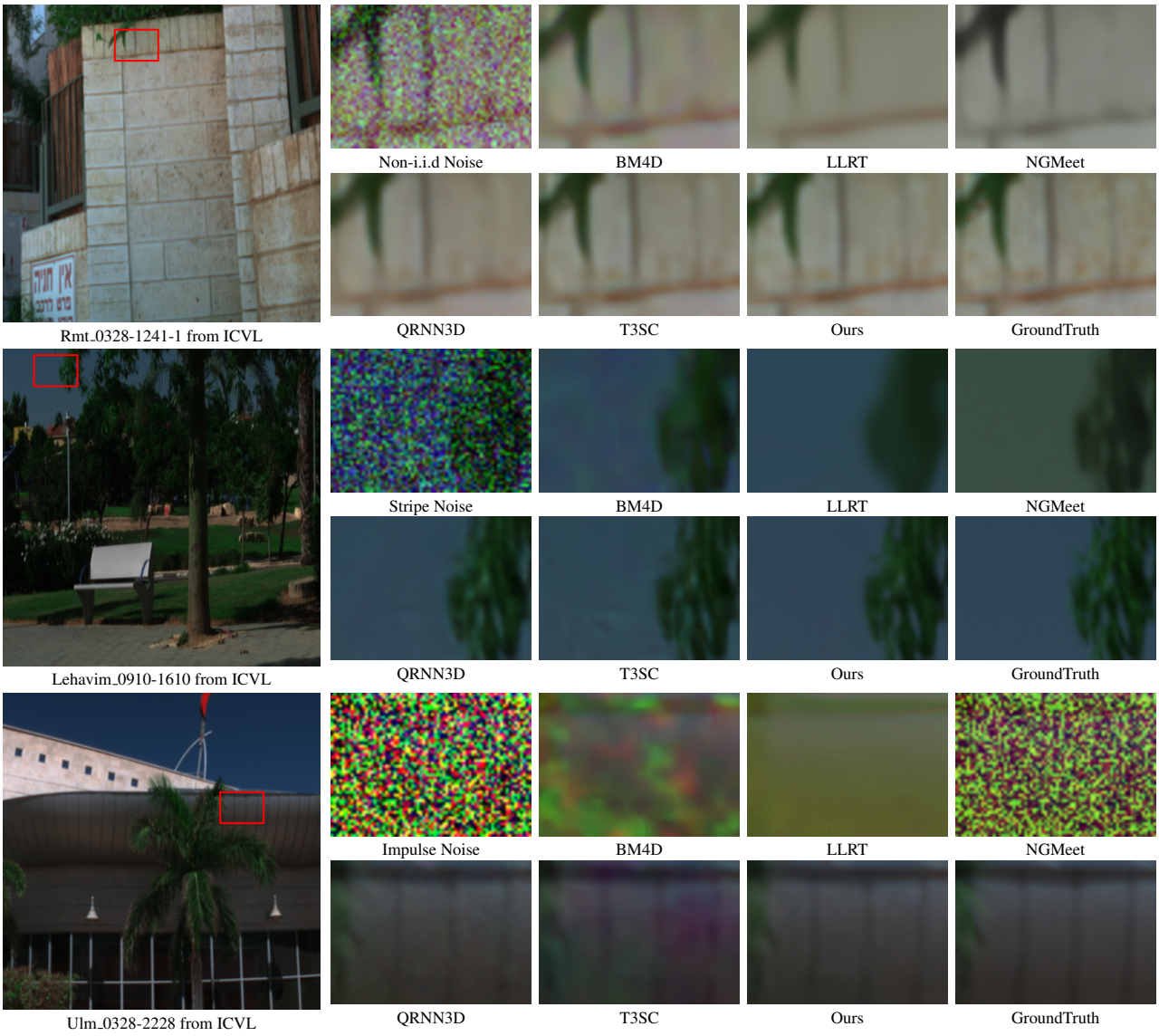

Figure 5: Visual quality comparison under complex noise using pesudo color image. Here, we use non-i.i.d Gaussian noise, Gaussian + stripe, and Gaussian + impulse, respectively.