Supplementary Materials: Spectral Enhanced Rectangle Transformer for Hyperspectral Image Denoising

Miaoyu Li [1*], Ji Liu[2*], Ying Fu[1†], Yulun Zhang[3], Dejing Dou[4]
[1]Beijing Institute of Technology, [2]Baidu Inc., [3]ETH Zürich, [4]BCG X
miaoyli@bit.edu.cn, liuji04@baidu.com, fuying@bit.edu.cn,
yulun100@gmail.com, dejingdou@gmail.com

In this supplementary material, we present more analysis and results of our proposed Spectral Enhanced Rectangle Transformer (SERT).

- We give a detailed analysis of the rectangle self-attention (RA) module in Section A.
- We show additional ablation study on different hyperparameter settings of our method in Section B.
- We provide the visualization of attention maps in Section C to further illustrate our proposed method.
- We show more visual comparisons between our method and the state-of-the-art methods in Section D.

A. Analysis of Rectangle Self-Attention

A.1. Self-Attention for Denoising

Hyperspectral image denoising is a classic inverse problem and has been studied for a long time. Taking a spatial average of neighboring pixels is the simplest way for denoising since most pixels have roughly the same value as their neighbor. A typical example is the Mean filtering. However, not all neighbors have the same value. Accordingly, it is important to consider neighbors that have similar values. Existing spatial domain methods [2] aim to remove noise by calculating the value of each pixel based on the correlation between pixels/image patches in the original image.

For our proposed rectangle self-attention, it also provide a re-weighting mechanism for noise removal. As stated in Eq. (6) in the main paper, attention matrix $\boldsymbol{A}_i^1$ of the i-th horizontal rectangle $\boldsymbol{Z}_i^1$ is calculated through

$$\boldsymbol{A}_i^1 = \text{SoftMax}(\boldsymbol{Q}_i^1 {\boldsymbol{K}_i^1}^T / \sqrt{d} + \boldsymbol{P}) \tag{1}$$

where $\boldsymbol{Q}_i^1$ and $\boldsymbol{K}_i^1$ are the projected matrix of input $\boldsymbol{Z}_i^1$. d and $\boldsymbol{P}$ are the feature dimension and the position embedding. The attention matrix $\boldsymbol{A}_i^1 \in \mathbb{R}^{hw \times hw}$ actually provides the similarity information in the i-th horizontal rectangle. Then, $\boldsymbol{V}_i^1$ is re-weighted by $\boldsymbol{A}_i^1$ as

$$\hat{\boldsymbol{Z}}_i^1 = \boldsymbol{A}_i^1 \boldsymbol{V}_i^1 \tag{2}$$

Through self-attention, the noise pixels can obtain information from their similar neighboring pixels. Normally, the traditional spatial filters eliminate noise to a reasonable extent but lose sharp edges. Our proposed SERT benefits from the power of deep learning and has a better model capacity for texture preservation and detail maintenance.

Architecture Hyperparameter	Settings
embed dim	96
size of rectangles	[16,1], [32,2], [32,4]
memory block	128
rank size	12
weight factor of SE	0.1

Table 1. Employed settings of hyperparameters in our SERT.

A.2. Non-local Rectangular Self-Attention

We apply two strategies to make interactions between non-overlapping rectangles that enable the network to obtain information beyond the rectangle. First, the shift operation is employed in spatial domain. Thus, in consecutive layers, the pixels included in a rectangle are different. Second, the spectral enhancement (SE) module aggregates the information from several neighboring rectangles. The acquired spectral characteristic contributes to the result of RA module by adding enhanced features to it. These two strategies allow our proposed SERT to obtain non-local features in the spatial domain for HSI denoising.

In addition to interactions between rectangles, we conduct the rectangle self-attention both horizontally and vertically by splitting the spectral domain into two parts. The noisy pixels can search pixels that are more similar to themselves from a larger number of neighbors compared to window self-attention [6], which is conducted on the whole spectral domain. Since neighboring pixels are more likely to belong to the same object or material, adjacent pixels are

*Equal Contribution, † Corresponding author

more likely to be similar in spectral characteristics. It implies that exploring the similarity in the spatial domain on the entire feature map is less efficient. Therefore, we conduct the self-attention in rectangles instead of using stripes [4] to achieve better results with comparable complexity.

B. Additional Ablation Studies

In this section, we first provide the detailed hyperparameters of our proposed SERT in Table 1. Then, we analyze the influence of different hyperparameters and network architectures on the denoising results. These experiments are conducted on ICVL dataset with random Gaussian noise.

Hyperparameters of SE Module. To study the effect of the weighting factor of SE module that contributes to the Transformer block, we report the denoising results of different values of the weighting factor in Table 2(a). It implies that the proposed SE module can facilitate the denoising process regardless of the weight factor. SERT with a weight factor of 0.1 achieves the best performance.

A key component of our proposed SE module is the memory unit, which restores the global low-rank vectors. Different settings of memory blocks and rank size have different impacts on the network. To study the effect, we set the number of memory block B to 64, 128, and 256 with different rank size K, respectively. As shown in Table 2(b), our employed setting slightly outperforms other choices.

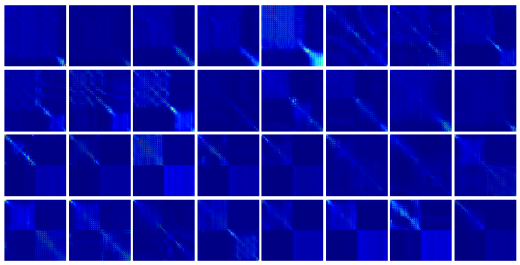
RA Module. The effectiveness of our proposed rectangle self-attention module is verified in Table 2(c). Without RA module, PSNR is decreased by 2dB. Using the SE module alone in Transformer block without RA module is not sufficient to extract spatial features.

Model Size. We compare four variants of SERT with other deep learning methods in Table 2(d). The variants include SERT-T (Tiny), SERT-S (Small) and SERT-B (Base). With much less time and complexity, our SERT-T still outperforms other deep learning methods.

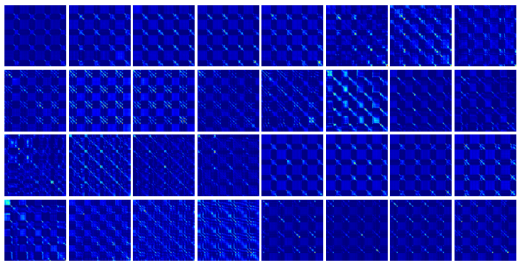
C. Visualization of Attention Maps.

We further conduct a visual analysis of the intermediate features obtained by our method. The attention maps of two branches in RA module are shown in Figure 1. For the example cropped figure, we show the partitions of our horizontal rectangle branch and vertical rectangle branch in Figure 1(a). The example figure is of size $64 \times 64 \times 31$. We show its false-color image that is generated using bands 10, 16, and 29. The feature map of the example image is split into 32 horizontal rectangles and 32 vertical rectangles. The corresponding attention maps of each horizontal rectangle and vertical rectangle are shown in Figures 1(b) and 1(c) respectively. The horizontal rectangle self-attention branch and vertical rectangle self-attention branch present different features to model correlations between pixels.

(a) The partition of self-attention in horizontally and vertically with image size $64 \times 64 \times 31$.



(b) Attention maps of horizontal rectangle self-attention.



(c) Attention maps of vertical rectangle self-attention.

Figure 1. Visualization of attention maps of horizontal RA module and vertical RA module.

D. Additional Visual comparisons

D.1. More Visual Results on Realistic Dataset

We provide the denoising results as well as it corresponding spectral density curves in Figures 2 and 3 to visually evaluate the spectral fidelity of our method on Realistic [11] dataset. It can be observed that our method shows a very similar spectral curve to the GroundTruth with a high correlation. Traditional model-based methods including BM4D and NGMeet lose high frequency details in the denoising process while other deep learning methods including T3SC and MACNet cannot suppress the real noise well. Our method achieves the best visual effect.

D.2. More Visual Results on ICVL Dataset

We show more visual results on the ICVL dataset under various complex noise, which are illustrated in Figures 4, 5, 6 and 7. Our method achieves the best denoising results under various noise.

Specifically, we illustrate the visual denoising results of all the methods under non-i.i.d Gaussian noise in Figure 4. While NGMeet and T3SC exhibit excessive smoothness,

Weight	0	0.1	0.5	0.9	1
PSNR/SSIM	42.06	42.82	42.69	42.76	42.74

(a) Effects of the weighting factor of SE module.

Memory Blocks B	Rank Size K	Params (M)	GFLOPS	PSNR
64	3	1.85	1018.8	42.73
	12	1.89	1018.9	42.72
	24	1.95	1018.9	42.67
128	3	1.85	1018.8	42.71
	12	1.91	1018.9	**42.82**
	24	1.97	1019.0	42.76
256	3	1.86	1018.9	42.73
	12	1.93	1018.9	42.72
	24	2.03	1019.0	42.66

(b) Quantitative comparison with different settings of blocks in memory unit and the rank size of SE module.

Method	Params (M)	GFLOPS	PSNR (dB)
w/o RA	0.82	681.6	40.55
w RA (Ours)	1.91	1018.9	42.82

(c) Ablation study on the effectiveness of RA module.

Method	Settings	Params (M)	GFLOPS	Time (s)	PSNR (dB)
QRNN3D [8]	-	0.83	2513.7	0.683	41.34
T3SC [1]	-	0.83	-	1.123	41.64
MACNet	-	0.43	-	3.627	41.31
SERT-T	(4, 4)	0.98	501.4	0.424	42.41
SERT-S	(4, 4, 4)	1.4	746.1	0.524	42.56
SERT-B	(6, 6, 6)	1.91	1018.9	0.717	42.82

(d) Quantitative comparison with four variants of SERT. Params, FLOPS, Inference Time and PSNR are reported.

Table 2. Ablation studies on architecture and hyperparameters.

our method not only removes noise, but also effectively restores the texture. Results under deadline noise are shown in Figure 5. Figure 6 provides the results under impulse noise. Figure 7 shows the results under mixture noise. Though QRNN3D, T3SC, and MAC-Net obtain relatively good denoising results compared to traditional model-based methods under deadline noise and impulse noise, their results under mixture noise are not promising. Our methods can effectively handle the complex mixture noise, showing its robustness and stronger model capacity.

References

[1] Théo Bodrito, Alexandre Zouaoui, Jocelyn Chanussot, and Julien Mairal. A trainable spectral-spatial sparse coding model for hyperspectral image restoration. In *NeurIPS*, volume 34, pages 5430–5442, 2021. 3, 4, 5

[2] Antoni Buades, Bartomeu Coll, and J-M Morel. A non-local algorithm for image denoising. In *CVPR*, volume 2, pages 60–65. Ieee, 2005. 1

[3] Xiangyong Cao, Xueyang Fu, Chen Xu, and Deyu Meng. Deep spatial-spectral global reasoning network for hyperspectral image denoising. *IEEE TGRS*, 2021. 4, 5

[4] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Weiming Zhang, Nenghai Yu, Lu Yuan, Dong Chen, and Baining Guo. Cswin transformer: A general vision transformer backbone with cross-shaped windows. In *CVPR*, pages 12124–12134, 2022. 2

[5] Wei He, Quanming Yao, Chao Li, Naoto Yokoya, and Qibin Zhao. Non-local meets global: An integrated paradigm for hyperspectral denoising. In *CVPR*, pages 6868–6877, 2019. 4, 5

[6] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 10012–10022, 2021. 1

[7] Matteo Maggioni, Vladimir Katkovnik, Karen Egiazarian, and Alessandro Foi. Nonlocal transform-domain filter for volumetric data denoising and reconstruction. *IEEE TIP*, 22(1):119–133, 2012. 4, 5

[8] Kaixuan Wei, Ying Fu, and Hua Huang. 3-d quasi-recurrent neural network for hyperspectral image denoising. *IEEE TNNLS*, 32(1):363–375, 2020. 3, 4, 5

[9] Fengchao Xiong, Jun Zhou, Qinling Zhao, Jianfeng Lu, and Yuntao Qian. Mac-net: Model-aided nonlocal neural network for hyperspectral image denoising. *IEEE TGRS*, 60:1–14, 2021. 4, 5

[10] Qiangqiang Yuan, Qiang Zhang, Jie Li, Huanfeng Shen, and Liangpei Zhang. Hyperspectral image denoising employing a spatial–spectral deep residual convolutional neural network. *IEEE TGRS*, 57(2):1205–1218, 2018. 4, 5

[11] Hongyan Zhang, Lu Liu, Wei He, and Liangpei Zhang. Hyperspectral image denoising with total variation regularization and nonlocal low-rank tensor decomposition. *IEEE TGRS*, 58(5):3071–3084, 2019. 2, 4

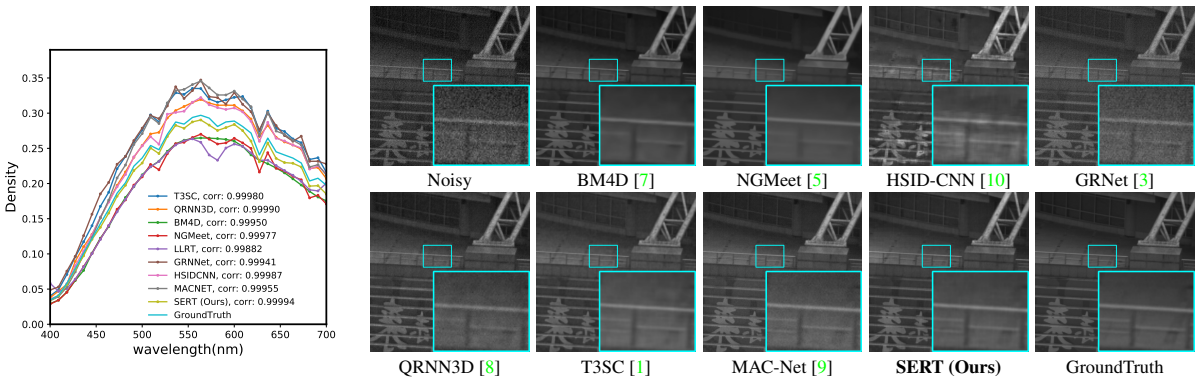


Figure 2. More Visual comparison on Realistic dataset [11] of scene 2 on band 27 with corresponding spectral density curves.

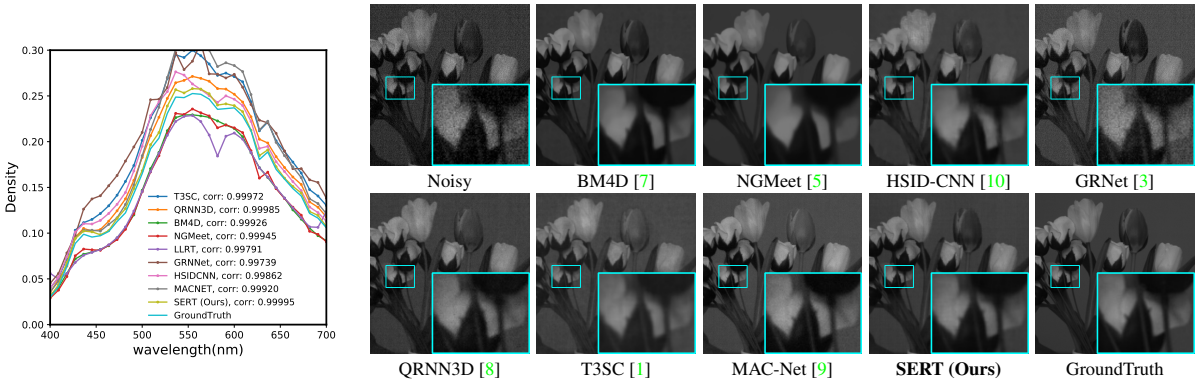


Figure 3. More Visual comparison on Realistic dataset [11] of scene 52 on band 27 with corresponding spectral density curves.

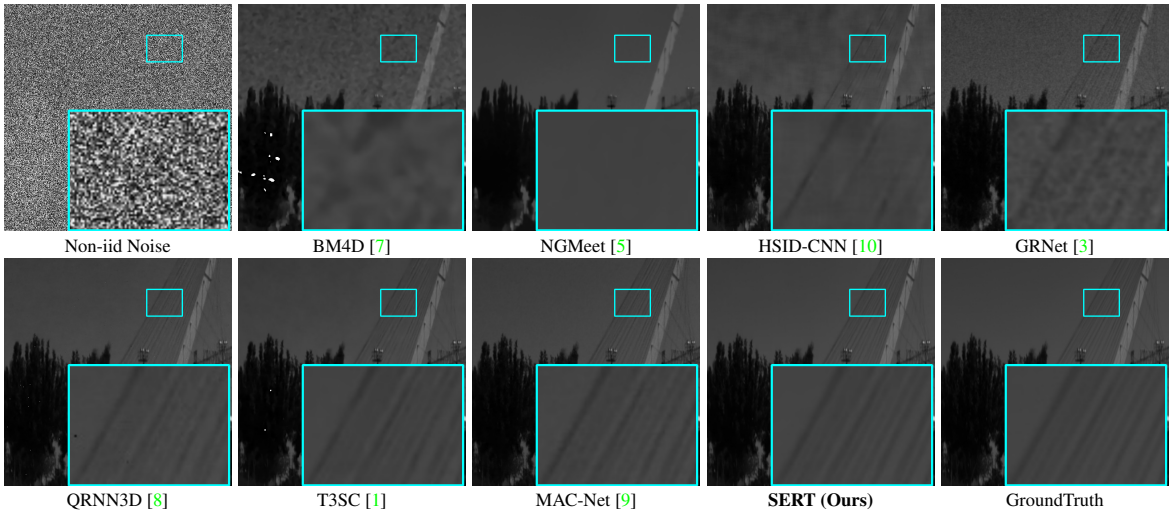


Figure 4. More Visual comparison on ICVL dataset on band 29 under non-iid Gausssina noise.

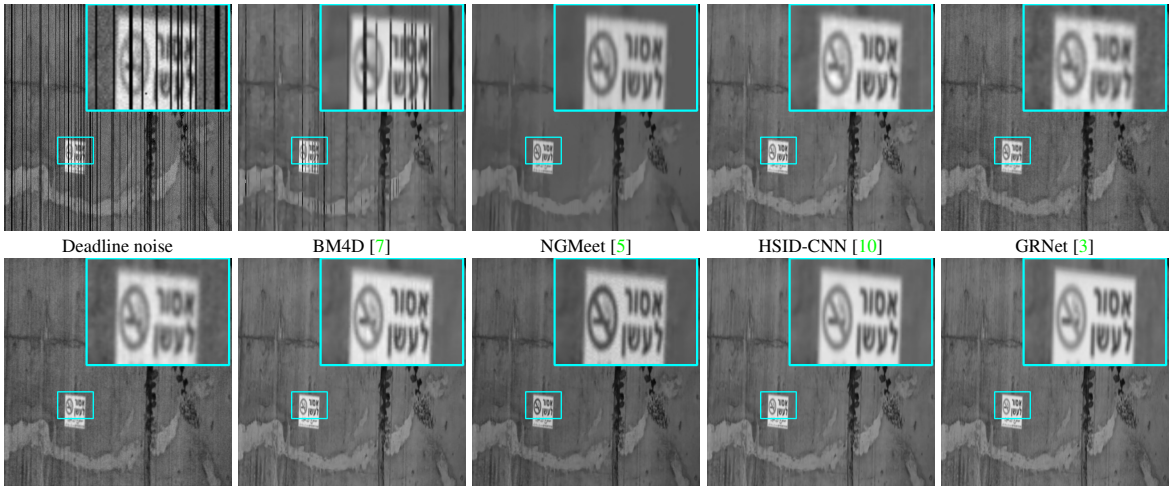


Figure 5. More Visual comparison on ICVL dataset on band 27 under deadline noise.

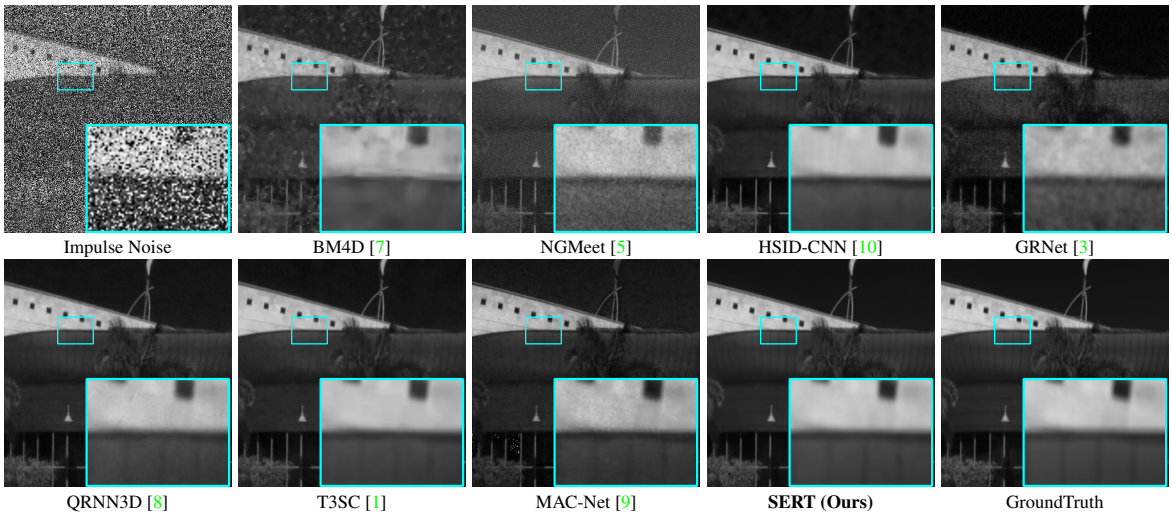


Figure 6. More Visual comparison on ICVL dataset on band 31 under impulse noise.

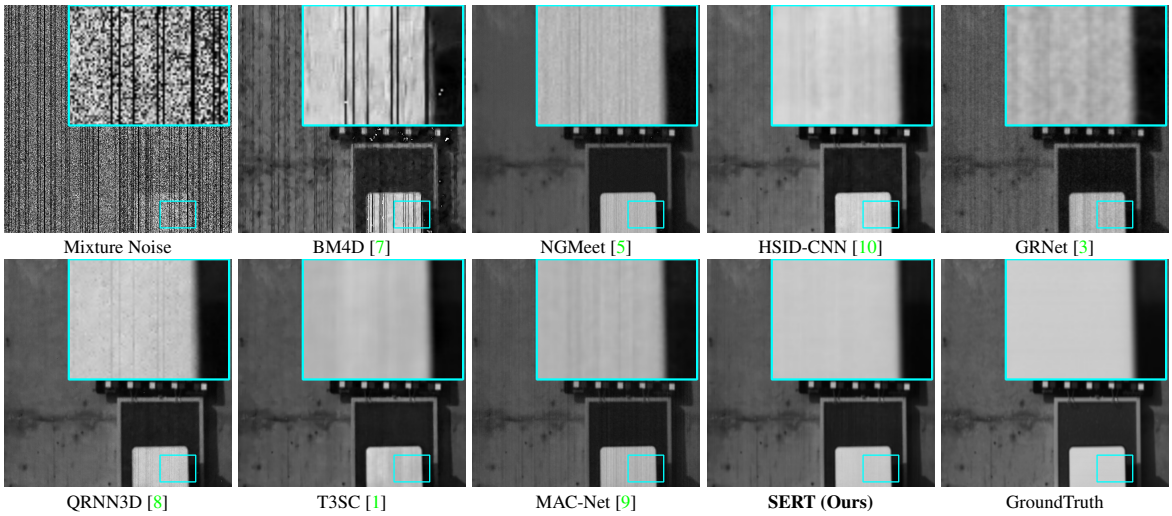


Figure 7. More Visual comparison on ICVL dataset on band 30 under mixture noise.