Supplementary Material for "Low-Light Image Enhancement with Multi-stage Residue Quantization and Brightness-aware Attention"

Yunlong Liu[1,*] Tao Huang[1,*] Weisheng Dong[1,†] Fangfang Wu[1] Xin Li[2] Guangming Shi[1]
[1] Xidian University [2] University at Albany
liuyunlong@stu.xidian.edu.cn thuang_666@stu.xidian.edu.cn wsdong@mail.xidian.edu.cn
wufangfang@xidian.edu.cn xli48@albany.edu gmshi_xidian@163.com

In this Supplementary Material, we provide details of the spectral attention block (SAB) [1, 2]. In addition, more visual comparison results for different scenes are provided.

1. Spectral attention block

As shown in Figure 1 (a), the spectral attention block (SAB) consists of two cascaded residual modules. Specifically, the two residual modules contain a layer norm layer, a short-cut connection, and a spectral-wise multi-head self-attention (S-MSA) or FFN layer, respectively. The FFN layer consists of three linear projections.

Spectral-wise Multi-head Self-Attention Figure 1 (b) demonstrates the S-MSA calculation process with a single head, some details are omitted for brevity.

Given the embedded feature $\mathbf{X}_{in} \in \mathbb{R}^{H\times W\times C}$ as input, $\mathbf{X}_{in}$ is first spatially flattened to tokens $\mathbf{X} \in \mathbf{R}^{HW\times C}$. Then $\mathbf{X}$ is linearly mapped to the query $\mathbf{Q} \in \mathbb{R}^{HW\times C}$, the key $\mathbf{K} \in \mathbb{R}^{HW\times C}$, and the value $\mathbf{V} \in \mathbb{R}^{HW\times C}$ as follows:

$$\mathbf{Q} = \mathbf{X}\mathbf{W}^{Q}, \mathbf{K} = \mathbf{X}\mathbf{W}^{K}, \mathbf{V} = \mathbf{X}\mathbf{W}^{V}, \tag{1}$$

where $\mathbf{W}^{Q}, \mathbf{W}^{K}, \mathbf{W}^{V} \in \mathbb{R}^{C\times C}$ are trainable parameters. Following that, $\mathbf{Q}$, $\mathbf{K}$ and $\mathbf{V}$ are split into N heads along the spectral dimension, respectively:

$$\begin{aligned}[\mathbf{Q}_1, \mathbf{Q}_2, ..., \mathbf{Q}_N] &= \mathbf{Q},\\ [\mathbf{K}_1, \mathbf{K}_2, ..., \mathbf{K}_N] &= \mathbf{K},\\ [\mathbf{V}_1, \mathbf{V}_2, ..., \mathbf{V}_N] &= \mathbf{V}.\end{aligned} \tag{2}$$

S-MSA treats each channel as a token and computes self-attention for head j as follows:

$$\mathbf{A}_j = \mathrm{Softmax}(\sigma_j \mathbf{K}_j^T \mathbf{V}_j), \tag{3}$$

$$head_j = \mathbf{V}_j \mathbf{A}_j, \tag{4}$$

where $\sigma_j \in \mathbb{R}^1$ is a trainable parameter to adapt self-attention $\mathbf{A}_j$. After that, all outputs of N heads are concatenated and fed into a linear projection layer and a position embedding (PE) layer as

$$\text{S-MSA}(\mathbf{X}) = \mathrm{Concat}_{j=1}^{N}(head_j) \cdot \mathbf{W} + \mathrm{PE}(\mathbf{X}) \tag{5}$$

where $\mathbf{W}$ is a trainable matrix and $\mathrm{PE}(\cdot)$ denotes a stack of two convolution layers.

2. More visual comparison results

Note that the input images shown in Figures 1, 2, and 5-8 of the main text are actually grayscale versions of the low-light images. At first, we provide the color version of the low-light inputs as shown in Figures **??-??**. Furthermore, we provide more visual comparison results of different scenes on the LOLv1 [3], LOLv2-Real [5], and LOLv2-Synthetic [5] dataset, as shown in Figures 2-4 for further demonstration of the superiority of our proposed methods.

MIR-Net [6] tends to produce noise and blurred artifacts, as shown in Figure 2 and 3 and a different color tone from the ground truth in Figure 4. DCC-Net [7] shows color inconsistency and blurred artifacts as shown in Figure 2 (middle) and (bottom), respectively. SNR [4] produces blurred artifacts when faced with complex details, as shown in Figure 2 and 3, and the color tends to be different from the ground truth as shown in Figure 3 and 4. Compared to the above methods, our proposed method reconstructs higher visual quality with better color tone and more image details and textures.

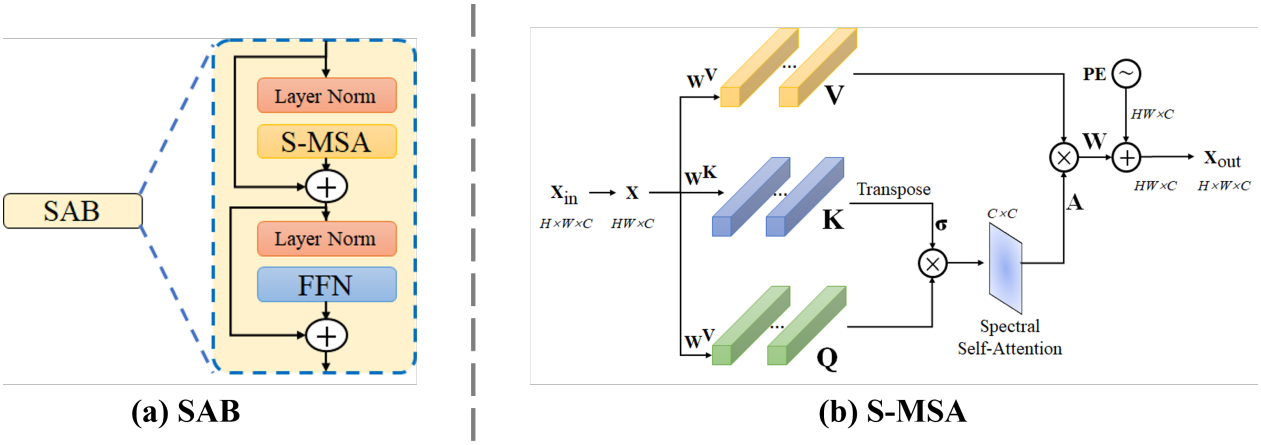

Figure 1: (a) Structure of spectral attention block (SAB), where S-MSA and FFN denote Spectral-wise Multi-head Self-Attention and Feed Forward Network, respectively. (b) The calculation process of Spectral-wise Multi-head Self-Attention.

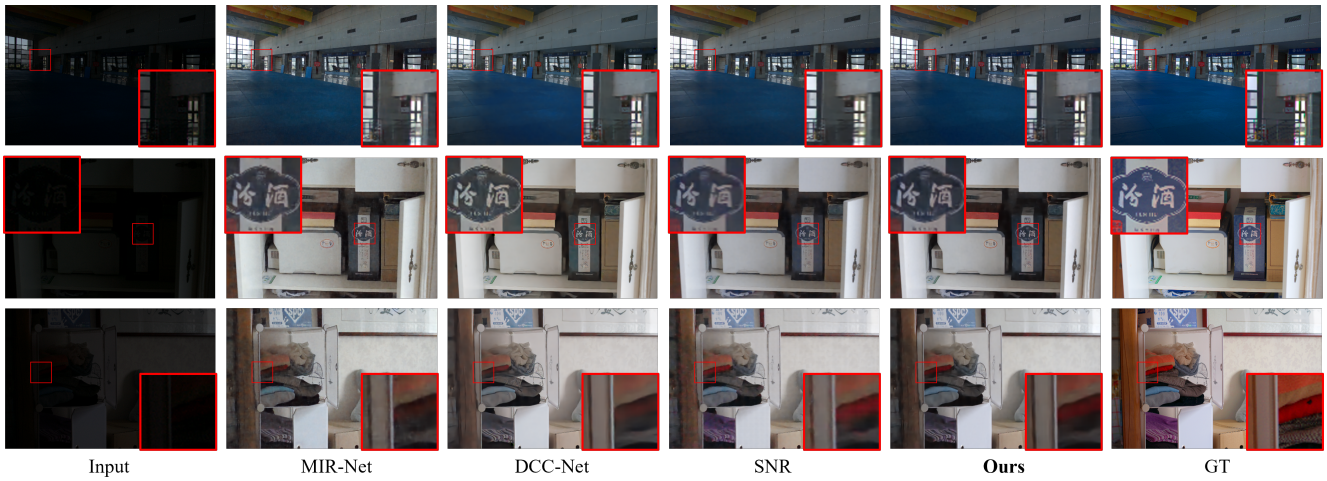

Figure 2: More visual quality comparisons of different low-light image enhancement methods on the LOLv1 dataset.

Figure 3: More visual quality comparisons of different low-light image enhancement methods on the LOLv2-Real dataset.

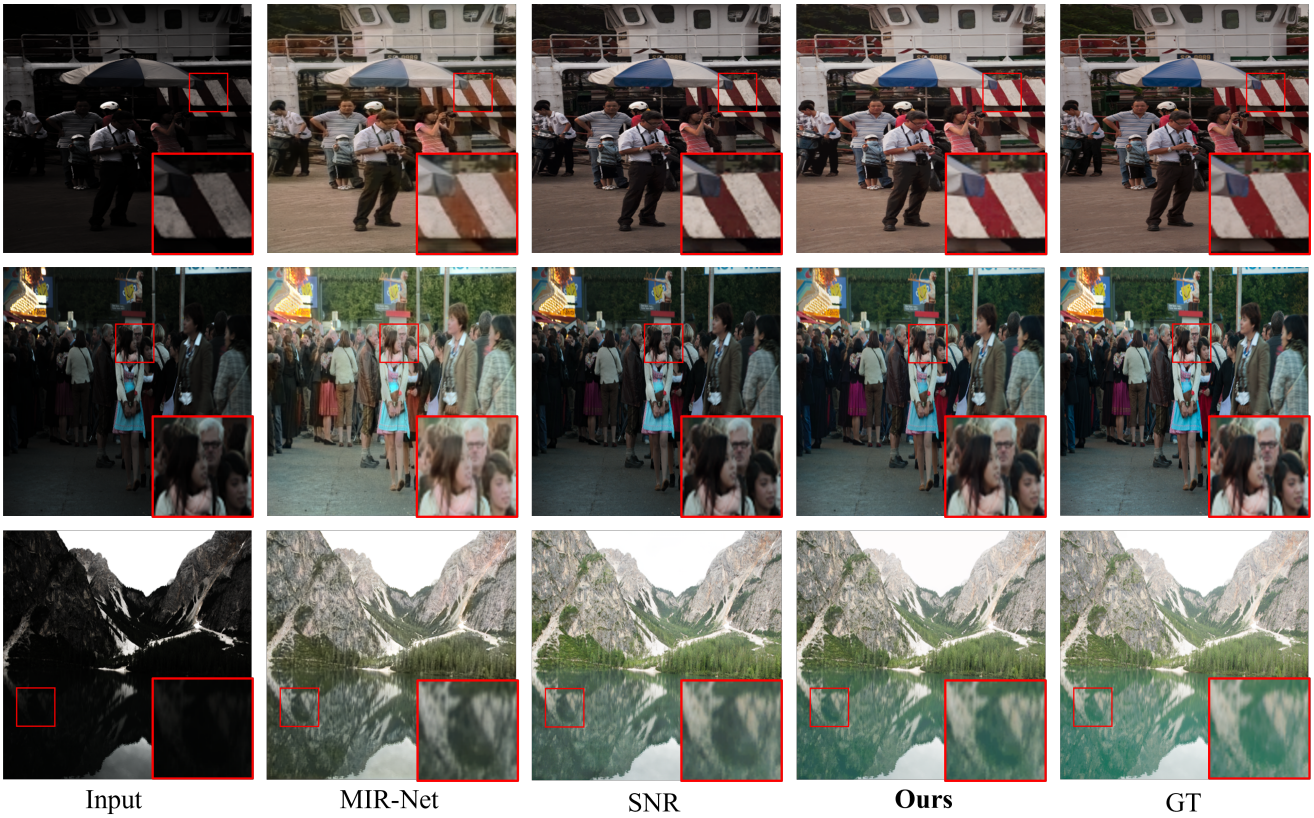

Figure 4: More visual quality comparisons of different low-light image enhancement methods on the LOLv2-Synthetic dataset.

References

[1] Yuanhao Cai, Jing Lin, Xiaowan Hu, Haoqian Wang, Xin Yuan, Yulun Zhang, Radu Timofte, and Luc Van Gool. Mask-guided spectral-wise transformer for efficient hyperspectral image reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17502–17511, 2022. 1

[2] Yuanhao Cai, Jing Lin, Zudi Lin, Haoqian Wang, Yulun Zhang, Hanspeter Pfister, Radu Timofte, and Luc Van Gool. Mst++: Multi-stage spectral-wise transformer for efficient spectral reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 745–755, 2022. 1

[3] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. *arXiv preprint arXiv:1808.04560*, 2018. 1

[4] Xiaogang Xu, Ruixing Wang, Chi-Wing Fu, and Jiaya Jia. Snr-aware low-light image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17714–17724, 2022. 1

[5] Wenhan Yang, Wenjing Wang, Haofeng Huang, Shiqi Wang, and Jiaying Liu. Sparse gradient regularized deep retinex network for robust low-light image enhancement. *IEEE Transactions on Image Processing*, 30:2072–2086, 2021. 1

[6] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Learning enriched features for real image restoration and enhancement. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*, pages 492–511. Springer, 2020. 1

[7] Zhao Zhang, Huan Zheng, Richang Hong, Mingliang Xu, Shuicheng Yan, and Meng Wang. Deep color consistent network for low-light image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1899–1908, 2022. 1