

Supplementary Materials: Wave-Mamba: Wavelet State Space Model for Ultra-High-Definition Low-Light Image Enhancement

Anonymous Authors

We first present the specific structural details of the Selective kernel feature fusion module that we used for the Network in Part A. We will further analyze the motivation for our study in Section B. In Section C we will provide more visual comparisons on the UHD LLIE dataset. Along with that, we will provide visual comparisons on LOLv1 and LOLv2Real datasets in Section D.

A THE ARCHITECTURE OF SKFF

On the basis of the fact that the receptive field of visual cortical neurons is variable, the application of this adaptive adjustment mechanism to neural networks can improve the robustness of the network [1]. Therefore, we use a selective kernel feature fusion (SKFF) module to fuse high-frequency features as shown in Figure 1. Instead of the existing simple feature aggregation methods of concatenation and sum, which are most commonly used, we adopt a cross-scale attention fusion mechanism to aggregate features from different resolution streams. Specifically, the CSFM adapts the receptive field dynamically through two operations: fusion and assignment. The fusion operation focuses on combining feature information from multi-resolutions to generate global feature descriptors. It can be expressed as:

$$F_{sum} = F_{LH} + F_{HL} + F_{HH}, \quad (1)$$

where F_{sum} indicates the features after initial fusion. F_{LH} , F_{HL} , and F_{HH} denote the input of different high-frequency features respectively. Then, we employ global average pool (GAP) in the spatial dimension of $F_{sum} \in \mathcal{R}^{H \times W \times C}$ to compute channel-wise statistics $F_c \in \mathcal{R}^{1 \times 1 \times C}$. Next, we exploit an 1×1 convolution to generate a compact feature representation $F_r \in \mathcal{R}^{1 \times 1 \times \frac{C}{r}}$, where $r = \frac{C}{8}$ for all our experiments. It can be formulated as:

$$F_r = \sigma(H_{conv}^{cd}(\text{GAP}(F_{sum}))), \quad (2)$$

where $\sigma(\cdot)$, $\text{GAP}(\cdot)$, and H_{conv}^{cd} denote the activation function, global average pooling, and channel downscaling convolution respectively. Then, we utilize the assignment operation to recalibrate the feature maps of different scales followed by their aggregation. To be specific, we further use a channel-upscaling convolution to extend the compact feature F_r to obtain the attention feature $A \in \mathcal{R}^{1 \times 1 \times 3C}$ for different scale features. Next, we adopt the split operation to separate attention features $V_1, V_2, V_3 \in \mathcal{R}^{1 \times 1 \times C}$. It can be written as:

$$V_1, V_2, V_3 = \text{split}(\text{Softmax}(H_{conv}^{cu}(F_r))), \quad (3)$$

where $\text{Softmax}(\cdot)$, $H_{conv}^{cu}(\cdot)$, and $\text{split}(\cdot)$ denote the softmax function, channel-upscaling convolution, and channel splitting respectively. Finally, we use attention features V_1, V_2 , and V_3 to adaptively recalibrate high-frequency feature maps F_{LH} , F_{HL} , and F_{HH} , respectively. It can be expressed as:

$$F_{out} = V_1 \otimes F'_{LH} + V_2 \otimes F'_{HL} + V_3 \otimes F'_{HH}, \quad (4)$$

Table 1: Quantitative results to support our motivation. The measurement metrics are the average luminance value (ALV) and MUSIQ value.

UHD-LL	Real		Exchange	
	normal-light	low-light	$I_{high}^{LF} + I_{low}^{HF}$	$I_{low}^{LF} + I_{high}^{HF}$
ALV	130.7895	47.8125	130.7895	47.889
MUSIQ	46.14	26.76	31.87	39.68

where \otimes denotes the operation of element-wise multiplication. F_{out} represents the output features. Compared with the concat method of aggregating features, our method can effectively reduce the network parameters and computational effort and deliver superior performance.

B FURTHER ANALYSIS OF MOTIVATION

Recall that in Section 3.1 of the manuscript, we discussed two observations that serve as the motivation to design our network. 1. In the wavelet domain, most image information resides in the low-frequency component, with only a minor portion of texture information in the high-frequency component. 2. High-frequency information has a minimal impact on the results of low-light image enhancement.

We first show more motivation cases in Figures 2, 3, and 4. These visual results suggest the same tendency as the motivations shown in the main paper.

To further analyze our first motivation, we compared the image histograms of frequency sub-bands obtained from real low-light images and normal-light after wavelet transform. The visualization of both image and histogram information from the Figure 2, 3, and 4 demonstrates that the low-frequency component contains most of the information of the image, however the high-frequency component contains only a small amount of texture information.

For the second motivation, we compare the luminance and noise of real low-light and normal-light images and exchange low-light and normal-light images. To compare the luminance similarity, we compute the average luminance. For real noise level measurement, we use the recent multi-scale Transformer MUSIQ [2] for image quality assessment. MUSIQ is not sensitive to luminance changes. Moreover, it can be used to measure the noise level as its training dataset contains noisy data and it shows state-of-the-art performance for assessing the quality of natural images. A large MUSIQ value reflects better image quality with less noise and artifacts. We compare the average scores of all images on the UHD-LL dataset and give the quantitative results in Table 1.

As shown in Table 1, $I_{high}^{LF} + I_{low}^{HF}$ has similar luminance values to the real normal-light image and has similar noise values to the real low-light image. Similarly, $I_{low}^{LF} + I_{high}^{HF}$ has similar luminance values to the real low-light image and similar noise values to the

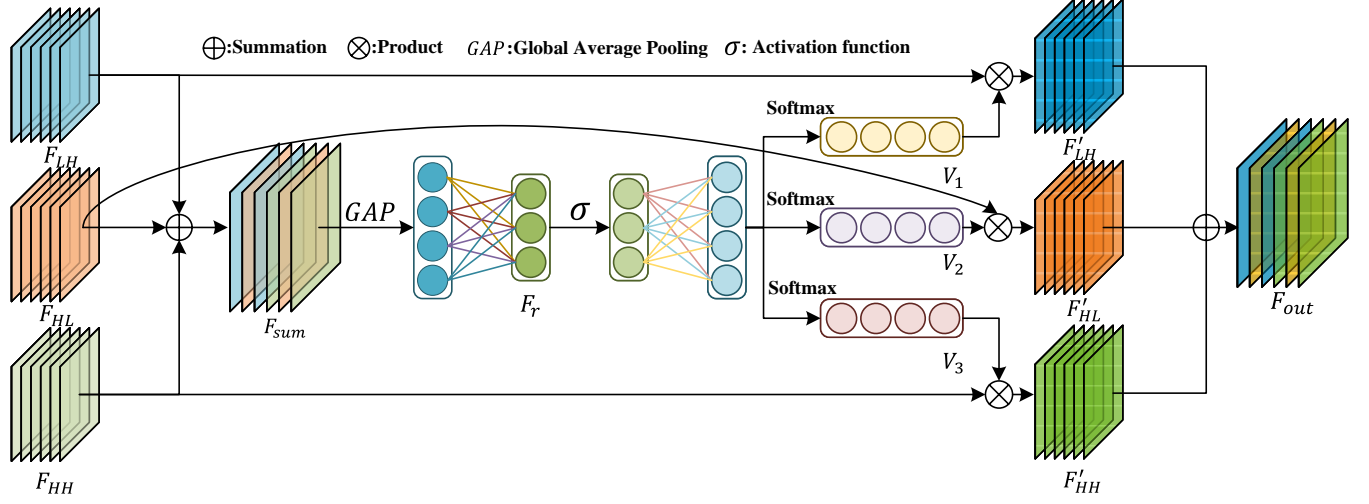


Figure 1: Schematic for selective kernel feature fusion (SKFF). It operates on features from multiple convolutional streams, and performs aggregation based on self-attention.

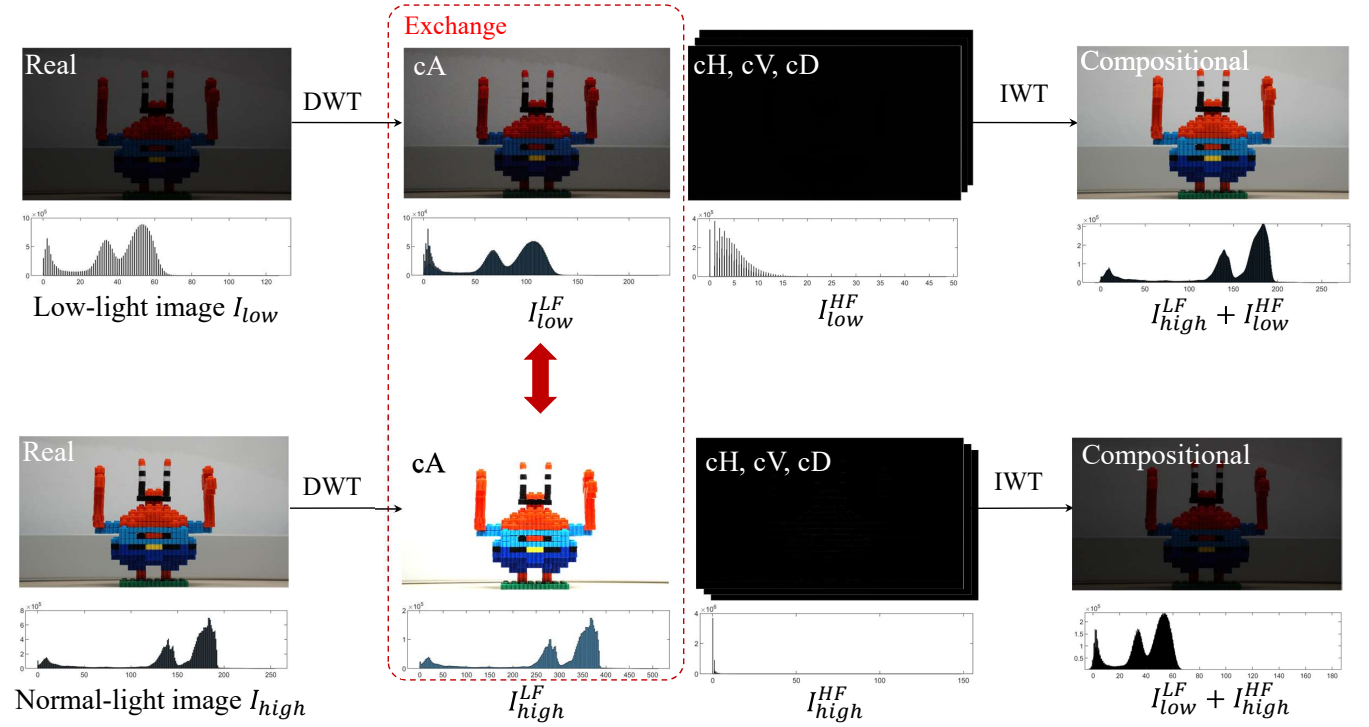


Figure 2: Examples of our motivations.

real normal-light image. The results further show that exchanging the high-frequency components has little effect on the overall information of the image.

C MORE RESULTS ON THE UHD LLIE DATASET

We provide more visual comparisons of our method with state-of-the-art methods on the UHD-LL and UHDLL4K dataset in Figures 5, 6, 7 and 8. As the results shown, for UHD low-light image enhancement, the retrained models on the UHD-LL dataset still cannot

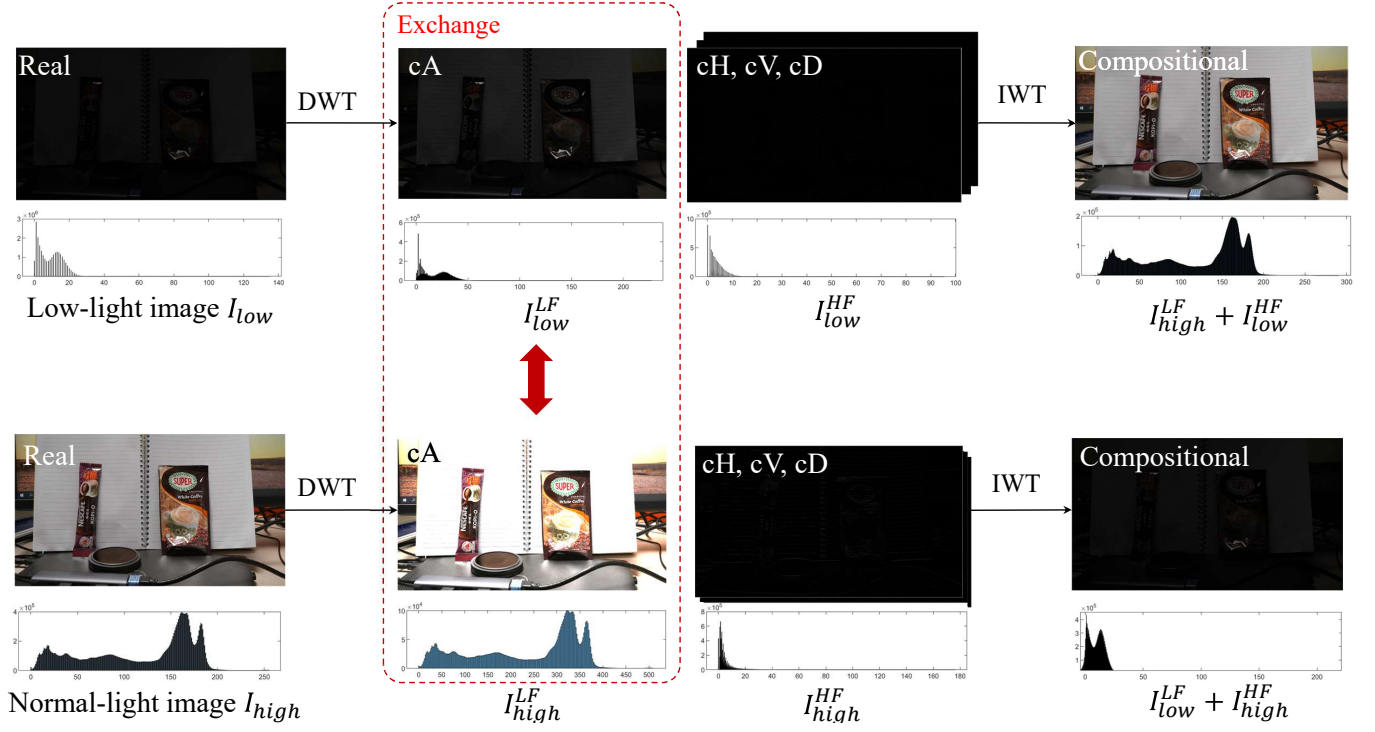


Figure 3: Examples of our motivations.

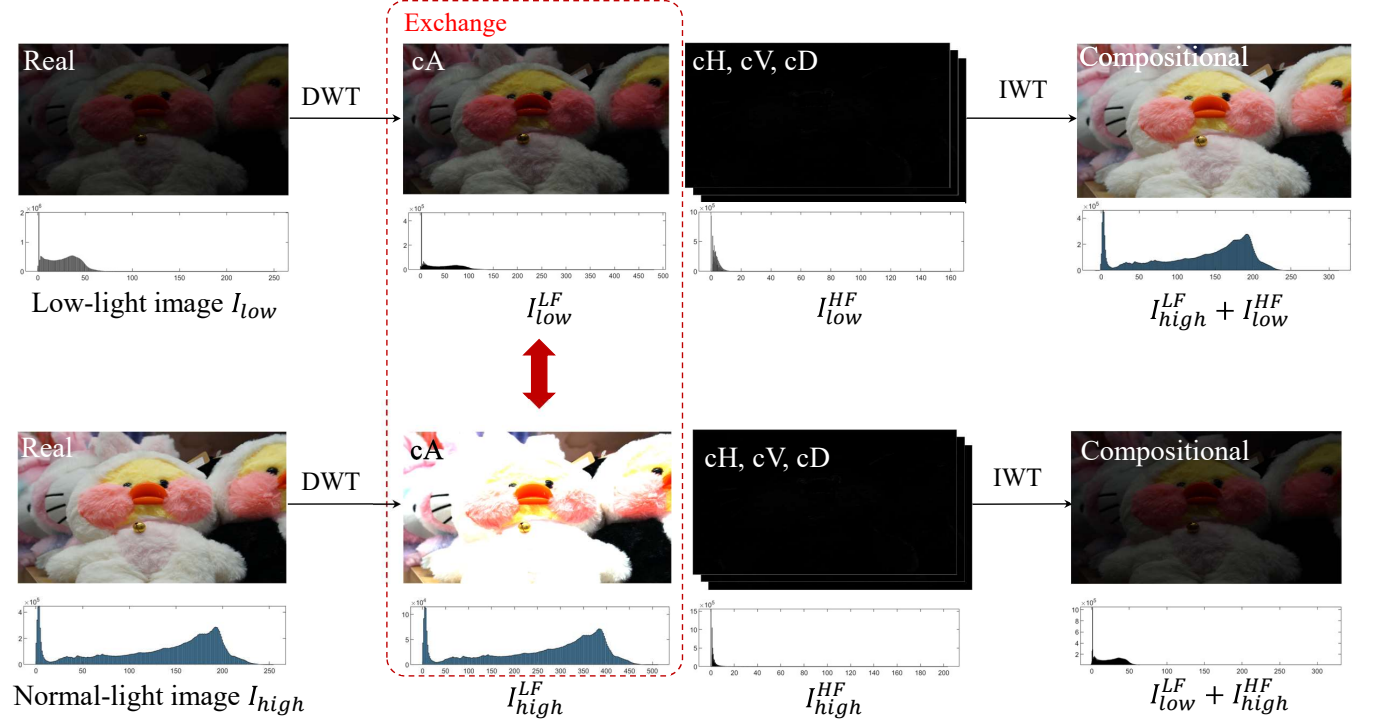


Figure 4: Examples of our motivations.

achieve satisfactory results. Noise and artifacts can still be found in their results. The results suggest that joint luminance enhancement and noise removal in the spatial domain are difficult. Our solution effectively handles this challenging problem by wavelet transform and state-space models, in which luminance and noise can be decomposed to a certain extent and are processed separately.

D MORE RESULTS ON LOLV1 AND LOLV2-REAL DATASETS

We also provide more visual comparisons of our method with the models that were pre-trained or fine-tuned on the LOLv1 datasets in Figures 9. As can be seen from the Figure 9, our method also achieves excellent performance and visual effects in low-light images at low resolutions. The results demonstrate the potential of our solution in different situations.

REFERENCES

- [1] 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.
- [2] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5148–5157, 2021.
- [3] Haoxiang Jie, Xinyi Zuo, Jian Gao, Wei Liu, Jun Hu, and Shuai Cheng. Llformer: An efficient and real-time lidar lane detection method based on transformer. In

- Proceedings of the 2023 5th international conference on pattern recognition and intelligent systems*, pages 18–23, 2023.
- [4] Chongyi Li, Chun-Le Guo, Man Zhou, Zhixin Liang, Shangchen Zhou, Ruicheng Feng, and Chen Change Loy. Embedding fourier for ultra-high-definition low-light image enhancement. *arXiv preprint arXiv:2302.11831*, 2023.
- [5] Cong Wang, Jinshan Pan, Wei Wang, Gang Fu, Siyuan Liang, Mengzhu Wang, Xiao-Ming Wu, and Jun Liu. Correlation matching transformation transformers for uhd image restoration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5336–5344, 2024.
- [6] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. *arXiv preprint arXiv:1808.04560*, 2018.
- [7] Feifan Lv, Feng Lu, Jianhua Wu, and Chongsoon Lim. Mbllen: Low-light image/video enhancement using cnns. In *BMVC*, volume 220, page 4, 2018.
- [8] Yonghua Zhang, Xiaojie Guo, Jiayi Ma, Wei Liu, and Jiawan Zhang. Beyond brightening low-light images. *International Journal of Computer Vision*, 129:1013–1037, 2021.
- [9] Chunle Guo, Chongyi Li, Jichang Guo, Chen Change Loy, Junhui Hou, Sam Kwong, and Runmin Cong. Zero-reference deep curve estimation for low-light image enhancement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1780–1789, 2020.
- [10] Yifan Jiang, Xinyu Gong, Ding Liu, Yu Cheng, Chen Fang, Xiaohui Shen, Jianchao Yang, Pan Zhou, and Zhangyang Wang. Enlightengan: Deep light enhancement without paired supervision. *IEEE transactions on image processing*, 30:2340–2349, 2021.
- [11] Yuanhao Cai, Hao Bian, Jing Lin, Haoqian Wang, Radu Timofte, and Yulun Zhang. Retinexformer: One-stage retinex-based transformer for low-light image enhancement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12504–12513, 2023.
- [12] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17683–17693, 2022.
- [13] Hai Jiang, Ao Luo, Haoqiang Fan, Songchen Han, and Shuaicheng Liu. Low-light image enhancement with wavelet-based diffusion models. *ACM Transactions on Graphics (TOG)*, 42(6):1–14, 2023.

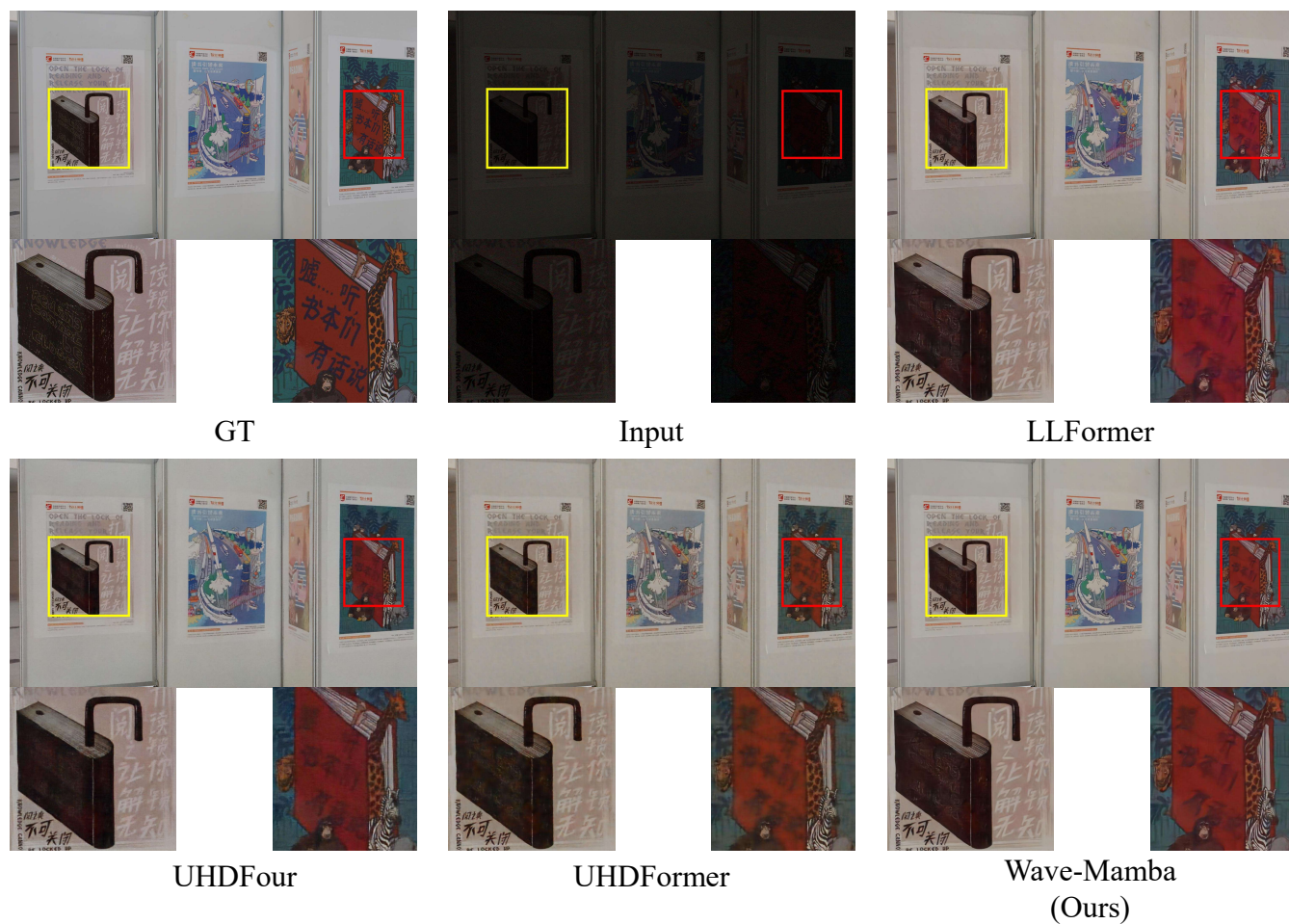


Figure 5: Visual comparison of the state of the arts methods on the UHD-LL dataset. The compared methods include LLFormer [3], UHDFour [4], UHDFormer [5].

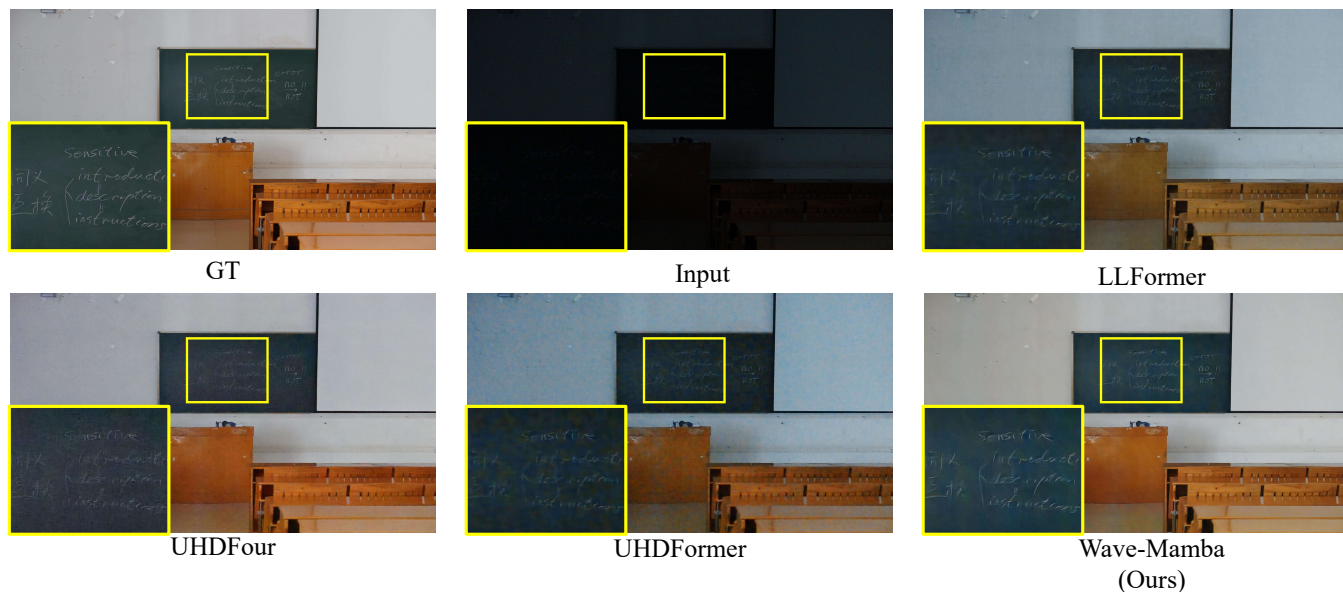


Figure 6: Visual comparison of the state of the arts methods on the UHD-LL dataset. The compared methods include LLFormer [3], UHDFour [4], UHDFormer [5].

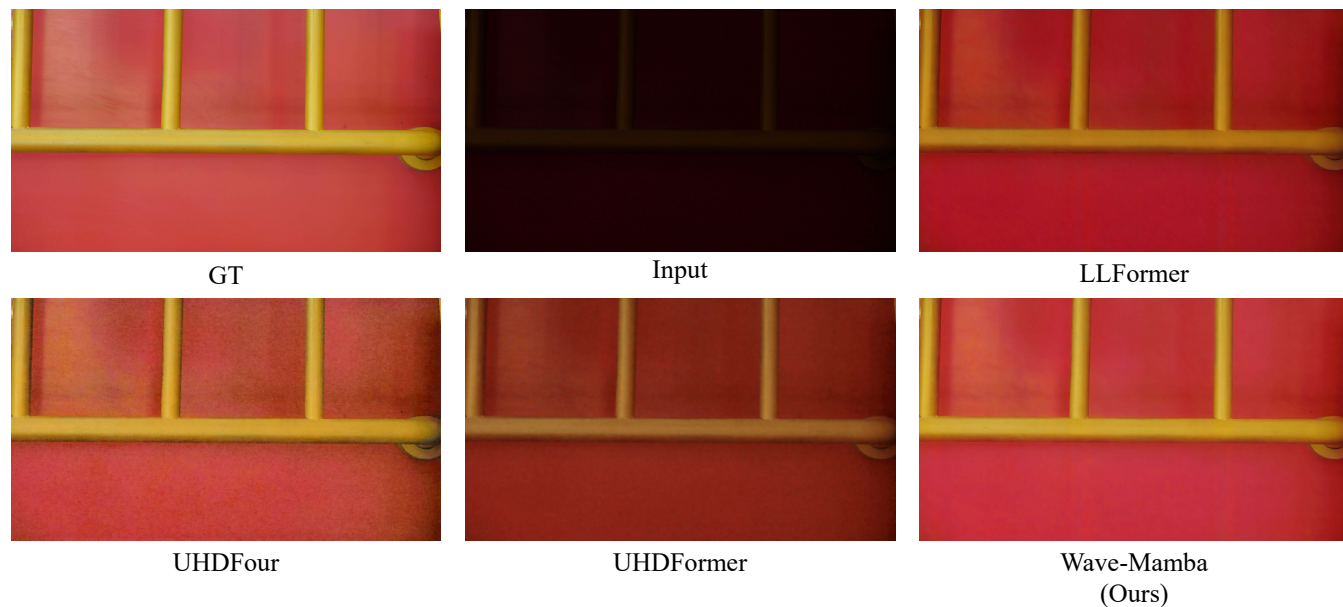


Figure 7: Visual comparison of the state of the arts methods on the UHD-LL dataset. The compared methods include LLFormer [3], UHDFour [4], UHDFormer [5].

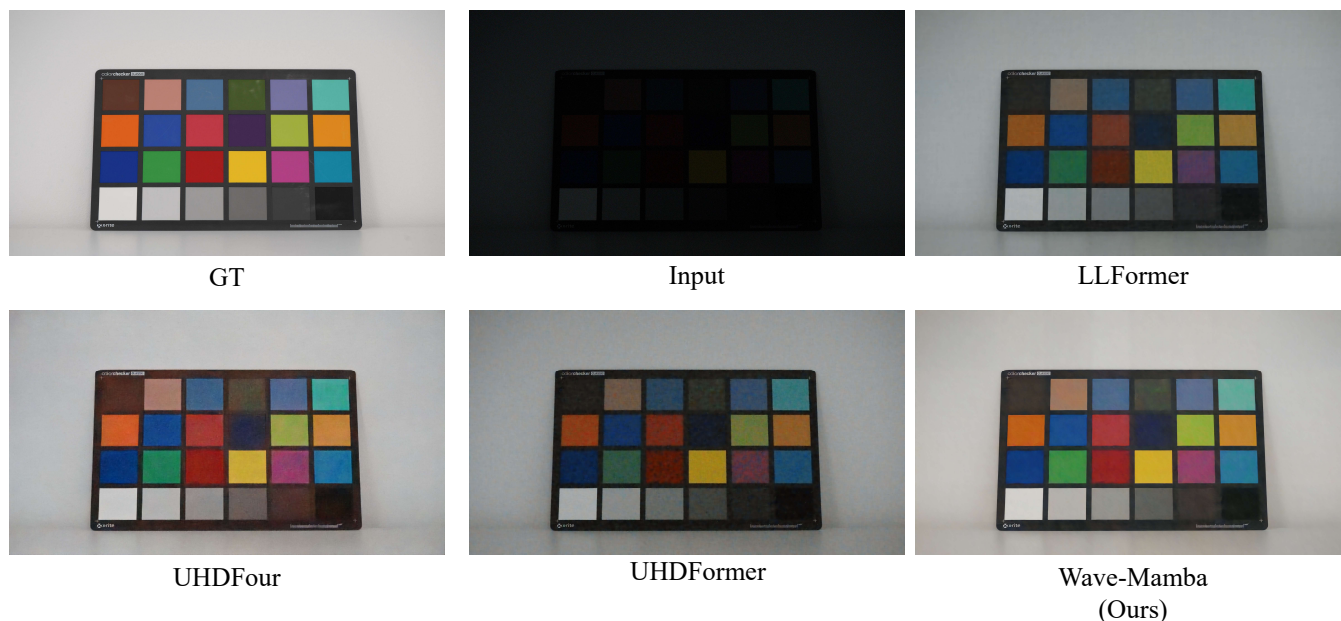


Figure 8: Visual comparison of the state of the arts methods on the UHD-LL dataset. The compared methods include LLFormer [3], UHDFour [4], UHDFormer [5].

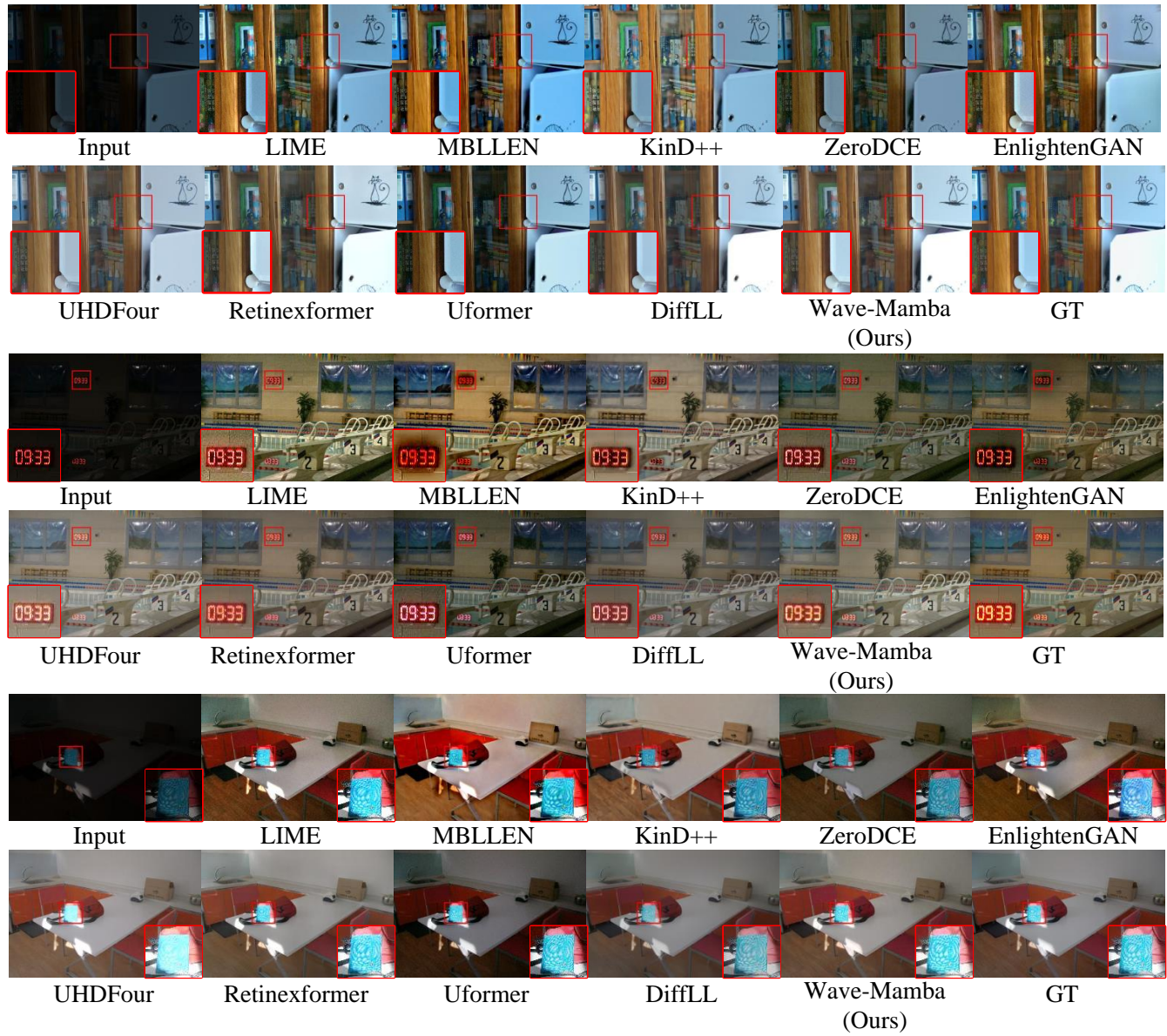


Figure 9: Visual comparison of the state of the arts methods on the LOLv1 and LOLv2-Real dataset. The compared methods include LIME [6], MBLLEN [7], KinD++ [8], ZeroDCE [9], EnlightenGAN [10], UHDFour [4], Retinexformer [11], Uformer [12], DiffLL [13].