Learning Pixel-adaptive Multi-layer Perceptrons for Real-time Image Enhancement

Supplementary Material

In Section A, we present detailed information about our network architecture and the process of generating our bilateral grids. In Section B, we describe the training configurations for each dataset. In Section C, we provide further experimental analyses. In Section D, we offer additional visual comparisons.

A. Generation of the Bilateral Grids

Our proposed method utilize a three-layer U-net-style NAFNet [2] as our backbone to generate the bilateral grids, with each layer in both the encoder and decoder comprising two NAFblocks (see Figure 1), and the number of blocks at the bottom level is the same. The width in NAF-block is set as 16. The feature map output from NAFNet maintains a channel numble equal to the model's width. We then apply the pixel unshuffle operation to reduce its resolution by a factor of 4 while expanding the channel count by 16 times. Next, two 1×1 convolutions are used to further adjust the channel dimensions, yielding two final feature maps.

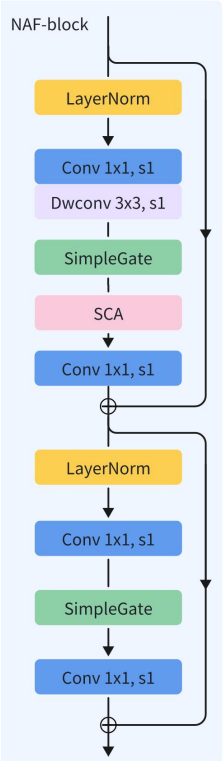


Figure 1. NAF-block

At this point, following HDRNet's [3] approach, we treat these feature maps as two bilateral grids whose third dimension has been unrolled

$$\mathcal{G}_{dc+z}[x, y] \leftrightarrow \mathcal{G}_c[x, y, z] \quad (1)$$

where d represents the depth of the grid .We slice each map along the channel dimension into d parts and expand a new dimension, resulting in a three-dimensional bilateral grid in which each grid cell contains the required parameters.

B. Experimental Details

B.1. FiveK

Experiments on the FiveK dataset [1] are conducted on two different resolutions (480p and full resolution) and two tasks (retouching and tone mapping). For two format images (8-bit sRGB and 16-bit CIE XYZ) at 480p resolution, we adopt the dataset released by [8]. For full-resolution images for tone mapping tasks, we utilize the dataset provided by [9].

The augmentations include random ratio cropping, random flipping and random rotating. For 480p and full-resolution images, we downsample by factors of 2 and 8, respectively, resulting in grids that are 1/8 and 1/32 of the original resolution. The training process consists of 225K iterations, we set the initial learning rate to 3e-4 and employ a cosine annealing schedule to gradually reduce it to 4e-6.

B.2. PPR10K

Experiments on the PPR10K dataset [6] are conducted on the 360p resolution for photo retouching task. The dataset also includes five augmented versions of the original training images, and the final training set comprises 53,250 images. For more details, please refer to [6].

The augmentations include random ratio cropping, random flipping. We do not perform downsampling on this dataset, as its native resolution is already low, resulting in grids that are 1/4 of the original resolution. 3D LUT-based approaches [5, 6, 8] on this dataset employ ResNet-18 [4](11.7M) as their backbone, but this network cannot be used to generate our bilateral grids. For a fair comparison, we increased the depth of NAFNet to 4 layers, increased the width to 32 channels, and increased the number of bottom blocks to 4, thereby raising the model's parameter count to a comparable level (11.7M). The training process consists of 500K iterations, We set the initial learning rate to 2e-4 and employ a cosine annealing schedule to gradually reduce it to 2e-6.

B.3. LCDP

Experiments on the LCDP dataset [7] are conducted on the original resolution for exposure correction task. The augmentations include random ratio cropping, random flipping and random rotating. We downsample the images by a factor of 2, ultimately generating grids that are 1/8 of the original resolution. The training process consists of 120K iterations,, We set the initial learning rate to 4e-4 and employ a cosine annealing schedule to gradually reduce it to 2e-6.

C. Ablation Studies

Detailed Explanation about Grid Decomposition of Setting 3 in Table ??. We divide the 12 coefficients of the affine transformation into 3 parts and generate a 3-channel guidance map, where each channel takes its corresponding subset indicated by different colors.

$$\begin{aligned} R' &= a_1 R + a_2 G + a_3 B + b_1 \\ G' &= a_4 R + a_5 G + a_6 B + b_2 \\ B' &= a_7 R + a_8 G + a_9 B + b_3 \end{aligned}$$

Selection of Guidance Maps. We adopted the grid decomposition strategy to obtain multiple subgrids, which makes it necessary to use different guidance maps to steer the slicing operation for each corresponding subgrid. We designed two selection schemes and compare their performance for the tone mapping task on the FiveK dataset (480p), as shown in Table 1. In Setting 1, the grid decomposition strategy is not used, while in Settings 2 and 3 the grids are decomposed and different guidance map selection schemes are applied. The first scheme uses each input channel as its own guidance map to extract the corresponding weights, with a convolutional network fusing the input into a single channel for bias extraction. The second scheme, which is the one currently adopted, feeds the input into a convolutional network to generate a multi-channel guidance map, with each channel corresponding to a subgrid.

Table 1. Comparison of two guidance maps selection schemes.

Setting	Scheme 1	Scheme 2	PSNR	SSIM
1			25.70	0.939
2	✓		25.78	0.940
3		✓	**25.83**	**0.941**

It can be seen that after applying grid decomposition, both guidance map selection schemes yield performance improvements, with the second scheme proving to be superior.

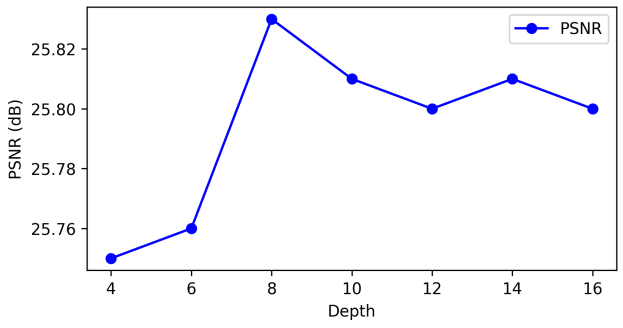


Figure 2. Ablation study on different numbers of bilateral grid depth.

Depth of Bilateral Grid. Depth in a bilateral grid controls the resolution along the intensity dimension. Generally speaking, higher depth captures finer details at the cost of efficiency, while lower depth improves speed but may lose subtle variations. However, since the information in our bilateral grids is generated by the network, increasing the depth adds more parameters, which could make the model harder to train and doesn't necessarily improve performance. Therefore, We conduct experiments for tone mapping on the FiveK dataset to determine the optimal settings, results can be seen in Figure 2. It can be observed that the model achieves its best performance when the grid depth is 8, so we ultimately adopt this setting.

Ablation study about MLP depth and channels. We train models with increased channel numbers and depth of MLP. According to the table 2, increasing the number of intermediate layers in the MLP degrades model performance. This occurs because the dramatic increase in internal parameters makes it difficult for the model to learn effectively. While increasing the depth of the MLP also contributes to this issue, it enhances the model's nonlinear capabilities, leading to some improvement in performance. Further increasing the complexity of MLPs is not able to significantly improve mapping ability but leads to more computational costs.

Table 2. Ablation study about MLP depth and channels.

Setting	Params	Time(4K)	PSNR
3-8-3	624K	27.8 ms	25.83
3-16-3	741K	42.8 ms	25.81
3-8-8-3	779K	48.3 ms	25.87

D. More Qualitative Results

We provide additional visual comparisons on the LCDP dataset in Figure 3, 4, 5, 6 and on the FiveK (4K) dataset in Figure 7, 8, 9, 10.

References

[1] Vladimir Bychkovsky, Sylvain Paris, Eric Chan, and Frédo Durand. Learning photographic global tonal adjustment with a database of input/output image pairs. In *CVPR 2011*, pages 97–104. IEEE, 2011. 1

[2] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *European conference on computer vision*, pages 17–33. Springer, 2022. 1

[3] Michaël Gharbi, Jiawen Chen, Jonathan T Barron, Samuel W Hasinoff, and Frédo Durand. Deep bilateral learning for real-time image enhancement. *ACM Transactions on Graphics (TOG)*, 36(4):1–12, 2017. 1

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings*

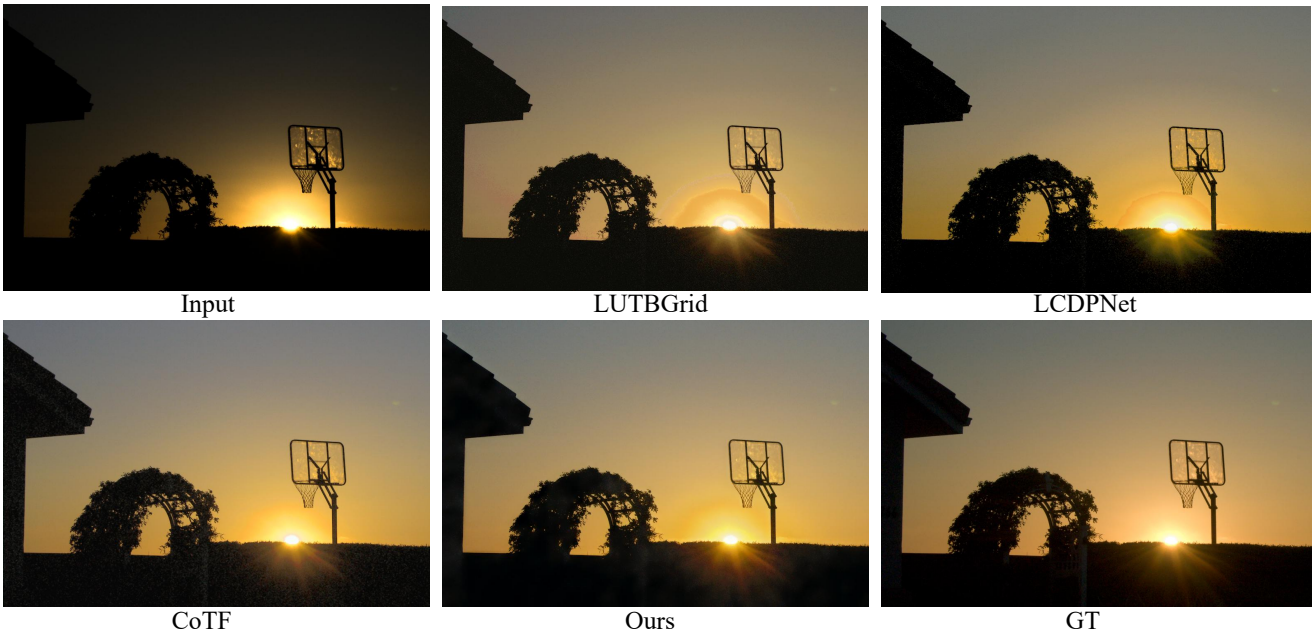


Figure 3. Visual comparison with state-of-the-art methods on the LCDP dataset for exposure correction.

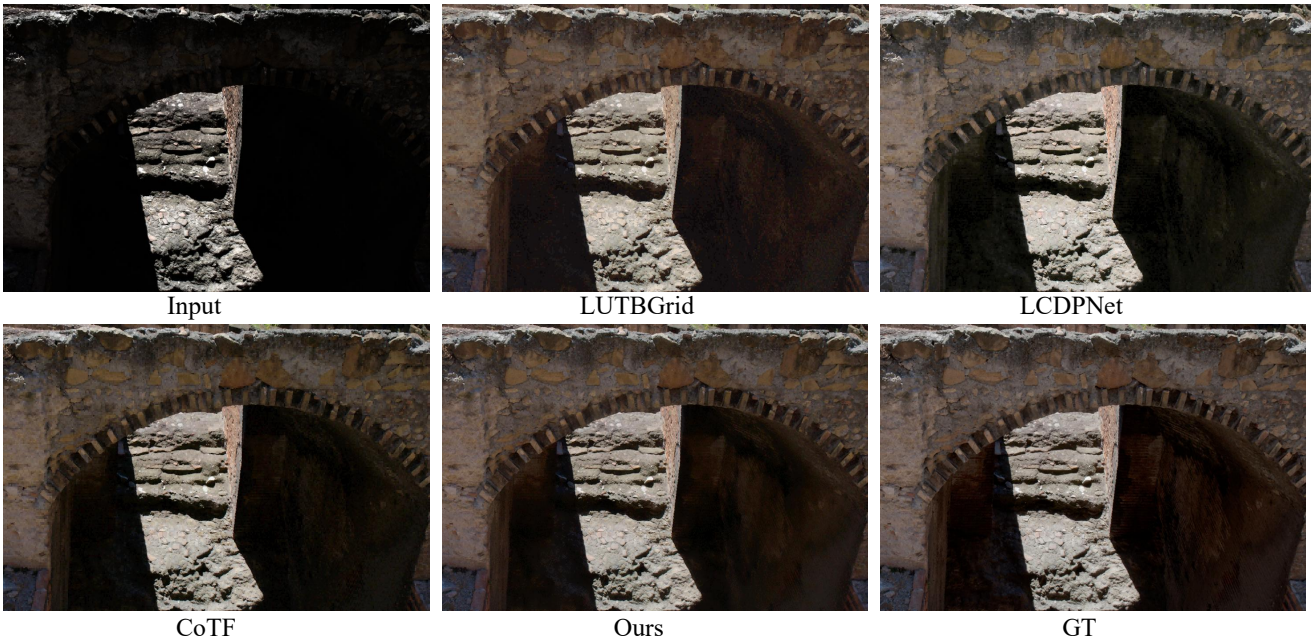


Figure 4. Visual comparison with state-of-the-art methods on the LCDP dataset for exposure correction.

of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016. 1

[5] Wontae Kim and Nam Ik Cho. Image-adaptive 3d lookup tables for real-time image enhancement with bilateral grids. In *European Conference on Computer Vision*, pages 91–108. Springer, 2024. 1

[6] Jie Liang, Hui Zeng, Miaomiao Cui, Xuansong Xie, and Lei Zhang. Ppr10k: A large-scale portrait photo retouching dataset with human-region mask and group-level consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 653–661, 2021. 1

[7] Haoyuan Wang, Ke Xu, and Rynson WH Lau. Local color distributions prior for image enhancement. In *European conference on computer vision*, pages 343–359. Springer, 2022.

Figure 5. Visual comparison with state-of-the-art methods on the LCDP dataset for exposure correction.

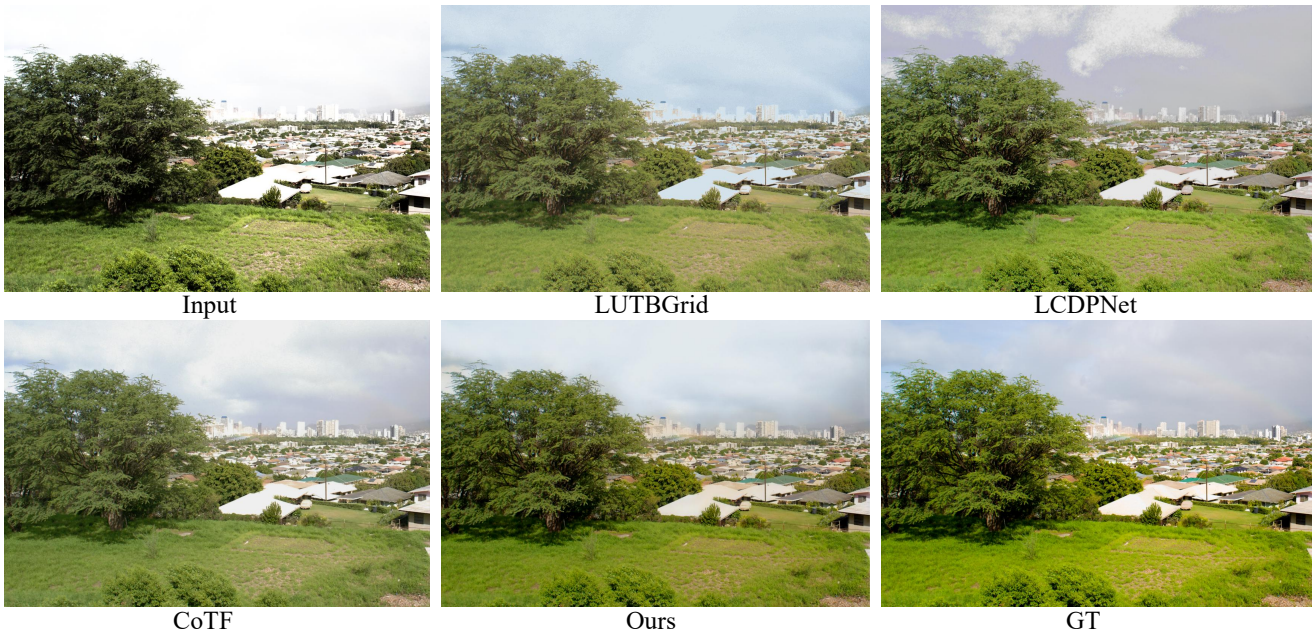


Figure 6. Visual comparison with state-of-the-art methods on the LCDP dataset for exposure correction.

1

[8] Hui Zeng, Jianrui Cai, Lida Li, Zisheng Cao, and Lei Zhang. Learning image-adaptive 3d lookup tables for high performance photo enhancement in real-time. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(4):2058–2073, 2020. 1

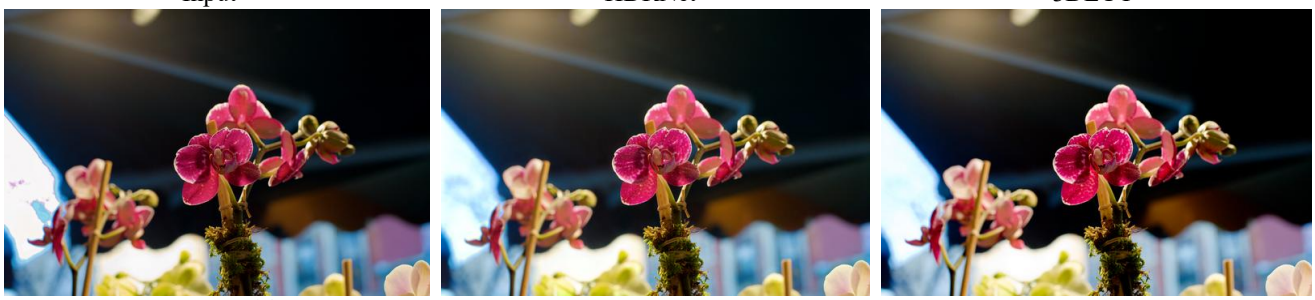
[9] Feng Zhang, Ming Tian, Zhiqiang Li, Bin Xu, Qingbo Lu, Changxin Gao, and Nong Sang. Lookup table meets local laplacian filter: pyramid reconstruction network for tone mapping. *Advances in Neural Information Processing Systems*, 36:57558–57569, 2023. 1

Input HDRNet 3DLUT

LUTBGrid Ours GT

Figure 7. Visual comparison with state-of-the-art methods on the FiveK dataset (4K) for tone mapping.

Input HDRNet 3DLUT



LUTBGrid Ours GT

Figure 8. Visual comparison with state-of-the-art methods on the FiveK dataset (4K) for tone mapping.

Figure 9. Visual comparison with state-of-the-art methods on the FiveK dataset (4K) for tone mapping.

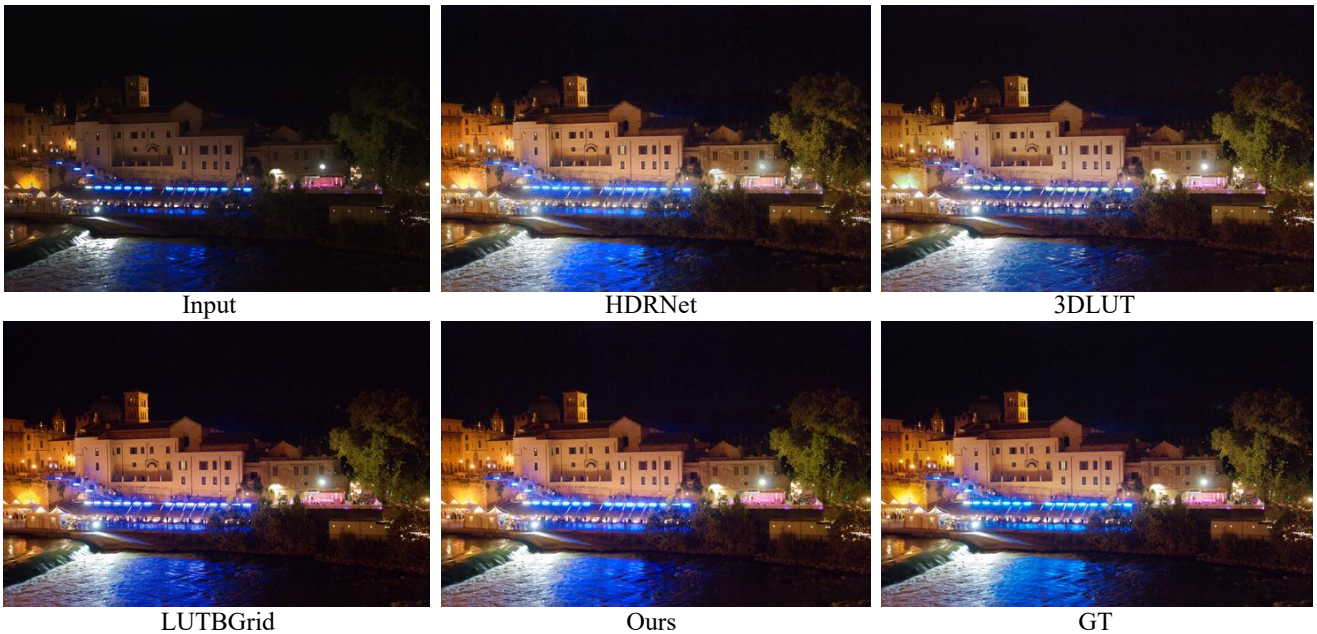


Figure 10. Visual comparison with state-of-the-art methods on the FiveK dataset (4K) for tone mapping.