Revisiting the Stack-Based Inverse Tone Mapping Supplementary Material

1 Overview

In this supplementary material, additional explanations and analyses have been provided. Firstly, the details of the proposed SICE-S dataset are given in Section 2. Secondly, details of the proposed network architectures are shown in Section 3. Thirdly, more experiments on the influences of different stack lengths are provided in Section 4. Additional objective and subjective results can be found in Section 5 and Section 6. The details of the user study are shown in Section 7. Finally, Section 8 analyzes the limitations of the proposed method.

2 SICE-S dataset

As specified in the manuscript, the SICE-S dataset is built based on the original SICE [2] dataset by selecting the optimal exposure-up image and exposure-down image. Specifically, the traditional HDRI technology takes many LDR photos (usually 7 photos such as in [2] [8]) with different exposure values to capture the details. However, according to [1], in most situations, 3 multi-exposure images can be enough to cover the whole dynamic range of the real scene. Therefore, for each MES, we take the normally captured LDR image as the input, and then we choose a darker image with a lower EV and a brighter image with a higher EV to form the candidate ground truth. There are N LDR images with different exposure values in each MES of the SICE dataset [2]. Consequently, there are total $\lfloor \frac{N}{2} \rfloor \cdot (N - \lceil \frac{N}{2} \rceil)$ kinds of candidates for chosen. We merge all of the LDR images in the MES into the ground truth HDR image I_{hdr} and for each combination, we merge them into the candidate HDR image $\hat{I}_{hdr}$ and calculate the image quality scores by the following equation:

$$S = \lambda_p \cdot P\left(\hat{I}_{tm}, I_{tm}\right) + \lambda_v \cdot V\left(\hat{I}_{hdr}, I_{hdr}\right) \tag{1}$$

where P, V denotes the metric PSNR and HDR-VDP-2 [14], and $\hat{I}_{tm}$, I_{tm} denotes the tone-mapped LDR image of the candidate HDR and ground truth HDR image. λ_p and λ_v are experimentally set to 0.3 and 0.7 separately. Then, we pick the combinations with the top-5 scores to further conduct the subjective image quality assessment. Specifically, 10 subjects are requested to rate all the picked combinations with the ground truth HDR I_{hdr} as reference. The integer opinion score ranges from 1 (the worst) to 5 (the best). Finally, we choose the combination with the highest mean opinion score as the supervised labels to train the adaptive exposure adjustment models. We will open the SICE-S in form of the optimal exposure value index in the SICE dataset [2] for the future study on the stack-based ITM methods. Note that there are only three images in each multiple exposure stack (MES) of SICE-S. Hence, compared with SICE, some details may be missing in SICE-S. Therefore, we conduct a double-stimulus-impairment-scale (DSIS) experiment to verify the quality of the SICE-S. Specifically, 10 subjects are requested to compare the visual quality between the constructed HDR images in SICE-S and SICE [2]. With the ground truth SICE image I_{hdr} as the reference, they need to choose one quality level for the SICE-S image $\hat{I}_{hdr}$ from: (a) level 0: can not tell the differences between the I_{hdr} and $\hat{I}_{hdr}$; (b) level 1: there are few differences between them, but they do not influence the overall visual experience; (c) level 2: there are obvious differences between them and $\hat{I}_{hdr}$ loses some details such as in over/under-exposed regions. There are 589 HDR images and the average results of each level are shown in Tab. 1. Fig. 1 shows an example of the comparison between the $\hat{I}_{hdr}$ and I_{hdr}.

Table 1: Average results of each level in the DSIS experiment.

	level 0	level 1	level 2
average number	579.2	9.8	0.0

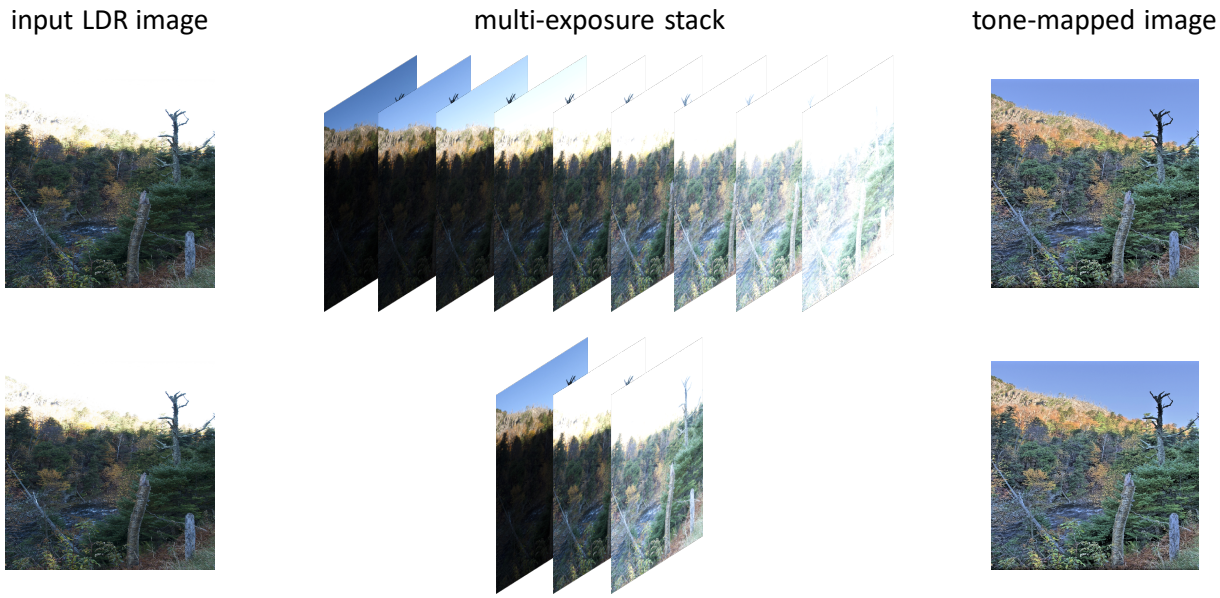

Figure 1: The visual comparison between the HDR images in SICE and SICE-S. The first row denotes the SICE dataset and the second row denotes the SICE-S dataset. Although the MES in SICE has 9 images and the MES in SICE-S only has three multi-exposure images, there are no noticeable differences between the constructed HDR images.

3 Details of the proposed models

The detailed architectures of the proposed adaptive exposure adjustment model and the mask-aware discriminator are shown in Fig. 2. We adopt the swim transformer block used in [13] [10]. As specified in [10], given an input of size $H \times W \times C$, swim transformer block first reshapes the input to a $\frac{HW}{M^2} \times M^2 \times C$ feature by partitioning the input into non-overlapping $M \times M$ local windows, where $\frac{HW}{M^2}$ is the total number of windows. Then, it computes the standard self-attention separately for each window. For a local window feature X, the query, key and value matrices Q, K and V are computed as:

$$Q = XP_Q, K = XP_K, V = XP_V, \tag{2}$$

where P_Q, P_K, and P_V are projection matrices that are shared across different windows. Generally, we have $Q, K, V \in \mathbb{R}^{M^2 \times d}$. The attention matrix is thus computed by the self-attention mechanism in a local window as:

$$Attention Q, K, V = SoftMax(QK^T/\sqrt{d} + B)V, \tag{3}$$

where B is the learnable relative positional encoding. In practice, following [10], we perform the attention function for h times in parallel and concatenate the results for multihead self-attention (MSA). Next, a multi-layer perceptron (MLP) that has two fullyconnected layers with GELU non-linearity between them is used for further feature transformations. The LayerNorm (LN) layer is added before both MSA and MLP, and the residual connection is employed for both modules. The whole process is formulated as:

$$X = MSA(LN(X)) + X, X = MLP(LN(X)) + X. \tag{4}$$

Furthermore, regular and shifted window partitioning are used alternately to enable cross-window connections [13], where shifted window partitioning means shifting the feature by $(\lfloor \frac{M}{2} \rfloor, \lfloor \frac{M}{2} \rfloor)$ pixels before partitioning.

4 Influences of different stack lengths

In this section, we conduct experiments to validate the influences of different stack lengths. More specified, we change the set of stack length during the inference of DrTMO [5], Deep Chain HDRI [8], and Deep Recursive HDRI [9], and then we analyze the influences on the performance. Tab. 2 shows the HDR-VDP-2.2 [14] scores with different stack lengths in the HDR-TEST dataset. As the results show, the length of the stack has a significant impact on the results. On the one hand, the quality of the HDR image is not proportional to the length of the stack. On the other hand, the larger the length of the stack, the longer the running time taken by the exposure adjustment model and fusion algorithm. On the contrary, the proposed method can achieve the best performance compared to these stack-based methods with a stack length of three. For the proposed method which can perform the optimal exposure adjustment, a longer exposure stack length does not yield better results while introducing extra cost and unexpected information.

Furthermore, we conduct experiment to validate that the proposed method can achieve impressive results with only estimating two exposure images by analyzing the histogram. For the output HDR images generated by different stack lengths of the proposed method, there are no obvious differences in the histogram because most pixels are relatively low in brightness, as shown in Fig. 2(b). Therefore, we show the histogram of tone mapped images in Fig. 2(c) where L=3 is much closer to ground truth. We also compare the relative brightness of the pixels indicated by the red lines in Fig. 2(d). L=3 and L=5 are closest to the ground truth and adding length to 7 yields more mismatches due to the introduction of unexpected information. L=9 and L=11 will not generate more information than L=7 and thus the relative luminance does not change.

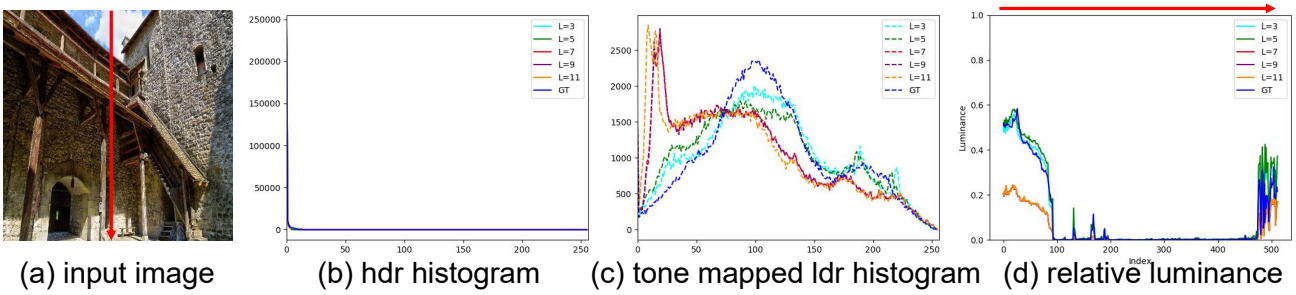

Figure 2: The influences of different lengths on the histogram. L denotes the length of the multi-exposure stack.

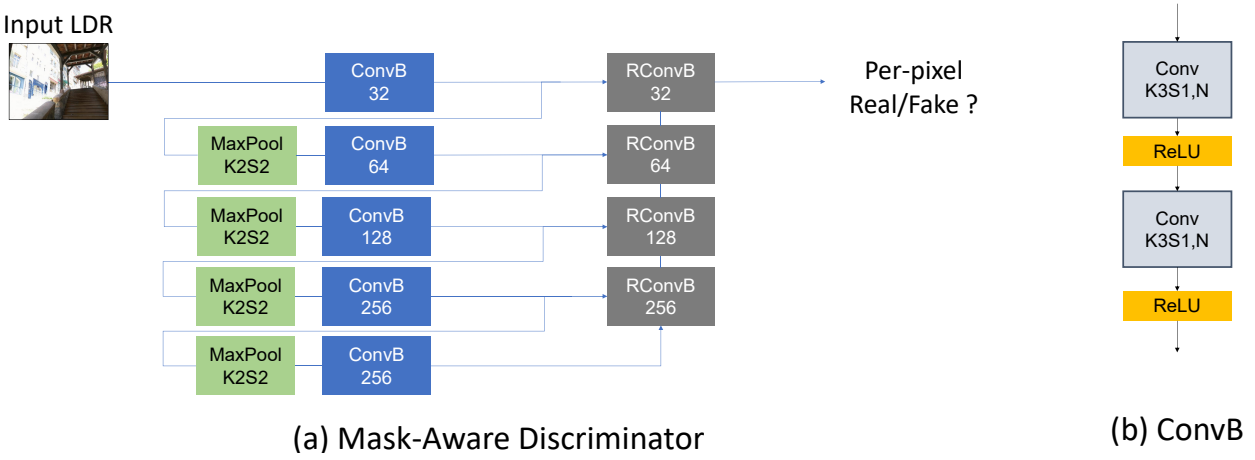

Figure 3: The detailed architectures of the proposed mask-aware discriminator. The "KxSy, N" denotes the convolution layer with a kernel size of x $\times$x, a stride of y, and the output channels of N. The "RConvB" means there is the up-sampling operator before the "ConvB".

Table 2: The influences of different stack lengths on the HDR-VDP-2 scores .

	length=3	length=5	length=7	length=9	length=11
DrTMO [5]	61.48	61.92	62.41	**62.84**	62.73
Deep Chain [8]	60.47	61.21	**61.89**	60.96	60.42
Deep Recursive [9]	62.72	**63.12**	62.94	61.38	59.63
Proposed	**65.20**	64.40	63.51	62.69	61.39

5 More quantitative comparisons

In this section, we conduct extra quantitative comparisons on the HDR images with the existing state-of-the-art ITM methods. Specifically, on the one hand, we show the results with a bigger resolution of 1024 $\times$ 1024 using model weights re-trained by ourselves with the same training dataset as specified in the main manuscript. On the other hand, we also provide the comparison results between the proposed method and the other methods with the pre-trained model weights released by the authors, as shown in the left half of Tab. 3. Note that the training dataset used in the pre-trained model is different from each method. We also compare the ITM methods on the HDR-Dynamic [6] and the result is shown in Tab. 4. Furthermore, we compare the average inference and multi-exposure fusion time for 512$\times$512 image: Deep Recursive [9] (79ms, 6743ms), Deep Synth [7] (294ms, 6743ms) and the proposed method (**54ms, 5ms**), which demonstrates that the proposed method can also reduce computational cost. Table 5 shows the quantitative comparison results on the model size (million) and running time (ms) to process a 512$\times$512 image in Tesla V100 GPU.

Table 3: Quantitative comparison on HDR images with existing methods. The training codes of [16] have not been released, so we show the pre-trained results of it on the resolution of 1024 $\times$ 1024.

	Pre-trained 512$\times$512			**Re-trained 1024$\times$1024**		
	TEST	**EYE**	**FAIR**	**TEST**	**EYE**	**FAIR**
HDRCNN [4]	64.14	55.29	60.25	63.57	55.57	57.85
DrTMO [5]	55.67	58.34	58.24	61.76	56.76	59.47
Deep Recur [9]	58.60	**59.10**	56.35	62.30	58.38	59.85
Deep Single [12]	63.70	57.56	60.01	64.14	57.55	60.58
Deep Synth [7]	60.56	58.59	59.98	62.95	57.29	59.23
Deep Mask [16]	64.01	55.30	58.40	63.99	55.62	58.77
Deep HDRUNet [3]	59.43	53.24	55.79	63.76	57.36	58.47
Proposed	**65.20**	58.92	**60.48**	**65.36**	**59.35**	**60.90**

Table 4: Quantitative comparison on HDR images with existing methods on the HDR-Dynamic [6] dataset.

	HDR-VDP-2.2	**TM-PSNR/SSIM**		**HDR-VDP-2.2**	**TM-PSNR/SSIM**
HDRCNN [4]	66.67	25.92/0.91	**Deep Synth [7]**	66.32	25.63/0.88
DrTMO [5]	66.21	25.61/0.87	**Deep Mask [16]**	67.08	25.75/0.89
Deep Recur [9]	66.45	25.58/0.87	**Deep HDRUNet [3]**	66.54	25.71/0.88
Deep Single [12]	67.29	25.90/0.91	**Proposed**	**68.01**	**26.74/0.94**

Table 5: Quantitative comparison on the model size (million) and running time (ms) to process a 512×512 image in Tesla V100 GPU.

	Model Size	Running Time		Model Size	Running Time
HDRCNN [4]	29.44	47	**Deep Synth [7]**	115.21	294
DrTMO [5]	48.09	274	**Deep Mask [16]**	51.54	40
Deep Recur [9]	24.82	79	**Deep HDRUNet [3]**	1.68	33
Deep Single [12]	29.01	102	**Proposed**	33.06	54

6 More visual comparisons

At first, we show the visual result of the training losses ablation study in Fig. 4. With only the L1 reconstruction loss, the details in the over-exposed regions can not be recovered, as shown in Fig. 4 (b). With the perceptual loss, the predicted textures in the highlights can be recovered to some extent. With the progressive reconstruction loss, which generates the small resolution image at first and then predicts the residual maps to get the large resolution results gradually, the lost detail can be restored with vivid color. However, there are still some unnatural artifacts in the over-exposed areas. The proposed mask-aware generative adversarial loss (MAGAN) can solve this problem and generate more realistic textures, as shown in Fig. 4 (e).

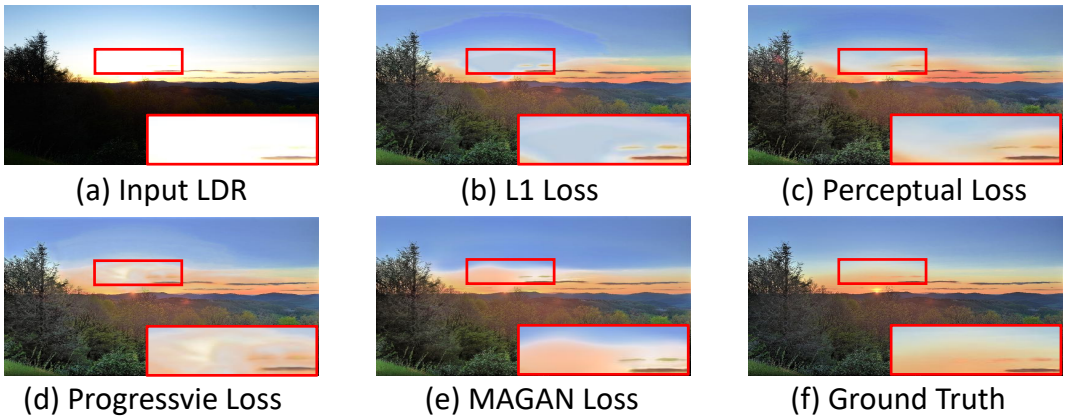

Figure 4: The ablation study of different training losses.

Fig. 5 - Fig. 8 show more visual comparisons between the proposed method and seven recent state-of-the-art CNN-based approaches: HDRCNN [4], DrTMO [5], Deep Recursive HDRI [9], Deep Single HDRI [12], Deep Synth HDRI [7], Deep Mask [16], and Deep HDRUNet [3]. All the HDR images are tone mapped by [11] for displaying on LDR screens. As specified in the main manuscript, the proposed method can recover more natural and accurate results in the under-exposed and over-exposed areas compared to these ITM methods. Meanwhile, we also test the proposed method on the real-world images and several results are shown in Fig. 9, where the input LDR images are shot by the CANON EOS 80D camera.

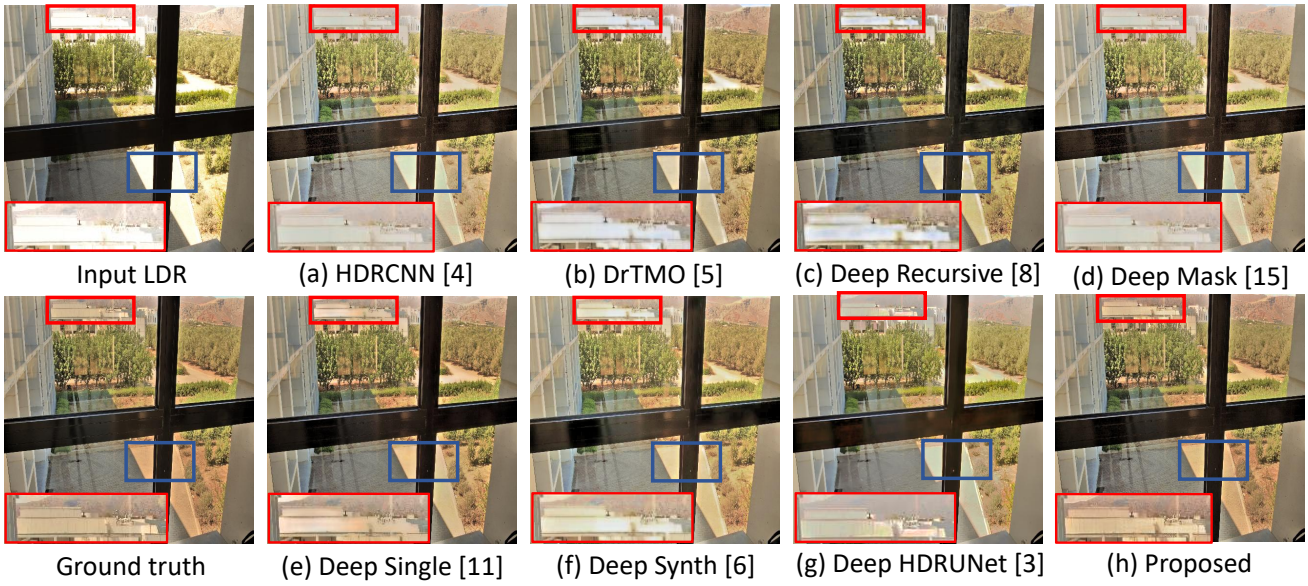

Figure 5: Visual comparison between the proposed method and the state-of-the-art ITM methods.

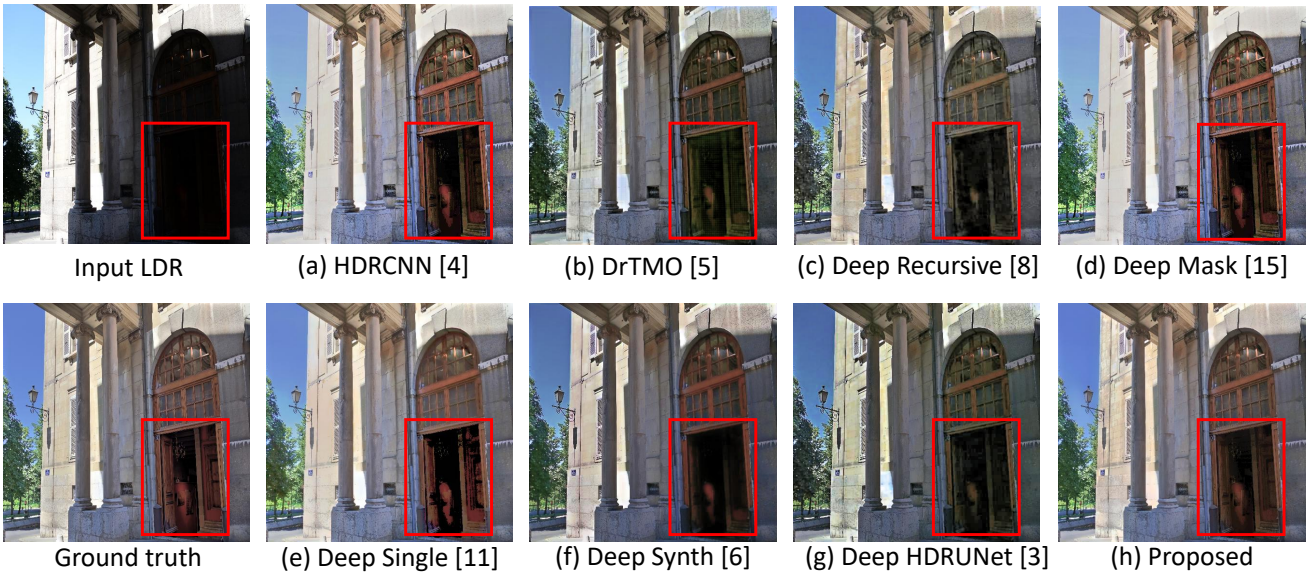

Figure 6: Visual comparison between the proposed method and the state-of-the-art ITM methods.

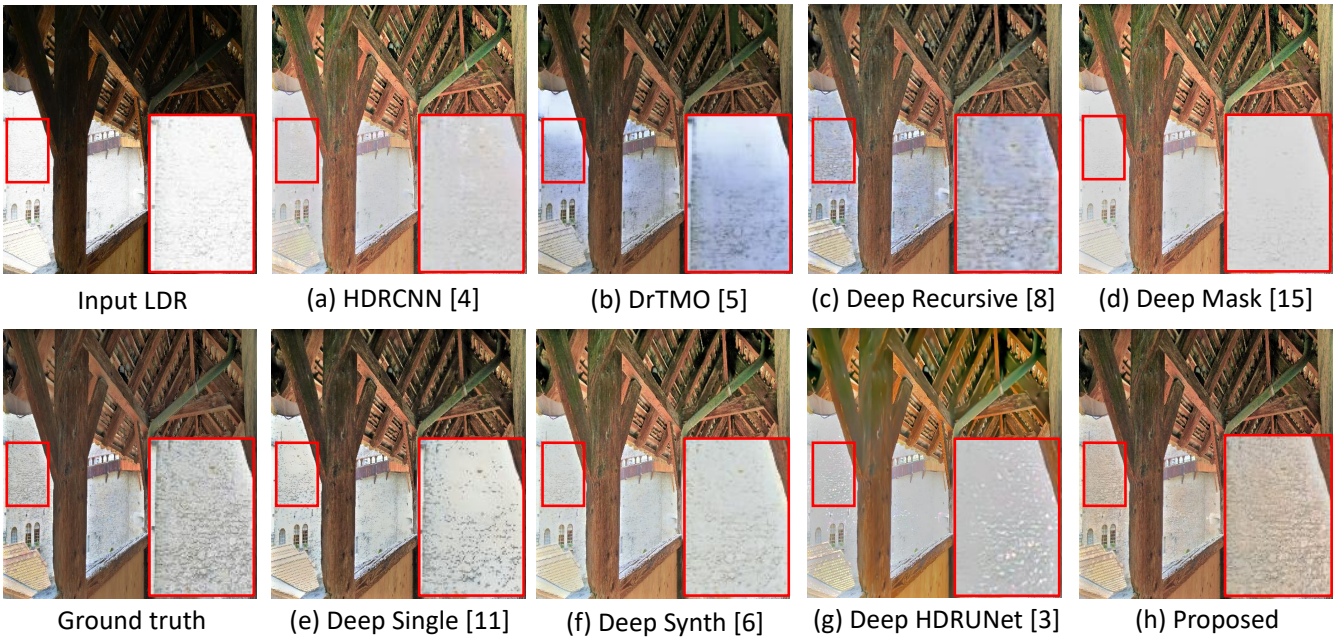

Figure 7: Visual comparison between the proposed method and the state-of-the-art ITM methods.

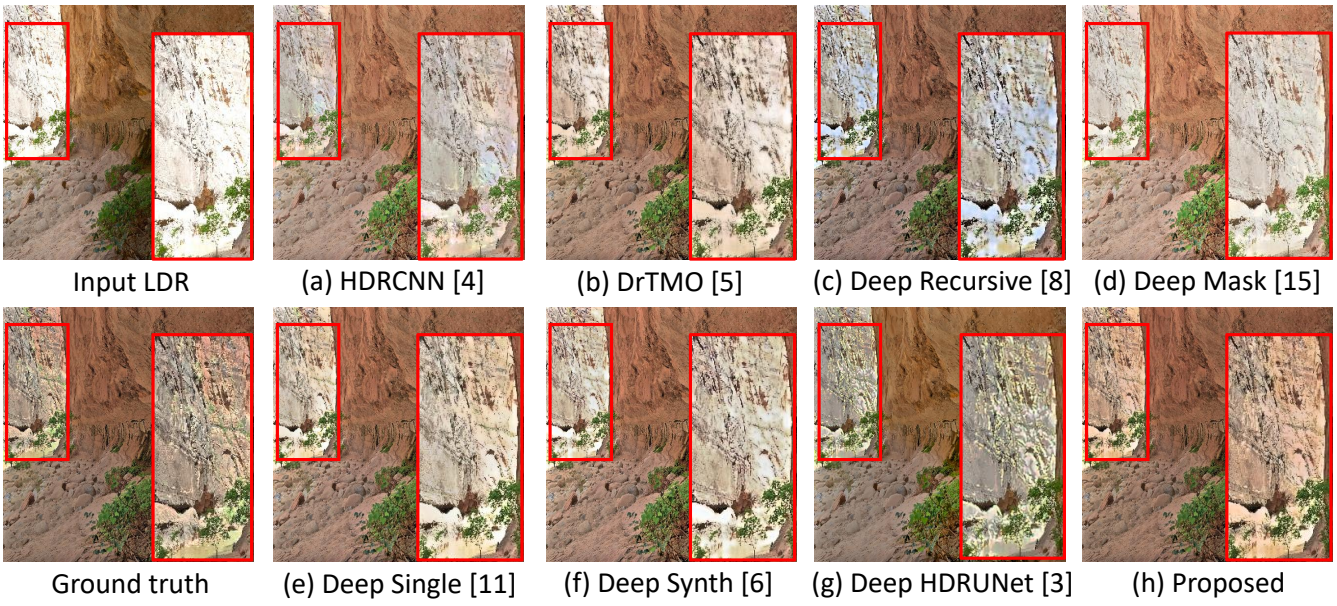

Figure 8: Visual comparison between the proposed method and the state-of-the-art ITM methods.

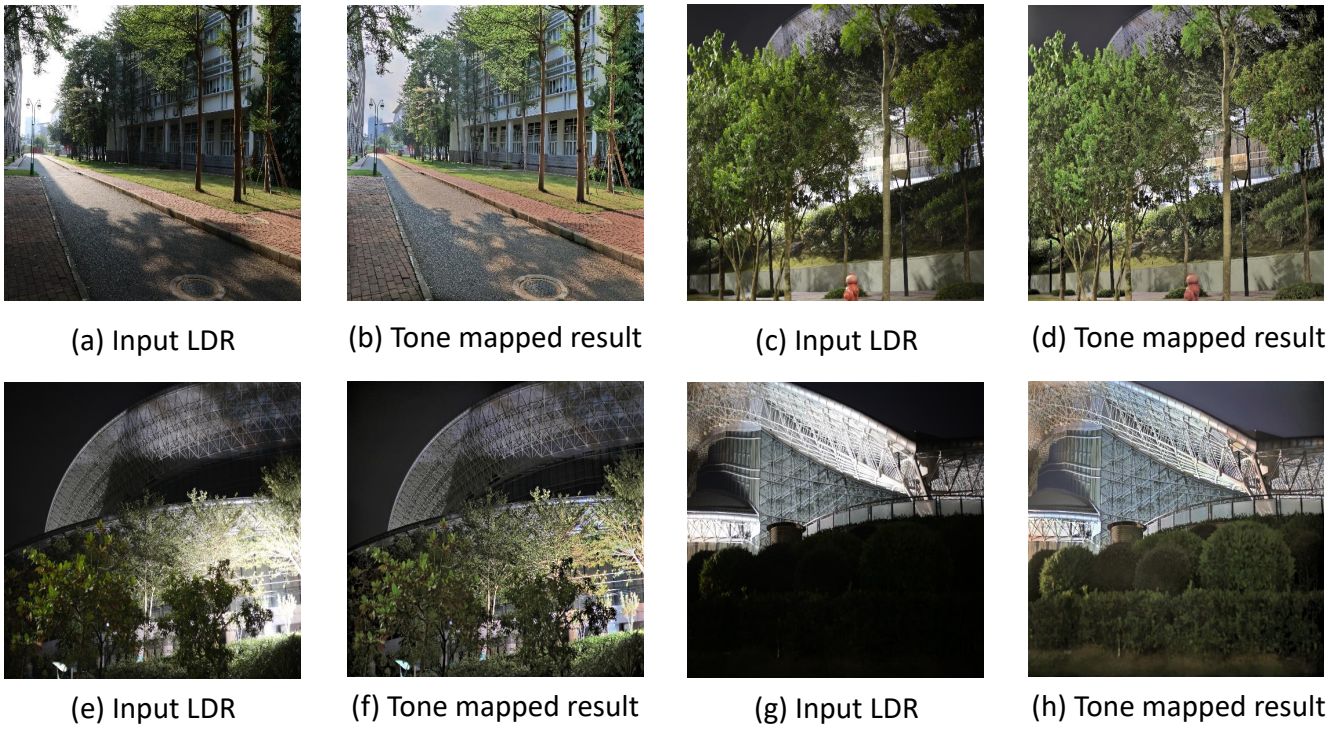

Figure 9: Visual results on the real-world images shot by the CANON EOS 80D camera.

7 User study

User study To further verify the performance of our ITM method, we conduct a user study to evaluate the human preference on the test HDR images. Specifically, 20 subjects are requested to rate 20 predicted HDR images by 6 methods with the input LDR image and ground truth HDR image as references. The participants of the user study mainly consist of college researchers in the image-restoration or video encoding field. The experiment is performed in the same room with the same monitor to rule out other factors such as viewing illumination The score ranges from 1 (the worst) to 8 (the best) spaced with 0.5. The mean opinion score statistics are illustrated in Fig. 10 and the proposed method achieves the highest mean scores (6.69) and the lowest standard deviation (0.83). The mean scores and standard deviations for other ITM methods are HDRCNN (4.95, 1.54), DrTMO (4.73, 1.47), Deep Recursive (4.28, 1.80), Deep Mask (4.875, 1.42), Deep Single (5.77, 1.20), Deep Synth (5.20, 1.63), and Deep HDRUNet (5.45, 0.82).

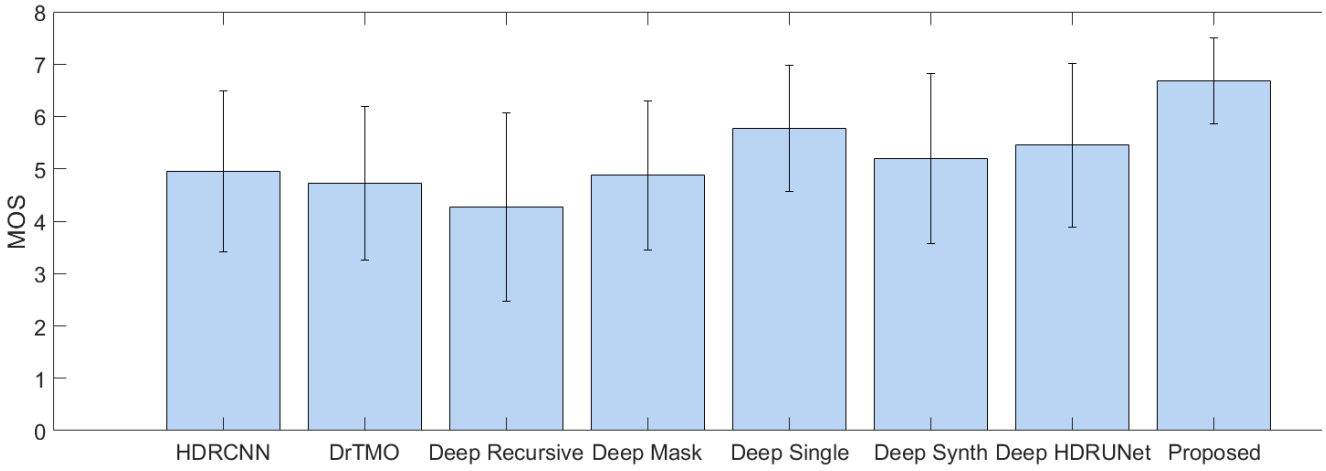

Figure 10: The mean opinion scores of the user study.

8 Limitation and future work

Estimating the lost details of the over-exposed regions is a challenging work, and the current ITM methods cannot handle it well especially when there is no available information from the LDR image to help the hallucination of the over-exposed regions. As Fig. 11 shows, the predicted exposure-down image can recover the lost textures to a certain extent. However, compared to the ground truth image, there lacks some high-frequency details, such as the shape of the clouds, even with the help of the GAN loss. Recently, the diffusion model [15] methods show impressive performance on generating images with vivid high-frequency details. In the future, we will explore how to utilize the diffusion model to help recover more realistic results.

Figure 11: Limitations of the proposed method.

References

[1] Neil Barakat, A Nicholas Hone, and Thomas E Darcie. Minimal-bracketing sets for high-dynamic-range image capture. *IEEE Transactions on Image Processing*, 17(10):1864–1875, 2008.

[2] Jianrui Cai, Shuhang Gu, and Lei Zhang. Learning a deep single image contrast enhancer from multi-exposure images. *IEEE Transactions on Image Processing*, 27(4):2049–2062, 2018.

[3] Xiangyu Chen, Yihao Liu, Zhengwen Zhang, Yu Qiao, and Chao Dong. Hdrunet: Single image hdr reconstruction with denoising and dequantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 354–363, June 2021.

[4] Gabriel Eilertsen, Joel Kronander, Gyorgy Denes, Rafał K Mantiuk, and Jonas Unger. Hdr image reconstruction from a single exposure using deep cnns. *ACM transactions on graphics (TOG)*, 36(6):1–15, 2017.

[5] Yuki Endo, Yoshihiro Kanamori, and Jun Mitani. Deep reverse tone mapping. *ACM Trans. Graph.*, 36(6):177–1, 2017.

[6] Nima Khademi Kalantari, Ravi Ramamoorthi, et al. Deep high dynamic range imaging of dynamic scenes. *ACM Trans. Graph.*, 36(4):144–1, 2017.

[7] Jung Hee Kim, Siyeong Lee, and Suk-Ju Kang. End-to-end differentiable learning to hdr image synthesis for multi-exposure images. *arXiv preprint arXiv:2006.15833*, 2020.

[8] Siyeong Lee, Gwon Hwan An, and Suk-Ju Kang. Deep chain hdri: Reconstructing a high dynamic range image from a single low dynamic range image. *IEEE Access*, 6:49913–49924, 2018.

[9] Siyeong Lee, Gwon Hwan An, and Suk-Ju Kang. Deep recursive hdri: Inverse tone mapping using generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 596–611, 2018.

[10] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1833–1844, 2021.

[11] Zhetong Liang, Jun Xu, David Zhang, Zisheng Cao, and Lei Zhang. A hybrid l1-l0 layer decomposition model for tone mapping. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4758–4766, 2018.

[12] Yu-Lun Liu, Wei-Sheng Lai, Yu-Sheng Chen, Yi-Lung Kao, Ming-Hsuan Yang, Yung-Yu Chuang, and Jia-Bin Huang. Single-image hdr reconstruction by learning to reverse the camera pipeline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1651–1660, 2020.

[13] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.

[14] Rafał Mantiuk, Kil Joong Kim, Allan G Rempel, and Wolfgang Heidrich. Hdr-vdp-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions. *ACM Transactions on graphics (TOG)*, 30(4):1–14, 2011.

[15] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

[16] Marcel Santana Santos, Tsang Ing Ren, and Nima Khademi Kalantari. Single image hdr reconstruction using a cnn with masked features and perceptual loss. *ACM Transactions on Graphics (TOG)*, 39(4):80–1, 2020.