

# Supplementary – LEDiff: Latent Exposure Diffusion for HDR Generation

Chao Wang<sup>1</sup>   Zhihao Xia<sup>2</sup>   Thomas Leimkühler<sup>1</sup>   Karol Myszkowski<sup>1</sup>   Xuaner Zhang<sup>2</sup>  
<sup>1</sup>MPI Informatik   <sup>2</sup>Adobe

{chaowang, karol, thomas.leimkuehler}@mpi-inf.mpg.de   {cezhangxer, zhihao.zach.xia}@gmail.com

## 1. Overview

This supplementary material offers further insights into the details and strengths of our method. Further details on the training process are provided in Sec. 2, computational analysis is discussed in Sec. 3, and the motivation for the learnable fusion module is presented in Sec. 4. In Sec. 5 we present additional results for Text-to-HDR, Image-to-Video, and Inverse Tone Mapping (ITM) tasks. The blending algorithm used for Inverse Tone Mapping is detailed in Sec. 6. Furthermore, the impact of utilizing fewer exposures (two) and additional exposures (five) is analyzed in Sec. 7. We also include the discussion of extreme dynamic range scenes with both highlight and shadow clipping in Sec. 8.

## 2. Training Details

We only show the highlight hallucination branch in Eq. (3) and Eq. (4), where the higher exposure latent code is used as a condition to infer the lower exposure latent codes, as explained earlier, with the optimized target being the noise. The shadow hallucination branch is quite symmetric to highlight handling.

Given the latent exposure bracket  $\{C_-, C_0, C_+\}$ . We then corrupt the overexposed latent  $C_+$  and  $C_0$  with Gaussian noise at a randomly sampled timestamp  $t$ . The task of the denoising network,  $\epsilon_{\theta_+}(\cdot)$  is to predict the noise  $\hat{\epsilon}_{C_+, C_0} = \epsilon_{\theta_+}(C_0, C_+^t, t)$  and  $\hat{\epsilon}_{C_0, C_-} = \epsilon_{\theta_+}(C_-, C_0^t, t)$  using the objective:

$$\mathcal{L}_C = \mathbb{E}_{C, \epsilon_{C_+, C_0} \sim \mathcal{N}(0, I), t \sim \mathcal{U}(T)} \|\epsilon_{C_+, C_0} - \hat{\epsilon}_{C_+, C_0}\|_2^2 \quad (1)$$

$$+ \mathbb{E}_{C, \epsilon_{C_0, C_-} \sim \mathcal{N}(0, I), t \sim \mathcal{U}(T)} \|\epsilon_{C_0, C_-} - \hat{\epsilon}_{C_0, C_-}\|_2^2 \quad (2)$$

During inference, given an LDR image, we iteratively apply the fine-tuned denoiser to generate its corresponding higher-exposure latent codes. The weight  $\lambda$  in loss function Eq. (2) of the main paper is set to  $10^{-4}$ , and a batch size of is used for finetuning both the VAE and the denoiser Unet.

## 3. Inference Time and Computational Cost

We also compare inference time and computational cost, both evaluated on an A100 at 512×512, as shown in Table 1.

While the regression-based methods are more efficient but yield poorer reconstructions (Fig. 5 in the main paper). Meanwhile, compared to the generation-based method GlowGAN [12], our method achieves orders-of-magnitude superiority in inference speed and computational efficiency.

Method	FLOPs	Time (s)
HDRCNN	60.73G	0.03
MaskHDR	75.36G	0.54
SingleHDR	570.79G	0.08
ExpandNet	53.78G	0.02
GlowGAN	7P	780
Ours	68T	8

Table 1. Comparisons of the inference time and computational cost.

## 4. Motivation of depth-wise convolution

We use depthwise convolution for its lightweight nature and efficiency. It extracts features from each channel of the latent code independently, while concatenation and softmax-based normalization enable cross-talk between latent codes of different exposures. In traditional image exposure fusion [7], weight maps are manually designed based on factors like color, exposure, and contrast. However, in the latent space, these variables are more challenging to define, which is why we adopt a learning-based approach.

## 5. More Results

If you have a device that supports HDR, for which you can test at: <https://gregbenzphotography.com/hdr/#tests>, we suggest that you also view our results with the html pages we provide in the supplement (i.e., `text-to-hdr.html`, `inverse-tone-mapping-highlight.html`, `inverse-tone-mapping-shadow.html`). Please use Chrome to ensure the images are displayed correctly.

### 5.1. Text to HDR

In this section, we present more results of text-to-HDR image generation, as shown in Figs. 1–3. We reduce the exposure and provide multiple exposure brackets to show a more comprehensive comparison of dynamic range and hallucination.

### 5.2. Image to Video

Our method can be seamlessly integrated into SD-based image-to-video models [2], enabling the synthesis of HDR video from a single LDR image. An example is shown in Fig. 4. Please also refer to the folder of `image-to-video` for all the generated HDR frames.

### 5.3. Inverse Tone Mapping

We provide additional comparisons of inverse tone mapping to evaluate highlight hallucination in Figs. 5–10 and shadow hallucination in Figs. 11–16. Highlight hallucination is demonstrated using tone-mapped images, while shadow hallucination is illustrated by increasing exposure brackets to enhance clarity. LEDiff demonstrates superior quality in reconstructing plausible content within clipping regions, whereas other approaches either fail to generate meaningful outputs or produce artifacts characterized by significant blurriness.

## 6. Blending Algorithm

For highlight hallucination, we follow [4, 9, 12] to blend the generated content with the inputs. The process models HDR luminance  $\hat{\mathcal{H}}$  in the non-overexposed regions of the LDR image  $I$ , aligning it with the corresponding HDR area.

$$I_{\text{non-overexposed}} = (1 - \mathcal{M}) \odot I \quad (3)$$

$$\mathcal{M} = \min \left( 1.0, \max \left( 0.0, \frac{\max(I) - 1.0 + \text{thr}}{\text{thr}} \right) \right), \quad (4)$$

where  $(\text{thr}) = 0.1$  is a threshold that controls the smoothness of the transition. Specifically, we estimate the HDR luminance as

$$\hat{\mathcal{H}}_{\text{estimated}} = (I_{\text{non-overexposed}})^{\gamma} \cdot 2^{\text{exp}} \quad (5)$$

where  $\gamma$  and  $\text{exp}$  are the parameters that control the transformation. Note that we use this simple inversion of the camera response function solely for image quality evaluation to meaningfully align input LDR pixel values to their reconstructed HDR counterpart. In our HDR generation and reconstruction tasks, we do not estimate the exposure or the camera response curve. The objective is to minimize the difference  $\Delta$  between  $\hat{\mathcal{H}}_{\text{estimated}}$  and the reconstructed by the decoder HDR image  $\hat{\mathcal{H}}$  for non-overexposed regions:

$$\Delta = \hat{\mathcal{H}}_{\text{estimated}} - \hat{\mathcal{H}}_{\text{non-overexposed}}. \quad (6)$$

To optimize  $\gamma$  and  $\text{exp}$ , we use a least-squares approach that minimizes the squared error across all non-overexposed pixels:

$$\min_{\gamma, \text{exp}} \sum (\hat{\mathcal{H}}_{\text{estimated}} - \hat{\mathcal{H}}_{\text{non-overexposed}})^2. \quad (7)$$

We initialize  $\gamma$  at 2.0 and  $\text{exp}$  at 0.0, constraining  $\gamma$  within the range  $1.8 \leq \gamma \leq 2.4$  for stability during optimization. Once the optimal  $\gamma$  and  $\text{exp}$  values are found, they are applied to the entire LDR image to adjust its luminance:

$$\hat{\mathcal{H}}_{\text{adjusted}} = (I)^{\gamma} \cdot 2^{\text{exp}}. \quad (8)$$

Next, we blend the adjusted image with the HDR image to create a seamless transition between the two. Using a blending mask that isolates the non-overexposed regions of the LDR image, we compute the blended image as follows:

$$\hat{\mathcal{H}}_{\text{blend}} = (1 - \mathcal{M}) \cdot \hat{\mathcal{H}}_{\text{adjusted}} + \mathcal{M} \cdot \hat{\mathcal{H}}. \quad (9)$$

## 7. Ablation of Exposures

In this paper, we use a three-exposure bracket, a common practice in image-based exposure fusion [3, 5, 6, 10, 11, 13]. Additionally, we study the effects of using fewer exposures (two) and more exposures (five). In this ablation, we modify the exposure interval during data creation and finetune both the VAE and the denoiser. We then use ITM to hallucinate highlight areas to assess its performance. While five exposures increases the dynamic range it can capture, it requires additional and accumulative denoising iterations, where errors add up.

In traditional multi-exposure HDR reconstruction methods, including the concurrent diffusion-based solution [1], exposure brackets are defined in the image space, with the number of brackets typically determining the extent of the reconstructed dynamic range. In contrast, our approach, LEDiff, introduces exposure brackets defined in the latent space. These exposure brackets represent an internal, intermediate HDR image encoding, which is further processed by a learnable decoder. As a result, drawing an analogy to image-space exposures is not straightforward. This explains why we found increasing the number of exposure levels leads to worse results, contrary to the conventional approach in image-space exposure bracketing.

Exposures	HDR-VDP $\uparrow$	PU21-PIQE $\downarrow$
Two	6.14 $\pm$ 0.86	48.70 $\pm$ 6.94
Five	5.73 $\pm$ 1.15	49.19 $\pm$ 6.80
Ours	6.16 $\pm$ 0.97	48.46 $\pm$ 7.04

Table 2. Ablation study on the number of exposures. The presented results indicate that using three exposures yields superior performance.



## 8. Extreme Dynamic Range Scenes

Our model addresses such cases by sequentially processing the underexposed regions followed by the overexposed regions. We provide one of such examples in Fig. 17 and will include more detailed discussions and examples in the supplement.



Figure 1. Results of Text-to-HDR image. Our method enables the generation of HDR images from text prompts, overcoming the limitation of Stable Diffusion [8], which is restricted to producing LDR images.

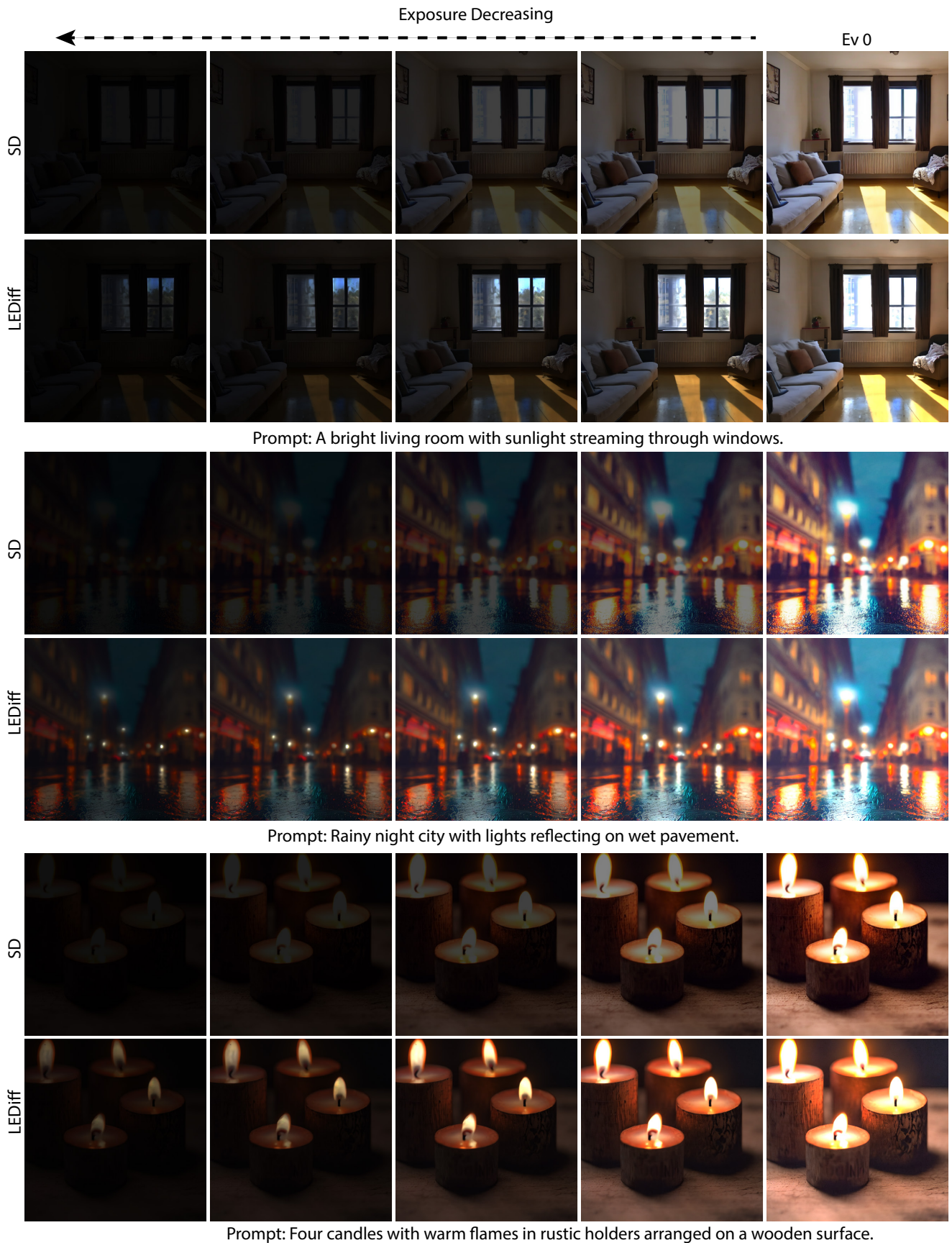


Figure 2. Results of Text-to-HDR image. Our method enables the generation of HDR images from text prompts, overcoming the limitation of Stable Diffusion [8], which is restricted to producing LDR images.





Figure 3. Results of Text-to-HDR image. Our method enables the generation of HDR images from text prompts, overcoming the limitation of Stable Diffusion [8], which is restricted to producing LDR images.

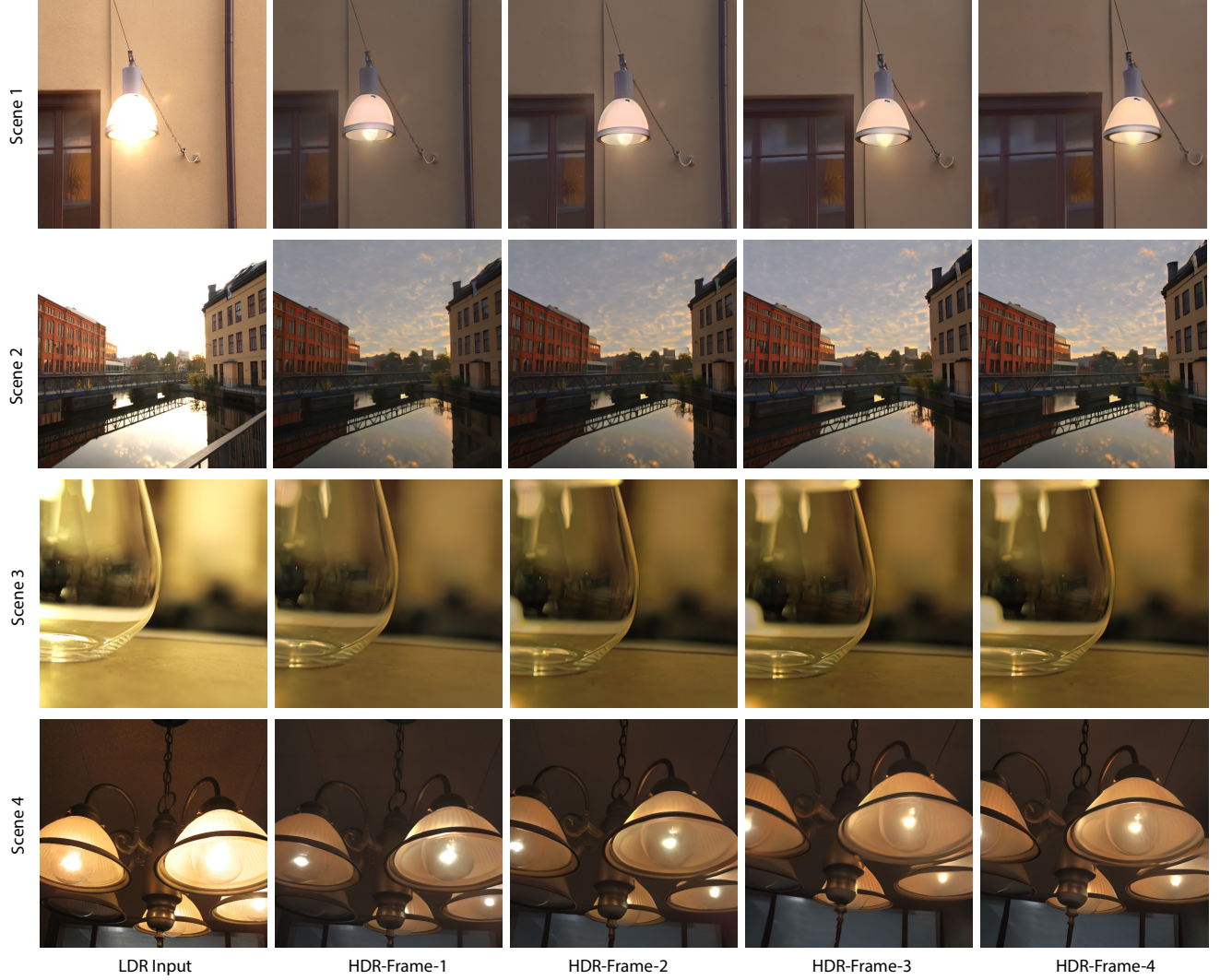


Figure 4. Results of image-to-HDR video. Our method enables the baseline model, SVD, to generate HDR video from a single LDR image.



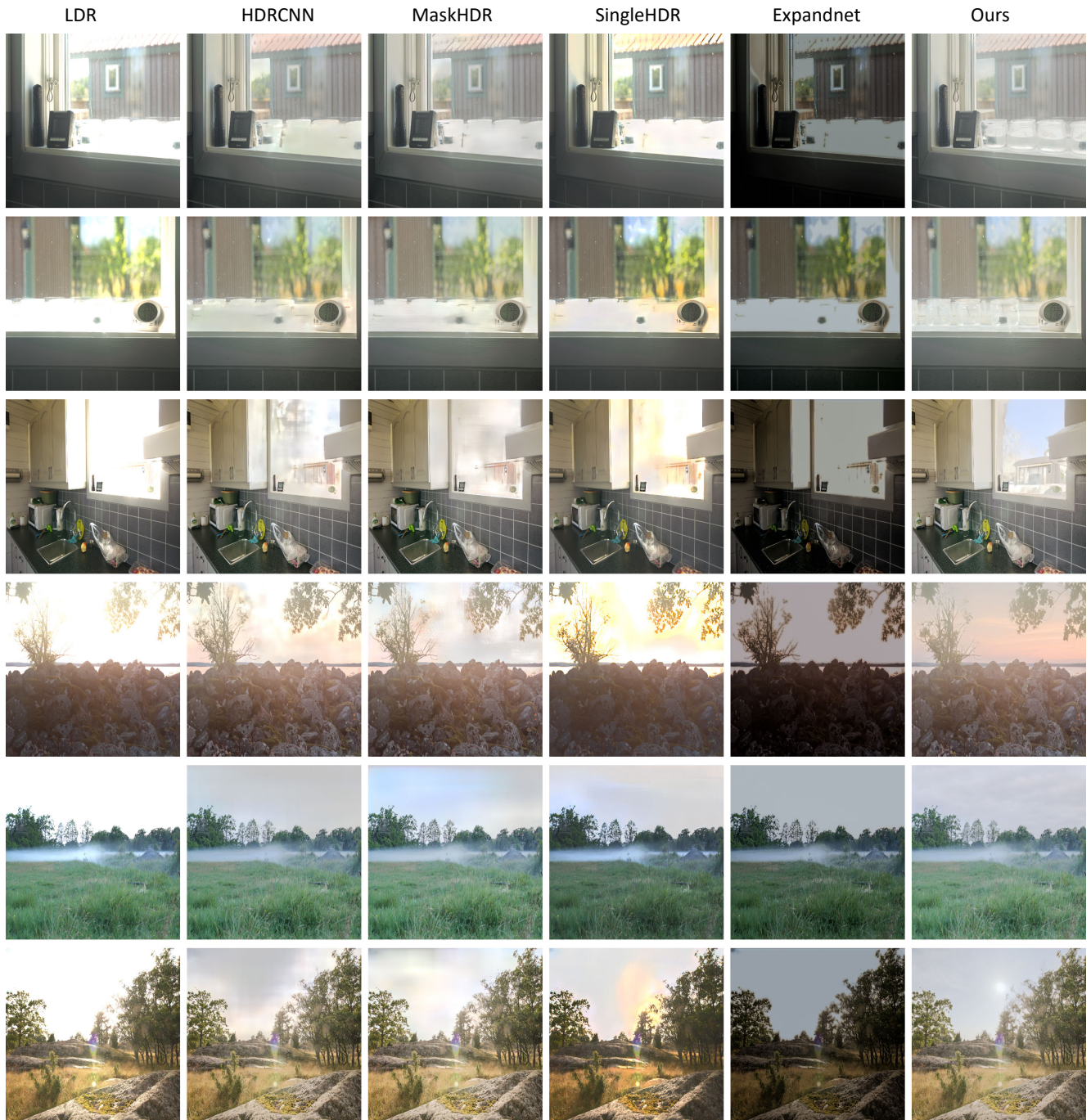


Figure 5. Visualizations of inverse tone mapping evaluations for hallucination are shown across several representative HDR scenes.



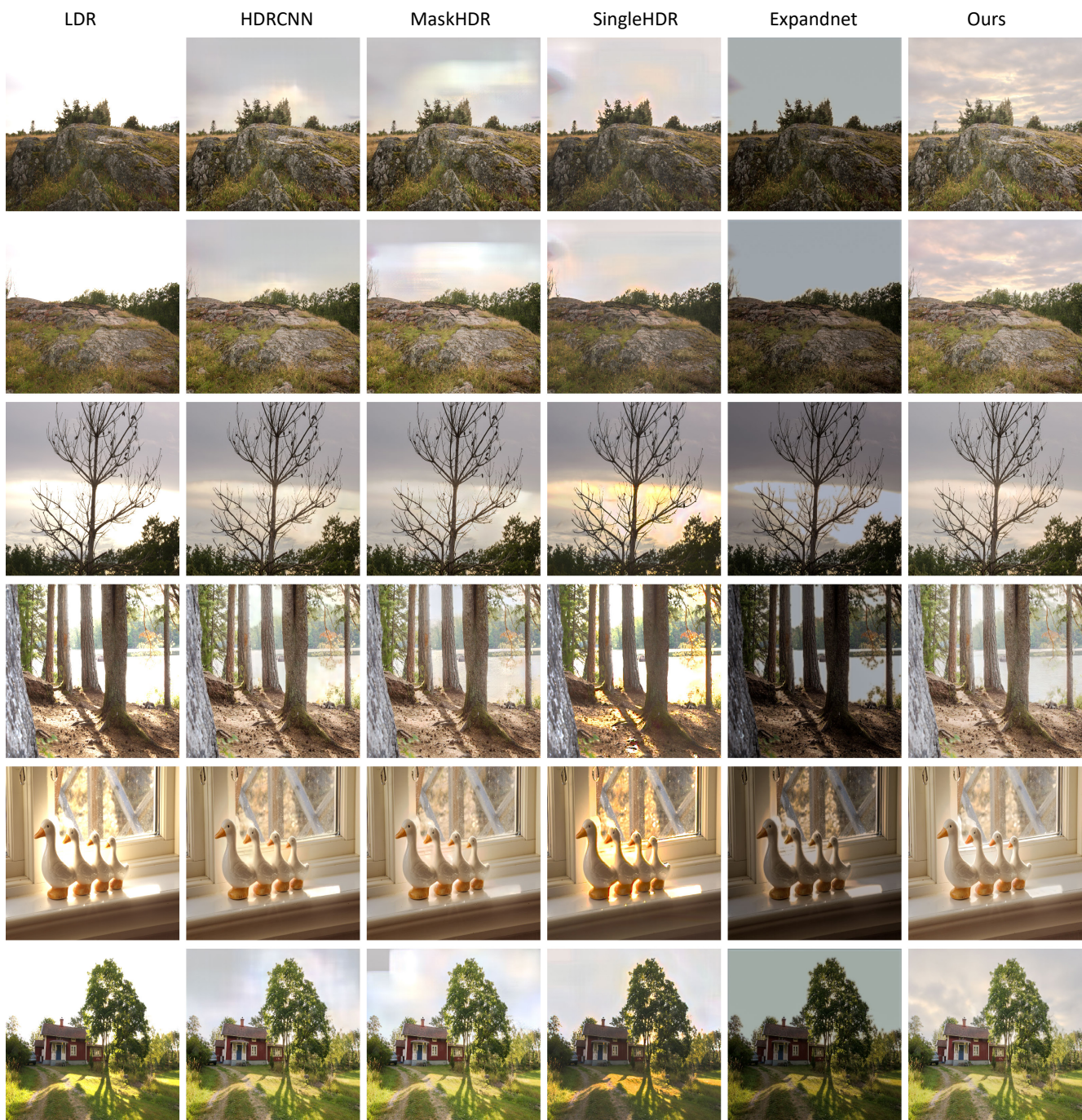


Figure 6. Visualizations of inverse tone mapping evaluations for hallucination are shown across several representative HDR scenes.



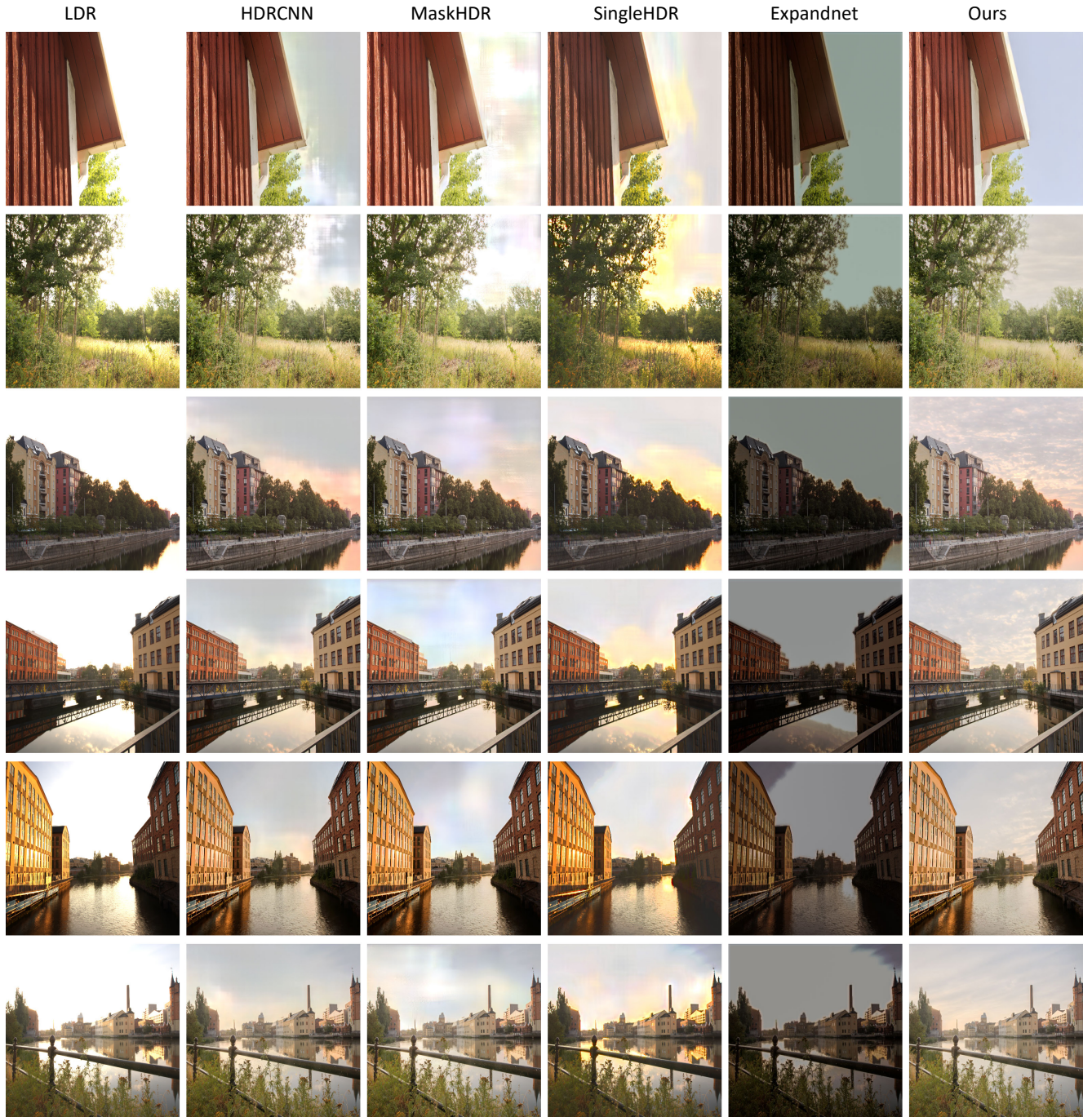


Figure 7. Visualizations of inverse tone mapping evaluations for hallucination are shown across several representative HDR scenes.



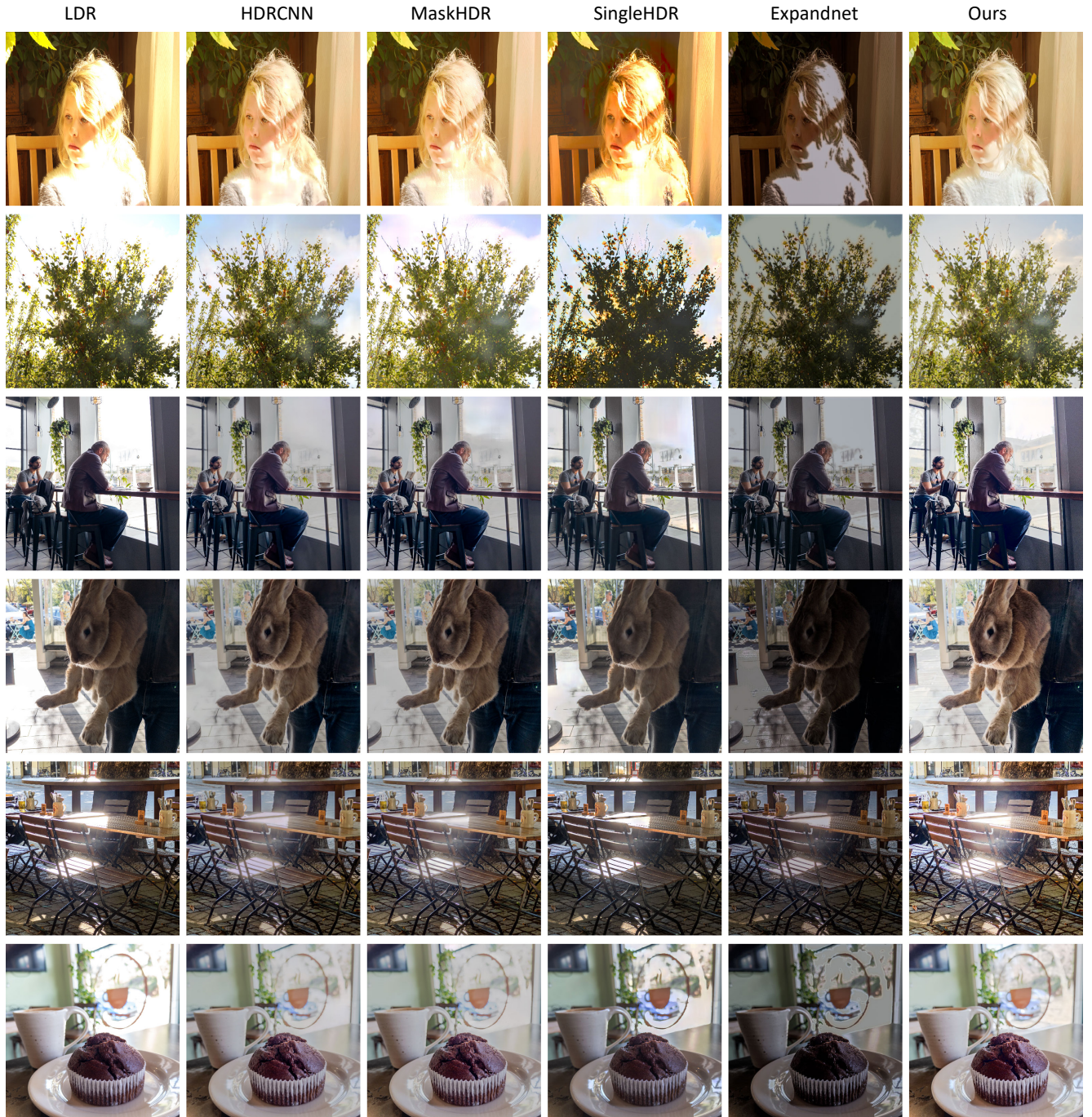


Figure 8. Visualizations of inverse tone mapping evaluations for hallucination are shown across several representative HDR scenes.



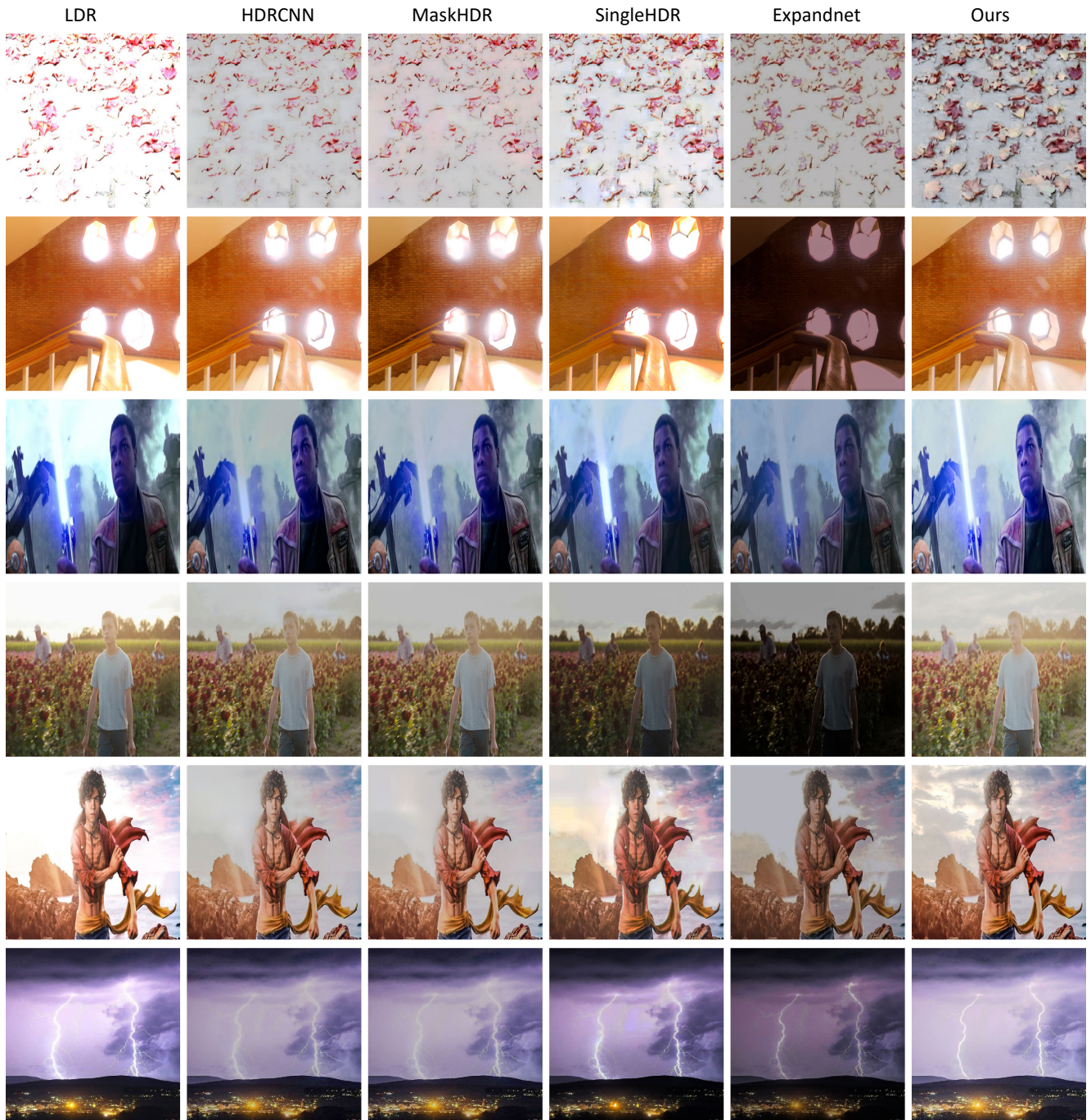


Figure 9. Visualizations of inverse tone mapping evaluations for hallucination are shown across several representative HDR scenes.



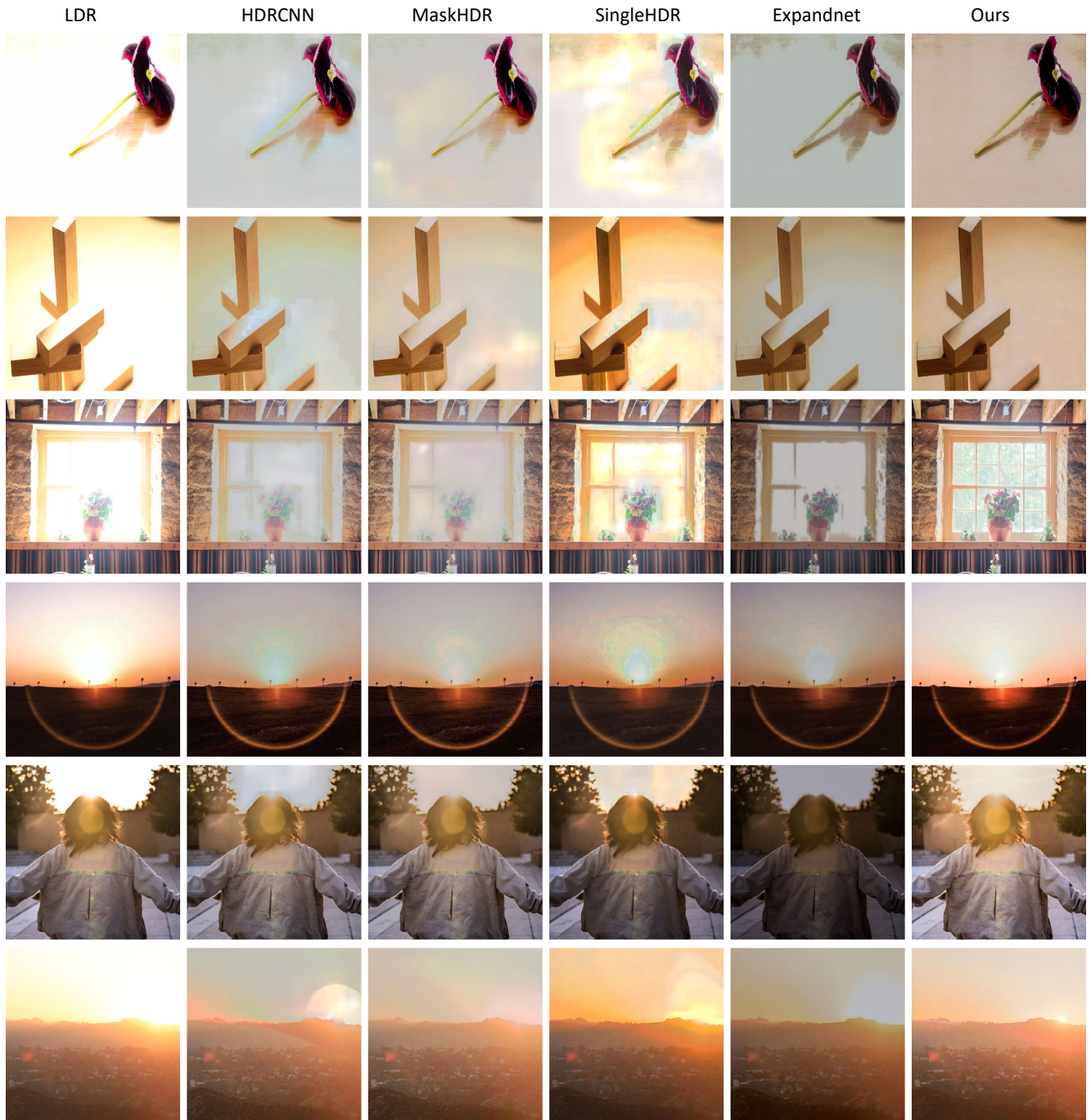


Figure 10. Visualizations of inverse tone mapping evaluations for hallucination are shown across several representative HDR scenes.

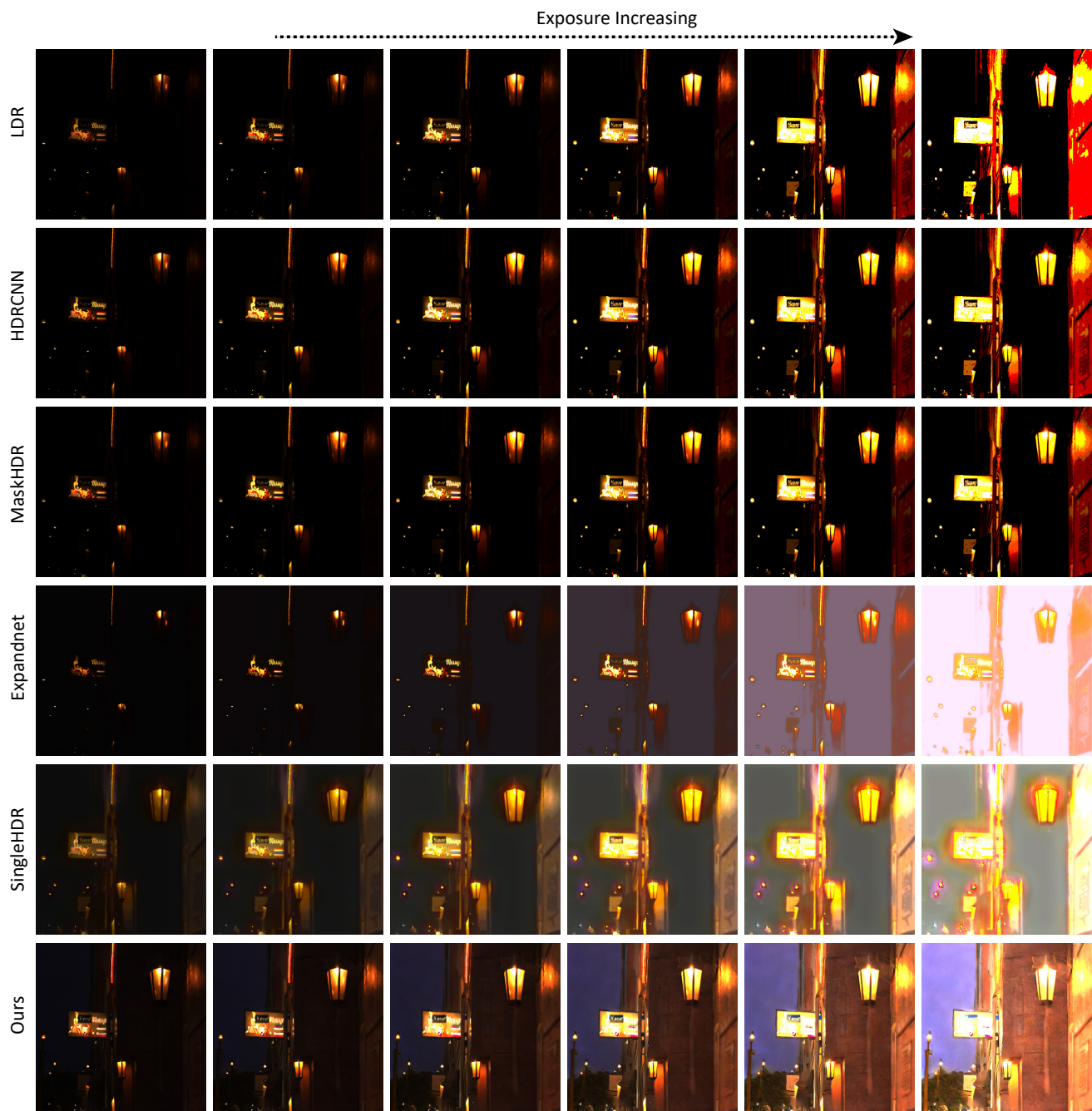


Figure 11. Visualizations of inverse tone mapping evaluations focusing on shadow hallucination. For improved illustration, the exposure levels are increased to enhance the visibility of shadow regions and show the hallucinated content produced by our method.



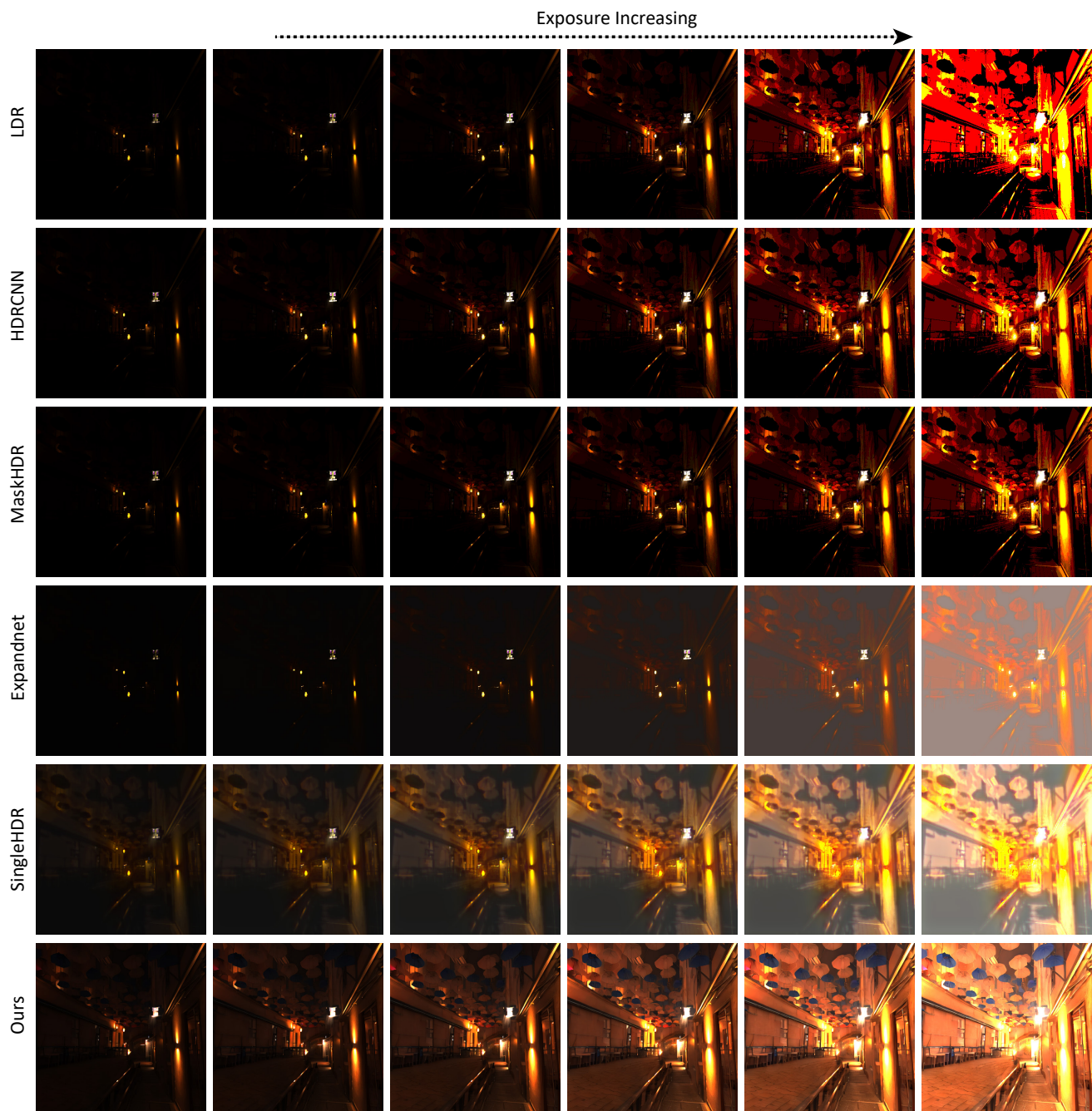


Figure 12. Visualizations of inverse tone mapping evaluations focusing on shadow hallucination. For improved illustration, the exposure levels are increased to enhance the visibility of shadow regions and show the hallucinated content produced by our method.

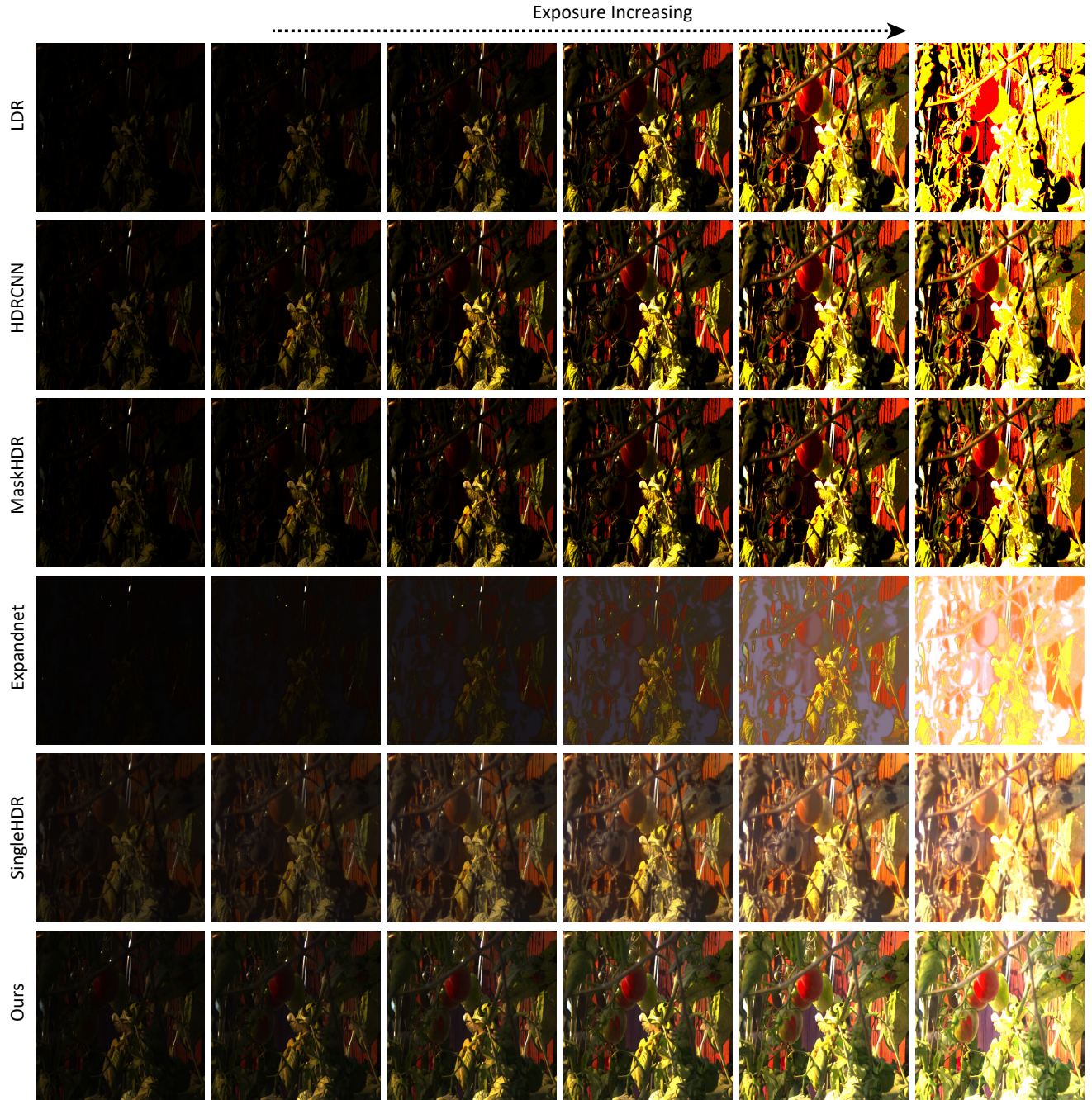


Figure 13. Visualizations of inverse tone mapping evaluations focusing on shadow hallucination. For improved illustration, the exposure levels are increased to enhance the visibility of shadow regions and show the hallucinated content produced by our method.





Figure 14. Visualizations of inverse tone mapping evaluations focusing on shadow hallucination. For improved illustration, the exposure levels are increased to enhance the visibility of shadow regions and show the hallucinated content produced by our method.



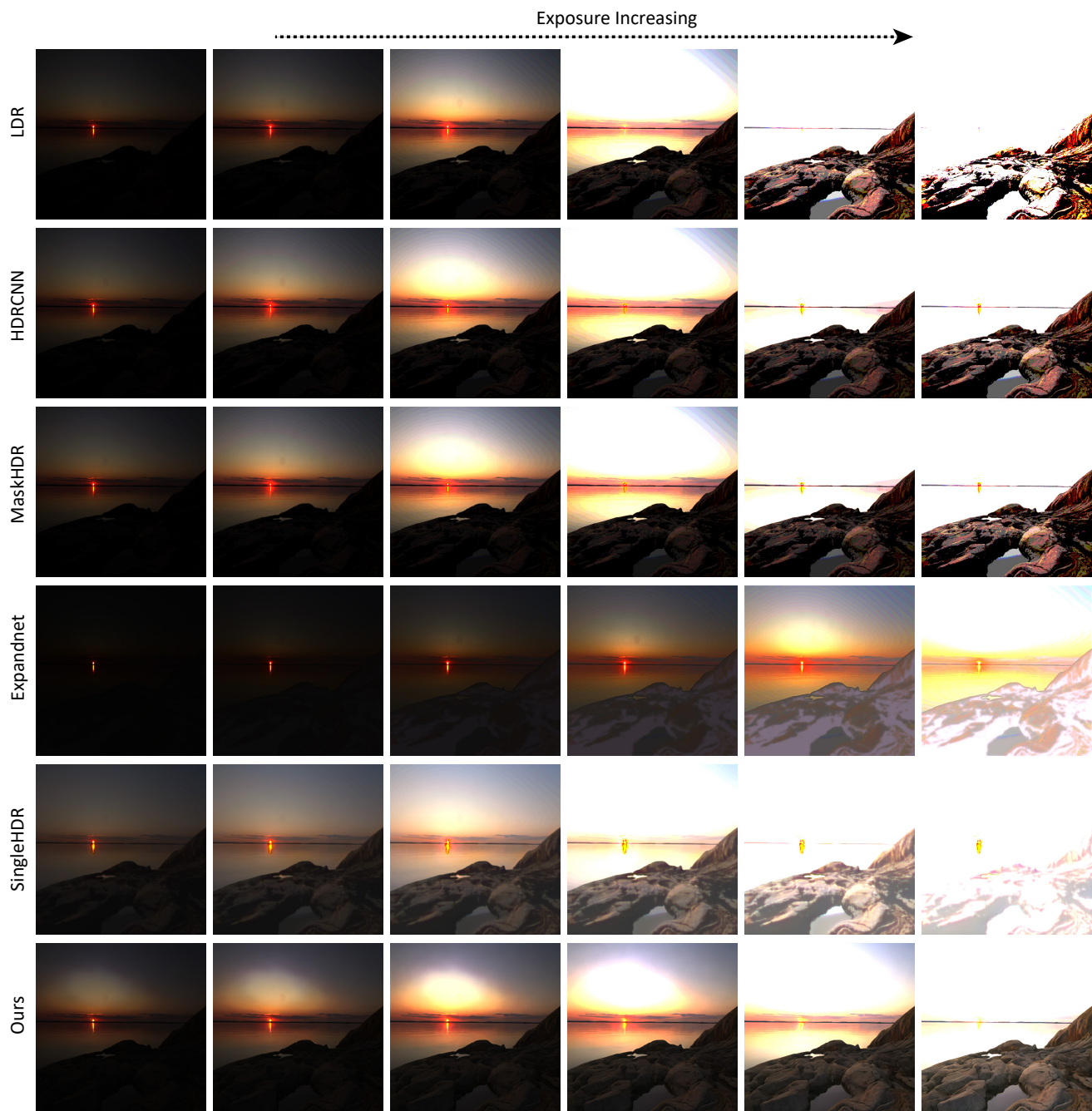


Figure 15. Visualizations of inverse tone mapping evaluations focusing on shadow hallucination. For improved illustration, the exposure levels are increased to enhance the visibility of shadow regions and show the hallucinated content produced by our method.

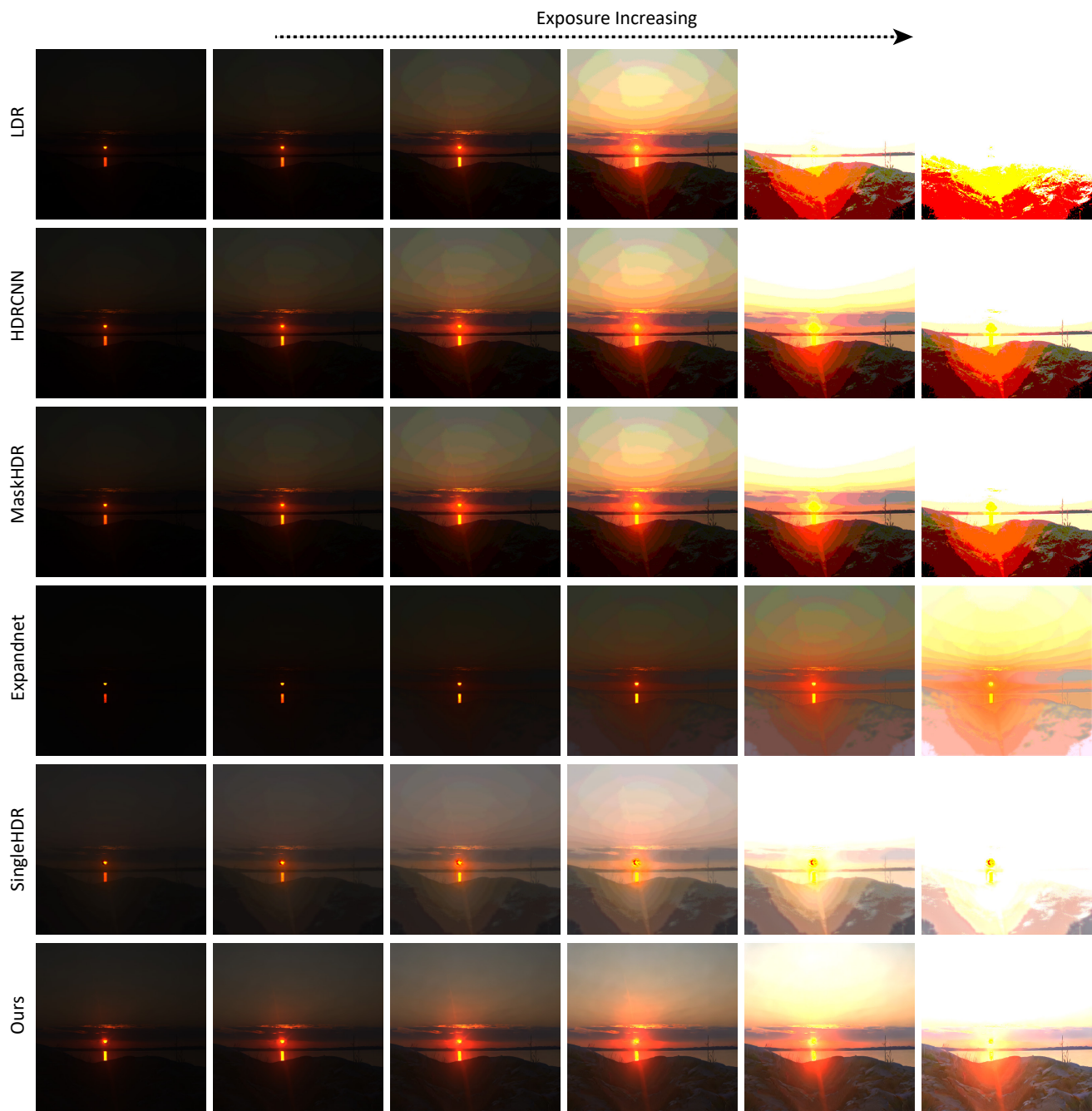


Figure 16. Visualizations of inverse tone mapping evaluations focusing on shadow hallucination. For improved illustration, the exposure levels are increased to enhance the visibility of shadow regions and show the hallucinated content produced by our method.

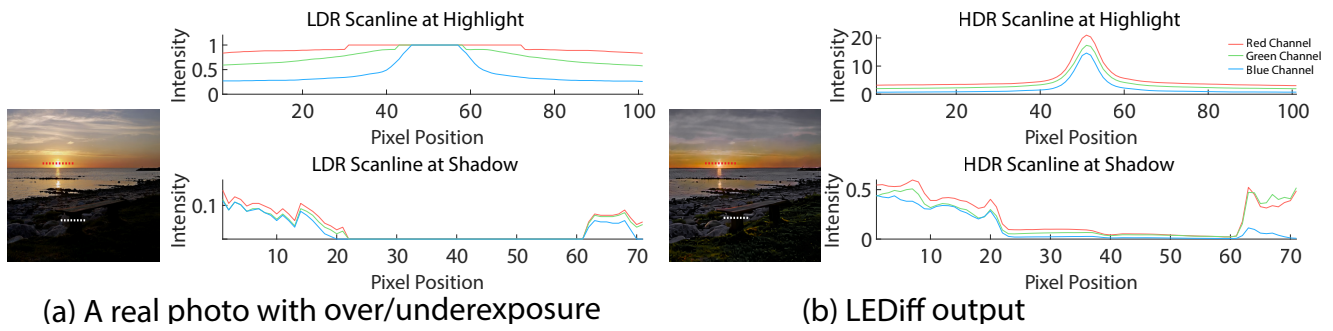


Figure 17. Reconstruction of an image with both highlight and shadow clipping. (a) A real photo that captures an extreme dynamic range scene, and its scanlines. (b) Our HDR reconstruction successfully recovers details in both shadow and highlight regions. Note the differing y-axis scales in the left and right plots.

## References

- [1] Mojtaba Bemana, Thomas Leimkühler, Karol Myszkowski, Hans-Peter Seidel, and Tobias Ritschel. Exposure diffusion: Hdr image generation by consistent ldr denoising. *arXiv preprint arXiv:2405.14304*, 2024. [2](#)
- [2] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. [2](#)
- [3] Rufeng Chen, Bolun Zheng, Hua Zhang, Quan Chen, Cheng-gang Yan, Gregory Slabaugh, and Shanxin Yuan. Improving dynamic hdr imaging with fusion transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 340–349, 2023. [2](#)
- [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. [2](#)
- [5] Nima Khademi Kalantari, Ravi Ramamoorthi, et al. Deep high dynamic range imaging of dynamic scenes. *ACM Trans. Graph.*, 36(4):144–1, 2017. [2](#)
- [6] Zhen Liu, Yinglong Wang, Bing Zeng, and Shuaicheng Liu. Ghost-free high dynamic range imaging with context-aware transformer. In *European Conference on Computer Vision*, pages 344–360. Springer, 2022. [2](#)
- [7] Tom Mertens, Jan Kautz, and Frank Van Reeth. Exposure fusion. In *15th Pacific Conference on Computer Graphics and Applications (PG’07)*, pages 382–390. IEEE, 2007. [1](#)
- [8] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. [4](#), [5](#), [6](#)
- [9] 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. [2](#)
- [10] Jou Won Song, Ye-In Park, Kyeongbo Kong, Jaeho Kwak, and Suk-Ju Kang. Selective transhdr: Transformer-based selective hdr imaging using ghost region mask. In *European Conference on Computer Vision*, pages 288–304. Springer, 2022. [2](#)
- [11] Steven Tel, Zongwei Wu, Yulun Zhang, Barthélémy Heyrman, Cédric Demonceaux, Radu Timofte, and Dominique Ginjac. Alignment-free hdr deghosting with semantics consistent transformer. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12790–12799. IEEE, 2023. [2](#)
- [12] Chao Wang, Ana Serrano, Xingang Pan, Bin Chen, Karol Myszkowski, Hans-Peter Seidel, Christian Theobalt, and Thomas Leimkühler. Glowgan: Unsupervised learning of hdr images from ldr images in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10509–10519, 2023. [1](#), [2](#)
- [13] Shangzhe Wu, Jiarui Xu, Yu-Wing Tai, and Chi-Keung Tang. Deep high dynamic range imaging with large foreground motions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 117–132, 2018. [2](#)