Generating Content for HDR Deghosting from Frequency View

Supplementary Material

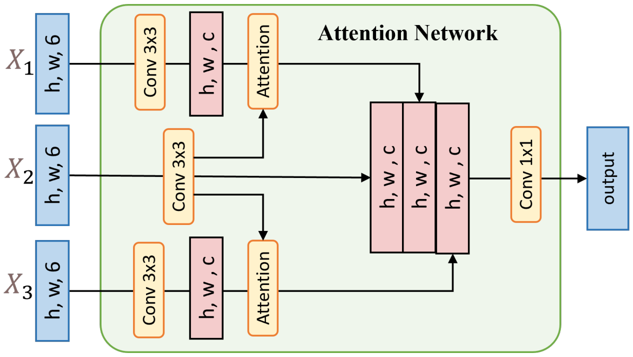

Figure 1. We using the attention module from AHDR [12] as the alignment module. Here, X_i represents 6-channel LDR images, and the output features have $C = 60$ channels.

1. Model Details

Denoising Network: The network architecture is a modified version of the U-Net found in DDPM [2]; As shown in Fig. 2, we replace the original DDPM residual blocks with slightly modified nonlinear activation free blocks (NAFBlocks)[9]. Nonlinear activation free means that we replace all nonlinear activation functions with the "SimpleGate", an element-wise operation that splits feature channels into two parts and then multiplies them together to produce the output. As illustrated in Fig.2, to make the model share parameters across time, we add an additional multilayer perceptron to process the time embedding to channel-wise scale and shift parameters γ and β, for both the attention layer and feed-forward layer. In practical applications, we select $C = 32$, and employ four resolution depths in the U-Net architecture, with channel multipliers of $\{1, 2, 4, 8\}$. In both the encoder and decoder, each stage comprises two NAFBlocks, while the middle stage specifically consists of one NAFBlock. This results in the entire denoising network containing $1.76M$ parameters.

DHRNet: DHRNet is comprised of multiple stacked Reconstruction Blocks, each including one Prior Integration Module (PIM) 3 (a) and several Feature Refinement Modules (FRM) 3 (b). In practical applications, the number of FRMs in each block is set to 3, and C is configured as 60. In PIM, N is set to 3. The downsampling kernel sizes for avgpool in PIM and FRM are set as 4 and 2, respectively. It is noteworthy that DHRNet does not directly utilize LDR images as input; instead, it employs an alignment module (AM) to obtain implicitly aligned features as input. As depicted in Fig. 1, we utilize the Attention Network from AHDR [12] to process various LDR images, extracting implicitly-aligned features for input to DHRNet.

2. Algorithm

Our LF-Diff consists of two training stages. After completing the first pretraining phase for LF-Diff, the algorithm for the second stage LF-Diff training is outlined in Algorithm 1. The algorithm for LF-Diff inference is summarized in Algorithm 2.

3. Perceptual Metrics

As indicated in Table 1, we additionally computed various common perceptual metrics, including FID [1], LPIPS [15], VSI [14], and AHIQ [6]. Due to the domain differences between HDR images and natural images, tonemapping was applied to both the generated results and ground truth (GT) for computing perceptual metrics. This approach enables a more accurate evaluation of the quality of the generated images. It can be observed that, compared to DNN-based methods, DDPM-based methods typically exhibit superior perceptual metrics. Our approach maintains excellent perceptual metric performance while being 10× faster than the previous diffusion-model-based method DiffHDR.

4. Additional Qualitative Results

In this section, we present additional qualitative results that we did not show in the main text due to the limited space of paper. In Figs. 4 show visual results for various motion cases in Kalantari's dataset [5] and Hu's dataset [4]. Fig. 5 and 6 provide additional qualitative results without ground truth.

References

[1] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 2

[2] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, pages 6840–6851. Curran Associates, Inc., 2020. 2

[3] Jun Hu, O. Gallo, K. Pulli, and Xiaobai Sun. HDR deghosting: How to deal with saturation? In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1163–1170, 2013. 4

[4] Jinhan Hu, Gyeongmin Choe, Zeeshan Nadir, Osama Nabil, Seok-Jun Lee, Hamid Sheikh, Youngjun Yoo, and Michael Polley. Sensor-realistic synthetic data engine for multi-frame high dynamic range photography. In *Proceedings of*

Algorithm 1 DM Training

```
Input:
    LDRs-HDR image pairs $(X_i, H)$,
    Total diffusion step $T$, implicit sampling step $S$, Noise schedule $\beta_t (t \in [1, T])$.
Initialize: $\alpha_t = 1 - \beta_t, \bar{\alpha}_T = \prod_{i=0}^{T} \alpha_i$
Initialize: The DHRNet (contains AM and Conv3 $\times$ 3 behind DHRNet) of LF-Diff copies the parameters of trained LF-Diff from stage one.
repeat
    $t \sim Uniform\{1, \cdots, T\}$
    $\epsilon_t \sim \mathcal{N}(0, \mathbf{I})$
    $z =$ LPENet (PixelUnshuffle ( Concat $(H, \mathcal{T}(H))))$
    $\mathbf{D} =$ LPENet$_{DM}$ $(AM($ PixelUnshuffle $(X_i)))$
    $z_0 = z$
    Take gradient descent step on $\mathbb{E}_{t, z_0, \epsilon_t} [\|\epsilon_t - \epsilon_\theta(\sqrt{\bar{\alpha}_t} z_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon_t, t, D)\|^2]$
    $\hat{\mathbf{z}}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
    for $i = S : 1$ do
        $t = (i - 1) \cdot T/S + 1$
        $t_{\text{next}} = (i - 2) \cdot T/S + 1$ if $i > 1$, else 0
        $\hat{\mathbf{z}}_{t_{next}} \leftarrow \sqrt{\bar{\alpha}_{t_{\text{next}}}} \left( \frac{\hat{\mathbf{z}}_t - \sqrt{1 - \bar{\alpha}_t} \cdot \epsilon_\theta(\hat{\mathbf{z}}_t, D, t)}{\sqrt{\bar{\alpha}_t}} \right) + \quad \sqrt{1 - \bar{\alpha}_{t_{\text{next}}}} \cdot \epsilon_\theta (\hat{\mathbf{z}}_t, D, t)$
    end for
    $\hat{z} = \hat{z}_0$
    Take gradient descent step on $\|\hat{z} - z\|_1$
    $\hat{H} = Conv3 \times 3(DHRNet(AM(X_i), \hat{z}))$
    Take gradient descent step on $\mathcal{L}_r$ (paper Eq. (11))
until converged
```

Algorithm 2 LF-Diff Inference

```
Input:
    LDRs images $X_i$, Total diffusion step $T$, implicit sampling step $S$,
    Noise schedule $\beta_t (t \in [1, T])$, Trained LF-Diff
Initialize: $\alpha_t = 1 - \beta_t, \bar{\alpha}_T = \prod_{i=0}^{T} \alpha_i$
Reverse Process:
Sample $\hat{\mathbf{z}}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
$\mathbf{D} =$ LPENet$_{DM}$ $(AM($ PixelUnshuffle $(X_i)))$
for $i = S : 1$ do
    $t = (i - 1) \cdot T/S + 1$
    $t_{\text{next}} = (i - 2) \cdot T/S + 1$ if $i > 1$, else 0
    $\hat{\mathbf{z}}_{t_{next}} \leftarrow \sqrt{\bar{\alpha}_{t_{\text{next}}}} \left( \frac{\hat{\mathbf{z}}_t - \sqrt{1 - \bar{\alpha}_t} \cdot \epsilon_\theta(\hat{\mathbf{z}}_t, D, t)}{\sqrt{\bar{\alpha}_t}} \right) + \quad \sqrt{1 - \bar{\alpha}_{t_{\text{next}}}} \cdot \epsilon_\theta (\hat{\mathbf{z}}_t, D, t)$
end for
$\hat{z} = \hat{z}_0$
$\hat{H} = Conv3 \times 3(DHRNet(AM(X_i), \hat{z}))$
Output reconstructed HDR image $\hat{H}$
```

the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pages 516–517, 2020. 2, 6

[5] Nima Khademi Kalantari and Ravi Ramamoorthi. Deep high dynamic range imaging of dynamic scenes. *ACM Transactions on Graphics*, 36(4):1–12, 2017. 2, 4, 6

[6] Shanshan Lao, Yuan Gong, Shuwei Shi, Sidi Yang, Tianhe Wu, Jiahao Wang, Weihao Xia, and Yujiu Yang. Attentions help cnns see better: Attention-based hybrid image quality assessment network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1140–1149, 2022. 2

[7] Zhen Liu, Wenjie Lin, Xinpeng Li, Qing Rao, Ting Jiang, Mingyan Han, Haoqiang Fan, Jian Sun, and Shuaicheng Liu. Adnet: Attention-guided deformable convolutional network for high dynamic range imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern*

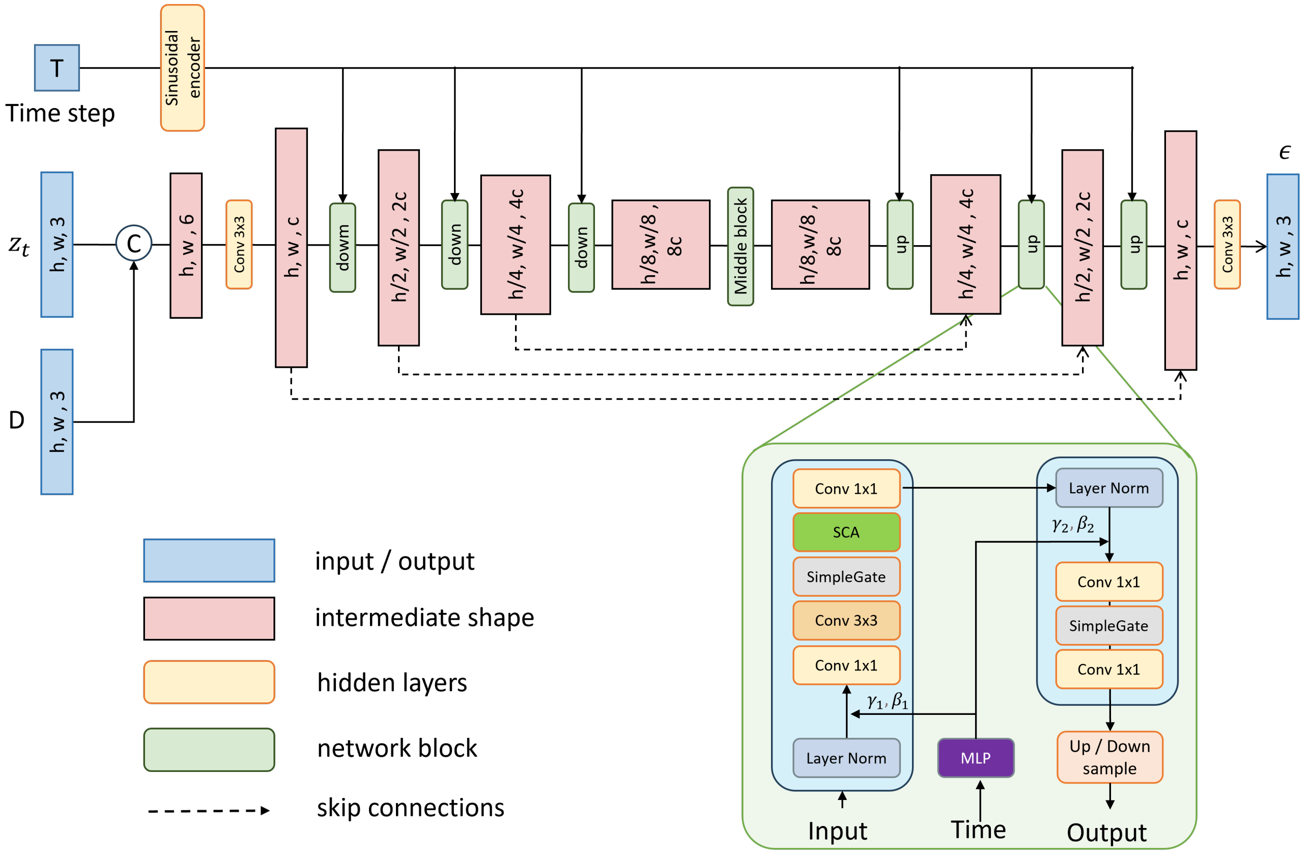

Figure 2. The diagram illustrates the U-Net architecture used for the Denoising Network. The dimensions W, H, and C correspond to the width, height, and number of channels of the features respectively.

Table 1. Quantitative comparison of proposed network with several state-of-the-art methods on Kalantari's [5] datasets.

Models	GT	Hu[3]	Kalantari[5]	AHDR[12]	HDRGAN[10]	ADNet[7]	CA-ViT[8]	DiffHDR[13]	Ours
FID ↓	0	37.27	33.3	9.43	9.32	12.42	5.91	6.20	**5.73**
LPIPS ↓	0	0.0302	0.0341	0.0166	0.0159	0.0169	0.0132	0.0109	**0.0099**
VSI ↑	100	96.38	98.27	99.13	99.3	98.97	99.36	99.48	**99.52**
AHIQ ↑	50	34.07	42.61	46.83	47.2	46.6	46.57	47.82	**47.84**

Recognition, pages 463–470, 2021. 4

[8] Zhen Liu, Yinglong Wang, Bing Zeng, and Shuaicheng Liu. Ghost-free high dynamic range imaging with context-aware transformer. pages 344–360, 2022. 4

[9] Ziwei Luo, Fredrik K Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas B Schön. Refusion: Enabling large-size realistic image restoration with latent-space diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1680–1691, 2023. 2

[10] Yuzhen Niu, Jianbin Wu, Wenxi Liu, Wenzhong Guo, and Rynson WH Lau. Hdr-gan: Hdr image reconstruction from multi-exposed ldr images with large motions. *IEEE Transactions on Image Processing*, 30:3885–3896, 2021. 4

[11] Okan Tarhan Tursun, Ahmet Oğuz Akyüz, Aykut Erdem, and Erkut Erdem. An objective deghosting quality metric for hdr images. In *Computer Graphics Forum*, pages 139–152. Wiley Online Library, 2016. 7, 8

[12] Qingsen Yan, Dong Gong, Qinfeng Shi, Anton van den Hengel, Chunhua Shen, Ian Reid, and Yanning Zhang. Attention-guided network for ghost-free high dynamic range imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1751–1760, 2019. 2, 4

[13] Qingsen Yan, Tao Hu, Yuan Sun, Hao Tang, Yu Zhu, Wei Dong, Luc Van Gool, and Yanning Zhang. Towards high-quality hdr deghosting with conditional diffusion models. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–1, 2023. 4

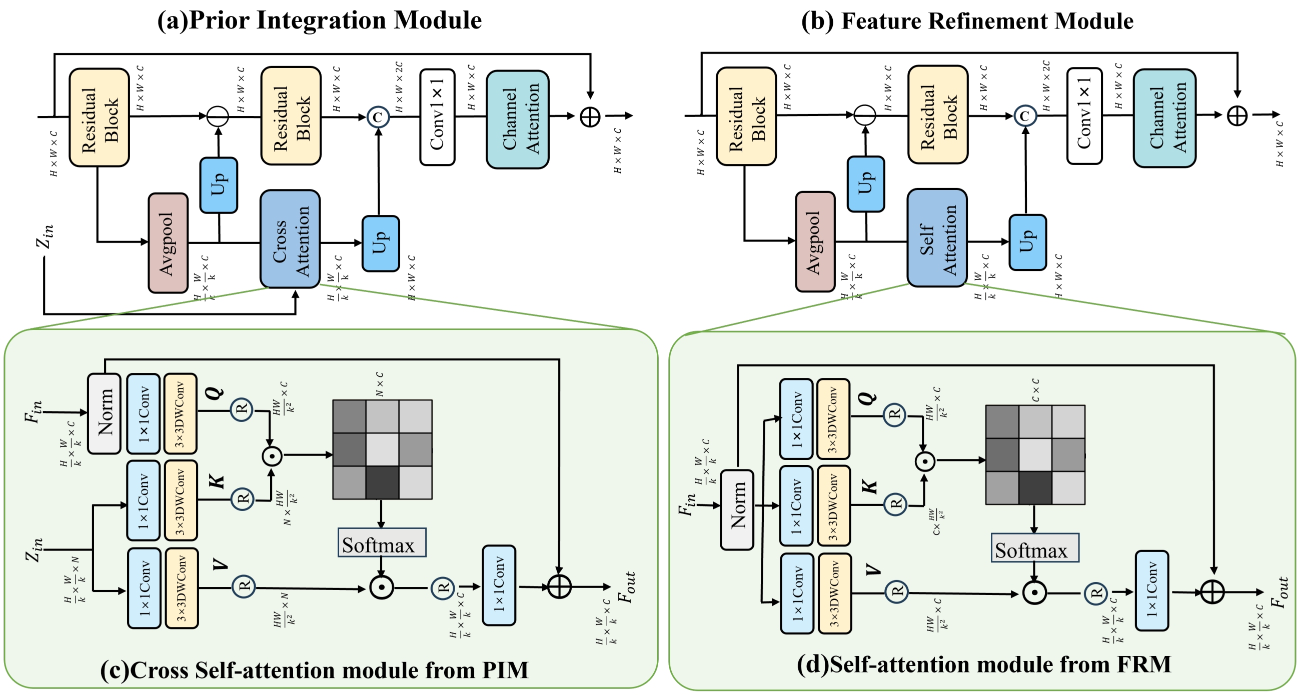
(a)Prior Integration Module
(b) Feature Refinement Module
Residual Block
Residual Block
Avgpool
Up
Cross Attention
Self Attention
Up
Conv1×1
Channel Attention
Z_{in}
$H \times W \times C$
$H \times W \times 2C$
$\frac{H}{k} \times \frac{W}{k} \times C$
Norm
1×1Conv
3×3DWConv
Q
K
V
R
Softmax
F_{in}
Z_{in}
F_{out}
$N \times C$
$C \times C$
$\frac{HW}{k^2} \times C$
$N \times \frac{HW}{k^2}$
$C \times \frac{HW}{k^2}$
$\frac{HW}{k^2} \times N$
$\frac{H}{k} \times \frac{W}{k} \times N$
(c)Cross Self-attention module from PIM
(d)Self-attention module from FRM

Figure 3. DHRNet consists of two modules: a Prior Integration Module (PIM) that fuses the LPR with intermediate features of DHRNet, and a Feature Refinement Module (FRM) that further processes the fused features to HDR image.

[14] Lin Zhang, Ying Shen, and Hongyu Li. Vsi: A visual saliency-induced index for perceptual image quality assessment. *IEEE Transactions on Image processing*, 23(10): 4270–4281, 2014. 2

[15] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 2

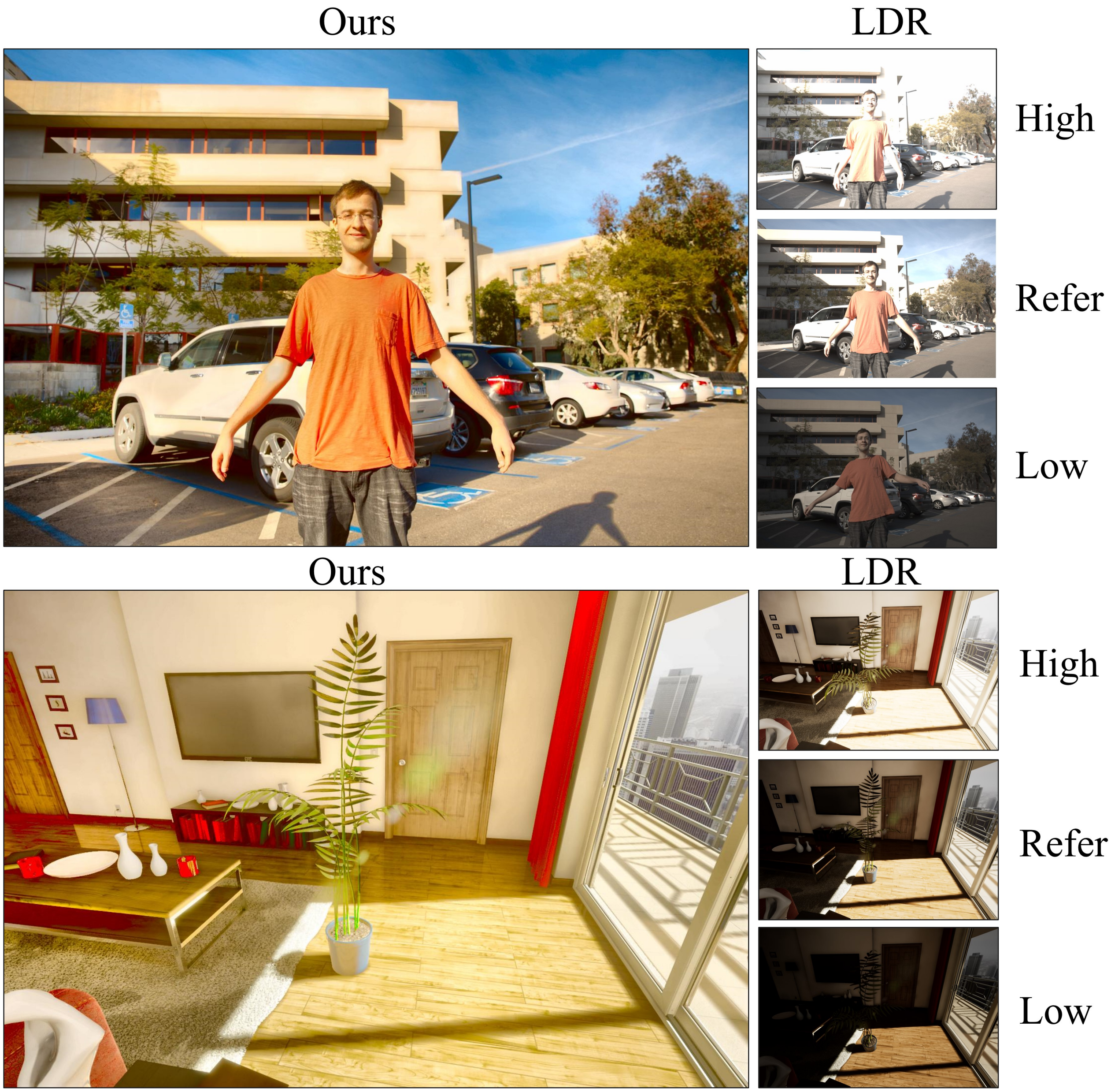

Figure 4. Qualitative results for various motion cases on the Kalantari's dataset [5] and Hu's dataset [4].

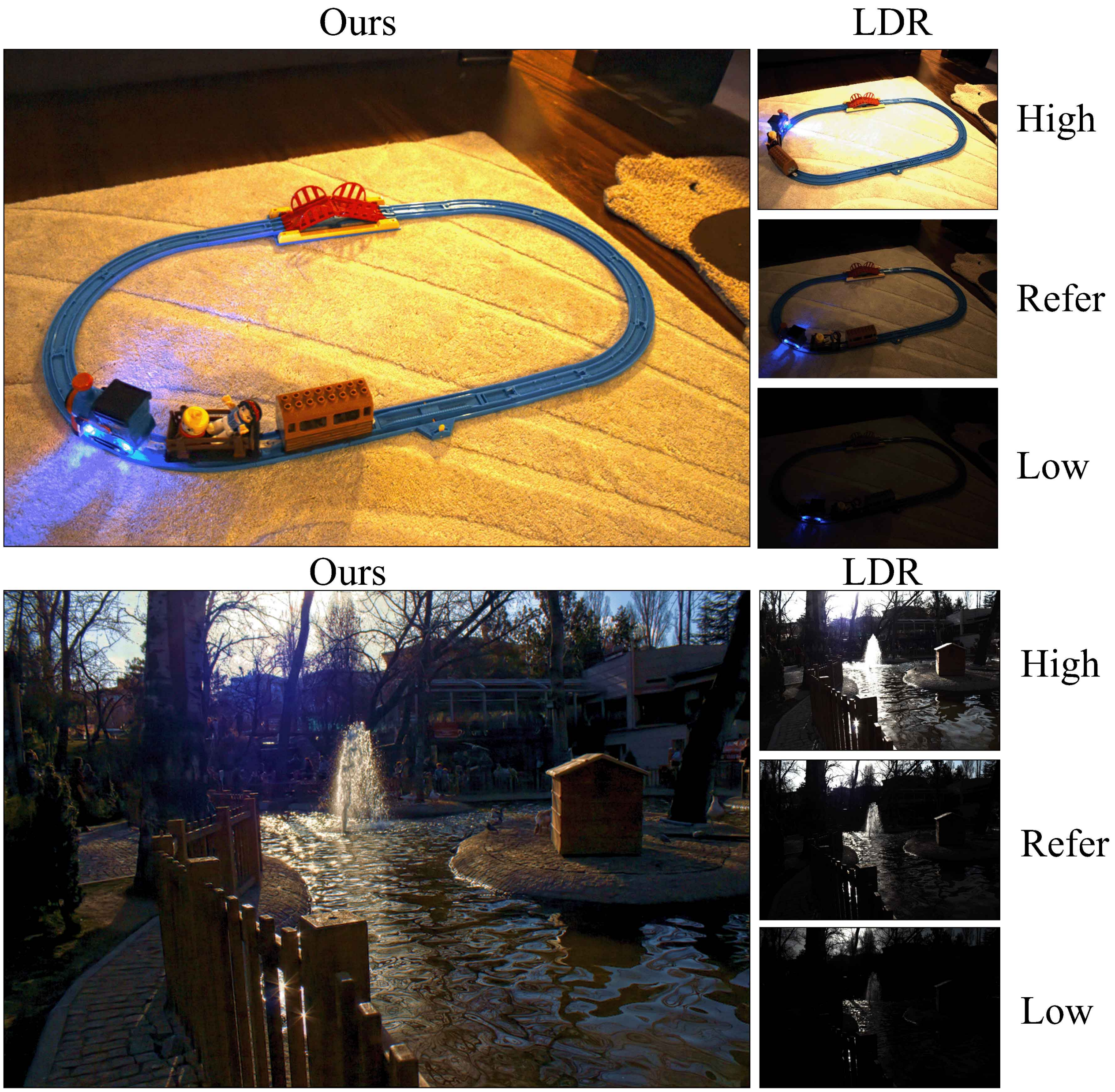

Figure 5. Qualitative results for various motion cases on the Tursun *et al*. [11] dataset.

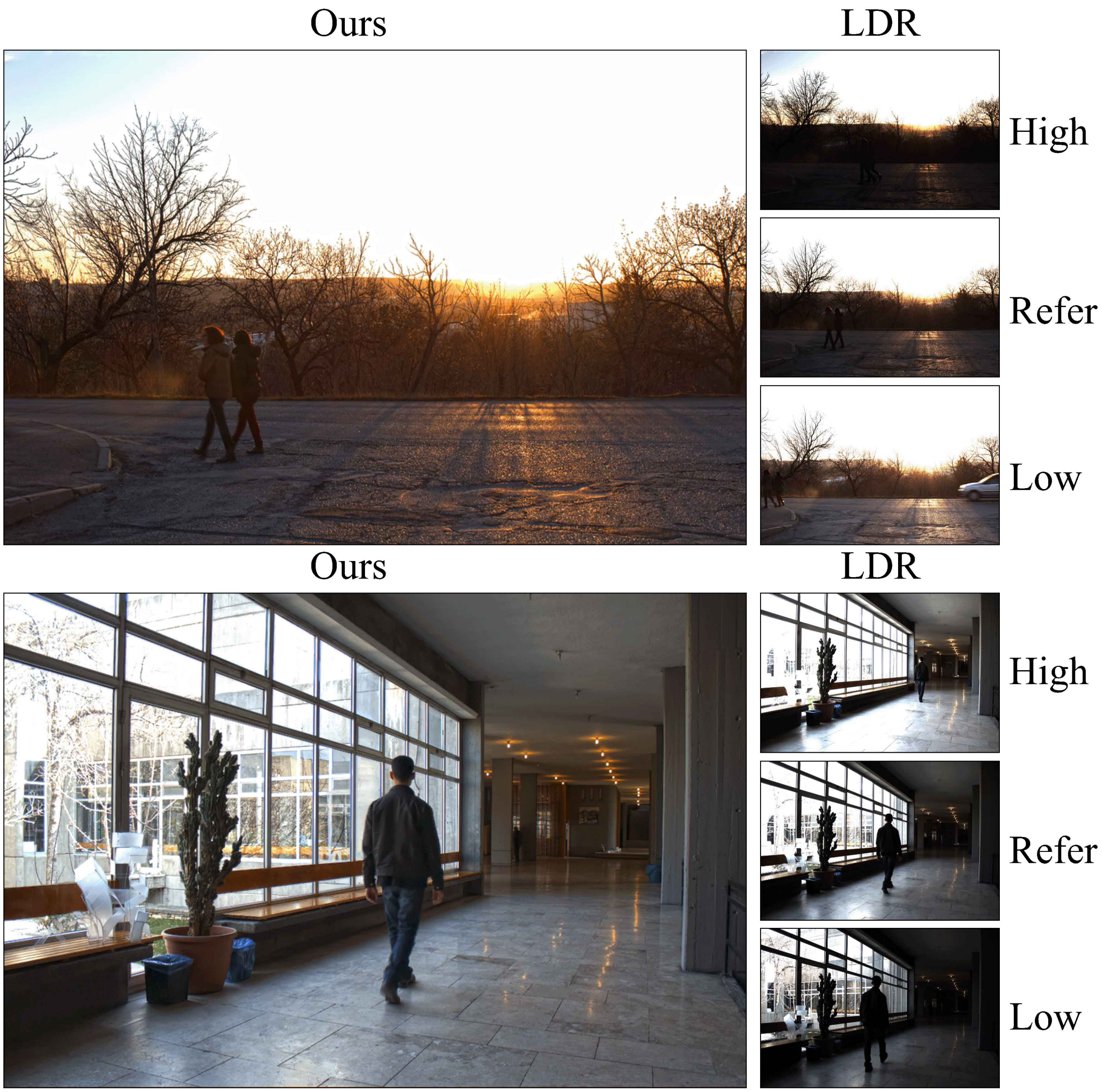

Figure 6. Qualitative results for various motion cases on the Tursun *et al*. [11] dataset.