Learning Event Guided High Dynamic Range Video Reconstruction

Yixin Yang[1,2] Jin Han[3,4] Jinxiu Liang[1,2] Imari Sato[3,4] Boxin Shi[*1,2]

[1] National Key Laboratory for Multimedia Information Processing, School of Computer Science, Peking University

[2] National Engineering Research Center of Visual Technology, School of Computer Science, Peking University

[3] Graduate School of Information Science and Technology, The University of Tokyo [4] National Institute of Informatics

{yangyixin93, cssherryliang, shiboxin}@pku.edu.cn {jinhan, imarik}@nii.ac.jp

6. Additional analysis

6.1. Influence from number of scales in HDRev-Net

We adopt U-Net as the base architecture in the proposed HDRev-Net for event guided HDR video reconstruction, which progressively downsamples and upsamples the feature maps at different scales. Specifically, we use a lightweight model with 3 scales of feature maps in the main submission. We provide comparisons of the proposed network using different scales of upsampling/downsampling operations in Table 3. The best performance can be obtained by the model with 4 scales. However, it has parameters 4 times more than the one with 3 scales. The model with 5 scales seems to overfit to the training data, which performs the worst among different models. For consideration of both effectiveness and efficiency, we adopt the model with 3 scales in the main submission.

Table 3. Quantitative results of using different numbers of scales in the proposed HDRev-Net. ↑ (↓) means higher (lower) is better.

#Scale	#Param	PSNR↑	SSIM↑	LPIPS↓	VDP↑	VQM↓
3	13.43M	24.071	0.928	0.110	**8.108**	**0.103**
4	57.93M	**24.768**	**0.929**	**0.104**	8.054	0.104
5	233.32M	24.426	0.927	0.113	7.884	0.117

6.2. Contributions of each loss

We employ different settings of loss functions during training process to evaluate the contributions of different losses. The quantitative results are shown in Table 4, which validates the effectiveness of each loss. In particular, without the color loss $\mathcal{L}_{\text{color}}$ for enforcing the reconstruction of color appearance, a significant performance drop in terms of all metrics appears. The perceptual quality becomes worse when the perceptual loss $\mathcal{L}_{\text{perc}}$ is removed, as indicated by the value of LPIPS. However, due to the distortion-perception tradeoff, introducing $\mathcal{L}_{\text{perc}}$ results in a slight drop in the value of PSNR.

Table 4. Quantitative results of using different loss functions for training, with the results of "LDR-first" training. ↑ (↓) means higher (lower) is better.

Setting	PSNR↑	SSIM↑	LPIPS↓	VDP↑	VQM↓
W/o $\mathcal{L}_{\text{color}}$	19.064	0.869	0.250	5.337	0.202
W/o $\mathcal{L}_{\text{perc}}$	**24.219**	0.907	0.170	7.417	0.160
W/o $\mathcal{L}_{\text{mse}}$	23.540	0.923	0.119	8.020	0.113
LDR-first	23.002	0.918	0.130	7.942	0.426
Ours	24.071	**0.928**	**0.110**	**8.108**	**0.103**

6.3. Analysis to the training strategy

In the main submission, we discuss the contribution of the proposed pretraining strategy (denoted as "Complete"). In this section, we compare the "Complete" model with two variants. One is joint training (denoted as "Joint"), which trains $\mathcal{F}_{\text{L}}$, $\mathcal{F}_{\text{E}}$, $\mathcal{F}_{\text{fusion}}$, and $\mathcal{F}_{\text{H}}$ jointly without pretraining. The other one is "LDR-first", where the LDR-to-HDR encoder $\mathcal{F}_{\text{L}}$ is pretrained at first, and the event-to-HDR encoder $\mathcal{F}_{\text{E}}$ is trained subsequently. The quantitative result of "LDR-first" is shown in Table 4. The qualitative results are shown in Fig. 10. The "Complete" model shows cleaner intermediate result H_E and a more natural appearance in the final result H, which demonstrates the effectiveness of our pretraining strategy.

6.4. Analysis to exposure ratio for two-exposure-based methods

We set different exposure ratios for two-exposure-based methods, Debevec *et al.* [8] and Li *et al.* [40], and compare them with our event guided approach in detail. The optimal exposure ratio is often scene-dependent, which cannot handle rapid changes in a scene when capturing HDR videos. The results are shown in Fig. 11. With a small exposure ratio, two exposure-based methods preserve fine details at

* Corresponding author

Project page: https://yixinyang-00.github.io/HDRev/

the cost of narrowing dynamic range covered (*e.g.*, the contour of the sun). With a large exposure ratio, high dynamic range of the reconstructed frames can only be achieved by sacrificing some fine details (*e.g.*, the color distortion in the sky). In contrast, the proposed method can achieve better performance by introducing event streams without considering the exposure ratio.

7. Efficiency Comparison

Efficiency comparisons between deep-learning-based methods, the proposed one, and the variants of ours are shown in Table 5. Debevec *et al.* [8] and Li *et al.* [40] are omitted since they are not learning-based methods. We calculate the parameters (#Param), the floating-point operations per second (FLOPs), and the average running time per seconds (Time) for all those methods on our test dataset described in Section 3.3. We test ten videos at first to warm them up and then calculated the total run time for each video to get the average run time per frame. As shown in Table 5, the proposed method achieves the fastest inference speed at 36 FPS compared to existing methods implemented with PyTorch. E2VID [57] reconstructs separately for 5-channel (RGBW + grayscale), and then merges them into a color image, which takes 5 times longer inference time. Liu *et al.* [42] based on TensorFlow is faster due to the static graph.

Table 5. Efficiency comparisons

Methods	#Param	FLOPs	Time	Framework
eSL-Net [68]	0.188M	147.470G	84.5ms	PyTorch
E2VID [57]	10.712M	41.392G	47.0ms	PyTorch
Liu *et al.* [42]	27.688M	164.066G	0.5ms	TensorFlow
Han *et al.* [23]	53.512M	106.776G	28.8ms	PyTorch
Ours	13.427M	119.283G	27.6ms	PyTorch
W/o Fusion	8.781M	78.770G	22.1ms	PyTorch
Joint training	13.427M	119.283G	27.7ms	PyTorch
W/o LSTM	9.452M	62.644G	22.9ms	PyTorch

8. Hybrid camera system

To capture LDR videos and events simultaneously in real-world scenarios, we build a hybrid-camera system. As shown in Fig. 9, we use a beam-splitter to divide the incident light equivalently into two cameras. Please refer to Section 4.2 for more details.

9. More visual quality comparisons

We provide more visual comparisons of HDR reconstruction results in Fig. 12 and Fig. 13. As shown in the results, the proposed method can produce HDR frames with higher visual quality than the comparing methods, especially in over-exposed and under-exposed regions of the LDR frames. For synthetic data, the Q-scores computed from VDP metrics are labeled in each image (except for Li *et al.* [40]), which demonstrate higher quantitative evaluation results of the proposed method.

Please refer to our supplementary video for HDR video reconstruction results of different methods.

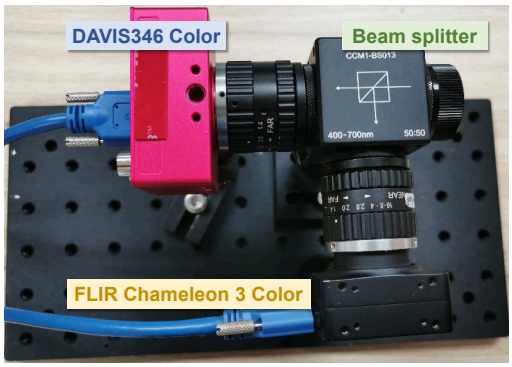

Figure 9. Hybrid camera system for capturing real data.

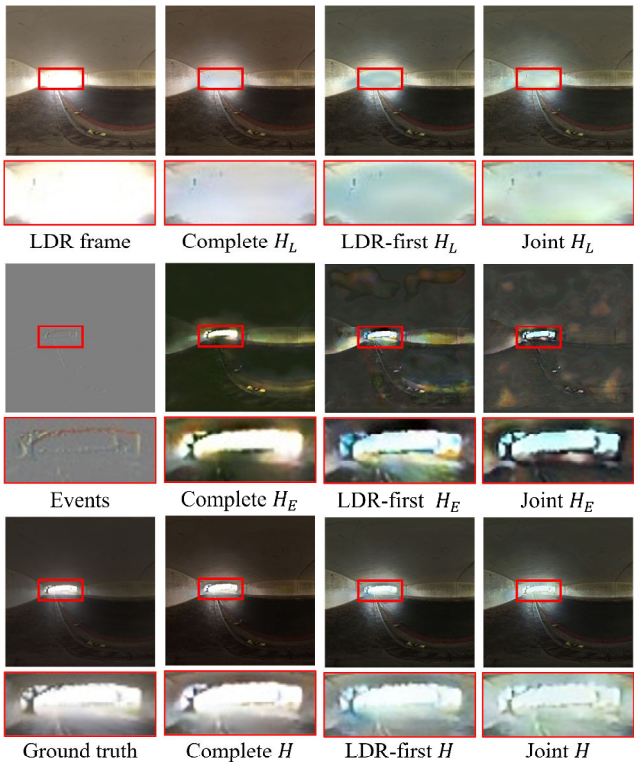

Figure 10. Comparisons between "Joint" training, "LDR-first" training, and "Complete" model. H_L (H_E) denotes the network output with encoder F_E (F_L) disabled, whose input are set to zero.

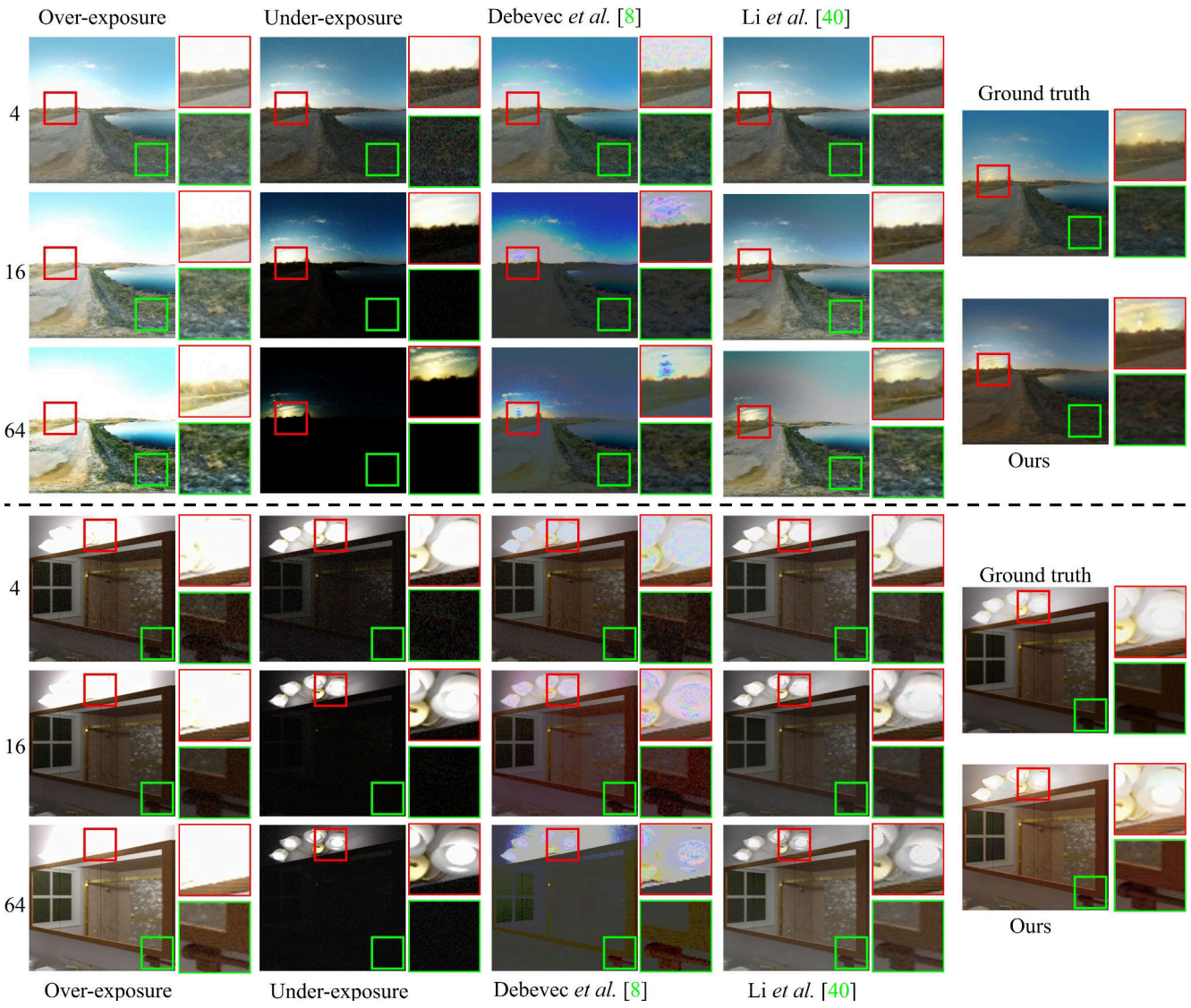

Figure 11. Comparison of two-exposure-based methods with ours. The numbers on the left denote the exposure ratio between "Over-exposure" and "Under-exposure" images, which is used as the input of the two-exposure-based methods. To demonstrate the generalization ability of the proposed event guided method, which is free of scene-dependent exposure ratio balancing, we show two scenes: one with a rather high dynamic range (above) and the other one with a relatively lower dynamic range (bottom).

Figure 12. Visual quality comparisons on synthetic data. The Q-scores (higher the better) computed from VDP metrics are labeled in each image. The results of our method have finer details and higher Q-scores than all the others.

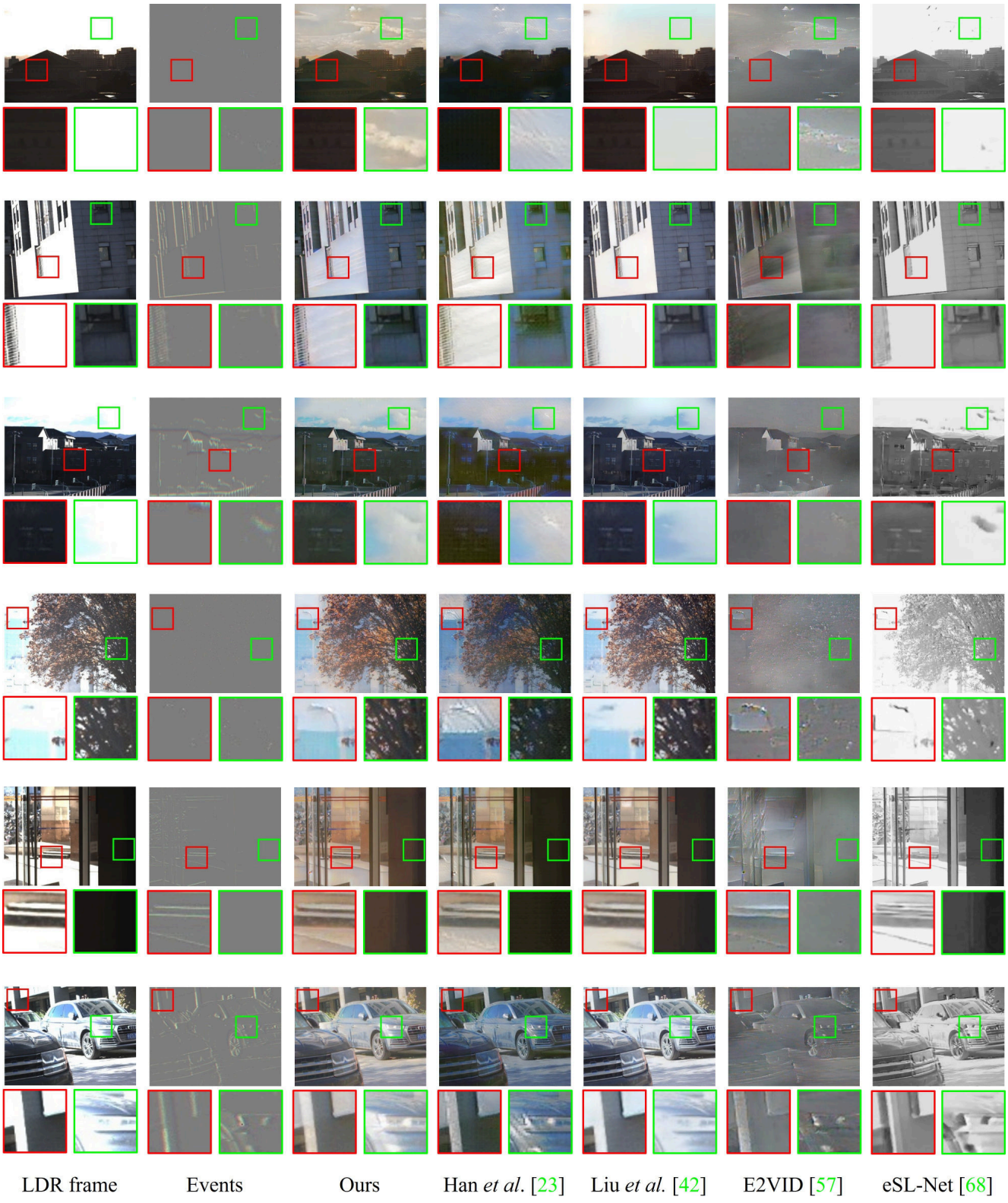

Figure 13. Visual quality comparisons on real data. By leveraging the information from the event stream, the proposed method can reconstruct HDR images with better visual quality than others.