Learning Event Guided High Dynamic Range Video Reconstruction

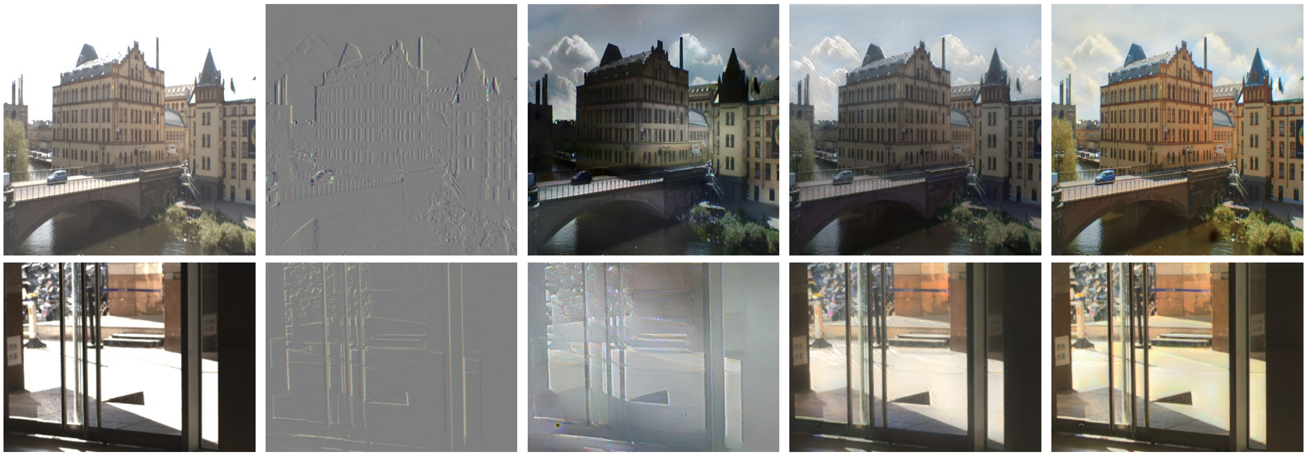
Yixin Yang[1,2] Jin Han[3,4] Jinxiu Liang[1,2] Imari Sato[3,4] Boxin Shi*[1,2]

[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



(a) LDR images (b) Events (c) E2VID [60] (d) Han *et al.* [24] (e) Ours

Figure 1. Given hybrid inputs of (a) LDR video and (b) stacked events, the HDR **video** can be reconstructed using different methods shown in (c) E2VID [60], (d) Han *et al.* [24], and (e) the proposed HDRev-Net. The samples here are tested on synthetic data (top row) and real data (bottom row) respectively. The proposed method is able to generate the HDR video with more details and less flickering effects. Please view our project page for corresponding animations.

Abstract

Limited by the trade-off between frame rate and exposure time when capturing moving scenes with conventional cameras, frame based HDR video reconstruction suffers from scene-dependent exposure ratio balancing and ghosting artifacts. Event cameras provide an alternative visual representation with a much higher dynamic range and temporal resolution free from the above issues, which could be an effective guidance for HDR imaging from LDR videos. In this paper, we propose a multimodal learning framework for event guided HDR video reconstruction. In order to better leverage the knowledge of the same scene from the two modalities of visual signals, a multimodal representation alignment strategy to learn a shared latent space and a fusion module tailored to complementing two types of signals for different dynamic ranges in different regions are proposed. Temporal correlations are utilized recurrently to suppress the flickering effects in the reconstructed HDR video. The proposed HDRev-Net demonstrates state-of-the-art performance quantitatively and qualitatively for both synthetic and real-world data.

1. Introduction

The dynamic range of the real world usually exceeds what a conventional camera and 8-bit image can record by a large margin. High dynamic range (HDR) imaging, which expands the luminance range limited by low dynamic range (LDR) images or videos, is a broadly used technique with extensive applications in photography/videography, video games, and high-end display.

Most HDR imaging methods for conventional cameras rely on capturing and merging multiple snapshots with different exposure times [9, 49], which is challenging for capturing videos. There have been enduring efforts for sophisticated modification on conventional frame based cameras to capture multi-exposure sequences (nearly) simultaneously, *e.g.*, beam splitting with three or more sensors [69, 70], temporally [7, 30, 32] or spatially [1, 8, 22, 28, 53, 54] varying exposure. Nevertheless, their abilities for HDR video reconstruction are limited by the trade-off between a higher frame rate (for a smooth viewing experience) and a higher dynamic range (for capturing details in dark regions with

* Corresponding author

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

prolonged exposure time). Moreover, the optimal exposure ratio between LDR frame sequences with different exposure settings is scene-dependent and temporally-varying, whose balancing is difficult for diverse scenes captured in videos. Even worse, moving objects or camera shaking during video capture can lead to ghosting effects in frames generated by long exposure shots. An HDR video could also be hallucinated from LDR inputs in a frame-by-frame manner by leveraging prior knowledge of tone-mapping operators [61] or data modeling powers of deep learning [11]. However, due to the highly ill-posed nature of the hallucination process, it inevitably leads to severe flickering effects.

In recent years, the event camera [16] has drawn increasing attention of researchers, due to its advantages over conventional frame based ones in sensing fast motions and extended dynamic ranges (*e.g.*, 120dB for DAVIS346). Unlike using multi-exposure frames, events recorded along with an LDR video encode HDR irradiance changes without sacrificing the frame rate/exposure time of the LDR video, which avoids ghosting artifacts as well, and is very promising as guidance for HDR video reconstruction.

However, integrating events with LDR video for HDR video reconstruction is challenging due to inconsistency between events and frames in three aspects: 1) *Modality misalignment*: Frames and events are completely *different representations of visual information*, and “fusing” them by first translating events into intensity values [60, 80] like [24] often includes artifacts from solving the ill-posed event integration problem. 2) *Dynamic range gap*: Performing image/video reconstruction under the guidance of events [75], *i.e.*, doubly integrating events as intensity changes within the exposure time [56, 57], ignores the dynamic range clipping in the capturing process of LDR frames, which leads to *uncertainties in under/over-exposed regions*. 3) *Texture mismatching*: Regions with smooth textures and slow motion hardly produce effective event observations, which results in inconsistent textures among consecutive event stacks and *flickering effects* in the reconstructed videos.

We propose **HDRev-Net**, a multimodal learning framework for event guided HDR video reconstruction to tackle the challenges by the following strategies: 1) To achieve *multimodal representation alignment* for the two modalities of the same scene, we propose a learning strategy to progressively project them onto *a shared representation space*. 2) To *reliably complement* information from the two modalities in over/under-exposed regions, the representations produced by the two modality-specific encoders are fused for an expressive joint representation using a *confidence guided multimodal fusion module*. 3) To effectively *suppress the flickering effects*, we utilize the temporal redundant information between consecutive frames and events via the proposed *recurrent convolutional encoders*.

As shown in Fig. 1, HDRev-Net can successfully fuse LDR frames and events to obtain HDR frames with more details and less flickering effects. It demonstrates state-of-the-art HDR video reconstruction performance on both synthetic and real data by making the following contributions:

- We design a multimodal alignment strategy to bridge the gap between events and frames by aligning their representation in a shared latent space.
- We develop a confidence guided fusion module to complement HDR information from events and finer details from well-exposed regions in LDR frames.
- We utilize the temporal correlation from consecutive events and LDR frames in a recurrent fashion to alleviate the flickering effects for recovered HDR videos.

2. Related Work

Frame based HDR reconstruction. A typical technique for HDR image reconstruction is to merge a stack of bracketed exposure LDR images into an HDR one [9, 49]. Such an idea is also feasible for HDR video reconstruction. Many works [38, 48, 69, 70] proposed to use multiple sensors to capture LDR images with different exposures simultaneously. Besides, there are methods designing spatially varying sensors [1, 8, 22, 28, 53, 54, 71] or using coded exposure by alternating the shutter speed [7, 29, 30, 32, 43]. However, HDR information is obtained at the cost of lowering spatial or temporal resolution in these methods. Moreover, for dynamic scenes, camera shake and moving objects during exposure can lead to ghosting artifacts. To tackle this, recent works [45, 68, 77] introduced attention mechanisms to perform implicit alignment between frames. However, they still heavily rely on fine-grained similarities across frames, which is hard to be captured in varying exposures.

To avoid ghosting artifacts, Banterle *et al.* [2] proposed inverse tone mapping to reconstruct the HDR image from a single LDR image, which attempted to hallucinate HDR information in over/under-exposure regions. More recently, many works [6, 11, 12, 19, 33, 39–41, 44, 47] attempted to better perform inverse tone mapping via convolutional neural networks (CNNs). However, the problem of hallucinating the lost information in over/under-exposed regions without any physical guidance is too ill-posed to be reliably solved in varying real-world scenarios by these methods.

Event based HDR reconstruction. Event cameras inherently have a higher dynamic range than conventional ones, which enable direct reconstruction of HDR frames from pure events. Early attempt [35] for intensity reconstruction from events was based on strict assumptions about the scene or motion, *e.g.*, brightness constancy and global rotational movement. Other methods [3, 4] optimized a variational model to reconstruct intensity images.With the help of deep learning, Mostafavi *et al.* [51, 73] and Wang *et al.* [74] used generative adversarial networks (GANs) [21] to reconstruct images from events. Rebecq *et al.* [59, 60] and

Zou *et al.* [80] used recurrent convolutional blocks [10, 25] to maintain temporal consistency inside long event streams for video reconstruction. However, these methods suffer from flickering effects and cannot provide stable colored video reconstruction because stream-like events only record the quantized logarithmic irradiance changes of each pixel. It is highly ill-posed to directly reconstruct frames from events without any intensity information as references.

Event guided HDR reconstruction. Conventional cameras record the irradiance within a limited dynamic range synchronously, while event cameras record the logarithmic irradiance change with a much higher dynamic range asynchronously. Event guided HDR reconstruction combines the best of the two worlds by considering hybrid input of events and frames for reconstructing HDR frames. Wang *et al.* [72] proposed a sparse learning method along with the integration method [56] to enhance the image quality. Han *et al.* [24] proposed to reconstruct HDR intensity maps from events first and then fused them with LDR frame in the intensity domain. It is also proposed to guide the fusion of bracketed exposure with events [50, 64] for plausible results. However, blurry artifacts are also introduced by the longer exposure of bracketed images. All these works focused on the reconstruction of HDR images.

For event guided video reconstruction, Wang *et al.* [75] integrated events to interpolate frames based on [56, 57] and handled the noise and artifacts under simplified assumption by a variant of Kalman filter. However, the integration model [56, 57] ignored the dynamic range clipping of LDR frames, which leads to uncertainties in under/over-exposed regions. In contrast, we propose to integrate the two modalities in a learned shared representation space by leveraging the data modeling power of deep learning.

3. Proposed Method

3.1. Formulation

Event generation and stacking. An event camera triggers an event $e = (t, p, \sigma)$ at pixel position $p = (p_x, p_y)^\top$ and time t with polarity $\sigma \in \{-1, +1\}$, when the logarithmic change of irradiance R since the last event at the pixel p and time $t - \delta t$ exceeds a dispatched threshold θ:

$$\|\log R_t^{(p)} - \log R_{t-\delta t}^{(p)}\| \geq \theta. \tag{1}$$

To process stream-like events using CNNs, we need to convert a stream of events into a fixed size tensor-like representation. We use voxel grid [79] that encodes temporal information in a C-channel 3D-volume. By discretizing duration $\Delta t = t_{K-1} - t_0$ spanned by K events into C temporal bins, each event $e_k = (t_k, p_k, \sigma_k)$ distributes its polarity σ_k to the two closest voxels as follows [60]:

$$E_t^{(p)} = \sum_{p_k = p} \sigma_k \max\left(0, 1 - |t - \tilde{t}_k|\right), \tag{2}$$

where $\tilde{t}_k := \frac{C-1}{\Delta t}(t_k - t_0)$ is the normalized timestamp.

LDR image formation. Given the irradiance R and exposure time Δt, an HDR image can be represented as $I_{\text{HDR}} = R \cdot \Delta t$. There are three steps to produce an LDR image: dynamic range clipping, tone mapping with a camera response function (CRF) f_{CRF}, and quantization:

$$I_{\text{LDR}} = \lfloor 255 \cdot f_{\text{CRF}}\left(\max(\min(I_{\text{HDR}}, 1), 0)\right)\rfloor / 255. \tag{3}$$

To align the LDR image and events in the linear irradiance domain, we first convert the LDR image by:

$$L = f_{\text{CRF}}^{-1}(I_{\text{LDR}}) / \Delta t, \tag{4}$$

where f_{CRF}^{-1} is the inverse CRF. To ease the learning difficulty in regression of values in linear domain spanning a wide range, the range of I_{HDR} are compressed by [31]:

$$H = \frac{\log(1 + \mu I_{\text{HDR}})}{\log(1 + \mu)}, \tag{5}$$

where μ controls the amount of compression.

3.2. HDRev-Net

Given the stacked event sequence $\mathbb{E} = \{E_t\}_{t=1}^T$ and the corresponding LDR frame sequence $\mathbb{L} = \{L_t\}_{t=1}^T$ of length T, our goal is to reconstruct the HDR frame sequence $\mathbb{H} = \{H_t\}_{t=1}^T$ through a multimodal learning network $\mathcal{F}$ dubbed HDRev-Net. As illustrated in Fig. 2, $\mathcal{F}$ consists of two modality-specific encoders $\mathcal{F}_{\text{L}}$ and $\mathcal{F}_{\text{E}}$ for both temporal context encoding and multimodal representation alignment, confidence guided multimodal fusion module $\mathcal{F}_{\text{fusion}}$, and decoder $\mathcal{F}_{\text{H}}$ for HDR video reconstruction from latent representation. Our pipeline can be described as:

$$\mathbb{H} = \mathcal{F}_{\text{H}}(\mathcal{F}_{\text{fusion}}(\mathcal{F}_{\text{L}}(\mathbb{L}), \mathcal{F}_{\text{E}}(\mathbb{E}))). \tag{6}$$

Multimodal representation alignment. Events and LDR frames are two quite different modalities for representing visual information, whose fusion is non-trivial due to distinct low-level statistical properties and highly non-linear relationships between them. Non-learning approaches based on the explicit fusion model [56, 57] may lead to uncertainties in over-exposed regions, while using CNNs to predict HDR frames from the two modalities by direct concatenation or performing event-to-frame translation firstly is ill-posed, which does not consider the differences and complementarities between the two modalities.

In this paper, we propose to perform implicit fusion by an encoder-decoder network: learning to represent events and LDR frames in a shared latent space by two modality-specific encoders and then integrating them by a fusion module in the bottleneck layer. A straightforward solution is training the two encoders simultaneously for jointly decoding the latent representation into HDR frames. However, due to the modality gap between frames and events,

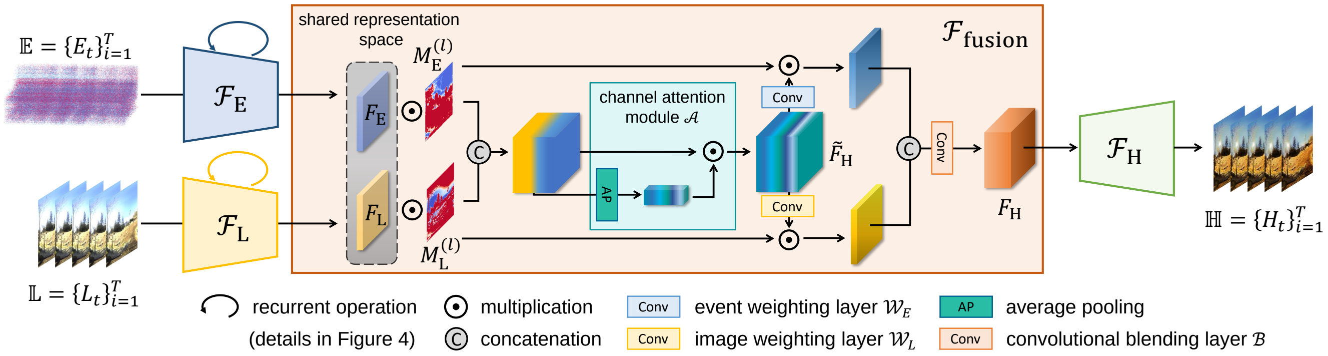

$\mathbb{E} = \{E_t\}_{i=1}^T$
$\mathcal{F}_\mathrm{E}$
$\mathbb{L} = \{L_t\}_{i=1}^T$
$\mathcal{F}_\mathrm{L}$
shared representation space
F_E
F_L
$M_\mathrm{E}^{(l)}$
$M_\mathrm{L}^{(l)}$
channel attention module $\mathcal{A}$
AP
Conv
$\tilde{F}_\mathrm{H}$
$\mathcal{F}_\mathrm{fusion}$
F_H
$\mathcal{F}_\mathrm{H}$
$\mathbb{H} = \{H_t\}_{i=1}^T$
recurrent operation (details in Figure 4)
⊙ multiplication
Ⓒ concatenation
Conv event weighting layer $\mathcal{W}_E$
Conv image weighting layer $\mathcal{W}_L$
AP average pooling
Conv convolutional blending layer $\mathcal{B}$

Figure 2. Overview of the proposed HDRev-Net. For multimodal representation alignment, the hybrid inputs of stacked events $\mathbb{E}$ and LDR frames $\mathbb{L}$ are mapped onto a shared representation space by the modality-specific encoders $\mathcal{F}_\mathrm{E}, \mathcal{F}_\mathrm{L}$, in which recurrent connections are applied for temporal context encoding. Under the guidance of confidence masks M_E and M_L extracted from $\mathbb{E}$ and $\mathbb{L}$, respectively, their latent representations $F_\mathrm{E}, F_\mathrm{L}$ are then integrated into a joint one F_H by the multimodal fusion module $\mathcal{F}_\mathrm{fusion}$. The HDR decoder $\mathcal{F}_\mathrm{H}$ decodes F_H into HDR frame sequence $\mathbb{H}$ eventually.

jointly performing intra-modality (LDR to HDR) and inter-modality (event to HDR) reconstruction can lead to a trivial solution: predicting HDR frames from the LDR frames mostly, leaving almost all the stacked events ignored in the training process and heavily relying on LDR frames.

To effectively utilize the properties of high dynamic range and high frame rate of events, we should find a shared representation space that can well express common and complementary information from events and LDR frames for the subsequent multimodal fusion. To achieve this, we propose a multi-stage approach: performing inter-modality reconstruction from events and then intra-modality reconstruction from LDR frames at first, then jointly training the two encoders and HDR decoder together.

Specifically, the event encoder $\mathcal{F}_\mathrm{E}$ is pretrained to map the input events to a latent representation, which is then decoded to HDR frames by an HDR decoder $\mathcal{F}_\mathrm{H}$:

$$F_\mathrm{E} = \mathcal{F}_\mathrm{E}(\mathbb{E}), \quad \mathbb{H} = \mathcal{F}_\mathrm{H}(F_\mathrm{E}). \tag{7}$$

In order to project the input LDR frames onto the same latent space for representation alignment, an LDR encoder $\mathcal{F}_\mathrm{L}$ is then trained with the parameters of $\mathcal{F}_\mathrm{H}$ fixed:

$$F_\mathrm{L} = \mathcal{F}_\mathrm{L}(\mathbb{L}), \quad \mathbb{H} = \mathcal{F}_\mathrm{H}(F_\mathrm{L}). \tag{8}$$

Thanks to the shared decoder $\mathcal{F}_\mathrm{H}$, the representation of events and LDR frames are aligned in a shared latent space and can then be effectively fused by the proposed event guided HDR video reconstruction pipeline Eq. (6).

Confidence guided multimodal fusion. As shown in Fig. 2, both manually designed confidence masks and learnable channel-wise attention are used for weighting the intermediate representation $F_\mathrm{L}, F_\mathrm{E}$ along two separate dimensions, spatial and channel. The manually designed confidence masks emphasize more reliable regions and robustness to inherent noise in data. Intuitively, events are triggered on high contrast edges. The confidence mask for events makes the fusion module pay more attention to such regions (an example is shown in Fig. 3 (e)). The confidence mask M_E is defined as:

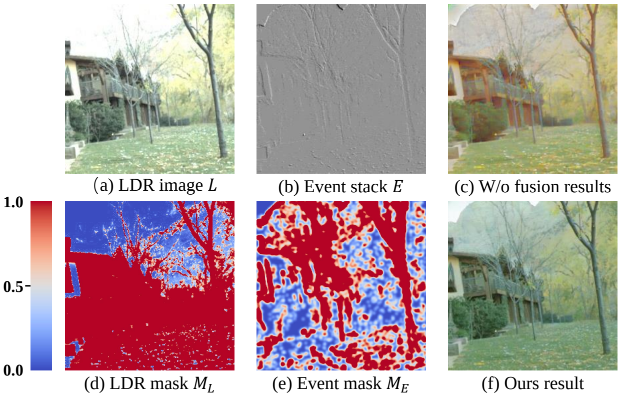

(a) LDR image L
(b) Event stack E
(c) W/o fusion results
(d) LDR mask M_L
(e) Event mask M_E
(f) Ours result

Figure 3. Visualization of the confidence masks (d) M_L and (e) M_E extracted from (a) LDR frame L and events (b) E, respectively. The HDR reconstruction results are (c) "W/o fusion" and (f) ours.

$$M_\mathrm{E} = \min\left(G * |E|, 1\right), \tag{9}$$

where $*$ denotes the convolution operator; G denotes a Gaussian kernel with variance 3 and support 21×21, which is used to propagate the influence of the sparsely distributed events to their neighboring regions.

For LDR frames, irradiance in well-exposed regions is more reliable. The confidence mask M_L for LDR frames (an example is shown in Fig. 3 (d)) is defined as:

$$M_\mathrm{L} = \frac{0.5 - \max(|L - 0.5|, \lambda - 0.5)}{1 - \lambda}, \tag{10}$$

where λ is a hyper-parameter (set as 0.8 in our experiments) that defines the range of values in which reliable pixels lie.

To guide the fusion of latent representation with a smaller spatial size, we propagate the values of the mask according to the receptive field of each pixel. Specifically,

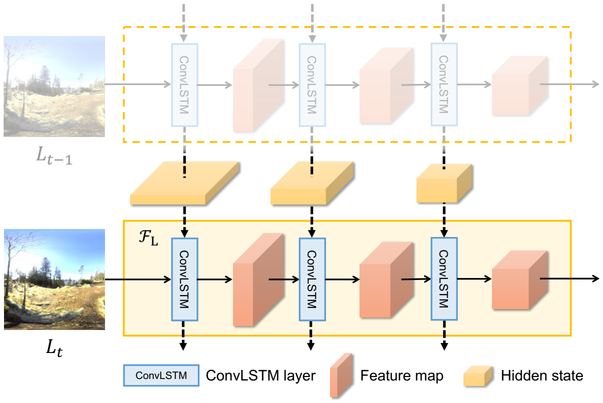


Figure 4. Illustration of the recurrent connections in encoder $\mathcal{F}_{\mathrm{L}}$ (the same structure is used for $\mathcal{F}_{\mathrm{E}}$).

the confidence mask $M^{(l)}$ for reweighing the chosen bottleneck feature $F_{\mathrm{E}}, F_{\mathrm{L}}$ outputted by the l-th layer is defined by the following recursive rule:

$$M^{(0)} = M, \quad M^{(l)} = G_{K_l} * M^{(l-1)}, \tag{11}$$

where G_{K_l} denotes a Gaussian kernel with variance 3 and support $K_l \times K_l$, whose size is the same as the l-th convolutional layer. Given the confidence masks $M_{\mathrm{L}}^{(l)}, M_{\mathrm{E}}^{(l)}$, the spatially weighted features are obtained by:

$$\tilde{F}_{\mathrm{L}} = F_{\mathrm{L}} \cdot M_{\mathrm{L}}^{(l)}, \quad \tilde{F}_{\mathrm{E}} = F_{\mathrm{E}} \cdot M_{\mathrm{E}}^{(l)}. \tag{12}$$

To adaptively exchange information between $\tilde{F}_{\mathrm{L}}$ and $\tilde{F}_{\mathrm{E}}$, they are concatenated and reweighted by a channel attention module [27] $\mathcal{A}$ to obtain an integrated one $\tilde{F}_{\mathrm{H}}$:

$$\tilde{F}_{\mathrm{H}} = \mathcal{A}([\tilde{F}_{\mathrm{L}}, \tilde{F}_{\mathrm{E}}]). \tag{13}$$

To better balance the contributions of different modalities, $\tilde{F}_{\mathrm{E}}, \tilde{F}_{\mathrm{L}}$ are reweighted along all dimensions with the weighting maps extracted from $\tilde{F}_{\mathrm{H}}$ by weighting layers $\mathcal{W}_{\mathrm{L}}$ and $\mathcal{W}_{\mathrm{E}}$. Then they are concatenated and fed into a convolutional blending layer $\mathcal{B}$ to obtain the joint representation for HDR reconstruction:

$$F_{\mathrm{H}} = \mathcal{B}([\tilde{F}_{\mathrm{L}} \odot \mathcal{W}_{\mathrm{L}}(\tilde{F}_{\mathrm{H}}), \tilde{F}_{\mathrm{E}} \odot \mathcal{W}_{\mathrm{E}}(\tilde{F}_{\mathrm{H}})]). \tag{14}$$

Temporal context encoding. To alleviate the flickering effects in the generated HDR videos caused by texture mismatching among consecutive event stacks, we exploit the temporal structure within the data via a recurrent convolutional encoder, as shown in Fig. 4. The temporal context encoders $\mathcal{F}_{\mathrm{E}}, \mathcal{F}_{\mathrm{L}}$ learn modality-specific latent representations $F_{\mathrm{E}}, F_{\mathrm{L}}$ by capturing their underlying temporal structures by a recurrent variant of the encoder in U-Net [62], in which each convolutional layer is replaced by the ConvLSTM [65]. At each timestamp, it outputs the internal state from its own input as well as the encoded past states from its previous input. It can provide more consistent global features of each modality and alleviates flickering effects in the recovered HDR videos.

The internal states not only preserve temporal correlation between consecutive frames to suppress the flickering effects but also reduce noise in both input modalities, especially in the under-exposed regions. Through the recurrent design of the encoders, the temporal complementary and redundant information through time can be well exploited in the learned latent representations of events and LDR frames for better HDR video reconstruction.

3.3. Training

Dataset preparation. For the training of HDRev-Net, a large amount of triplets $(\mathbb{E}, \mathbb{L}, \mathbb{H})$ are required. There are no readily available datasets containing these triplets nor HDR video datasets with satisfactory scale and quality. To synthesize such a dataset, we collect 733 HDR images from [13, 15, 17, 18, 26, 37, 55, 66, 76] and 110 videos with local motion from [13, 14, 34, 37]. For HDR images, 733 HDR videos with global motion are simulated by generated random camera motion trajectories [5]. From each frame sequence of these videos, the hybrid input of events and LDR frames are simulated by [58] and [12], respectively, like capturing with a virtual camera. To better simulate real LDR images, the LDR frames are generated with randomly sampled CRF, exposure time and Poisson-Gaussian noise [23] for each iteration during training. The event generator [58] also considered the normal distributed thresholds to bridge the gap between real and synthetic. There are a total of 26520 (global motion) + 63754 (local motion) = 90274 frames with a resolution of 240 × 180 for training and 5600 frames with a resolution of 256 × 256 for testing.

Loss functions. To train the proposed framework, the loss function to be minimized is defined as:

$$\mathcal{L} = \alpha_1 \mathcal{L}_{\mathrm{mse}} + \alpha_2 \mathcal{L}_{\mathrm{perc}} + \alpha_3 \mathcal{L}_{\mathrm{color}}, \tag{15}$$

where $\alpha_{1,2,3}$ are hyper-parameters for balancing different terms. Our goal is to reconstruct HDR frame sequence $\{H_t\}_{t=1}^{T}$, while the prediction of the network $\mathcal{F}$ is $\{O_t\}_{t=1}^{T}$. The mean square error (MSE) loss $\mathcal{L}_{\mathrm{mse}}$ is defined as:

$$\mathcal{L}_{\mathrm{mse}} = \sum_t (\|H_t - O_t\|_2^2). \tag{16}$$

The perceptual loss $\mathcal{L}_{\mathrm{perc}}$ is defined based on the feature maps extracted by the VGG-16 [67] network pre-trained on ImageNet [63]:

$$\begin{aligned}\mathcal{L}_{\mathrm{perc}} = \sum_t \Bigg(& \sum_l (\|\phi_l(H_t) - \phi_l(O_t)\|_2^2 \\ & + \|\mathcal{G}_l^{\phi}(H_t) - \mathcal{G}_l^{\phi}(O_t)\|_2^2) \Bigg),\end{aligned} \tag{17}$$

where ϕ_l denotes the feature map extracted from l-th layer of VGG-16, $\mathcal{G}_l^{\phi}$ calculates the Gram matrix of ϕ_l. The layer we use in our experiments is "relu4_3". To enforce the color appearance of the reconstructed HDR frames to match that in the ground truth, a color loss is defined as:

$$\mathcal{L}_{\text{color}} = \sum_t (\|\widetilde{H}_t - \widetilde{O}_t\|_2^2), \tag{18}$$

where $\widetilde{H} = G * H$ and $\widetilde{O} = G * O$ are blurred from H, O to eliminate texture and content.

Other implementation details. We adopt a lightweight variant of recurrent U-Net with 2 downsampling layers in all experiments. The network is unrolled for $T = 6$ steps during training. It is implemented on the PyTorch framework and runs on an NVIDIA GeForce RTX 3080 GPU. We use ADAM [36] with default parameter setting for optimization. The batch size is set as 1. The learning rate was fixed to 10^{-5} in the earlier epochs and reduced to 0 in the last 40% epochs by the linear decay strategy. The number of epochs for pretraining Eq. (7), representation alignment Eq. (8), and training the whole pipeline Eq. (6) are set as $40, 5, 5$, respectively. Xavier initialization [20] is used for network initialization. The hyper-parameters in the loss function Eq. (15) are set as $\alpha_1 = 30, \alpha_2 = 5, \alpha_3 = 10^{-3}$. The hyper-parameter μ for dynamic range compression Eq. (5) is set as 5000.

4. Experiments

We compare the proposed HDRev-Net to several state-of-the-art HDR imaging methods, including a frame based HDR image reconstruction method: Liu *et al.* [44]; the colored variant of an event based HDR reconstruction method: E2VID [60]; two event guided HDR image reconstruction methods: Han *et al.* [24] and eSL-Net [72]; and two-exposure-based methods [9, 42].

4.1. Evaluation on synthetic data

Metrics. We adopt the HDR-VDP-3 ("VDP" for short), the updated version of HDR-VDP [46], to evaluate the quality of reconstructed HDR images frame by frame. For evaluating the quality of HDR videos, we use HDR-VQM [52] ("VQM" for short), which is computed based on a spatio-temporal analysis related to human eye fixation behavior during video viewing. We also adopt commonly used peak signal-to-noise ratio (PSNR), structural similarity (SSIM), and the perceptual error with learned perceptual image patch similarity (LPIPS) [78] metrics as reference.

Results. The quantitative results are reported in Table 1. For the two-exposure-based method Li *et al.* [42], we only calculate the PSNR, SSIM, and LPIPS since it only output tone-mapped HDR images which can not be used to calculate VDP and VQM in the linear domain. We generate two-exposure frame sequences for all the two-exposure-based methods with an exposure ratio setting that takes a reasonable balance between the dynamic range covered and the details preserved. The proposed method outperforms existing methods in terms of all metrics.

Table 1. Comparisons on synthetic data. ↑ (↓) means higher (lower) is better.

Methods	PSNR↑	SSIM↑	LPIPS↓	VDP↑	VQM↓
eSL-Net [72]	16.575	0.713	0.413	5.903	0.467
E2VID [60]	13.734	0.589	0.451	4.143	0.343
Liu *et al.* [44]	23.159	0.901	**0.104**	7.543	0.107
Han *et al.* [24]	20.697	0.861	0.208	6.709	0.243
Debevec *et al.* [9]	23.596	0.877	0.264	6.192	0.264
Li *et al.* [42]	20.673	0.890	0.151	-	-
Ours	**24.071**	**0.928**	0.110	**8.108**	**0.103**

The visual quality comparisons are shown in Fig. 5[1]. In over/under-exposed regions of the LDR input, most of the details have been lost due to dynamic range clipping. Both frame-based [44] and event-based [60] approaches cannot restore lost information to produce high-quality HDR images. In particular, E2VID [60] produces severe artifacts and distorted color in the reconstructed HDR frames. Han *et al.* [24] converted one modality into another explicitly (*i.e.*, events to intensity) at first, which leads to stripe-like artifacts and cannot restore natural color appearance. eSL-Net [72] mainly focuses on LDR input and cannot fuse the two modalities well. The performance of Li *et al.* [42] heavily depends on the selected exposure ratio. As shown in Fig. 5[2], even with the optimal exposure ratio, the results of Li *et al.* presents insufficient details (*e.g.*, on the stones and windows). In comparison, the HDRev-Net effectively fuses the two modalities of visual information and successfully recovers the lost details, with no need to balance the exposure ratio for varying scenes captured by videos.

4.2. Evaluation on real data

Fig. 6 shows the reconstructed HDR frames on real-world scenarios captured by a hybrid-camera system. We build the hybrid camera system by using a beam-splitter to divide the incident light equivalently into two sensors, *i.e.*, an event camera (DAVIS346 Color) and an industrial camera (FLIR Chameleon 3 Color). Both of them share a similar field of view (FoV). To achieve spatial alignment between the two sensors, we crop the corresponding part of RGB frames w.r.t. the event sensor. With synchronized triggering, the hybrid camera system is able to capture aligned LDR frames and events simultaneously.

As shown in the HDR frame reconstruction results of

[1] We provide results of one method per category due to the space limit. Please refer to the supplementary materials to complete reconstruction results on synthetic data.

[2] The detailed analysis to exposure ratios are shown in supplementary materials.

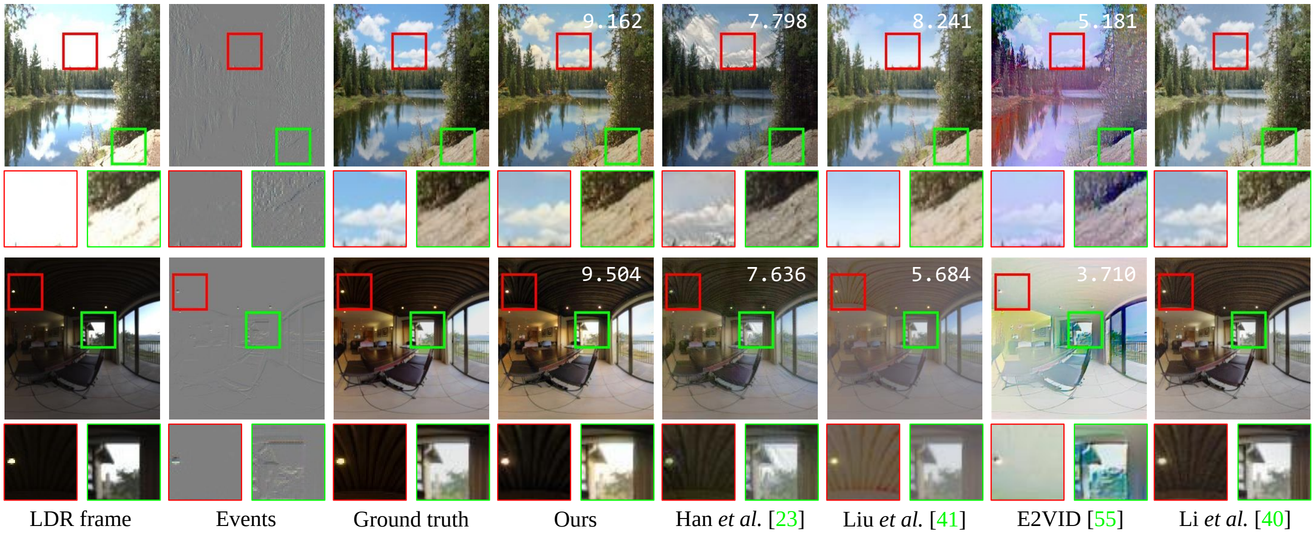


Figure 5. Visual quality comparisons on synthetic data. The Q-scores (higher the better) of VDP metrics are labeled in each image.

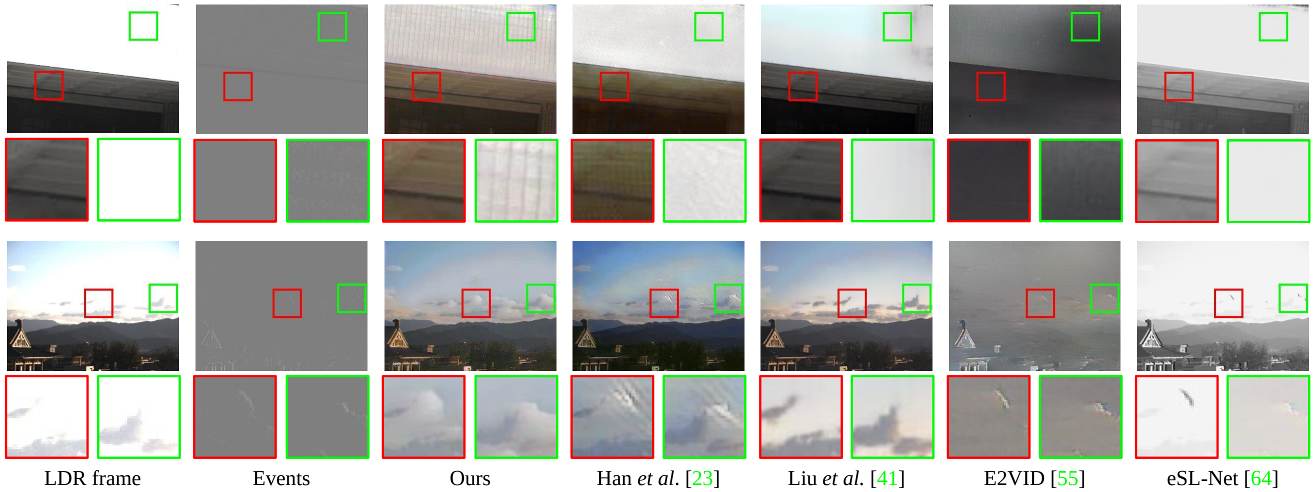


Figure 6. Visual quality comparisons on real-world data. The test samples here are under- and over-exposure cases respectively.

real-world data in Fig. 6, HDRev-Net can reconstruct HDR frames with a wider dynamic range and more details.

4.3. Results of HDR video reconstruction

The information lost in over/under-exposed regions of LDR frames is recorded in events. However, it is unstable for HDR video reconstruction since events are triggered with irradiance changes, rather than absolute values as in LDR frames. HDR video reconstruction in a frame-by-frame manner suffers from flickering effects since temporal connections between consecutive frames are not considered. In the proposed method, recurrent architectures are utilized in the encoders to better maintain the model temporal correlation between frames. Thanks to these components, the proposed HDRev-Net effectively suppresses the flickering effects in the reconstructed HDR videos[3].

4.4. Ablation study

To validate the effectiveness and necessity of the three components of the proposed HDRev-Net, we compare it with its three variants. The quantitative results are shown in Table 2, and the qualitative results are shown in Fig. 3, Fig. 7, Fig. 8, respectively.

Effectiveness of multimodal representation alignment technique. We compare the proposed multimodal representation alignment strategy with joint training the pipeline Eq. (6) in an end-to-end manner, which is denoted by "Joint training" in Table 2. It can be seen that a performance drop occurs when the three networks are jointly trained, especially in terms of PSNR. As shown in Fig. 7 , there are severe halo artifacts in over-exposed regions of the LDR in the results of "Joint" training, especially the contours of the cables and the pylons. Due to the domain gap between LDR frames and stacked events, jointly training

[3] Please refer to the supplementary video for details.

Table 2. Results of ablation study. ↑ (↓) means higher (lower) is better.

Setting	PSNR↑	SSIM↑	LPIPS↓	VDP↑	VQM↓
Joint training	22.185	0.911	0.162	7.871	0.120
W/o fusion	19.436	0.867	0.214	7.024	0.175
W/o LSTM	21.463	0.891	0.197	7.166	0.173
Complete model	**24.071**	**0.928**	**0.110**	**8.108**	**0.103**

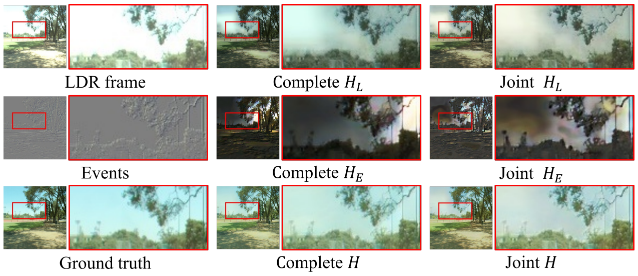


Figure 7. Comparisons between "Joint" training and "Complete" model. H_L (H_E) denotes the network output with encoder F_E (F_L) disabled by setting input to zero.

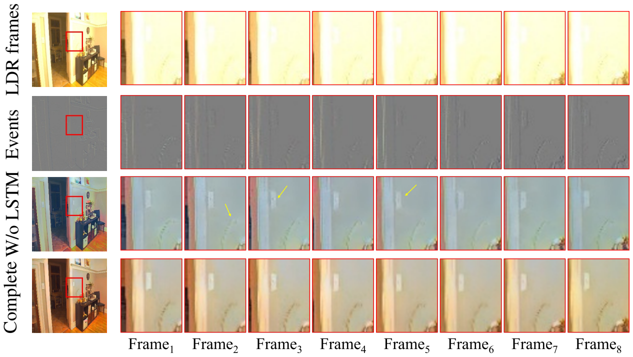


Figure 8. Comparison between "W/o LSTM" and "Complete" model. The results of consecutive frames are shown in eight columns. The animations are shown on our project page.

the whole pipeline cannot well utilize the HDR information in events, which is recorded in a different way from the intensity values in LDR frames.

Effectiveness of the confidence guided multimodal fusion module. To verify the effectiveness of the proposed confidence guided multimodal fusion module, we replace the fusion module with a convolution layer with kernel size 3 × 3 and "relu" activation fed by concatenated F_L, F_E in the "W/o fusion" variant. As shown in Table 2, its performance is worse than the proposed one in terms of all metrics. As shown in Fig. 3, the proposed module can better leverage the information from both modalities and reconstruct HDR images with more details (*e.g.*, the details of the shrub at the lower-left and the texture of the mountain) and less artifact (*e.g.*, the top-left corner). In comparison, the lost details in over-exposed regions cannot be well restored without the fusion module. The convolution layer lacks the ability to dynamically highlight local regions and feature maps that provide more information from one of the two modalities, which makes the recovering of global characteristics such as color and brightness worse than the proposed one.

Effectiveness of the temporal context encoding module. We evaluate the contribution of the recurrent connections in the proposed temporal context encoders by removing them in a "W/o LSTM" variant. As shown in Table 2, it performs the worst due to the absence of recurrent connections, especially in terms of VQM that reflects the temporal quality of the reconstructed HDR videos. As shown in Fig. 8, information lost in the over-exposed switch (*e.g.*, the flickering effects pointed out by the yellow arrows) can hardly be restored since there is no event at this timestamp. However, the complete model with recurrent connections can recover fine details and textures from the previous frame, which suppresses flickering effects effectively.

5. Conclusion

In this paper, we present a multimodal learning framework for reconstructing HDR videos from hybrid inputs of LDR videos and events. There are three components in the proposed method: a multimodal representation alignment strategy for aligning LDR frames and events in a learned shared latent space, a confidence guided multimodal fusion module for adaptively integrating complementary information from the two modalities, and a temporal context encoding module for suppressing the flickering effects by exploiting the temporal correlation between consecutive frames and stacked events. The effectiveness of the proposed method in the task of event guided HDR video reconstruction is verified by extensive experiments on both synthetic and real-world data.

Limitations. The proposed method aims at HDR video reconstruction under the guidance of HDR information provided by events. However, events contain inherent noise distributed differently from that of frames, which can lead to artifacts in the reconstructed HDR video, especially in low-light conditions. Moreover, due to the requirement of aligned spatial resolution of the hybrid inputs and low resolution of the chosen event camera (DAVIS346C), the proposed method has a limited ability for high-resolution HDR video reconstruction. Replacing the chosen event camera with a higher resolution one(*e.g.*, Prophesee EVK4-HD with a spatial resolution of 720 × 1280) and better low-light perception might further improve the performance.

Acknowledgement

This work was supported by National Key R&D Program of China (2021ZD0109803) and National Natural Science Foundation of China (Grant No. 62088102, 62136001). Jinxiu Liang was also supported by China Postdoctoral Science Foundation (Grant No. 2022M720236).

References

[1] Cecilia Aguerrebere, Andrés Almansa, Yann Gousseau, Julie Delon, and Pablo Musé. Single shot high dynamic range imaging using piecewise linear estimators. In *Proc. of International Conference on Computational Photography*, 2014. 1, 2

[2] Francesco Banterle, Patrick Ledda, Kurt Debattista, and Alan Chalmers. Inverse tone mapping. In *Proc. of International Conference on Computer Graphics and Interactive Techniques*, 2006. 2

[3] Patrick Bardow, Andrew J. Davison, and Stefan Leutenegger. Simultaneous optical flow and intensity estimation from an event camera. In *Proc. of Computer Vision and Pattern Recognition*, 2016. 2

[4] Souptik Barua, Yoshitaka Miyatani, and Ashok Veeraraghavan. Direct face detection and video reconstruction from event cameras. In *Proc. of Winter Conference on Applications of Computer Vision*, 2016. 2

[5] Giacomo Boracchi and Alessandro Foi. Modeling the performance of image restoration from motion blur. *IEEE Transactions on Image Processing*, 2012. 5

[6] Gaofeng Cao, Fei Zhou, Kanglin Liu, Anjie Wang, and Leidong Fan. A decoupled kernel prediction network guided by soft mask for single image hdr reconstruction. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2023. 2

[7] Guanying Chen, Chaofeng Chen, Shi Guo, Zhetong Liang, Kwan-Yee K Wong, and Lei Zhang. HDR video reconstruction: A coarse-to-fine network and a real-world benchmark dataset. In *Proc. of International Conference on Computer Vision*, 2021. 1, 2

[8] Hojin Cho, Seon Joo Kim, and Seungyong Lee. Single-shot high dynamic range imaging using coded electronic shutter. In *Computer Graphics Forum*, 2014. 1, 2

[9] Paul E. Debevec and Jitendra Malik. Recovering high dynamic range radiance maps from photographs. In *ACM Transactions on Graphics (Proc. of ACM SIGGRAPH)*, 1997. 1, 2, 6

[10] Rahul Dey and Fathi M Salem. Gate-variants of gated recurrent unit (GRU) neural networks. In *International Midwest Symposium on Circuits and Systems*, 2017. 3

[11] 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 (Proc. of ACM SIGGRAPH Asia)*, 2017. 2

[12] Yuki Endo, Yoshihiro Kanamori, and Jun Mitani. Deep reverse tone mapping. *ACM Transactions on Graphics (Proc. of ACM SIGGRAPH Asia)*, 2017. 2, 5

[13] Jan Froehlich, Stefan Grandinetti, Bernd Eberhardt, Simon Walter, Andreas Schilling, and Harald Brendel. Creating cinematic wide gamut HDR-video for the evaluation of tone mapping operators and HDR-displays. In *Digital photography X*, 2014. 5

[14] Jan Froehlich, Stefan Grandinetti, Bernd Eberhardt, Simon Walter, Andreas Schilling, and Harald Brendel. Creating cinematic wide gamut HDR-video for the evaluation of tone mapping operators and HDR-displays. In *Proc. of Digital photography X*, 2014. 5

[15] Funt *et al.* HDR dataset. `https://www2.cs.sfu.ca/~colour/data/funt_hdr/#DESCRIPTION`. 5

[16] Guillermo Gallego, Tobi Delbruck, Garrick Michael Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J. Davison, Jorg Conradt, Kostas Daniilidis, and Davide Scaramuzza. Event-based vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 2

[17] Marc-André Gardner, Kalyan Sunkavalli, Ersin Yumer, Xiaohui Shen, Emiliano Gambaretto, Christian Gagné, and Jean-François Lalonde. Learning to predict indoor illumination from a single image. *ACM Transactions on Graphics (Proc. of ACM SIGGRAPH)*, 2017. 5

[18] Mathieu Garon, Kalyan Sunkavalli, Sunil Hadap, Nathan Carr, and Jean-Francois Lalonde. Fast spatially-varying indoor lighting estimation. In *Proc. of Computer Vision and Pattern Recognition*, 2019. 5

[19] Michaël Gharbi, Jiawen Chen, Jonathan Barron, Samuel Hasinoff, and Frédo Durand. Deep bilateral learning for real-time image enhancement. *ACM Transactions on Graphics (Proc. of ACM SIGGRAPH)*, 2017. 2

[20] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proc. of International Conference on Artificial Intelligence and Statistics*, 2010. 6

[21] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Proc. of International Conference on Neural Information Proceeding Systems*, 2014. 2

[22] Jinwei Gu, Yasunobu Hitomi, Tomoo Mitsunaga, and Shree K. Nayar. Coded rolling shutter photography: Flexible space-time sampling. In *Proc. of International Conference on Computational Photography*, 2010. 1, 2

[23] Shi Guo, Zifei Yan, Kai Zhang, Wangmeng Zuo, and Lei Zhang. Toward convolutional blind denoising of real photographs. *Proc. of Computer Vision and Pattern Recognition*, 2019. 5

[24] Jin Han, Chu Zhou, Peiqi Duan, Yehui Tang, Chang Xu, Chao Xu, Tiejun Huang, and Boxin Shi. Neuromorphic camera guided high dynamic range imaging. In *Proc. of Computer Vision and Pattern Recognition*, 2020. 1, 2, 3, 6

[25] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 1997. 3

[26] Jun Hu, Orazio Gallo, Kari Pulli, and Xiaobai Sun. HDR deghosting: How to deal with saturation? In *Proc. of Computer Vision and Pattern Recognition*, 2013. 5

[27] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proc. of Computer Vision and Pattern Recognition*, 2018. 5

[28] Yitong Jiang, Inchang Choi, Jun Jiang, and Jinwei Gu. HDR video reconstruction with tri-exposure quad-bayer sensors. *arXiv preprint arXiv:2103.10982*, 2021. 1, 2

[29] Nima Khademi Kalantari and Ravi Ramamoorthi. Deep high dynamic range imaging of dynamic scenes. *ACM Transactions on Graphics (Proc. of ACM SIGGRAPH)*, 2017. 2

[30] Sing Bing Kang, Matthew Uyttendaele, Simon Winder, and Richard Szeliski. High dynamic range video. *ACM Transactions on Graphics (Proc. of ACM SIGGRAPH)*, 2003. 1, 2

[31] Nima Khademi Kalantari and Ravi Ramamoorthi. Deep high dynamic range imaging of dynamic scenes. *ACM Transactions on Graphics (Proc. of ACM SIGGRAPH)*, 2017. 3

[32] Nima Khademi Kalantari and Ravi Ramamoorthi. Deep HDR video from sequences with alternating exposures. In *Computer Graphics Forum*, 2019. 1, 2

[33] Zeeshan Khan, Mukul Khanna, and Shanmuganathan Raman. FHDR: HDR image reconstruction from a single LDR image using feedback network. In *Proc. of Global Conference on Signal and Information Processing*, 2019. 2

[34] Hamed Kiani Galoogahi, Ashton Fagg, Chen Huang, Deva Ramanan, and Simon Lucey. Need for speed: A benchmark for higher frame rate object tracking. In *Proc. of Computer Vision and Pattern Recognition*, 2017. 5

[35] Hanme Kim, Ankur Handa, Ryad Benosman, Sio-Hoi Ieng, and Andrew J. Davison. Simultaneous mosaicing and tracking with an event camera. In *Proc. of British Machine Vision Conference*, 2014. 2

[36] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[37] Joel Kronander, Stefan Gustavson, Gerhard Bonnet, and Jonas Unger. Unified HDR reconstruction from raw CFA data. In *Proc. of International Conference on Computational Photography*, 2013. 5

[38] Joel Kronander, Stefan Gustavson, Gerhard Bonnet, Anders Ynnerman, and Jonas Unger. A unified framework for multi-sensor HDR video reconstruction. *Signal Processing: Image Communication*, 2014. 2

[39] Phuoc-Hieu Le, Quynh Le, Rang Nguyen, and Binh-Son Hua. Single-image hdr reconstruction by multi-exposure generation. In *Proc. of Winter Conference on Applications of Computer Vision*, 2023. 2

[40] 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*, 2018. 2

[41] Siyeong Lee, Gwon Hwan An, and Suk-Ju Kang. Deep recursive HDRI: Inverse tone mapping using generative adversarial networks. In *Proc. of European Conference on Computer Vision*, 2018. 2

[42] Hui Li, Kede Ma, Hongwei Yong, and Lei Zhang. Fast multi-scale structural patch decomposition for multi-exposure image fusion. *IEEE Transactions on Image Processing*, 29:5805–5816, 2020. 6

[43] Yuelong Li, Chul Lee, and Vishal Monga. A maximum a posteriori estimation framework for robust high dynamic range video synthesis. *IEEE Transactions on Image Processing*, 2016. 2

[44] 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 *Proc. of Computer Vision and Pattern Recognition*, 2020. 2, 6

[45] Zhen Liu, Yinglong Wang, Bing Zeng, and Shuaicheng Liu. Ghost-free high dynamic range imaging with context-aware transformer. In *Proc. of European Conference on Computer Vision*, 2022. 2

[46] 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 (Proc. of ACM SIGGRAPH)*, 2011. 6

[47] Demetris Marnerides, Thomas Bashford-Rogers, Jonathan Hatchett, and Kurt Debattista. ExpandNet: A deep convolutional neural network for high dynamic range expansion from low dynamic range content. In *Computer Graphics Forum*, 2018. 2

[48] Morgan McGuire, Wojciech Matusik, Hanspeter Pfister, Billy Chen, John F. Hughes, and Shree K. Nayar. Optical splitting trees for high-precision monocular imaging. *IEEE Computer Graphics and Applications*, 2007. 2

[49] Tom Mertens, Jan Kautz, and Frank Van Reeth. Exposure fusion. In *Proc. of Conference on Computer Graphics and Applications*, 2007. 1, 2

[50] Nico Messikommer, Stamatios Georgoulis, Daniel Gehrig, Stepan Tulyakov, Julius Erbach, Alfredo Bochicchio, Yuanyou Li, and Davide Scaramuzza. Multi-bracket high dynamic range imaging with event cameras. In *Proc. of Computer Vision and Pattern Recognition*, 2022. 3

[51] Mohammad Mostafavi, Lin Wang, and Kuk-Jin Yoon. Learning to reconstruct hdr images from events, with applications to depth and flow prediction. *International Journal of Computer Vision*, 2021. 2

[52] Manish Narwaria, Matthieu Perreira Da Silva, and Patrick Le Callet. HDR-VQM: An objective quality measure for high dynamic range video. *Signal Processing: Image Communication*, 2015. 6

[53] Shree K. Nayar and Vlad Branzoi. Adaptive dynamic range imaging: Optical control of pixel exposures over space and time. In *Proc. of International Conference on Computer Vision*, 2003. 1, 2

[54] Shree K. Nayar and Tomoo Mitsunaga. High dynamic range imaging: Spatially varying pixel exposures. In *Proc. of Computer Vision and Pattern Recognition*, 2000. 1, 2

[55] Hiromi Nemoto, Pavel Korshunov, Philippe Hanhart, and Touradj Ebrahimi. Visual attention in LDR and HDR images. In *International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, 2015. 5

[56] Liyuan Pan, Richard Hartley, Cedric Scheerlinck, Miaomiao Liu, Xin Yu, and Yuchao Dai. High frame rate video reconstruction based on an event camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 2, 3

[57] Liyuan Pan, Cedric Scheerlinck, Xin Yu, Richard Hartley, Miaomiao Liu, and Yuchao Dai. Bringing a blurry frame alive at high frame-rate with an event camera. In *Proc. of Computer Vision and Pattern Recognition*, 2019. 2, 3

[58] Henri Rebecq, Daniel Gehrig, and Davide Scaramuzza. ESIM: An open event camera simulator. 2018. 5

[59] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. Events-to-video: Bringing modern computer

vision to event cameras. In *Proc. of Computer Vision and Pattern Recognition*, 2019. 2

[60] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. High speed and high dynamic range video with an event camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 1, 2, 3, 6

[61] Allan G. Rempel, Matthew Trentacoste, Helge Seetzen, H. David Young, Wolfgang Heidrich, Lorne Whitehead, and Greg Ward. LDR2HDR: On-the-fly reverse tone mapping of legacy video and photographs. *ACM Transactions on Graphics (Proc. of ACM SIGGRAPH)*, 2007. 2

[62] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Proc. of International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2015. 5

[63] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 2015. 5

[64] Richard Shaw, Sibi Catley-Chandar, Ales Leonardis, and Eduardo Pérez-Pellitero. Hdr reconstruction from bracketed exposures and events. In *Proc. of British Machine Vision Conference*, 2022. 3

[65] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In *Proc. of International Conference on Neural Information Proceeding Systems*, 2015. 5

[66] sIBL archive. http://www.hdrlabs.com/sibl/archive.html. 5

[67] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5

[68] 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 *Proc. of European Conference on Computer Vision*, 2022. 2

[69] Michael D Tocci, Chris Kiser, Nora Tocci, and Pradeep Sen. A versatile HDR video production system. *ACM Transactions on Graphics (Proc. of ACM SIGGRAPH)*, 2011. 1, 2

[70] Marc Comino Trinidad, Ricardo Martin Brualla, Florian Kainz, and Janne Kontkanen. Multi-view image fusion. In *Proc. of International Conference on Computer Vision*, 2019. 1, 2

[71] An Gia Vien and Chul Lee. Exposure-aware dynamic weighted learning for single-shot HDR imaging. In *Proc. of European Conference on Computer Vision*, 2022. 2

[72] Bishan Wang, Jingwei He, Lei Yu, Gui-Song Xia, and Wen Yang. Event enhanced high-quality image recovery. In *Proc. of European Conference on Computer Vision*, 2020. 3, 6

[73] Lin Wang, Yo-Sung Ho, Kuk-Jin Yoon, et al. Event-based high dynamic range image and very high frame rate video generation using conditional generative adversarial networks. In *Proc. of Computer Vision and Pattern Recognition*, 2019. 2

[74] Lin Wang, Tae-Kyun Kim, and Kuk-Jin Yoon. EventSR: From asynchronous events to image reconstruction, restoration, and super-resolution via end-to-end adversarial learning. In *Proc. of Computer Vision and Pattern Recognition*, 2020. 2

[75] Ziwei Wang, Yonhon Ng, Cedric Scheerlinck, and Robert Mahony. An asynchronous kalman filter for hybrid event cameras. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 448–457, 2021. 2, 3

[76] Feng Xiao, Jeffrey M. DiCarlo, Peter B. Catrysse, and Brian A. Wandell. High dynamic range imaging of natural scenes. In *Proc. of Color and Imaging Conference*, 2002. 5

[77] 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 *Proc. of Computer Vision and Pattern Recognition*, 2019. 2

[78] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proc. of Computer Vision and Pattern Recognition*, 2018. 6

[79] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Unsupervised event-based learning of optical flow, depth, and egomotion. In *Proc. of Computer Vision and Pattern Recognition*, 2019. 3

[80] Yunhao Zou, Yinqiang Zheng, Tsuyoshi Takatani, and Ying Fu. Learning to reconstruct high speed and high dynamic range videos from events. In *Proc. of Computer Vision and Pattern Recognition*, 2021. 2, 3

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.

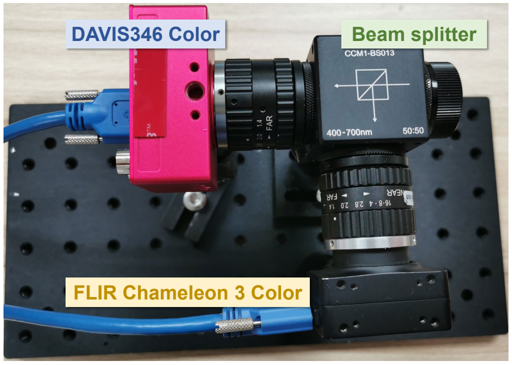
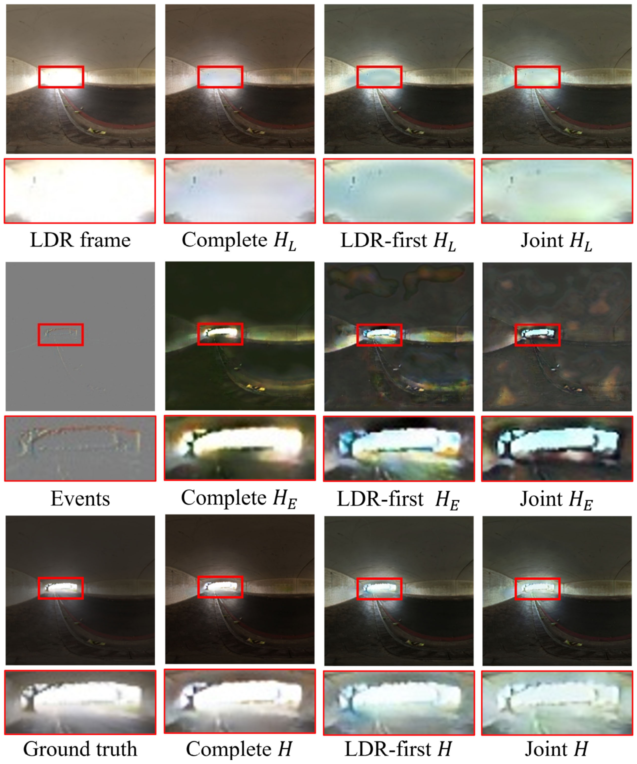


Figure 9. Hybrid camera system for capturing real data.


LDR frame | Complete H_L | LDR-first H_L | Joint H_L

Events | Complete H_E | LDR-first H_E | Joint H_E

Ground truth | Complete H | LDR-first H | Joint H

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.

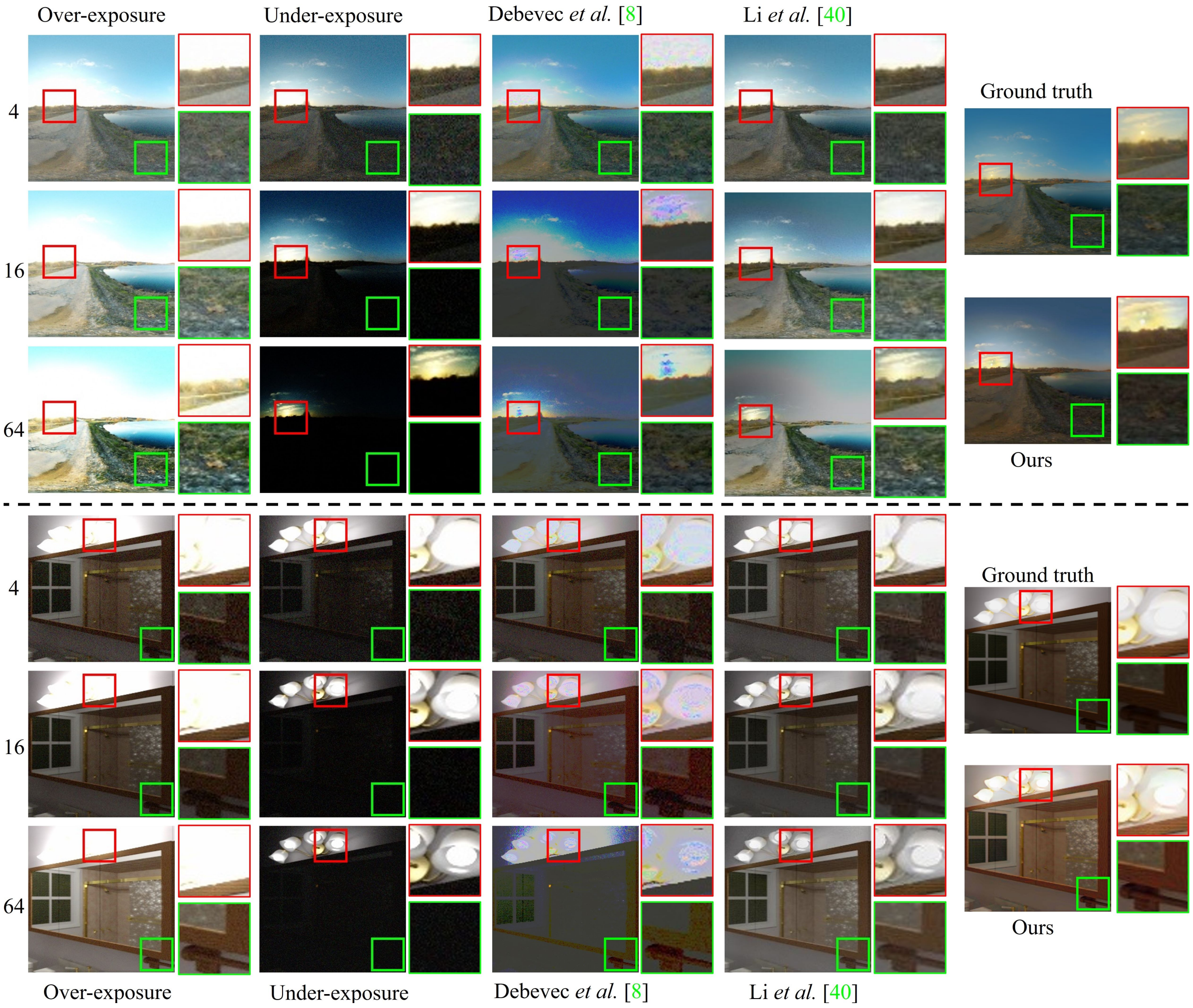


Figure 11. Comparison of two-exposure-based methods with ours. The numbers on the left denote the exposure ratio between "Over-exposure" and "Under-exposure" images, which is used as the input of the two-exposure-based methods. To demonstrate the generalization ability of the proposed event guided method, 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).

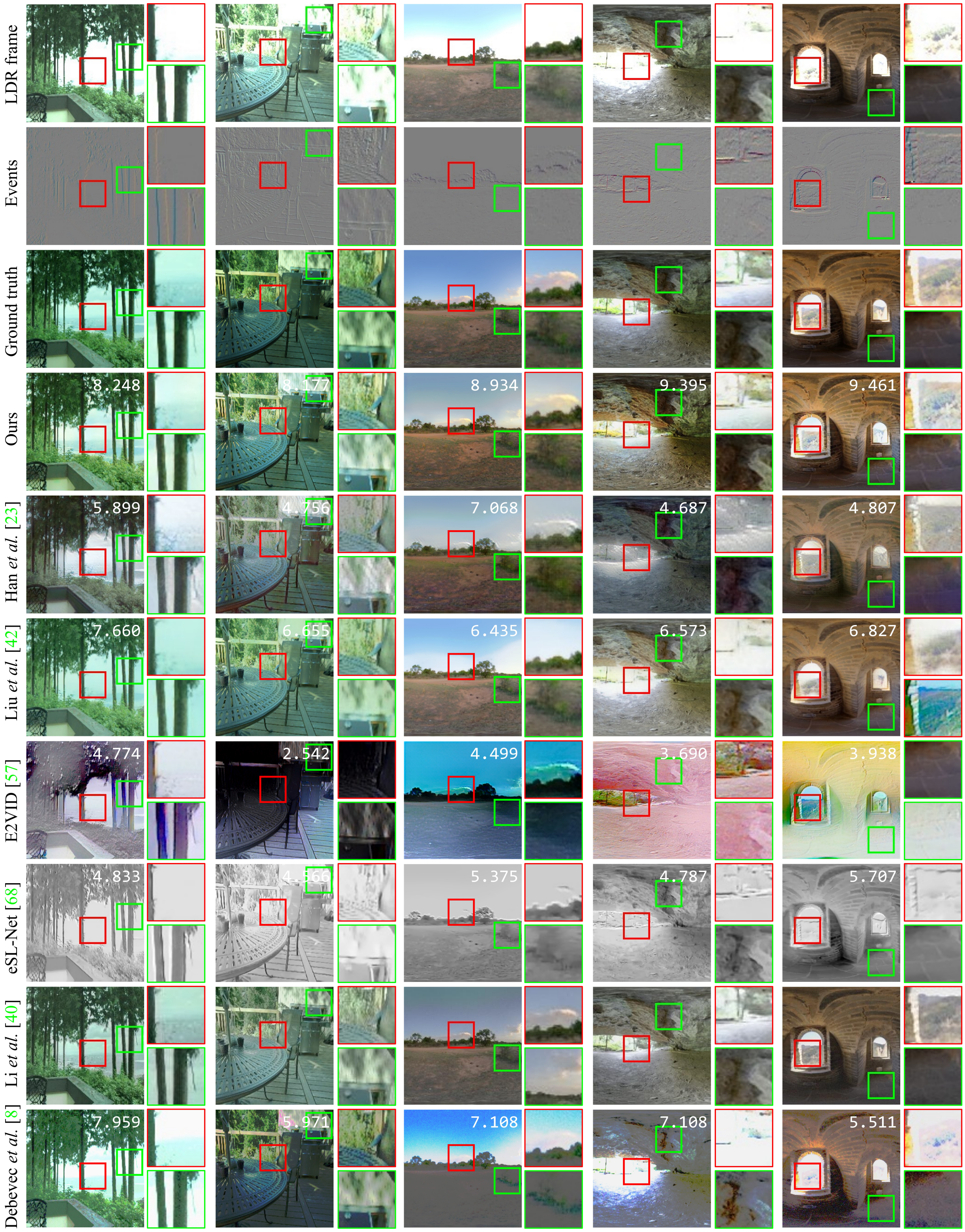


Figure 12. Visual quality comparisons on synthetic data. The Q-scores (higher the better) computed from VDP metrics are labeled in each image. The results of our method have finer details and higher Q-scores than all the others.

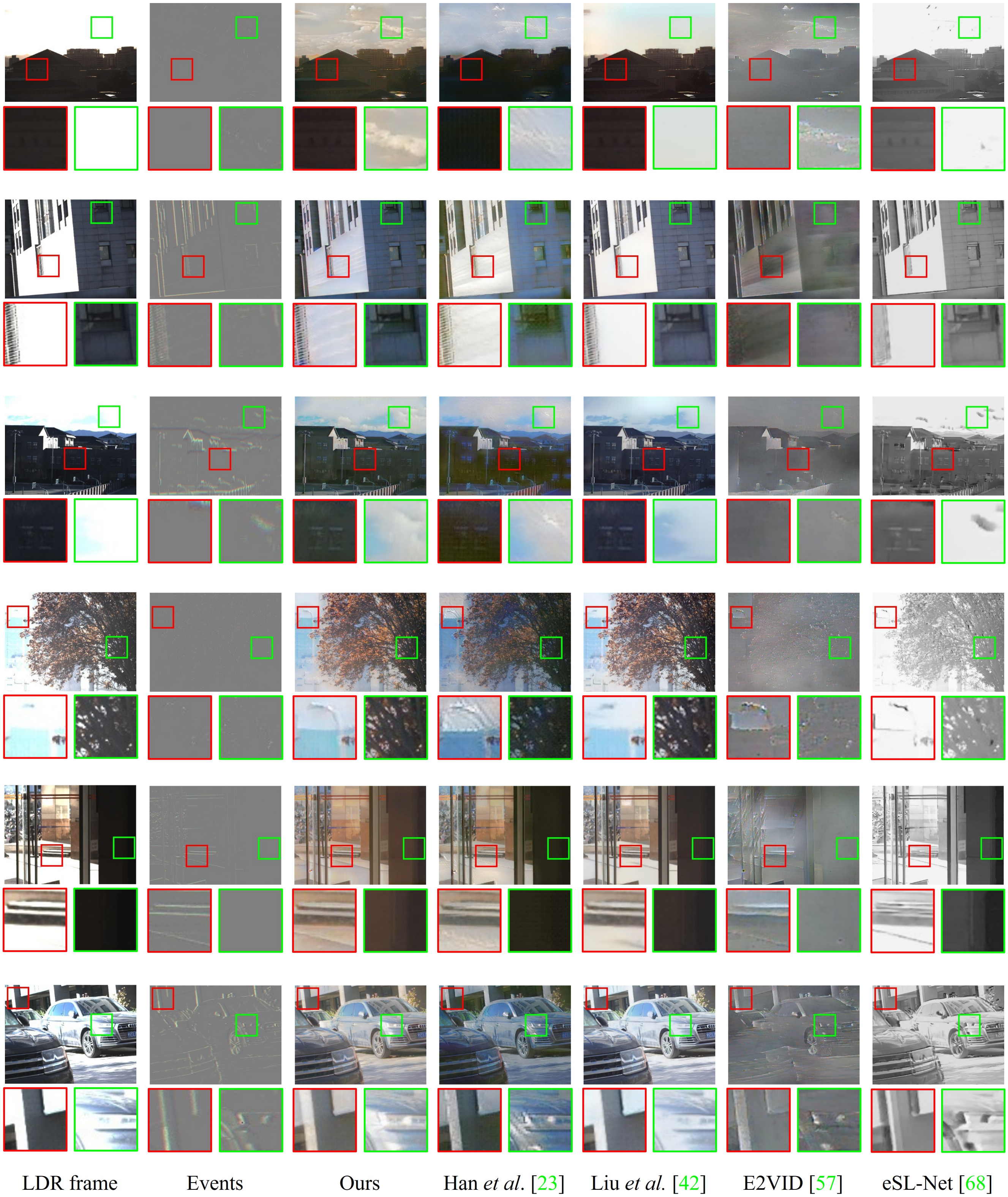


Figure 13. Visual quality comparisons on real data. By leveraging the information from the event stream, the proposed method can reconstruct HDR images with better visual quality than others.