Event-guided HDR Reconstruction with Diffusion Priors

Yixin Yang[1,2,†] Jiawei Zhang[3] Yang Zhang[1,2]
Yunxuan Wei[3] Dongqing Zou[4] Jimmy S. Ren[3,5] Boxin Shi[1,2,*]
[1] State Key Lab of Multimedia Info. Processing, School of Computer Science, Peking University
[2] Nat'l Eng. Research Ctr. of Visual Technology, School of Computer Science, Peking University
[3] SenseTime Research [4]PBVR [5] Hong Kong Metropolitan University
{yangyixin93, shiboxin}@pku.edu.cn, Github Page: github.com/YixinYang-00/HDRev-Diff

Abstract

Events provide High Dynamic Range (HDR) intensity change which can guide Low Dynamic Range (LDR) image for HDR reconstruction. However, events only provide temporal intensity differences and it is still ill-posed in over-/under-exposed areas due to missing initial reference brightness and color information. With strong generation ability, diffusion models have shown their potential for tackling ill-posed problems. Therefore, we introduce conditional diffusion models to hallucinate missing information. Whereas, directly adopting events and LDR image as conditions is complicated for diffusion models to sufficiently utilize their information. Thus we introduce a pre-trained events-image encoder tailored for HDR reconstruction and a pyramid fusion module to provide HDR conditions, which can be efficiently and effectively utilized by the diffusion model. Moreover, the generation results of diffusion models usually exhibit distortion, particularly for fine-grained details. To better preserve fidelity and suppress distortion, we propose a fine-grained detail recovery approach using a histogram-based structural loss. Experiments on real and synthetic data show the effectiveness of the proposed method in terms of both detail preservation and information hallucination.

1. Introduction

Extending the dynamic range of an image (usually with low dynamic range, LDR) to record the brightness of real scenes plausibly is particularly useful for daily photography and computer vision tasks. High dynamic range (HDR) imaging can be, in the most direct way, achieved by utilizing neural networks to hallucinate missing information in over-/under-exposed areas [4, 5, 22, 24], as shown in Figure 1 (a), which is challenging to faithfully recover scene details without any auxiliary information. Neuromorphic cameras, such as event cameras, are proven to be helpful in providing additional details for HDR purpose [28, 32, 53]. Event cameras capture intensity changes

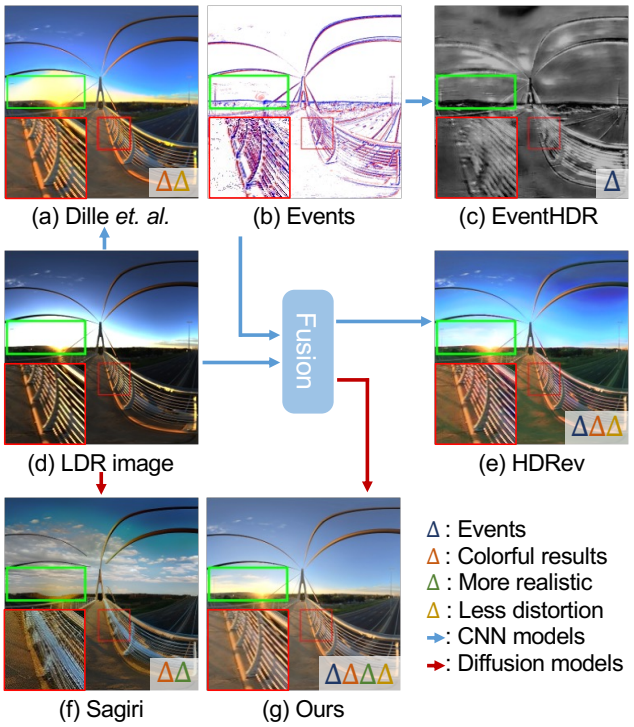

Figure 1. HDR image can be recovered by only using a single LDR image (d) as input (*e.g.*, Dille *et al.* [4]), which is a difficult problem as shown in (a). With only events (b) as input, event-to-image reconstruction (*e.g.*, EventHDR [53]) (c) recovers an HDR scene without color. Event-guided HDR methods (*e.g.*, HDRev [48]) (e) take both events and the LDR image as input, which recovers colorful HDR scenes with moderate details. Diffusion-based methods (*e.g.*, Sagiri [22]) (f) take the LDR image (and optionally the text prompt) as input showing visually pleasing results but may have challenges in maintaining consistent contents with the original scene. The proposed method (g) takes events and the LDR image as input and leverages the diffusion priors to recover a plausible and realistic HDR image with rich details and less distortion.

[†]This work is done during Yixin's internship at SenseTime.
[*]Corresponding author.

by triggering events asynchronously, which inherently contains HDR information (*e.g.*, 120 dB for DAVIS340), and events can be directly converted to HDR images, as demonstrated by EventHDR [53] in Figure 1 (c). The event-to-image reconstruction [28, 32, 53] can recover details in over-/under-exposed regions, but fail to provide colorful results due to lacking color information.

To generate colorful HDR images, event-guided HDR reconstructions [13, 14, 48] take both LDR image and events as inputs, complementarily learning to predict color and intensity in over-/under-exposed areas, as demonstrated in Figure 1 (e). They can reconstruct HDR information, *e.g.*, the cloud in the green box, and provide colorful results. Although regression models are adopted to build the relationship between the predicted pixel values and its neighboring pixels, those methods cannot recover proper intensity, especially when the LDR image has large over-/under-exposed areas. Because absolute HDR intensity (events only contain thresholded radiance changes) and color information are still missing in these regions, it is difficult for regression models to appropriately hallucinate their values.

Diffusion models [16, 37–39] are emerging generative models that show great potential for generating plausible and high-quality images. Even if for ill-posed problems, they can provide visually pleasing results thanks to the expressive power of diffusion priors. Diffusion models have also been adopted for image restoration [23, 43, 49]. Recent attempts, *e.g.*, Sagiri [22], have shown the potential of diffusion priors for HDR reconstruction by providing more visually realistic details in over-exposed areas, *e.g.*, the cloud in the sky of Figure 1 (f). However, the generated HDR images still cannot appropriately keep the fidelity of the original scene, and additional distortion are brought into well-exposed regions, as shown by the red box in Figure 1 (f).

To reconstruct visually pleasing and faithful HDR images with less distortion, we introduce events to provide differential HDR intensity and adopt conditional diffusion models [50] to compensate for the ill-posed nature of hallucinating missing information and constrain the generative model to be consistent with the original scene. Specially, we focus on solving the following two key problems via employing diffusion models in event-guided HDR reconstruction: **1) Condition extraction**: Since events and LDR image are two different modalities of visual representation, directly adopting them as conditions is non-compatible for the conditional diffusion models to conduct effective fusion. Inspired by the modality alignment of HDRev [48], we introduce an event-image encoder and a pyramid fusion module to fuse events and LDR image in different resolutions and provide conditions to guide the generation process. **2) Details preservation**: Applying diffusion models to HDR reconstruction [22, 23] suffers from fine-grained detail distortion, as shown in Figure 1 (f). Such artifacts become more obvious when LDR images are high quality and have large well-exposed areas. To refine the fine-grained details and make them consistent with LDR images and events, we design a refinement module with histogram-based structure loss. The proposed method is validated on both synthetic and real data and outperforms existing methods by providing fidelity-preserving HDR reconstruction, as illustrated in Figure 1 (g), with the following technical contributions:

- integrating events and the conditional diffusion models to recover missing information faithfully;
- adopting the pretrained event-image encoder and pyramid fusion to effectively apply conditions; and
- designing the refinement module and histogram-based structure loss to further strengthen fine-grained details.

2. Related works

Image-based HDR reconstruction Many approaches leverage the capabilities of neural networks to directly hallucinate HDR information from training data, such as CNNs [5, 6, 9], Generative Adversarial Networks (GANs) [40], and diffusion models [7, 22]. Liu *et al.* [24] incorporates the LDR image formulation pipeline into its hallucination process to perform step-by-step prediction. For a more comprehensive view, we refer to GTA-HDR [2]. Since information is unavoidably missing in over-/under-exposed areas, the above methods only rely on hallucination, which may not correspond to real scenes.

Event-based HDR reconstruction Benefiting from the HDR property of event cameras, previous works reconstruct HDR images from events with recurrent neural networks [32, 33, 53] or GANs [28, 41]. Since events only record logarithmic irradiance changes, it's difficult to reconstruct color images. To reconstruct color HDR images, Han *et al.* [13, 14] adopted a hybrid camera system to capture events and frame, enabling event-guided HDR reconstruction. Yang *et al.* [48] treated events and frame as different modalities and proposed multi-modalities alignment and fusion to provide HDR videos. Other methods [26, 35] incorporate events and bracketed exposure for HDR reconstruction, while long-exposed images bring blurry artifacts. The proposed method takes events and LDR image as input, incorporating diffusion models to compensate for realistic and faithful information in over-/under-exposed areas.

Generative priors Generative Priors capture the underlying data distribution, enabling the generated images to follow real-world image distribution. GAN [10] has shown its potential to leverage generative priors for image restoration [20, 21, 30]. Recently, diffusion models [16, 37–39] have widely adopted as effective generative priors for producing realistic images [29, 34, 42, 50], which also demonstrate strong potential for image restorations [23, 43, 49], specifically, for HDR image reconstruction from bracketed exposure [12, 17, 46]. There are some attempts [11, 22] to

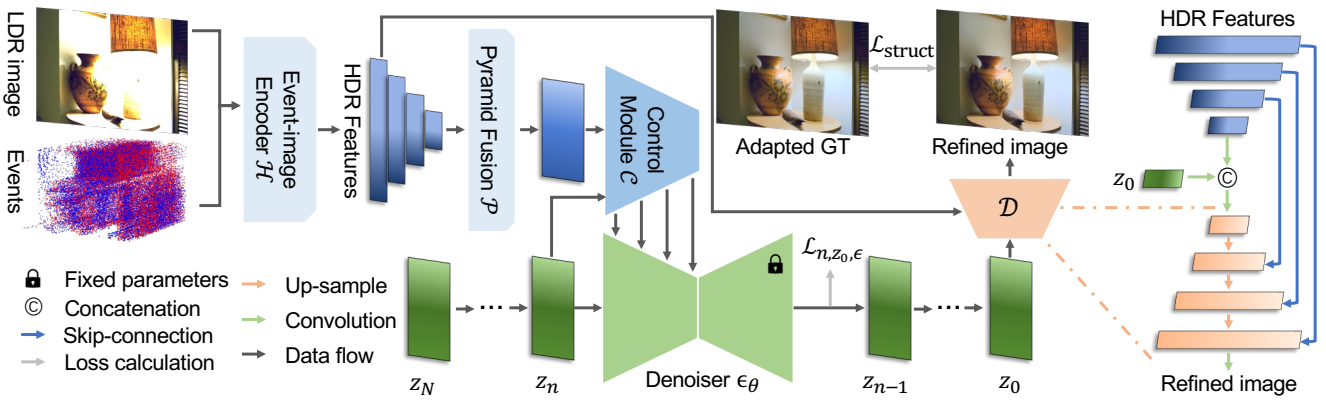

Figure 2. Pipeline of the proposed method consists of three parts: diffusion model and denoising process (green part), event-image conditioning (blue part), and detail refinement (orange part). The blue part fuses LDR image and events to generate injected features for Denoiser as described in Section 3.2. The green part adopts the injected features as conditions to estimate noise from latent z_t step by step with fixed pretrained parameters. To refine the distortion of denoised images, the refinement module in the orange part, whose architecture is shown on the right, is proposed to leverage the HDR features and denoised latent z_0. Considering the uncertainty of diffusion results in color and brightness, we introduce structure loss, which adopts adaptive Ground Truth (GT) as the supervision target, as described in Section 3.3. After refinement, we obtain a plausible and realistic refined image without distortion.

take advantage of diffusion models to generate HDR images from single LDR image, while it cannot provide faithful reconstruction due to the inevitable dynamic clipping.

3. Proposed Method

Section 3.1 introduces basic concepts about diffusion models and event cameras. The pipeline of the proposed method is illustrated in Figure 2. The green part is the pretrained latent diffusion model, which performs denoising in latent space step by step with the injected feature from the blue part. The blue part is the proposed event-guided conditioning and generation process, which adopts an event-image encoder and a pyramid fusion to utilize events and the LDR image efficiently and effectively, as described in Section 3.2. The orange part is the fine-grained detail enhancement based on the Variant Auto Encoder (VAE) decoder, which learns to refine fine-grained details by the proposed histogram-based structure loss as described in Section 3.3. Training details are described in Section 3.4.

3.1. Preliminary

Diffusion models Diffusion models [16, 34, 38] can generate realistic high-quality images by diffusing and denoising processes. Latent diffusion models [34] are a type of diffusion models that perform those processes on latent space. During training, the latent diffusion models project the clean image into the latent space with a VAE encoder as z_0, then diffuse z_0 by adding Gaussian noise for n steps to get the noisy latent z_n. Adding noise in one step is as:

$$z_n = \sqrt{a_n} z_{n-1} + \sqrt{1 - a_n}\epsilon_{n-1}, \tag{1}$$

where $n \in [1, N]$ means the n-th diffusion step, N is the total diffusion steps, $\epsilon_n \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is the added noise, a_n is the pre-defined parameter of the noise scheduler. Defining $\overline{a}_n = \prod_{i=1}^{n} a_i$, the closed form equation of Eq. (1) [16] is:

$$z_n = \sqrt{\overline{a}_n} z_0 + \sqrt{1 - \overline{a}_n}\epsilon, \tag{2}$$

where $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The denoising process trains a denoiser ϵ_θ to estimate the noise ϵ usually by minimizing Mean Square Error (MSE) $\mathcal{L}_{n,z_0,\epsilon} = \|\epsilon - \epsilon_\theta(z_n, n)\|$. As the variable z_n will be an approximately standard Gaussian distribution when n is large enough, during inference, the denoising process often starts from an i.i.d. noise z_N and generates a clean image z_0 using ϵ_θ over N steps.

Event triggering and stacking Event cameras trigger an event (t, x, y, p) when illuminance changes of pixel (x, y) in the logarithm domain over a predefined threshold η in time t, where the polarity $p \in \{-1, 1\}$ indicates the decrease and increase of the illuminance. Mathematically, an event is generated when the following inequation occurs:

$$\|\log I_{x,y}(t) - \log I_{x,y}(t_{\text{ref}})\| \geq \eta, \tag{3}$$

where $I_{x,y}(t)$ is the illuminance at time t, $I_{x,y}(t_{\text{ref}})$ is the reference illuminance level, t_{ref} is the last event triggered time of pixel (x, y), η is a pre-defined event threshold. We convert the stream-like events to tensors using voxel grid [52], which encodes temporal information in a C-channel 3D volume by discretizing event timestamps into C temporal bins. Following Rebecq *et al.* [32, 33], k-th event (t_k, x_k, y_k, p_k) distributes its polarity p_k to the two closest temporal bins related to its normalized timestamp by:

$$E_j^{x,y} = \sum_{x_k = x, y_k = y} p_k \max\left(0, 1 - |\tilde{t}_k - j|\right), \tag{4}$$

where $\tilde{t}_k = \frac{C-1}{\Delta t}(t_k - t_0)$ is the normalized timestamp, $\Delta_t = \max_k(t_k) - t_0$ is the time span of events, t_0 is the start timestamp, and $j \in [0, C-1]$ is the index of the bin.

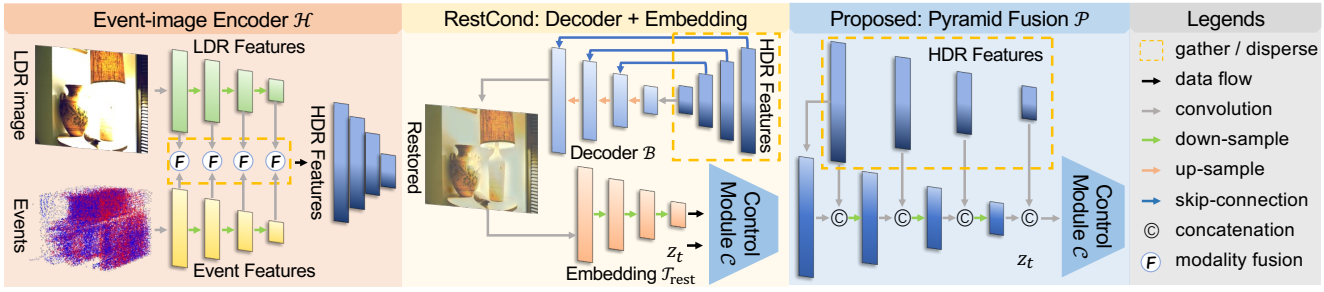

Figure 3. Pipeline of the event-guided conditioning process, with comparison to conditioning using HDRev [48] restored image. Directly concatenating events and LDR image cannot fully utilize events, thus we introduce an event-image encoder to fuse them as shown in the orange part. The HDRev [48] restored images by decoding HDR features can serve as input to the embedding module, as shown in the yellow part. The proposed pyramid fusion module, as shown in the blue part, fuses different levels HDR features to provide conditions.

3.2. Event-guided Conditioning and Generation

To hallucinate missing information for HDR reconstruction, previous methods mainly come in two ways: adopting better hallucination models (*e.g.*, diffusion models [22]) or introducing HDR sensors (*e.g.*, event cameras [13, 14, 48]). Diffusion models generate high-quality and realistic images by modeling real-world image distributions with diffusion priors. Event cameras provide differential HDR intensity, making events well-suited as conditions for diffusion models to generate more accurate details in over-/under-exposed areas, thereby compensating the dynamic range of LDR images. Thus we propose to leverage both event cameras and diffusion models to hallucinate realistic HDR images to be more consistent with real scenes.

To generate images corresponding to conditions, conditional models [18, 50] are proposed to generate features injected into specific layers of the well-trained denoiser. ControlNet [50] is one of the widelyconditional models for image restoration [23, 43, 44], which generates and injects features for each down-sample and middle layers of the denoiser. As illustrated in the interaction between the blue and green parts of Figure 2, we perform feature injection similar to ControlNet.In the green part, we utilize Stable Diffusion [34] V1.5 as the pretrained denoiser ϵ_θ to denoise from the initial latent z_N step by step. The generated image is obtained by decoding the estimated initial latent $\hat{z}_0$ with VAE decoder. With an image-like input, ControlNet employs an embedding module to generate conditions, then adopts a control module to obtain injected features. However, it is challenging for the embedding module to effectively fuse LDR image with events and extract useful information.

Events record temporal intensity changes, making them distinct from LDR images. Directly concatenating stacked events E and LDR images I_{LDR} together as the input for the embedding module $\mathcal{T}_{\text{con}}$ introduces HDR information, where the conditioning process[1] is:

$$\mathcal{E}_{\text{ConCond}}(I_{\text{LDR}}, E) = \mathcal{T}_{\text{con}}([I_{\text{LDR}}, E]), \tag{5}$$

$[,]$ denotes concatenation. But it cannot fully utilize HDR information encoded in events, *e.g.*, decoration and cable in over-exposed areas cannot be accurately reconstructed in Figure 4 (d). Therefore, a specifically designed HDR information extraction model is needed. HDRev [48] performs event-guided HDR reconstruction through representation alignment and feature fusion, which addresses the modality gap between events and LDR images. It performs representation alignment through pre-training to obtain modality-specific encoders as shown in the orange part of Figure 3. The LDR image and events are encoded into modality-specific features, represented by green features (LDR image) and yellow features (events), respectively. Those two features are fused through the fusion module to obtain HDR features. The restoration process can be written into:

$$I_{\text{restored}} = \mathcal{B}(\mathcal{H}(I_{\text{LDR}}, E)), \tag{6}$$

where $\mathcal{B}$ and $\mathcal{H}$ are the decoder and encoder in HDRev [48]. The restoration process utilizes events to recover more details, as shown by Figure 4 (e), which offers a better reconstruction of the decoration and cable than (d). The intuitive way to better utilize events for conditioning is adopting I_{restored} as input for the embedding module $\mathcal{T}_{\text{rest}}$ as shown in the yellow part of Figure 3. The conditioning process[1] is:

$$\mathcal{E}_{\text{RestCond}}(I_{\text{LDR}}, E) = \mathcal{T}_{\text{rest}}(\mathcal{B}(\mathcal{H}(I_{\text{LDR}}, E))). \tag{7}$$

Diffusion models improve image quality as shown by the improved shape and color of the table in Figure 4 (f) compared to (e). Adopting I_{restored} as conditions ("RestCond") better utilizes events and provides more details than "ConCond", *e.g.*, the decoration in Figure 4 (f) is better than (d).

Although HDRev is an encoder-decoder architecture, the fusion between LDR image and events is performed only in the encoder. Applying I_{restored} as conditions not only suffers from a decoder-encoder redundant computation $\mathcal{T}_{\text{rest}}(\mathcal{B}(\cdot))$ as shown in the yellow part of Figure 3, but also fails to reconstruct details from events, *e.g.*, the shape of table in the red box of Figure 4, and is difficult to adjust the proper brightness of over-exposed areas, *e.g.*, the cloud in Figure 1 (e). The embedding module $\mathcal{T}_{\text{rest}}$ may misunderstand

[1]More details are provided in the supplementary materials.

Figure 4. Comparison with different conditioning ways: directly concatenating events and LDR image ("ConCond") and HDRev [48] restored image ("RestCond"). "ConCond" (d) cannot fully utilize events in over-exposed areas to reconstruct cable. Illustrated by "RestCond" (f), the unsatisfactory results of decoration and table in "Restored" (e) failed to provide HDR information for conditioning in over-exposed areas. The proposed conditioning module can restore missing information as depicted in (g).

these unsatisfactory areas as shown by the shape of the decoration in the red box of Figure 4 (f).

To mitigate issues of poor brightness adjustment in HDRev, and leverage its advantage in event-image fusion, we adopt HDR features from each layer of its event-image encoder to provide conditions. Directly adopting HDR features as conditions is not only more efficient, because of removing the redundant decoder-encoder process, but also more effective, because of avoiding introducing unsatisfactory reconstruction from the decoder $\mathcal{B}$. However, the HDR features are pyramidal, in which different levels of visual information are encoded separately. To fuse different level HDR features, we propose a pyramid fusion module $\mathcal{P}$ as shown in the blue part of Figure 3, which fuses different level features with down-sample layers and feature concatenation to provide conditions. More accurate details and colors can be recovered with it as shown in Figure 4 (g). The proposed conditioning process $\mathcal{E}$ consists of an event-image encoder and pyramid fusion module, denoted by:

$$\mathcal{E}(I_{\text{LDR}}, E) = \mathcal{P}(\mathcal{H}(I_{\text{LDR}}, E)). \tag{8}$$

Then the predicted noise is $\epsilon_\theta(z_n, n, \mathcal{C}(\mathcal{E}(I_{\text{LDR}}, E))$, where $\mathcal{C}$ is the control module derived from ControlNet [50]. With the pretrained denoiser ϵ_θ, we are able to train the control module $\mathcal{C}$ and conditioning process $\mathcal{E}$ to hallucinate HDR images by minimizing the noise loss:

$$\mathcal{L}_{n,z_0,\epsilon_n} = \|\epsilon_n - \epsilon_\theta(z_n, n, \mathcal{C}(\mathcal{E}(I_{\text{LDR}}, E)))\|^2. \tag{9}$$

3.3. Fine-grained Detail Refinement

Although diffusion models can reconstruct realistic HDR images, it suffers from distortion, especially for fine-grained details as shown in Figure 1 (f) and Figure 5 "W/o Refinement". Therefore, we aim to introduce information from events and LDR image to refine those distortion. Typically, we adopt HDR features from the event-image encoder for refinement. As shown in the orange part of Figure 2, our refinement module is a decoder $\mathcal{D}$ in which the up-sample layers are initialized from the VAE decoder. It fuses HDR features $\mathcal{H}(I_{\text{LDR}}, E)$ with latent z_0 to provide the refined HDR image H:

$$H = \mathcal{D}(\mathcal{H}(I_{\text{LDR}}, E), z_0). \tag{10}$$

By training the refinement module with the Ground Truth (GT), we can refine some distortion as shown in Figure 5 "W/o Structure". Due to the inconsistent color between GT and diffusion results as illustrated in Figure 5, applying GT as a supervision target brings color artifacts similar to regression-based methods [24, 48] as depicted in the red box of Figure 5 "W/o Structure". Those artifacts come from the discrepancy between the deterministic GT and the uncertainty in diffusion results during the training process of refinement. [2] Additionally, it also depicts that the details in diffusion results have been wiped (the road curb in the red box) after this refinement. Because it is easier for the refinement to minimize the total loss by reducing the brightness and color difference than the detail difference.

Meanwhile, although the diffusion results suffer from distortion, they already have visually pleasant brightness and color consistent with LDR images. Therefore, we tend to maintain the brightness and color properties of diffusion results and only focus on detail refinement by introducing local histogram matching [45] to adjust the GT to serve as a new supervision target. Local histogram matching [45] is able to adjust the pixel intensity distribution to match the histogram of a target image for each local area, which is suitable for adjusting only color and brightness, and maintaining details of GT. Adopting local histogram matching to adjust ground truth image O to diffusion result H_{diff} is as:

$$O_{\text{adj}} = \text{Hist}(O, H_{\text{diff}}). \tag{11}$$

Our histogram-based structure loss adopts the adjusted image O_{adj} as ground truth and is composed of two losses. The first one is MSE loss:

$$\mathcal{L}_{\text{MSE}}(H, O_{\text{adj}}) = \|H - O_{\text{adj}}\|^2, \tag{12}$$

and the perceptual loss $\mathcal{L}_{perc}$ based on the feature maps extracted by the pretrained VGG-16 [36] network:

$$\mathcal{L}_{\text{perc}} = \sum_l \|\phi_l(H) - \phi_l(O_{\text{adj}})\|^2 + \|\mathcal{G}_l^\phi(H) - \mathcal{G}_l^\phi(O_{\text{adj}})\|^2, \tag{13}$$

[2] Diverse diffusion results are shown in the supplementary materials.

Figure 5. Validation for the effectiveness of the structure loss. With only conditioning and diffusion process, the results of "W/o Refinement" suffer from distortion as shown in the green box. Training the refinement module with ground truth, some distortions are refined, but introducing unnatural saturation adjustment and unsatisfactory detail prediction as shown in the red box of "W/o Structure". We introduce local histogram matching [45] to adjust the brightness and color of ground truth to diffusion results, making the refinement module focus on refining the details. The adapted ground truth is shown in "Adapted GT". As shown in "Proposed", by adopting adapted ground truth as a supervision target, the refinement module with structure loss refines distortion with more natural colors and details.

where ϕ_l is the extracted feature from l-th layer of VGG-16, $\mathcal{G}_l^\phi$ is the Gram matrix of ϕ_l. Our structure loss is:

$$\mathcal{L}_{\text{struct}} = \alpha \mathcal{L}_{\text{MSE}} + \beta \mathcal{L}_{\text{perc}}, \tag{14}$$

where $\alpha = 0.01, \beta = 0.001$ are the balance weights of different terms. Using O_{adj} as the supervision target enables the refinement module to focus on fine-grained detail refinement as illustrated in Figure 5 "Complete", in which the building is undistorted and the road curb is natural.

3.4. Training details

As a large number of paired events, LDR image, and GT is needed for training, we adopt 733 HDR images collected by Yang *et al.* [48] and follow the same data generation process to generate training and testing data with resolution $(512, 512)$. We first simulate 733 HDR videos with random global motion by generating random camera motion trajectories [3]. We utilize the event simulator [31] and virtual camera [6] to simulate the events and LDR images, respectively. The dataset generated from 733 images is separated into training and testing following the setting of Yang *et al.* [48], while 663 for training and 70 for testing. During training, we randomly over- or under-exposed 20% to 50% pixels of the HDR images to generate LDR images.

Other implementation details. The proposed method is implemented with the Pytorch framework and runs on a single NVIDIA GeForce RTX 3090 GPU. For optimization, we use the ADAM optimizer [19] with a linear decay learning rate scheduler starting from step 0 and an initial learning rate of $\gamma = 10^{-5}$. DDIM noise scheduler is adopted both in training and sampling, with training steps $N = 1000$ and linear-scaled beta schedule. We adopt $C = 5$ same to HDRev [48] to apply their pretrain model. During sampling, we adopt classifier-free guidance with guidance scale $u = 1.5$ and inference steps $n = 9$ to obtain our results.

Table 1. Quantitative evaluation of synthetic data on dataset collected by Yang *et al.* [48]. The red metric shows the best performance. ↑ (↓) means higher (lower) is better.

	PSNR↑	SSIM↑	LPIPS↓	CIEDE↓	FID↓	NIQE↓
Liu *et al.* [24]	18.35	0.771	0.276	15.33	78.41	4.02
EventHDR [53]	11.04	0.334	0.447	23.44	182.11	4.58
Sagiri [22]	12.50	0.453	0.414	17.67	83.46	5.35
HDRev [48]	14.05	0.619	0.238	17.69	46.23	3.88
NeurImg [14]	18.53	0.621	0.338	17.42	105.07	4.04
Dille *et al.* [4]	19.73	0.820	0.243	9.76	76.83	4.22
Ours	25.67	0.926	0.099	6.01	27.09	3.86

4. Experiments

The comparison with existing methods including four categories: event-based EventHDR [53], single-image CNN-based Liu *et al.* [24] and Dille *et al.* [4], single-image diffusion-baed Sagiri [22], and event-guided NeurImg-HDR [14] and HDRev [48]. Experiments are conducted both on synthetic data and real data, as described in Section 4.1 and Section 4.2 respectively.

Metrics For fidelity preserving, we adopt reference metrics including peak signal-to-noise ratio (PSNR), structural similarity (SSIM), the perceptual error with learned perceptual image patch similarity (LPIPS) [51], the Frechet Inception Distance (FID) [15], and the Color Difference Evaluation 2000 (CIEDE) [25], respectively. A non-reference metric named the Natural Image Quality Evaluator (NIQE) [27] is adopted to evaluate the image quality.

4.1. Evaluation on Synthetic Data

The quantitative evaluation is reported in Table 1. The proposed method outperforms the previous methods in terms of fidelity and structure preservation as shown by reference metrics. Also, ours image quality outperforms other methods, as indicated by NIQE [27]. The qualitative evaluation is illustrated in Figure 6. EventHDR [53] reconstructs faithful grayscale HDR images without color. Liu *et al.* [24]

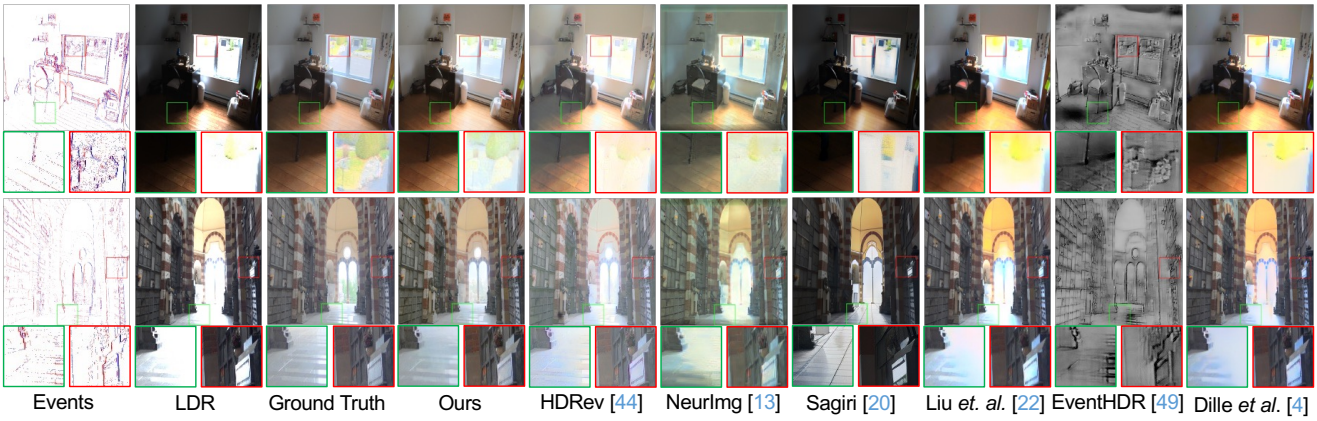

Figure 6. Qualitative evaluation of synthetic data on dataset collected by Yang *et al.* [48].

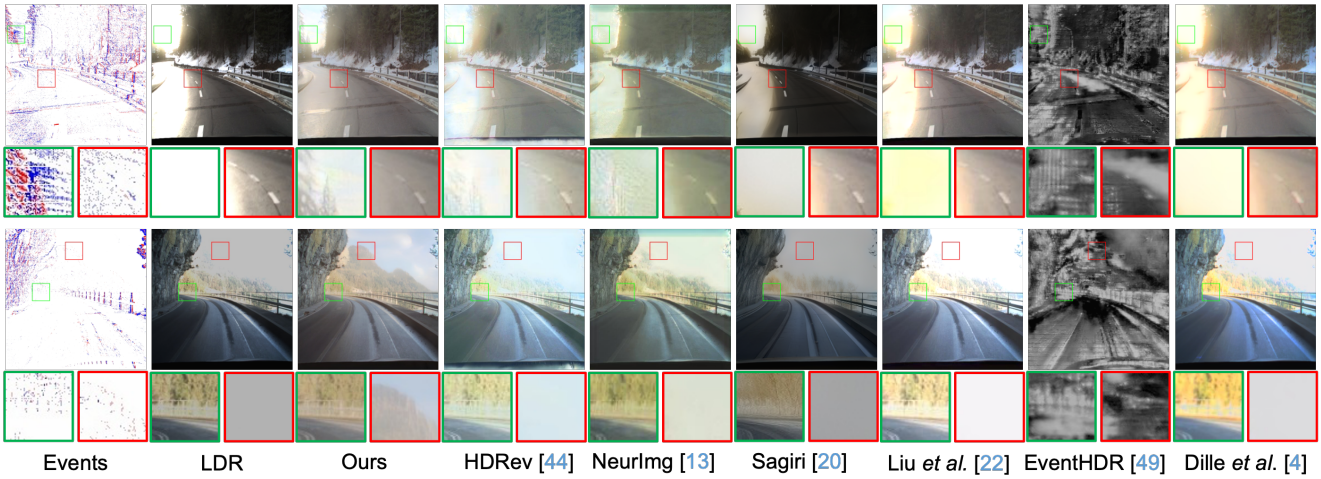

Figure 7. Qualitative evaluation of real data on DSEC [8].

Table 2. Quantitative evaluation (NIQE [27]↓/MANIQA [47]↑/QualiClip [1]↑) of real data. Blue metrics indicate the second-best.

Dataset	Liu *et al.* [24]	EventHDR [53]	Sagiri [22]	HDRev [48]	NeurImg [14]	Dille *et al.*	Ours
DSEC [8]	(3.65/0.466)	(4.00/0.270)	(5.87/0.617)	(3.41/0.402)	(3.80/0.570)	(3.63/0.467)	(3.12/0.693)
HES-HDR [14]	(5.67/0.335)	(5.52/0.203)	(6.40/0.576)	(5.01/0.322)	(4.61/0.442)	(5.72/0.367)	(4.81/0.530)

maintains details from LDR images but is hard to compensate for missing information. Sagiri [22] adopts diffusion models to provide high-quality images but fails to preserve consistency with inputs. NeurImg [14] has difficulty combining events with LDR image to predict over-exposed areas and has noise-like artifacts. HDRev [48] recovers HDR information, while it is difficult to deal with brightness and color adjustment in over-/under-exposed areas. Our method provides natural results and outperforms other methods both on missing information generation and fidelity preservation.

4.2. Evaluation on Real Data

Comparison on DSEC dataset DSEC [8] is a stereo event camera dataset for driving scenarios. We choose parts[3] of data recording HDR scenes as our testing data. The quantitative evaluation using non-reference metrics is shown in Table 2, indicating that the proposed method achieves higher image quality than others. The qualitative evaluation of the DSEC dataset is depicted in Figure 7. EventHDR [53] fails to restore grayscale results. It's challenging for Liu *et al.* [24] and Sagiri [22] to hallucinate missing information. NeurImg [14] and HDRev [48] have problems utilizing the HDR information provided by events. The proposed method not only provides realistic and faithful results compared with other methods but also exhibits generation ability even with fewer events.

Comparison on HES-HDR dataset HES-HDR [14] is a hybrid event and spike HDR dataset. Its hybrid event dataset is suitable for our task. Although the non-reference metrics are not the best, we show better recovery ability as shown by qualitative evaluation[3]. The qualitative evalua-

[3]More details are provided in the supplementary materials.

Figure 8. Qualitative evaluation of real data on HES-HDR [14].

Table 3. Ablation studies of synthetic data on the dataset collected by Yang *et al.* [48]. "Proposed*" is calculated with adjusted GT O_{adj}. The red metric shows the best performance among different ablation parts. ↑ (↓) means higher (lower) is better.

	PSNR↑	SSIM↑	LPIPS↓	CIEDE↓	FID↓	NIQE↓
ConCond	22.60	0.749	0.200	7.48	48.14	4.23
RestCond	20.79	0.725	0.247	9.33	61.74	4.20
W/o Pyramid	21.62	0.728	0.223	8.67	46.99	3.89
W/o Refinement	23.29	0.77	0.170	7.29	38.13	3.63
W/o structure	26.65	0.933	0.093	5.67	25.90	3.90
Proposed	25.67	0.926	0.099	6.01	27.09	3.86
Proposed*	29.77	0.939	0.080	3.79	22.62	3.86

tion of the HES-HDR dataset is depicted in Figure 8, which shows similar conclusions as DSEC [8]. Besides, even with unnatural color-shifted LDR image, our results have a more natural appearance and better fidelity.

4.3. Ablation Study

Our ablation study includes two parts, different conditioning processes, and different refinement targets. The quantitative evaluations of different conditioning processes are shown in the upper part of Table 3. "'ConCond" and "RestCond" are as described in Section 3.2. "W/o Pyramid" represents removing the pyramid fusion module and adopting the lowest HDR features as conditions. The proposed one ("W/o Refinement") outperforms others. The qualitative evaluations in Figure 4 show that the proposed one can better utilize events and LDR information to provide faithful and realistic images. The quantitative evaluations of refinement and structure loss are depicted in the lower parts of Table 3. The ablation of removing the refinement module is shown by "W/o Refinement". Compared with diffusion-only method "W/o Refinement", the refinement module improves the fidelity in terms of supervised metrics. Employing GT as a supervision target ("W/o structure") makes the reference metrics a little better than "Proposed". However, due to the different supervision targets, directly comparing the proposed method with GT cannot show our structure-preserving capacity. Therefore, we include "Proposed*", which is calculated with the adjusted GT O_{adj}. The image quality reflected by NIQE of "Proposed" is better than "W/o structure". As shown by Figure 5, more natural and higher-quality images without distortion can be generated with the proposed refinement module.

5. Conclusion

In this paper, we present a fidelity-preserving HDR reconstruction method based on diffusion models. Our method consists of two parts, event-guided conditioning generation and fine-grained detail refinement. We utilize an event-image encoder and a pyramid fusion module to efficiently and effectively fuse events and images, providing HDR conditions to guide the diffusion process and generate missing information. A refinement module with histogram-based structure loss is proposed to tackle distortion without changing the brightness and contrast. Experiments on both synthetic and real data demonstrate the generation and fidelity preservation ability of our method.

Limitations. Since diffusion models have uncertainty in the generation process, it is difficult to apply our method for video generation, for the challenge of maintaining consistency over consecutive frames.

Acknowledgement

This work was supported by National Natural Science Foundation of China (Grant No. 62088102, 62136001), Beijing Natural Science Foundation (Grant No. L233024), and Beijing Municipal Science & Technology Commission, Administrative Commission of Zhongguancun Science Park (Grant No. Z241100003524012). PKU-affiliated authors thank openbayes.com for providing computing resource.

References

[1] Lorenzo Agnolucci, Leonardo Galteri, and Marco Bertini. Quality-aware image-text alignment for opinion-unaware image quality assessment. *arXiv preprint arXiv:2403.11176*, 2024. 7

[2] Hrishav Bakul Barua, Kalin Stefanov, KokSheik Wong, Abhinav Dhall, and Ganesh Krishnasamy. GTA-HDR: A large-scale synthetic dataset for HDR image reconstruction. *arXiv preprint arXiv:2403.17837*, 2024. 2

[3] Giacomo Boracchi and Alessandro Foi. Modeling the performance of image restoration from motion blur. *IEEE TIP*, 2012. 6

[4] Sebastian Dille, Chris Careaga, and Yağız Aksoy. Intrinsic single-image hdr reconstruction. In *ECCV*, 2024. 1, 6

[5] 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. 1, 2

[6] Yuki Endo, Yoshihiro Kanamori, and Jun Mitani. Deep reverse tone mapping. *ACM Transactions on Graphics (Proc. of ACM SIGGRAPH Asia)*, 2017. 2, 6

[7] Ben Fei, Zhaoyang Lyu, Liang Pan, Junzhe Zhang, Weidong Yang, Tianyue Luo, Bo Zhang, and Bo Dai. Generative diffusion prior for unified image restoration and enhancement. In *CVPR*, 2023. 2

[8] Mathias Gehrig, Mario Millhäusler, Daniel Gehrig, and Davide Scaramuzza. E-RAFT: Dense optical flow from event cameras. In *International Conference on 3D Vision (3DV)*, 2021. 7, 8

[9] 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

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 2

[11] Abhishek Goswami, Aru Ranjan Singh, Francesco Banterle, Kurt Debattista, and Thomas Bashford-Rogers. Semantic aware diffusion inverse tone mapping. *arXiv preprint arXiv:2405.15468*, 2024. 2

[12] Yuanshen Guan, Ruikang Xu, Mingde Yao, Ruisheng Gao, Lizhi Wang, and Zhiwei Xiong. Diffusion-promoted HDR video reconstruction. *arXiv preprint arXiv:2406.08204*, 2024. 2

[13] 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 *CVPR*, 2020. 2, 4

[14] Jin Han, Yixin Yang, Peiqi Duan, Chu Zhou, Lei Ma, Chao Xu, Tiejun Huang, Imari Sato, and Boxin Shi. Hybrid high dynamic range imaging fusing neuromorphic and conventional images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2, 4, 6, 7, 8

[15] 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. *NeurIPS*, 2017. 6

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 2, 3

[17] Tao Hu, Qingsen Yan, Yuankai Qi, and Yanning Zhang. Generating content for hdr deghosting from frequency view. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 2

[18] Xuan Ju, Xian Liu, Xintao Wang, Yuxuan Bian, Ying Shan, and Qiang Xu. Brushnet: A plug-and-play image inpainting model with decomposed dual-branch diffusion. In *Proceedings of the European Conference on Computer Vision (ECCV)*. ECCV, 2024. 4

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

[20] Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin, and Jiří Matas. Deblurgan: Blind motion deblurring using conditional adversarial networks. In *CVPR*, 2018. 2

[21] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, 2017. 2

[22] Baiang Li, Sizhuo Ma, Yanhong Zeng, Xiaogang Xu, Youqing Fang, Zhao Zhang, Jian Wang, and Kai Chen. Sagiri: Low dynamic range image enhancement with generative diffusion prior. *arxiv*, 2024. 1, 2, 4, 6, 7

[23] Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Bo Dai, Fanghua Yu, Wanli Ouyang, Yu Qiao, and Chao Dong. DiffBIR: Towards blind image restoration with generative diffusion prior. *arXiv preprint arXiv:2308.15070*, 2024. 2, 4

[24] 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 *CVPR*, 2020. 1, 2, 5, 6, 7

[25] M Ronnier Luo, Guihua Cui, and Bryan Rigg. The development of the cie 2000 colour-difference formula: Ciede2000. *Color Research & Application: Endorsed by Inter-Society Color Council, The Colour Group (Great Britain), Canadian Society for Color, Color Science Association of Japan, Dutch Society for the Study of Color, The Swedish Colour Centre Foundation, Colour Society of Australia, Centre Français de la Couleur*, 26(5):340–350, 2001. 6

[26] 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 *CVPR*, 2022. 2

[27] Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2013. 6, 7

[28] Mohammad Mostafavi, Lin Wang, and Kuk-Jin Yoon. Learning to reconstruct hdr images from events, with applications to depth and flow prediction. *IJCV*, 2021. 1, 2

[29] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and

Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *ICLR*, 2024. 2

[30] Yohan Poirier-Ginter and Jean-François Lalonde. Robust unsupervised stylegan image restoration. In *CVPR*, 2023. 2

[31] Henri Rebecq, Daniel Gehrig, and Davide Scaramuzza. ESIM: An open event camera simulator. In *Proc. of Conference on Robotics Learning*, 2018. 6

[32] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. Events-to-video: Bringing modern computer vision to event cameras. In *CVPR*, 2019. 1, 2, 3

[33] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. High speed and high dynamic range video with an event camera. *IEEE TPAMI*, 2019. 2, 3

[34] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2, 3, 4

[35] Richard Shaw, Sibi Catley-Chandar, Ales Leonardis, and Eduardo Pérez-Pellitero. HDR reconstruction from bracketed exposures and events. In *BMVC*, 2022. 2

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

[37] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, 2015. 2

[38] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 3

[39] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021. 2

[40] 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 *ICCV*, 2023. 2

[41] 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 *CVPR*, 2019. 2

[42] Zejia Weng, Xitong Yang, Zhen Xing, Zuxuan Wu, and Yu-Gang Jiang. Genrec: Unifying video generation and recognition with diffusion models. *arXiv preprint arXiv:2408.15241*, 2024. 2

[43] Rongyuan Wu, Tao Yang, Lingchen Sun, Zhengqiang Zhang, Shuai Li, and Lei Zhang. Seesr: Towards semantics-aware real-world image super-resolution. In *CVPR*, 2024. 2, 4

[44] Wenhan Wu, Xian Liu, and Jie Chen. Image restoration with controlnet. *IEEE Transactions on Image Processing*, 2023. 4

[45] Garima Yadav, Saurabh Maheshwari, and Anjali Agarwal. Contrast limited adaptive histogram equalization based enhancement for real time video system. In *2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 2014. 5, 6

[46] 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*, 2023. 2

[47] Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. Maniqa: Multi-dimension attention network for no-reference image quality assessment. In *CVPR*, 2022. 7

[48] Yixin Yang, Jin Han, Jinxiu Liang, Imari Sato, and Boxin Shi. Learning event guided high dynamic range video reconstruction. In *CVPR*, 2023. 1, 2, 4, 5, 6, 7, 8

[49] Fanghua Yu, Jinjin Gu, Zheyuan Li, Jinfan Hu, Xiangtao Kong, Xintao Wang, Jingwen He, Yu Qiao, and Chao Dong. Scaling up to excellence: Practicing model scaling for photo-realistic image restoration in the wild. In *CVPR*, 2024. 2

[50] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 2, 4, 5

[51] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 6

[52] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Unsupervised event-based learning of optical flow, depth, and egomotion. In *CVPR*, 2019. 3

[53] Yunhao Zou, Yinqiang Zheng, Tsuyoshi Takatani, and Ying Fu. Learning to reconstruct high speed and high dynamic range videos from events. In *CVPR*, 2021. 1, 2, 6, 7