Color Event Enhanced Single-Exposure HDR Imaging

Mengyao Cui [1, 2*], Zhigang Wang [2*], Dong Wang [2†], Bin Zhao [2, 3†], Xuelong Li [2, 3]

[1]The University of Hong Kong
[2]Shanghai AI Laboratory
[3]Northwestern Polytechnical University
cuimengyao@connect.hku.hk, {wangzhigang, zhaobin, wangdong}@pjlab.org.cn, li@nwpu.edu.cn

Abstract

Single-exposure high dynamic range (HDR) imaging aims to reconstruct the wide-range intensities of a scene by using its single low dynamic range (LDR) image, thus providing significant efficiency. Existing methods pay high attention to restoring the luminance by inversing the tone-mapping process, while the color in the over-/under-exposed area cannot be well restored due to the information loss of the single LDR image. To address this issue, we introduce color events into the imaging pipeline, which record asynchronous pixel-wise color changes in a high dynamic range, enabling edge-like scene perception under challenging lighting conditions. Specifically, we propose a joint framework that incorporates color events and a single LDR image to restore both content and color of an HDR image, where an exposure-aware transformer (EaT) module is designed to propagate the informative hints, provided by the normal-exposed LDR regions and the event streams, to the missing areas. In this module, an exposure-aware mask is estimated to suppress distractive information and strengthen the restoration of the over-/under-exposed regions. To our knowledge, we are the first to use color events to enhance single-exposure HDR imaging. We also contribute corresponding datasets, consisting of synthesized datasets and a real-world dataset collected by a DAVIS346-color camera. The datasets can be found at **https://www.kaggle.com/datasets/mengyaocui/ce-hdr**. Extensive experiments demonstrate the effectiveness of the proposed method.

Introduction

High dynamic range (HDR) imaging techniques are widely studied, aiming at reconstructing wide-range intensities in real scenes from low dynamic range (LDR) images. LDR images are produced in large quantities by conventional cameras, which barely keep details of both bright and dark parts at the same time. In this way, effectively restoring the intensity of LDR images is of great significance.

Multi-exposure HDR imaging is the most common approach, where multiple-exposed LDR images are aligned and merged to obtain an HDR image(Niu et al. 2021; Choi et al. 2020). Although achieving great successes, such methods cannot handle scenes with large motions and usually need more computations(Wang and Yoon 2022). In contrast, single-exposure HDR imaging (Wang et al. 2022; Liu et al. 2020) uses a single LDR image for restoration, avoiding the above drawbacks. However, as shown in Figure. 1, single-exposure HDR imaging methods lack sufficient information to reconstruct details in over-/under-exposed regions, making it an ill-posed problem. Mono event camera (Chen et al. 2021b; Han and et al 2020) has been used to help restore such saturated regions, but it merely provides grayscale data to complement contour information, the color distortion problem still remains unresolved, *i.e.* , the color of the tower.

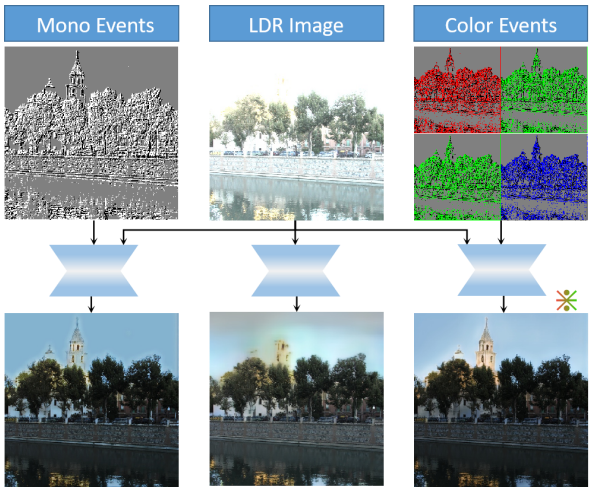

Figure 1: Our task (⁎) vs traditional and mono-event-based single-exposure HDR imaging tasks.

Fortunately, color event cameras are exploited recently that can sense the intensity changes of each color in a high dynamic range and output ”event” data. As can be seen from Figure. 1, color event cameras provide massive information in the over-exposed image, appearing in a color-classified edge-like form. Motivated by these intrinsic advantageous characteristics, we propose to introduce color events as auxiliary data to restore more realistic HDR images, especially in over-/under-exposed regions. Considering different intrinsics between the image data and the event stream, we use different encoders that extract the multi-scale features of the LDR image and the color events, respectively. The multi-

*Equal contribution. † Corresponding author.

scale multi-modality features are then fused at each scale using our image-event fusion block.

Although color events compensate for some missing details, we find that fully restoring the color in severely over-/under-exposed areas is still challenging. Specifically, when aggregating a single LDR image and color events into a plain joint framework, the optimizing procedure tends to simply restore the structure line of the missing regions, while the color is ignored. This is because event streams are edge-like data that provide little information in texture-less regions. To mitigate this issue, we design an exposure-aware transformer (EaT) module to carry out color propagation for the missing regions. By feeding the fused image-event feature into the EaT module, the color hints from both the color events and the normal-exposed LDR regions are expected to propagate to other regions due to the large receptive field of Transformer (Vaswani et al. 2017). However, directly using the transformer to implement propagation causes some "mottled" results, since the distractive color from over-/under-exposed regions is mistakenly propagated to normal regions. Thus, we further introduce an exposure-aware mask in EaT to suppress the disturbing color and concentrate the propagation on the missing regions.

To our knowledge, this is the first work to explore color event enhanced single-exposure HDR imaging. To validate the proposed method, we first generate paired data using existing datasets and a color event simulator, *i.e.* , ESIM (Scheerlinck et al. 2019). A new dataset is also collected utilizing a DAVIS346 color event camera which can synchronically capture LDR images and corresponding color events. Extensive experiments on these datasets demonstrate the effectiveness of our method.

To summarize, our contributions can be listed as follows.

- We propose a novel task setting for HDR imaging where color event data is first used to enhance single-exposure HDR restoration.
- We design a joint framework to incorporate LDR images and color events. An exposure-aware transformer module is further explored that propagates color hints to strengthen the restoration of severely missing regions.
- We establish both simulated and real-world datasets for color-event-based HDR imaging tasks. Experiments on these datasets demonstrate the priority of our design.

Related Work

High dynamic range (HDR) imaging has many branches, we briefly review the two most related sub-fields to our work.

Single-Exposure HDR Imaging This branch of approaches aims to restore high dynamic range scenes by using a single LDR image, thus avoiding handling intricate misalignment among multiple LDR images in other methods. Expandnet (Marnerides et al. 2018) proposes a multi-scale network to extract different levels of content and merges multi-scale features to reconstruct more details in an HDR image. To alleviate the ill-posed mechanism of pure single-exposure HDR imaging, Kim, Lee, and Kang (2021) design a network to create LDR stacks while preserving correlations among multi-exposure images during the HDR imaging process. Zeng *et al.* (2022) explore the widely used 3-dimensional lookup tables (3D LUTs) and design an image-adaptive method to achieve fast manipulation of color and tone. Liu *et al.* (2020) point out that the detail missing from HDR to LDR images is mainly caused by dynamic range clipping, non-linearization, and quantization during the camera imaging pipeline, and propose a reverse procedure to fulfill LDR-to-HDR transformation. Metzler *et al.* (2020) propose an optical encoder to involve HDR information of the camera lens and use it as attached hardware during inference. Wang *et al.* (2022) introduce knowledge that formulates the HDR image generation process to traditional UNet architecture. Although various methods are exploited, reconstructing an HDR image only from a single LDR image remains a challenging problem.

HDR Imaging with Novel Sensors To introduce more realistic information into HDR imaging, this branch focuses on employing extra sensors and extracting useful features. Infrared (IR) or thermal cameras are able to capture the contours of thermal objects without losing information dramatically in HDR scenes. Several approaches (Ma et al. 2020; Li and Wu 2019) propose specially designed networks to merge heterogeneous data, *i.e.* , LDR images and IR images, for better HDR imaging. Besides, event cameras, a type of bio-inspired sensor, have attracted increasing attention, which can record intensity changes as the "event" in very high dynamic range scenes. Han *et al.* (2020) design a hybrid camera system consisting of synchronized event and conventional cameras. They use an off-the-shelf intensity reconstruction method to obtain an HDR intensity map and fuse it with a single LDR image for restoration. Wang *et al.* (2020) consider that event cameras may bring noises into image recovery, and propose an event-enhanced joint framework for denoising and super-resolution. Liu *et al.* (2022) fuse single-photon cameras with traditional cameras to recover high-resolution images in extreme situations. Yang *et al.* (2023) develop a representation alignment strategy to fuse events with LDR video. Similarly, Samra, Mitra, and Shedligeri (2023) facilitate HDR video generation by event-guided interpolation. These approaches achieve better results, while they all depend on grayscale auxiliary information, thus cannot ensure to solve color distortion problem. By comparison, we are the first to utilize color events in single-exposure HDR imaging, and further deployed global texture and color via designing transformer and fusion blocks.

Proposed Method

Our work assumes two photosensitive devices which work simultaneously. One of the devices works as a conventional camera capturing an LDR image $I_{LDR} \in \mathbb{R}^{C\times H\times W}$, where C is the channel number, H and W are the height and width, respectively. Another device works as a color event camera, which produces event streams $\{e_j\}_{j=1}^{N}$ during the exposure time of the LDR image. Each event e_j contains a four-attribute tuple $(x_j, y_j, t_j, \omega_j)$, which means an event occurs

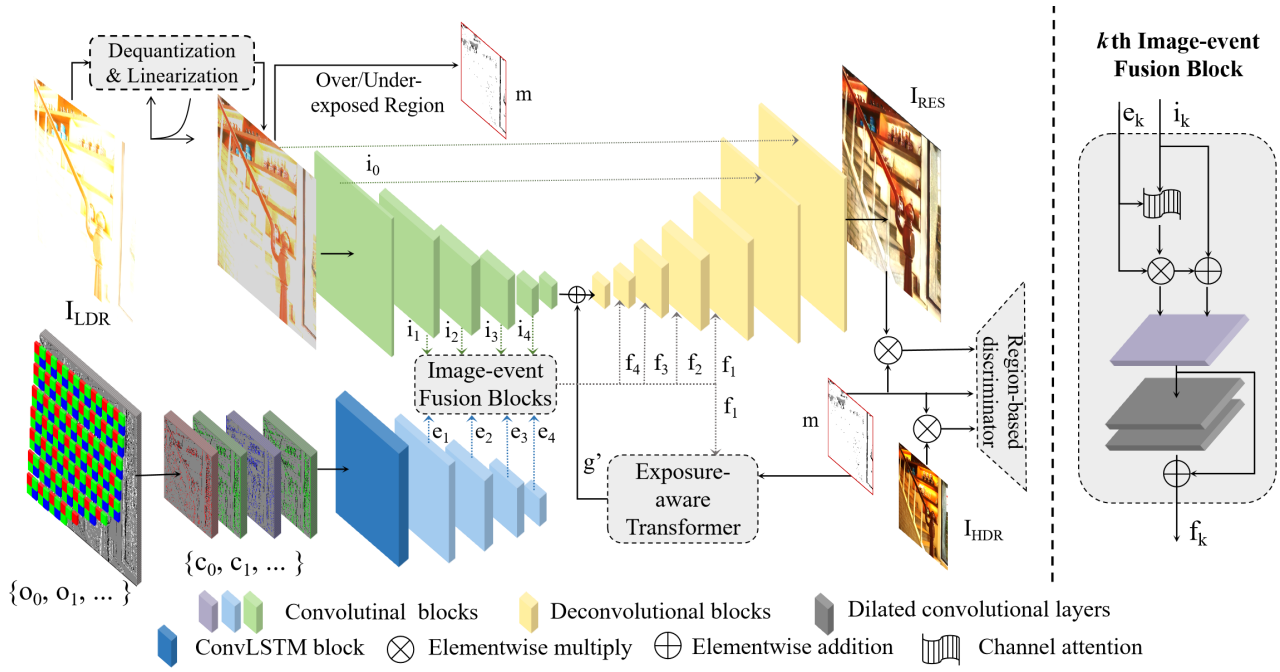

Figure 2: The architecture of our color event enhanced single-exposure HDR imaging method.

at the time $t_j \in \mathbb{R}^+$ and the location $(x_j, y_j) \in \mathbb{Z}^{H\times W}$ with polarity ω_j on the corresponding LDR image. The polarity is a binary value that indicates whether the intensity change is increased or decreased. It is notable that the resolutions of conventional cameras and color event cameras are not necessary to be the same. Accordingly, our work aims at recovering the HDR image from its corresponding LDR image and color event streams.

Preliminary of the Color Event Camera

In extreme lighting scenarios like nighttime or midday wilderness, conventional cameras are prone to failure. Since the luminance exceeds the maximum value or does not reach the minimum value that the sensor can capture, the gradients of over-/under-exposed areas disappear, which leads to the loss of both structure and color information. Such missing information may be restored by the color event camera since its sensor works in a much higher dynamic range (approx. 120dB). In the scenarios of our task, a pixel-wise color event is triggered when a slight motion of the scene or the camera occurs and the corresponding color change exceeds the threshold. This process can be formulated as:

$$\Gamma(p_{t'} - p_t) \approx C \times \omega, \tag{1}$$

where C is the threshold, ω is the polarity of the event, p'_t is the new pixel value at time t', p_t is the original pixel value at time t, and Γ is the filter on the pixel. The color array filter in the DAVIS346-color consists of an RGBG filter pattern (Scheerlinck et al. 2019). Thus, there are three kinds of filters Γ that generate three types of color events: red filter Γ_r, green filter Γ_g, and blue filter Γ_b. In this way, the color information of the events is encoded in the locations of the 2D plane with the filters.

It should be noted that "event" data is recorded continuously and asynchronously. To preserve the temporal information as well as improve efficiency, we convert raw events into voxel grid (Bardow, Davison, and Leutenegger 2016). The raw events are first split into a set of voxels $\{o_0, o_1, ...\}$ with a particular time interval τ. Then each voxel is equally divided into B bins by a smaller time interval. Each bin consists of the added polarity of the event during the time interval on the 2D plane. Here we get $o_i \in \mathbb{R}^{H\times W\times B}$. As shown in lower left corner of the Figure. 2, we transform raw voxels into RGBG color voxels $\{c_0, c_1, ...\}$, where $c_i \in \mathbb{R}^{4\times \frac{H}{2}\times \frac{W}{2}\times B}$ to extract color information.

Color Event Enhanced HDR Reconstruction

Overview As shown in Figure. 2, the process of HDR reconstruction can be divided into several steps: events & LDR feature extractions, image-event feature fusion, and image restoration. Since the HDR image is linear and the LDR image is nonlinear, we first map I_{LDR} to the linear space by estimating the inverse CRF as in (Liu et al. 2020). We also use a ConvLSTM block (Shi et al. 2015) to initially encode the color event data $\{c_0, c_1, ...\}$ which consists of the time series. Then multi-scale image features $\{i_0, i_1, ...\}$ and event features $\{e_1, e_2, ...\}$ are extracted with stacked convolutional and pooling layers. Notably, the event features start from e_1 because the spatial scale of the color events is half of that of the LDR image. After extracting multi-scale features, we fuse the features with image-event fusion blocks, which locally merge color events with the LDR image at different scales. The fused features are then fed into the HDR decoder. For normal-exposed regions, the HDR color of an LDR pixel can be roughly estimated by the adjacent pixel set

on the LDR image and the color changes encoded in color events. However, when it comes to the case of seriously missing regions, it will be difficult to find valid neighbors. As we can see in Figure. 1, in seriously missing regions, the LDR image barely contains color information, while the color events accumulated in voxels contain clear but sparse structural information. Thus we introduce the global context by adding an exposure-aware transformer module EaT, where the valid color hints from normal-exposed LDR regions and color events are propagated to the missing regions adaptively to assist in color restoration. Finally, a region-based discriminator is introduced to enhance the generation of severely missing regions.

Image-Event Fusion Block We design a set of image-event fusion blocks that merge the multi-modality features at different scales (see Figure. 2, right column). With input features (i_k, e_k), we highlight events that are more relevant to image color changes with channel attention (Woo et al. 2018). Then the intensity changes are merged with adjacent pixels by being added with the image feature i_k. The merged feature and the selected event are concatenated and fused by convolutional layers. The dilated convolutional layers with residual connections are also used to increase the receptive field of the network. Finally, we get the fused feature f_k.

The Region-Based Discriminator We use a region-based discriminator to facilitate chrominance generation in the extremely over-/under-exposed region. We extract a binary mask m of the over-/under-exposed region from the linearized LDR image. Specifically, with the threshold θ, the image region with the minimum chrominance value less than θ is viewed as an under-exposure area, while the region with the maximum chrominance value larger than $1-\theta$ is viewed as an over-exposure area. The generated image I_{RES} is multiplied by m to select the target region. I_{RES} and m are then fed into the region-based discriminator as in (Nazeri et al. 2019). The corresponding output scores are calculated as follows:

$$\begin{cases} s(x) = D([I_{HDR} \times m, m]), \\ s(\hat{x}) = D([I_{RES} \times m, m]), \end{cases} \tag{2}$$

where $[\cdot,\cdot]$ means the channel-wise concatenation, D denotes the discriminator, x represents the real image distribution, and $\hat{x}$ represents the model distribution.

Exposure-Aware Transformer Module By introducing the exposure-aware transformer module, we model the long-range color dependency from the fused feature to facilitate the restoration of missing regions. We notice that image regions with different exposures provide color hints in different ways. For the normal-exposed regions, fused image and event features contain abundant and convincing color details. Comparatively, for over-/under-exposed regions, color events $\{c_0, c_1, ...\}$ provide sparse points which reflect the structure line of the scene, while the image provides little useful information.

To exploit different characters in different regions, we estimate a hint-selected mask to guide the color propagation, which aims to select regions with convincing color hints

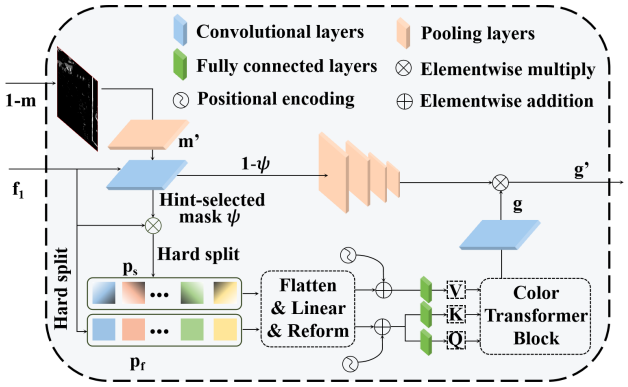

Figure 3: The exposure-aware transformer.

from fused feature f_1 and leave other regions. As shown in Figure. 3, we input the raw mask of $1-m$ and the fused feature f_1 into the EaT module. $1-m$ provides a preliminary knowledge of the normal-exposed regions, which is scaled to the same size with f_1 by the pooling layer. The pooled binary mask is m'. The hint-selected mask ψ is then estimated from m' and fused feature f_1:

$$\psi = \epsilon([m', f_1]), \tag{3}$$

where $\epsilon(\cdot)$ denotes a conv1 $\times$ 1 block and $[\cdot,\cdot]$ means the channel-wise concatenation.

Before the propagation, we calculate the masked fused feature s by multiplying the fused feature with ψ, which suppresses the disturbing color from the over-/under-exposed regions. The sequentialization operation is then performed on both the fused feature and the selected fused feature by dividing f_1 and s into smaller patch series p_f and p_s:

$$\begin{cases} pf = \{pf_0, pf_1, ..., pf_{\frac{HW}{P^2}-1}\}, pf_i \in \mathbb{R}^{C\times P\times P}, \\ ps = \{ps_0, ps_1, ..., ps_{\frac{HW}{P^2}-1}\}, ps_i \in \mathbb{R}^{C\times P\times P}. \end{cases} \tag{4}$$

In this equation, P is the height and width of the patches. The token matrix $T_f, T_s \in \mathbb{R}^{C\times M}$ are then generated as follows:

$$\begin{cases} T_f = g_1(pf) + E_{pos}, \\ T_s = g_2(ps) + E_{pos}, \end{cases} \tag{5}$$

where $g_*(\cdot)$ denotes the actions of flattening, linear projection, and reformulation, and E_{pos} denotes the sinusoidal positional encoding matrix (Dosovitskiy et al. 2020).

When fed into the color transformer block, we use T_f to generate the query and key which compute the feature similarities in self-attention and use T_s to generate the value. In this way, we remain the patch similarity when calculating the similarity matrix while only propagating the selected color. The generated feature g is reshaped to the same size as the first layer of the image decoder. To concentrate the propagation on the over/under-regions, we process the generated feature g with hint-selected mask ψ again:

$$g' = g \times (1 - \psi). \tag{6}$$

Finally, we propagate the color hints into the entire image by adding the generated features g' into the first layer of the image decoder.

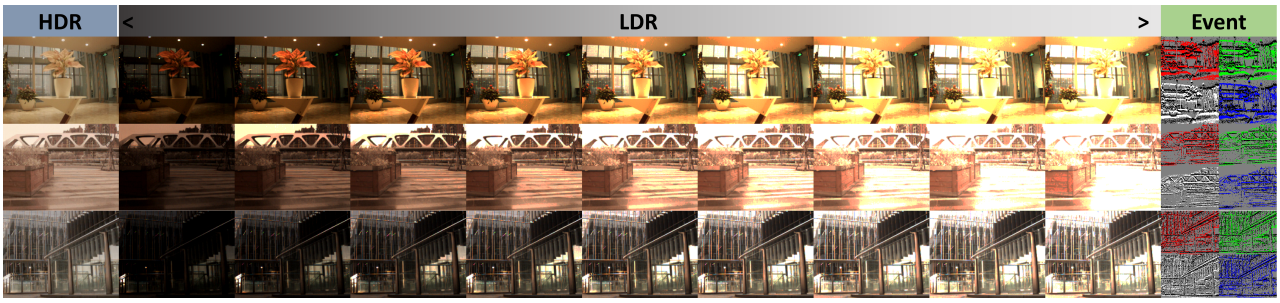

Figure 4: Samples of our HDR and color event (HDR-CE) dataset.

In the decoding stage, the image feature is upsampled with stacked deconvolutional layers. The last two deconvolutional layers are fed image feature i_0 and linearized I_{LDR} with skip connections, respectively. While the other layers are fed the fused feature f_* with skip connections.

Objective Function

We train the HDR reconstruction model in an end-to-end style. We take the logarithm for I_{RES}, I_{HDR}, then apply the L_1 loss and the total variation loss L_{tv}. Logarithmization is found can stabilize training (Liu et al. 2020). The adversarial loss is defined with the following equation:

$$\begin{cases} L_G = -\mathbb{E}_{\hat{x}\sim p_G}[log(s(\hat{x}))], \\ L_D = -\mathbb{E}_{x\sim p_{data}}[log(s(x))] - \mathbb{E}_{\hat{x}\sim p_G}[log(1-s(\hat{x}))]. \end{cases} \tag{7}$$

To enhance perceptual similarity, we further introduce perceptual loss L_{percep}, which can be calculated as:

$$L_{percep} = \mathbb{E}\left[\sum_i \frac{1}{N_i}||\Phi_i(\Lambda(I_{HDR})) - \Phi_i(\Lambda(I_{RES}))||_1\right], \tag{8}$$

where $\Phi_*(\cdot)$ denotes layers of the pre-trained VGG-19 network (Russakovsky et al. 2015), $\Lambda(\cdot)$ denotes a differentiable global tone-mapping operator (Wu et al. 2018).

Thus we get the overall generating loss function:

$$L = \lambda_1 L_1 + \lambda_2 L_{tv} + \lambda_3 L_G + \lambda_4 L_{percep}, \tag{9}$$

where λ_* are hyper-parameters that balance the losses.

Experiment

Experiment Setups

Here we introduce the most important content of the implementation details, datasets, and metrics. In the supplementary material, we provide more details of implementation, datasets, and more comparison results (on PSNR, SSIM).

Implementation Details In our work, all of the images and their corresponding voxel bins are randomly cropped, flipped, and rotated to prevent overfitting. The size of input images is $(256, 256)$, while the size of input voxel bins is $(128, 128)$. For over-/under-exposed image region selection, we empirically generate the raw mask with the threshold $\theta = 0.12$. In the exposure-aware transformer, f_k, s_k are split into 8×8 patches, and further processed as token matrix $T_f, T_s \in \mathbb{R}^{512\times 64}$. In the multi-head self-attention, the number of heads is 8 and the dropout rate is 0.1. The loss weight hyper-parameters are empirically set as: $\lambda_1 = 1, \lambda_2 = 0.1, \lambda_3 = 0.1, \lambda_4 = 0.001$.

Synthetic Datasets We make use of HDR-SYNTH and HDR-REAL datasets from the work of (Liu et al. 2020) to build our synthetic datasets. In detail, HDR-SYNTH collects 562 HDR images, which are split into 503 HDR training images and 60 HDR evaluating images. The corresponding LDR images are synthesized using the camera imaging pipeline (Liu et al. 2020). HDR-REAL consists of 480 HDR images and 4893 LDR images, where each HDR image corresponds to several LDR images. We simulated the color event for each HDR image in the above datasets using the extension of ESIM (Scheerlinck et al. 2019). In total, we simulated **20527** sets of color event data.

Real-World Datasets We build the first real-world dataset for the color event enhanced single-exposure HDR imaging task, which is named the HDR color event (HDR-CE) dataset. A capturing system is built to collect the dataset, which combines a color event camera (Color Davis 346 (Taverni et al. 2018)), a bracket, and a vibration platform. In total, we collect 1000 HDR images, 9000 LDR images, and 1000 sets of color events, a tenth of them are split into evaluating images while the rest are split into training images. Figure. 4 shows examples of our collected data.

Metrics We mainly adopt HDR-VDP-3 (Mantiuk et al. 2011) to evaluate the HDR reconstruction. HDR-VDP is a dedicated and widely used quality metric for HDR images (Wang et al. 2022; Liu et al. 2020), which predicts the visual differences between the reconstructed and the ground-truth HDR images for the observer [1].

Comparison with State-of-Art Methods

We compare our method with six deep learning-based approaches, *i.e.* , HDRCNN (Eilertsen et al. 2017), DrTMO (Endo, Kanamori, and Mitani 2017), ExpandNet (Marnerides et al. 2018), SingleHDR (Liu et al. 2020), Diff (Kim, Lee, and Kang 2021), KUNET (Wang et al. 2022), and four conventional approaches, *i.e.* , AEO (Akyüz

[1] https://github.com/gfxdisp/FovVideoVDP#predicted-quality-scores

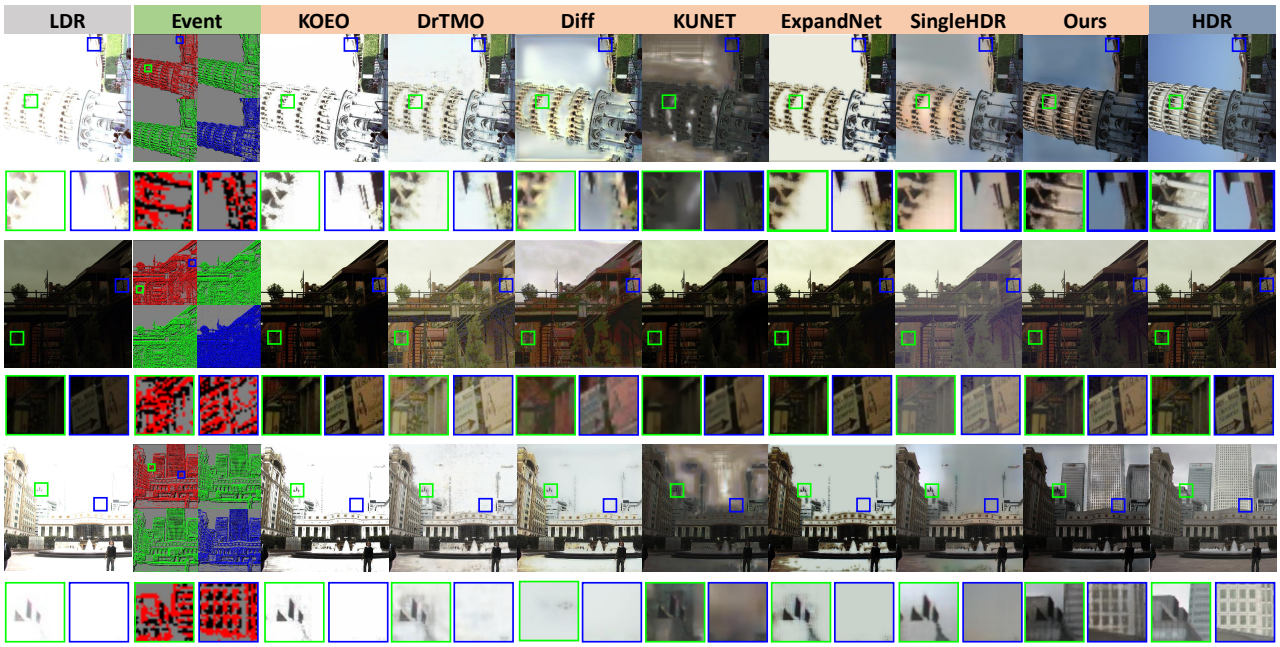

Figure 5: Qualitative comparison on the HDR-SYNTH dataset.

Method	Training dataset	HDR-SYNTH	HDR-REAL
AEO	-	5.12 ± 2.45	4.56 ± 2.18
HPEO	-	4.56 ± 2.11	3.49 ± 2.24
KOEO	-	5.58 ± 2.22	4.93 ± 2.12
MEO	-	5.12 ± 2.35	3.83 ± 2.36
HDRCNN	SYNTH	6.04 ± 1.75	-
HDRCNN	SYNTH, REAL	-	5.04 ± 1.76
DrTMO	SYNTH	6.55 ± 1.49	-
DrTMO	SYNTH, REAL	-	5.10 ± 1.96
Diff	Pre-trained	6.49 ± 1.62	5.24 ± 2.19
KUNET	Pre-trained	5.57 ± 1.37	4.51 ± 1.84
ExpandNet	Pre-trained	5.35 ± 1.90	4.44 ± 1.98
ExpandNet	SYNTH, REAL	6.00 ± 1.95	5.28 ± 2.16
SingleHDR	SYNTH	6.92 ± 1.67	5.32 ± 1.98
SingleHDR	SYNTH, REAL	-	5.93 ± 2.01
Ours	SYNTH	$\mathbf{7.39 \pm 1.16}$	5.61 ± 1.90
Ours	SYNTH, REAL	-	$\mathbf{6.04 \pm 1.68}$

Table 1: Quantitative comparison with ten-point scale HDR-VDP-3 on HDR-SYNTH and HDR-REAL datasets.

et al. 2007), HPEO (Huo et al. 2014), KOEO (Kovaleski and Oliveira 2014), MEO (Masia, Serrano, and Gutierrez 2017). Among those methods, we use pre-trained models of KUNET since it produces 16-bit non-linear HDR images, while HDR-SYNTH, HDR-REAL, and HDR-CE use 32-bit linear HDR images. We also use pre-trained models of Diff since it requires a multi-exposure stack training dataset that has several exposure values, but HDR-CE, HDR-SYNTH, and HDR-REAL are not organized in this way. Pre-trained model is commonly used in similar situations. For instance, Chen *et al.* (2021a) use the pre-trained model of SingleHDR during comparison. All of the other models are compared after re-training. Table. 1 shows the HDR-VDP-3 results on HDR-SYNTH, HDR-REAL datasets, while Table. 2 shows

Method	Training dataset	HDR-CE
Diff	Pre-trained	6.06 ± 1.16
KUNET	Pre-trained	4.93 ± 1.03
ExpandNet	Pre-trained	5.20 ± 1.25
ExpandNet	SYNTH	5.09 ± 1.14
ExpandNet	SYNTH, CE	6.66 ± 1.59
SingleHDR	SYNTH	6.12 ± 1.06
SingleHDR	SYNTH, CE	7.35 ± 1.32
Ours	SYNTH	6.59 ± 1.03
Ours	SYNTH, CE	$\mathbf{7.56 \pm 1.14}$

Table 2: Quantitative comparison with ten-point scale HDR-VDP-3 on HDR-CE dataset.

the HDR-VDP-3 results on HDR-CE datasets.

Results on Synthetic Datasets As shown in Table. 1, the HDR-VDP-3 of our work outperforms the state-of-the-art works by 0.47 on the HDR-SYNTH dataset, and 0.11 on the HDR-REAL dataset, respectively. Compared with other methods, our work produces higher-quality images in both color and structures under various light conditions. As shown in Figure. 5, our method is resilient to over-exposure situations. For instance, while other methods lose the details of the tower in row 1, our method produces a result that is closer to the ground truth. In row 3, other methods fail to reconstruct the over-exposure buildings of the LDR picture. Comparatively, our method reconstructs convincing details of the building with slight color perturbation. In under-exposure situations, our method also has encouraging results. As shown in Figure. 5, row 2, our method generates the image with the most consistent overall tone to the ground truth image. To sum up, our method performs better

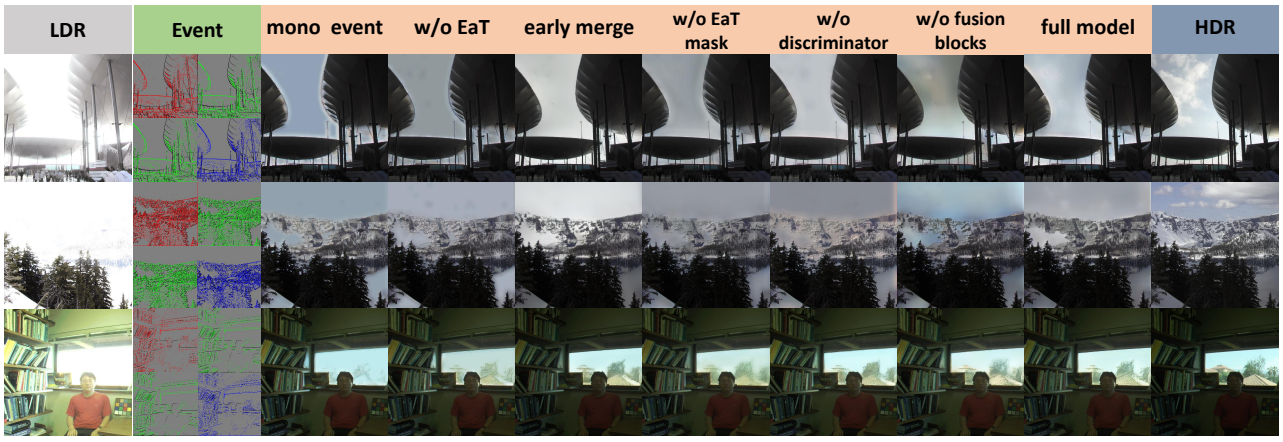

Figure 6: Ablation studies on the HDR-SYNTH dataset.

Method	HDR-VDP-3
SingleHDR	6.92 ± 1.67
mono event	6.98 ± 1.51
w/o EaT	7.14 ± 1.33
w/o EaT mask	7.29 ± 1.20
w/o fusion blocks	7.34 ± 1.14
early merge	7.29 ± 1.21
w/o discriminator	7.32 ± 1.22
full model	$\mathbf{7.39 \pm 1.16}$

Table 3: Ablation studies with ten-point scale HDR-VDP-3 on HDR-SYNTH dataset.

than others in single-exposure HDR imaging tasks.

Results on HDR-CE Dataset We compare our method with the four newest HDR-Imaging methods (ExpandNet, Diff, KUNET, and SingleHDR) on the HDR-CE dataset. Specifically, we fine-tune the methods of ExpandNet, and SingleHDR on the HDR-CE dataset. On the HDR-CE dataset, the HDR-VDP-3 of our work outperforms the state-of-the-art works by 0.21 (see Table. 2). The visual results can be found in the supplementary materials.

Ablation Study

We validate the effectiveness of our design on the HDR-SYNTH dataset with HDR-VDP3.

Ablation Study for the Color Events To validate the effectiveness of the color event, we train our model with only monochrome events and remove the exposure-aware transformer. As shown in Figure. 6, row 3, the result of "mono event" only restores part of the contours outside the window, but no convincing colors or textures. It also achieves a relatively low HDR-VDP-3 (see Table. 3). Comparatively, in Figure. 6, row 3, the result of "w/o EaT", after adding the color events and remaining the removal of the exposure-aware transformer, the window area begins to show more color (like the roof and the tree). The HDR-VDP-3 of the "w/o EaT" is also higher than "mono event" (see Table. 3).

Effectiveness of Exposure-Aware Transformer Based on the model of "w/o EaT", we first add the exposure-aware transformer but remove the hint-selected mask. The test shows that the HDR-VDP-3 of the method increases (see Table. 3, "w/o EaT mask"). In Figure. 6, row 3, the result of "w/o EaT mask", the color and texture of the window area are also further restored. We then add the hint-selected mask to the model, as shown in the result of the "full model". The color and texture of the window area become closer to the ground truth (see Figure. 6, row 3).

Ablation Study for the Architecture To further illustrate the validity of the fusion design, we replace the image-event fusion block with the concatenation of the color event features and image features in the full model. In Figure. 6, row 1 and 2, the results of "w/o fusion blocks" is dirtier than the results of the full model, and the HDR-VDP-3 is also slightly worse in Table. 3. In addition, we test another architecture that feeds the fused features into the encoding stage instead of the decoding stage of the model, as many colorization works merge the color hints and gray images in an early stage (Zhang et al. 2017; Guo, Yang, and Huang 2021). This approach also shows a lower HDR-VDP-3 than the full model (see Table. 3, "early merge").

Effectiveness of the Region-Based Discriminator In this part, we remove the region-based discriminator from the full model. The results of "w/o discriminator" show a less convincing color of the mountain (see Figure. 6, row 2). The HDR-VDP-3 of the "w/o discriminator" is also worse than the full model.

Conclusions

In this paper, we have proposed a novel color event enhanced single-exposure HDR imaging task. We constructed the first simulated and the first real-world dataset for this task and designed a color event enhanced network that effectively fuses the LDR image and the color event in HDR reconstruction. Extensive ablation studies and comparisons to other state-of-the-art methods demonstrate the effectiveness of the proposed method.

Acknowledgements

This work is partially supported by the Shanghai AI Laboratory, National Key R&D Program of China (2022ZD0160100) and the National Natural Science Foundation of China (62106183).

References

Akyüz, A. O.; Fleming, R.; Riecke, B. E.; Reinhard, E.; and Bülthoff, H. H. 2007. Do HDR displays support LDR content? A psychophysical evaluation. *ACM Trans. Graph.*, 26(3): 38–es.

Bardow, P.; Davison, A. J.; and Leutenegger, S. 2016. Simultaneous Optical Flow and Intensity Estimation From an Event Camera. In *Proc. CVPR*, 884–892.

Chen, X.; Liu, Y.; Zhang, Z.; Qiao, Y.; and Dong, C. 2021a. Hdrunet: Single image hdr reconstruction with denoising and dequantization. In *Proc. CVPR*, 354–363.

Chen, Z.; Zheng, Q.; Niu, P.; Tang, H.; and Pan, G. 2021b. Indoor Lighting Estimation Using an Event Camera. In *Proc. CVPR*, 14760–14770.

Choi, S.; Cho, J.; Song, W.; Choe, J.; Yoo, J.; and Sohn, K. 2020. Pyramid Inter-Attention for High Dynamic Range Imaging. *Sensors*, 20(18): 5102.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Eilertsen, G.; Kronander, J.; Denes, G.; Mantiuk, R. K.; and Unger, J. 2017. HDR image reconstruction from a single exposure using deep CNNs. *ACM Trans. Graph.*, 36(6): 1–15.

Endo, Y.; Kanamori, Y.; and Mitani, J. 2017. Deep reverse tone mapping. *ACM Trans. Graph.*, 36(6): 1–10.

Guo, X.; Yang, H.; and Huang, D. 2021. Image inpainting via conditional texture and structure dual generation. In *Proc. ICCV*, 14134–14143.

Han; and et al. 2020. Neuromorphic camera guided high dynamic range imaging. In *Proc. CVPR*, 1730–1739.

Huo, Y.; Yang, F.; Dong, L.; and Brost, V. 2014. Physiological inverse tone mapping based on retina response. *The Visual Computer*, 30: 507–517.

Kim, J.; Lee, S.; and Kang, S.-J. 2021. End-to-end differentiable learning to hdr image synthesis for multi-exposure images. In *Proc. AAAI*, volume 35, 1780–1788.

Kovaleski, R. P.; and Oliveira, M. M. 2014. High-quality reverse tone mapping for a wide range of exposures. In *SIBGRAPI Conference on Graphics, Patterns and Images*, 49–56.

Li, H.; and Wu, X. 2019. DenseFuse: A Fusion Approach to Infrared and Visible Images. *IEEE Trans. Image Process.*, 28(5): 2614–2623.

Liu, Y.; Gutierrez-Barragan, F.; Ingle, A.; Gupta, M.; and Velten, A. 2022. Single-photon camera guided extreme dynamic range imaging. In *Proc. WACV*, 1575–1585.

Liu, Y.-L.; Lai, W.-S.; Chen, Y.-S.; Kao, Y.-L.; Yang, M.-H.; Chuang, Y.-Y.; and Huang, J.-B. 2020. Single-image HDR reconstruction by learning to reverse the camera pipeline. In *Proc. CVPR*, 1651–1660.

Ma, J.; Liang, P.; Yu, W.; Chen, C.; Guo, X.; Wu, J.; and Jiang, J. 2020. Infrared and visible image fusion via detail preserving adversarial learning. *Inf. Fusion*, 54: 85–98.

Mantiuk, R.; Kim, K. J.; Rempel, A. G.; and Heidrich, W. 2011. HDR-VDP-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions. *ACM Trans. Graph.*, 30(4): 1–14.

Marnerides, D.; Bashford-Rogers, T.; Hatchett, J.; and Debattista, K. 2018. Expandnet: A deep convolutional neural network for high dynamic range expansion from low dynamic range content. In *Comput. Graph. Forum.*, volume 37, 37–49.

Masia, B.; Serrano, A.; and Gutierrez, D. 2017. Dynamic range expansion based on image statistics. *Multimedia Tools and Applications*, 76: 631–648.

Metzler, C. A.; Ikoma, H.; Peng, Y.; and Wetzstein, G. 2020. Deep Optics for Single-Shot High-Dynamic-Range Imaging. In *Proc. CVPR*, 1372–1382.

Nazeri, K.; Ng, E.; Joseph, T.; Qureshi, F. Z.; and Ebrahimi, M. 2019. Edgeconnect: Generative image inpainting with adversarial edge learning. *arXiv preprint arXiv:1901.00212*.

Niu, Y.; Wu, J.; Liu, W.; Guo, W.; and Lau, R. W. H. 2021. HDR-GAN: HDR Image Reconstruction From Multi-Exposed LDR Images With Large Motions. *IEEE Trans. Image Process.*, 30: 3885–3896.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115: 211–252.

Samra, R.; Mitra, K.; and Shedligeri, P. 2023. High-Speed HDR Video Reconstruction from Hybrid Intensity Frames and Events. In *Proc. CVMI*, 179–190. Springer.

Scheerlinck, C.; Rebecq, H.; Stoffregen, T.; Barnes, N.; Mahony, R.; and Scaramuzza, D. 2019. CED: Color event camera dataset. In *Proc. CVPRW*, 1684–1693.

Shi, X.; Chen, Z.; Wang, H.; Yeung, D.-Y.; Wong, W.-K.; and Woo, W.-c. 2015. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28: 802–810.

Taverni, G.; Moeys, D. P.; Li, C.; Cavaco, C.; Motsnyi, V.; Bello, D. S. S.; and Delbruck, T. 2018. Front and back illuminated dynamic and active pixel vision sensors comparison. *IEEE Trans. Circuits and Systems II: Express Briefs*, 65(5): 677–681.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, B.; He, J.; Yu, L.; Xia, G.; and Yang, W. 2020. Event Enhanced High-Quality Image Recovery. In *Proc. ECCV*, volume 12358, 155–171.

Wang, H.; Ye, M.; Zhu, X.; Li, S.; Zhu, C.; and Li, X. 2022. KUNet: Imaging Knowledge-Inspired Single HDR Image Reconstruction. In *IJCAI-ECAI*.

Wang, L.; and Yoon, K. 2022. Deep Learning for HDR Imaging: State-of-the-Art and Future Trends. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(12): 8874–8895.

Woo, S.; Park, J.; Lee, J.-Y.; and Kweon, I. S. 2018. Cbam: Convolutional block attention module. In *Proc. ECCV*, 3–19.

Wu, S.; Xu, J.; Tai, Y.-W.; and Tang, C.-K. 2018. Deep high dynamic range imaging with large foreground motions. In *Proc. ECCV*, 117–132.

Yang, Y.; Han, J.; Liang, J.; Sato, I.; and Shi, B. 2023. Learning event guided high dynamic range video reconstruction. In *Proc. CVPR*, 13924–13934.

Zeng, H.; Cai, J.; Li, L.; Cao, Z.; and Zhang, L. 2022. Learning Image-Adaptive 3D Lookup Tables for High Performance Photo Enhancement in Real-Time. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(4): 2058–2073.

Zhang, R.; Zhu, J.-Y.; Isola, P.; Geng, X.; Lin, A. S.; Yu, T.; and Efros, A. A. 2017. Real-time user-guided image colorization with learned deep priors. *ACM Trans. Graph.*, 36(4): 1–11.