# Hybrid High Dynamic Range Imaging fusing Neuromorphic and Conventional Images

Jin Han, Yixin Yang, Peiqi Duan, Chu Zhou, Lei Ma, Chao Xu,
Tiejun Huang, *Senior Member, IEEE,* Imari Sato, and Boxin Shi, *Senior Member, IEEE*

**Abstract**—Reconstruction of high dynamic range image from a single low dynamic range image captured by a conventional RGB camera, which suffers from over- or under-exposure, is an ill-posed problem. In contrast, recent neuromorphic cameras like event camera and spike camera can record high dynamic range scenes in the form of intensity maps, but with much lower spatial resolution and no color information. In this paper, we propose a hybrid imaging system (denoted as NeurImg) that captures and fuses the visual information from a neuromorphic camera and ordinary images from an RGB camera to reconstruct high-quality high dynamic range images and videos. The proposed NeurImg-HDR+ network consists of specially designed modules, which bridges the domain gaps on resolution, dynamic range, and color representation between two types of sensors and images to reconstruct high-resolution, high dynamic range images and videos. We capture a test dataset of hybrid signals on various HDR scenes using the hybrid camera, and analyze the advantages of the proposed fusing strategy by comparing it to state-of-the-art inverse tone mapping methods and merging two low dynamic range images approaches. Quantitative and qualitative experiments on both synthetic data and real-world scenarios demonstrate the effectiveness of the proposed hybrid high dynamic range imaging system. Code and dataset can be found at: https://github.com/hjynwa/NeurImg-HDR

**Index Terms**—High dynamic range imaging, neuromorphic sensor, hybrid camera, image fusion.

## 1 Introduction

High Dynamic Range (HDR) images are desired in modern cameras (or camera phones) because they capture a much wider range of scene radiance variation. A lot of HDR imaging techniques have been developed in recent decades by the computer vision and graphics community, as summarized in [1]. Traditional methods include taking multiple Low Dynamic Range (LDR) images under different exposures, then merging them with different weights to reproduce an HDR image [7], [28]. Another approach is inverse tone mapping (iTMO) [2], which hallucinates texture details from a single LDR image. iTMO is obviously an ill-posed problem, which relies on predicting badly exposed regions from neighboring areas [47] or priors learned through deep neural networks [9], [10], [27], [42].

In recent years, some specially designed neuromorphic cameras, such as DAVIS [3] and Vidar [15], have drawn increasing attention of researchers. Neuromorphic cameras have unique features different from conventional frame-based cameras, they are particularly good at sensing very fast motion and high dynamic range scenes ($1\mu s$ and 120 dB for DAVIS346). The latter characteristic can be utilized to form an *intensity map*, which encodes useful information lost in conventional imaging by a dynamic range capped camera due to over- and/or under-exposure. Despite the distinctive advantages in dynamic range, neuromorphic cameras generally bear low spatial resolution ($260 \times 346$ for DAVIS346) and do not record color information, resulting in intensity maps less aesthetically pleasing than LDR photos captured by a modern camera. It is therefore interesting to study the fusion of LDR images and intensity maps with mutual benefits being combined for high-quality HDR imaging.

To realize the fusion of hybrid images for HDR reconstruction, a *"NeurImg"* fusion pipeline was firstly introduced in [14], which merged visual information from a Neuromorphic camera and a conventional camera (usually as RGB Image) by the intensity map guided HDR network. We denote such an approach as "NeurImg-HDR". It successfully took two types of images as input and bridged the great domain gaps on spatial resolution, dynamic range, color representation and so on to reconstruct a high-quality HDR image. A hybrid camera was built to demonstrate that NeurImg-HDR [14] is applicable to real cameras and scenes. Although the NeurImg-HDR [14] naturally supports HDR video, directly applying it in a frame-by-frame manner shows "flickering" artifacts due to the lack of temporal constraint. The simulation of intensity maps in previous NeurImg-HDR [14] tried to integrate different types of neuromorphic signals (*e.g.*, events and spikes) into one type of data, which affected the performance of the proposed method when applying to real data. Due to the mismatching on spatial resolution of two input images fed for training

---

- *Jin Han and Imari Sato are with Graduate School of Information Science and Technology, the University of Tokyo, and National Institute of Informatics, Tokyo, Japan.*
- *Yixin Yang, Peiqi Duan, Tiejun Huang, and Boxin Shi (corresponding author: shiboxin@pku.edu.cn) are with National Engineering Research Center of Visual Technology, School of Computer Science, Peking University, Beijing, China.*
- *Chu Zhou and Chao Xu are with School of Intelligence Science and Technology, Peking University, Beijing, China.*
- *Lei Ma is with National Biomedical Imaging Center, Peking University, and Life Simulation Research Center, Beijing Academy of Artificial Intelligence, Beijing, China.*
- *Part of this research was done when J. Han was with B. Shi's lab at Peking University.*
- *This work was supported in part by the National Key R&D Program of China under Grant 2021ZD0109803, and the National Natural Science Foundation of China under Grant No. 62136001 and 62088102.*

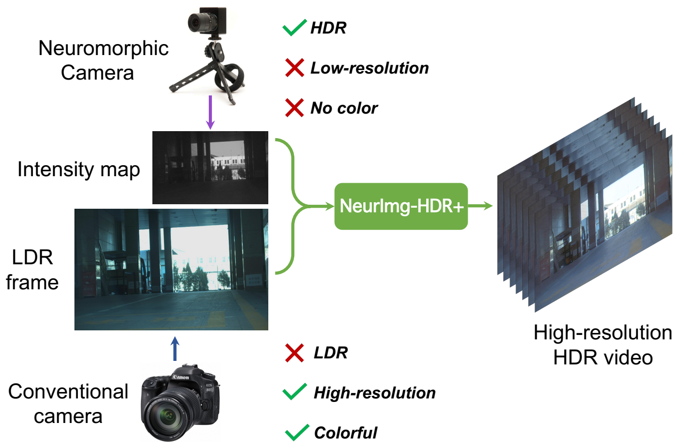


Figure. 1. The "NeurImg" hybrid images fusion framework merges a low-resolution, grayscale intensity map captured by a neuromorphic camera with a high-resolution, colorful LDR image from a conventional camera. Previous NeurImg-HDR [14] only supports HDR images at the resolution of $512 \times 512$. The improved NeurImg-HDR+ can produce HDR videos with higher spatial resolution up to $3200 \times 2000$.

the network, the previous approach easily encounters bottleneck when reconstructing high-resolution HDR images (*e.g.*, only supports $512 \times 512$ resolution).

In this paper, we extend the NeurImg-HDR pipeline in [14] from several aspects including HDR video generation and higher-resolution reconstruction, denoted as "NeurImg-HDR+" shown in Fig. 1. For HDR video reconstruction, we design a new chrominance compensation network with implicit color fusion and recurrent architecture to improve the quality of color restoration and the smoothness of HDR video over time. We introduce an upsampling network for intensity maps to match the spatial resolution of LDR images and achieve high-resolution reconstruction. At last, we analyze the limitations of merging two LDR images on dynamic range recovery and detailed information preservation to demonstrate the superiority of the proposed NeurImg fusion strategy. Our major contributions of this paper are summarized as follows:

1) We improve the chrominance compensation network and achieve temporal consistent HDR video reconstruction. We use the hybrid camera to capture a dataset named Hybrid Events & Spikes HDR (HES-HDR) dataset for testing, which consists of hybrid neuromorphic signals and ordinary LDR frames with spatial alignment and temporal synchronization.
2) We propose an improved architecture with three sub-networks according to the NeurImg fusion pipeline. The new upsampling network supports $8\times$ super-resolution of intensity maps and achieves high-resolution HDR reconstruction up to $3200 \times 2000$ on real data. The new chrominance compensation network implicitly compensates $U$, $V$ channels and converts them to RGB frames in feature space.
3) We compare the proposed method to the state-of-the-art approach of merging two LDR images with different exposures. It demonstrates the limitations on dynamic range recovery and details preservation of the LDR-only approach, while fusing neuromorphic images can overcome such bottlenecks.

## 2 Related Work

**Image-based HDR reconstruction.** The classic HDR imaging method was proposed by Mann and Picard [28], which merges several photographs under different exposures. However, aligning different LDR images may lead to ghosting in the HDR results due to misalignment caused by camera movement or changes in the scene. This problem incurs a lot of research on deghosting in HDR images [20], [36]. Instead of merging multiple images, inverse tone mapping was proposed by Banterle *et al*. [2], whose intent is to reconstruct visually convincing HDR images from a single LDR image. This ill-posed problem was attempted to be solved by several optimized approaches [25], [30].

In recent years, Convolutional Neural Networks (CNNs) have been applied to plenty of HDR image reconstruction tasks. Several works [17], [40] merged images under different exposure by feeding them to a neural network to reconstruct an HDR image. As for iTMO, Eilertsen *et al*. [9] used a U-Net like network to predict the saturated areas, and applied a mask to reserve non-saturated pixels in LDR images, then fused the masked image with predicted image to get the HDR results. Some approaches [10], [24] predicted the LDR images under multiple exposures, then merged these LDR images using classic method [7]. Liu *et al*. proposed the single-image HDR reconstruction method [27] by learning the reverse camera pipeline. Santos *et al*. [42] conditionally applied convolutional layers on the saturated pixels by using a feature masking mechanism to get the HDR results.

For HDR video reconstruction, Kalantari *et al*. [19] proposed a patch-based optimization method to reconstruct the missing details in HDR videos. Li *et al*. [26] treated this problem as a maximum posterior estimation. They split background and foreground via a multi-scale adaptive kernel regression to tackle misalignment. Learning-based methods [5], [18] generated HDR video using convolutional networks that merge a sequence of frames with alternating exposures.

**Computational HDR imaging.** HDR imaging problem would become less ill-posed by using computational approaches or even unconventional cameras that implicitly or explicitly encode expanded dynamic range of the scene. Nayar *et al*. [35] added an optical mask on a conventional camera sensor to get spatially varying pixel exposures. Some approaches [22], [46] modified the inner structure of cameras to implement an HDR-video system, which used beam splitters to simultaneously capture multiple images with different exposure levels, then merged them into an HDR image. Zhou *et al*. [54] used a modulo camera [51] that wrapped the high radiance of an HDR scene periodically and saved as modulo information, then proposed UnModNet to unwrap and predict the HDR scene radiance pixel-wisely. Merzler *et al*. [31] optimized the image signal processor (ISP) by placing a diffractive optical element (DOE) that encoded the saturated pixel values into nearby pixels. They used the information from the encoded measurements to recover clipped information.

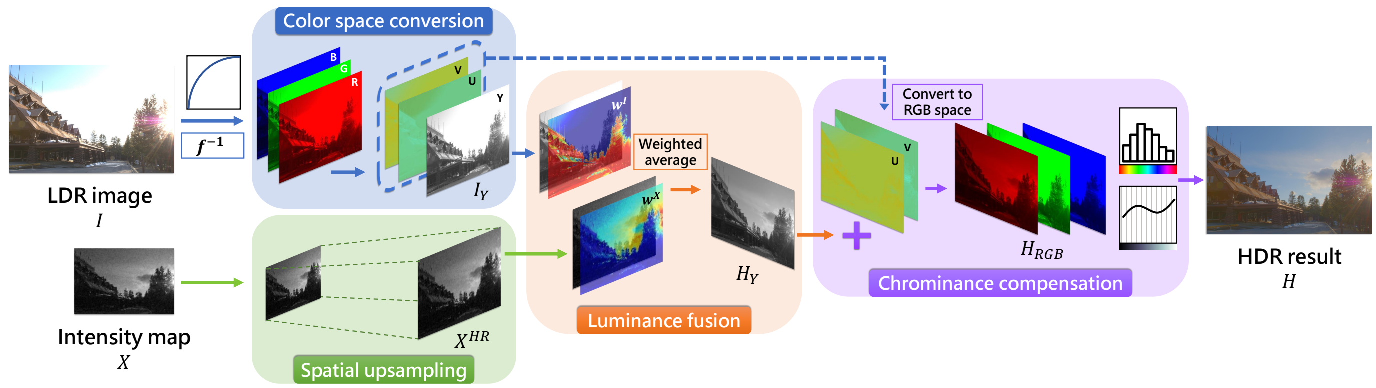


Figure. 2. The conceptual pipeline of NeurImg fusion process, which consists of four steps: color space conversion of the LDR image, spatial upsampling of the intensity map, luminance fusion to produce HDR image in luminance domain, and chrominance compensation that refills the color information to get a colorful HDR result.

There are bio-inspired neuromorphic sensors such as DAVIS [3] (Dynamic and Active Pixel Vision Sensor), ATIS [39] (Asynchronous Time-based Image Sensor), and spike camera (Vidar) [15] detecting the scene radiance asynchronously. This series of non-conventional sensors surpass conventional frame-based cameras in various aspects [11] including high dynamic range. Images reconstructed from raw event data have shown great potential in recovering very high dynamic range of the scene [41], [55]. Different from event data that is generated in a differential manner, spike data directly reflects the scene radiance by integrating asynchronously in each pixel [15]. Images reconstructed from spikes [52] can also recover high dynamic range due to the different densities of spike generation.

**Hybrid fusion for HDR reconstruction.** Combining neuromorphic data with conventional images to produce more visually pleasing HDR photos with higher resolution and realistic color appearance is becoming an interesting topic in recent years. Images captured by different types of sensors provide distinctive information of the scene. The fusion of hybrid signals can compensate each other for HDR reconstruction. The guided event filtering (GEF) [8] unified RGB images and event data via a motion compensation model to achieve high-resolution, noise-robust imaging. Wang *et al*. [48] integrated events based on event double integral (EDI) model [37] and merged to intensity frames for interpolation, then dealt with noise and artifacts using a variant of Kalman filter.

## 3 Proposed Method

### 3.1 NeurImg Fusion Pipeline

As illustrated in Fig. 1, our goal is to reconstruct HDR frames given the input of LDR frames $I$ from a conventional camera and intensity maps $X$ captured by a neuromorphic camera. We assume that the LDR frames do not suffer from the blurry artifact. Such a fusion pipeline can be conceptually illustrated using Fig. 2, which contains four key steps:

**Color space conversion.** Most conventional cameras record color images in RGB format and each channel contains pixel values represented by 8-bit integers. There exists a nonlinear mapping between scene radiance and the pixel values in the camera pipeline, so we have to firstly map LDR images to the linear domain via the inverse camera response function (CRF) $f^{-1}$. To fuse with the one-channel intensity map, we then convert the color space of LDR image from RGB to YUV. The $Y$ channel $I_Y$ indicates the luminance of $I$ which is in the same domain of $X$, and $U$, $V$ channels contain the color information. We use $I_Y$ to fuse with intensity map and reserve $U$, $V$ channels as chrominance information to be added back later.

**Spatial upsampling.** To bridge the resolution gap between $X$ and $I_Y$, we need to enlarge the spatial resolution of the intensity map to make it have the same size as $I_Y$. The upsampling operation $\mathcal{S}(\cdot)$ is defined as follows:

$$X^{SR} = \mathcal{S}(X), \tag{1}$$

where $X^{SR}$ is the upsampled intensity map. $\mathcal{S}(\cdot)$ can be any upsampling operator such as the nearest neighbor or bicubic interpolation, or a pre-trained neural network for super-resolution.

**Luminance fusion.** To expand the dynamic range of $I_Y$ under the guidance of $X^{SR}$, an intuitive solution is to define a weighting function, which indicates the pixels that should be retained for fusion and those should be discarded. This can be implemented by adopting a similar merging strategy proposed by Debevec and Malik [7]. The fused value of $H_Y$ is calculated as follows:

$$H_Y = \mathcal{W}(I_Y, X^{SR}) = \frac{w^I I_Y + w^X X^{SR}}{w^I + w^X}, \tag{2}$$

where $w^I$ and $w^X \in [0, 1]$ indicate corresponding weights for different types of input signals. A straightforward way to determine the weight values is to set a threshold $\tau$ (*e.g.*, $\tau > 0.5$) manually. Pixel values (normalized to $[0, 1]$) lying in the effective range $[1 - \tau, \tau]$ are given larger weights to retain the information, while values out of the range are either too dark (under-exposed) or too bright (over-exposed), hence smaller weights are given to discard such information. A binary mask could be calculated based on the threshold, which is the simplest way to get a weight map. Another option is to set weights as a linear ramp, which is similar

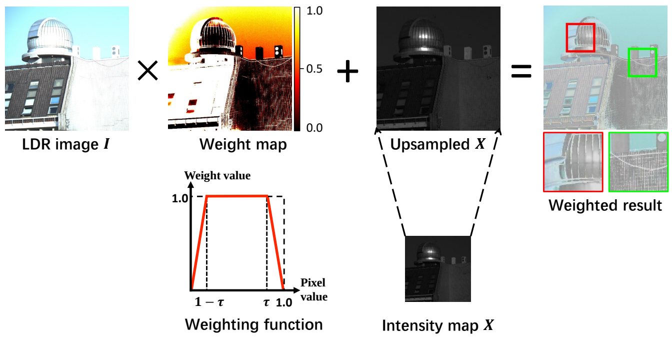


Figure. 3. A real example of fusing an intensity map and an LDR image using a linear ramp as the weighting function. Such a straightforward fusion strategy results in various unpleasant artifacts, such as color distortion in the insets.

to the pixel-wise blending in [9]. Such a weighting function can be expressed as

$$w_i = \frac{0.5 - \max(|I_i - 0.5|, \tau - 0.5)}{1-\tau}. \tag{3}$$

We use a real-captured sample to illustrated the weighting function in Eq. (3). The fusion result is shown in Fig. 3.

**Chrominance compensation.** After fusion in the luminance domain, $H_Y$ now contains HDR information in high-resolution, but without color information. The color can be compensated from $U$, $V$ channels of $I$, (*i.e.*, $I_U$, $I_V$). Denote $\mathcal{C}(\cdot)$ as the color compensation operator, this procedure can be represented as

$$H = \mathcal{C}(H_Y, I_U, I_V), \tag{4}$$

which combines $H_Y$ with $I_U$, $I_V$, and converts it back to RGB color space. Due to the dynamic range gap between $H_Y$ and $I_U(I_V)$, directly combining them leads to unnatural color appearance, as shown in the weighted result in Fig. 3. We should use some color correction methods to recover the realistic color appearance.

The example in Fig. 3 demonstrates that simply applying the conceptual pipeline in Fig. 2 may not achieve a satisfying HDR image. The dynamic range gap between two images and limited color information lead to unrealistic HDR results.

To address these issues, we translate the pipeline in Fig. 2 as an end-to-end network $\mathcal{F}(\cdot)$:

$$H = \mathcal{C}\left(\mathcal{W}\left(f^{-1}(I_Y), \mathcal{S}(X)\right), I_U, I_V\right) = \mathcal{F}(I, X; \theta), \tag{5}$$

where $\theta$ denotes parameters of the network. We will next introduce the specific concerns in realizing each of the four steps using deep neural networks.

### 3.2 NeurImg-HDR+ Network

In this subsection, we describe the details of the proposed network, whose architecture is shown in Fig. 4. Our model takes LDR frame $I$ and intensity map $X$ as the input and contains three consecutive sub-networks: upsampling network, luminance fusion network, and chrominance compensation network.

First of all, inverse CRF and color space conversion are conducted offline as a pre-processing to $I$. Then for each pixel in the input $I_Y$, the proposed network learns to extend the bit-width under the guidance of the information encoded in $X$. We design specific modules in the network in accordance with the remaining three steps described in Sec. 3.1. Spatial upsampling of $X$ is realized by the newly added upsampling network, instead of concatenating multi-scale feature maps as the preliminary work [14]. It learns to super-resolve $X$ to match the resolution of $I$ with multiple scales. The luminance fusion process can be realized by attention gates and skip connections in the luminance fusion network. Therefore, we design the network with U-Net architecture that consists of double encoders (encoder of $I_Y$ and $X^{SR}$) and one decoder. Finally, the chrominance information is compensated from $I_U, I_V$ by the chrominance compensation network. The detailed architecture of three sub-networks will be introduced as follows.

**Upsampling network.** Compared with ordinary RGB cameras with tens of millions of pixels, the intensity maps captured by neuromorphic cameras are in low spatial resolution due to the restriction from currently available sensors. In order to fuse with LDR frame $I$, we should firstly upsample the intensity map $X$ to the same size of $I$. We perform an upsampling operation by a super-resolution network with residual dense connections. Dense connections can preserve detailed information from shallow to deep layers and fuse features in different scales for image reconstruction. The dense block outputs residuals between SR result $X^{SR}$ and the interpolated input $X$. Thus, the final $X^{SR}$ is the summation of intensity residuals and interpolated $X$. Compared to a naive upsampling operation $\mathcal{S}(\cdot)$, the convolutional layers learn a comprehensive representation from image context to realize upsampling operation by end-to-end back propagation, rather than simply rely on interpolation from nearby pixels. Considering that LDR frames have much higher resolution than the intensity maps (*e.g.*, $2448 \times 2048$ *vs.* $346 \times 260$), for different resolution of $I$, we add different number of pixel shuffle layers [44] to the network for $2\times$, $4\times$ and $8\times$ SR.

**Luminance fusion with attention masks.** The fusion of pixel values in the luminance domain is the key step for dynamic range expansion. The proposed architecture applies skip-connections, which transfer feature maps between encoders and decoder to incorporate both rich textures in $I_Y$ and HDR information in $X^{SR}$. However, simply concatenating feature maps from two encoders is expected to be influenced by the dynamic range gap between the two input images. So we fuse the concatenated tensor by a $1 \times 1$ convolution before passing it to the next layer.

As stated in the luminance fusion part of Sec. 3.1, a weighting function $\mathcal{W}(\cdot)$ is added to determine the weight of each pixel, which can be implemented by introducing attention mechanism in the network. We choose to use the self-attention gate [43] as a mask added on $I_Y$ that assigns different importance to different pixels of an image. The attention mask is computed by $1 \times 1$ convolution on the skip-connected feature maps from $I_Y$ encoder, and the feature maps from the last layer in the decoder. Then the convolved feature maps are activated by a non-linear function. The element-wise multiplication of attention mask and input feature map from $I_Y$ helps to filter the badly exposed pixels

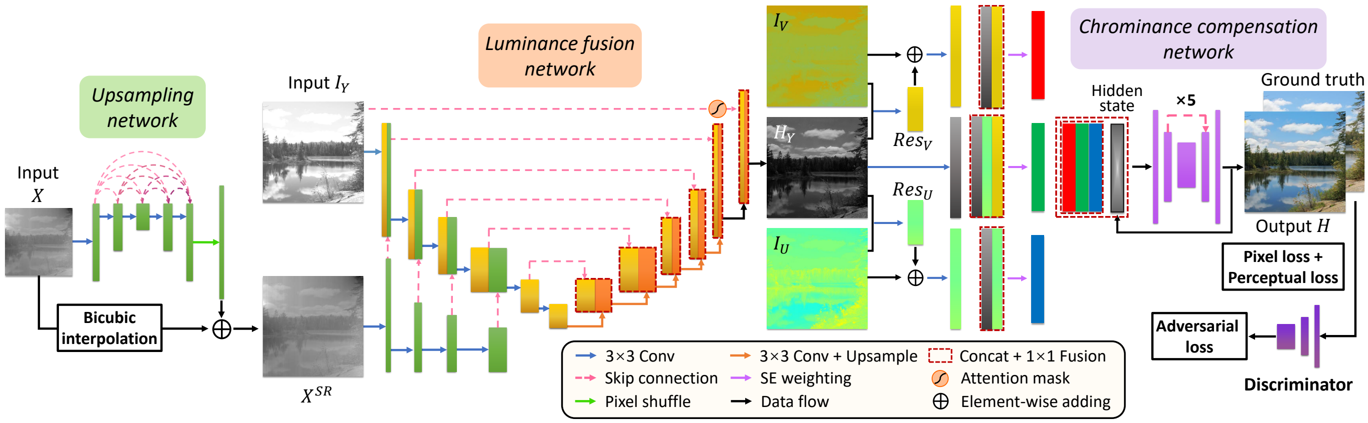


Figure. 4. Overview of NeurImg-HDR+ network architecture. It contains upsampling network, luminance fusion network, and chrominance compensation network.

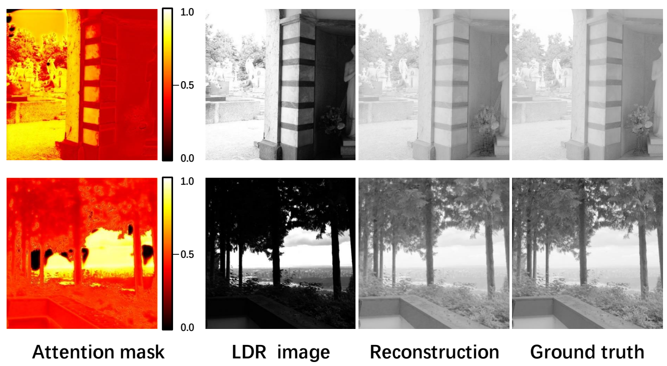


Figure. 5. Examples of attention mask calculated from self-attention module, which assigns different weights for each pixel and reserves useful information for fusion.

and preserve areas with valid information for reconstruction. Compared to assigning weights intuitively like Eq. (2), our attention mask is computed from feature maps of two input images, and the learnable parameters can be trained end-to-end to find suitable weights for different input images.

As Fig. 4 indicates, we add attention gates only to the first skip-connections, instead of to all of them. We find that removing attention masks from the inner skip-connections results in better reconstruction. Figure 5 shows two examples of attention mask and the fusion results in luminance domain.

**Chrominance compensation network.** Given the HDR image in luminance domain $H_Y$, we combine it with chrominance information $I_U$ and $I_V$ from the LDR image, then convert it to RGB color space to recover color appearance. Figure 5 shows that almost perfect reconstructions in the luminance domain can be obtained by the luminance fusion network, while the chrominance compensation process with prior of $U$, $V$ channels from $I$ is more difficult when converted back to RGB color space. In our preliminary work [14], we concatenate $[H_Y, I_U, I_V]$ and convert it to RGB color space using the following function:

$$\begin{bmatrix} R \\ G \\ B \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1.140 \\ 1 & -0.394 & -0.581 \\ 1 & 2.032 & 0 \end{bmatrix} \begin{bmatrix} Y \\ U \\ V \end{bmatrix}. \quad (6)$$

However, the directly converted $H_{RGB}$ may suffer from color distortion due to the dynamic range gap between $H_Y$ and $I_U(I_V)$. Because the luminance values of $Y$ channel are stored in high precision format (*e.g.*, float), while the values in $U$,$V$ channels directly inherited from $I$ are still in the 8-bit integer format. Thus, the converted $H_{RGB}$ tends to be colorless especially after tone mapping. And it becomes more difficult for chrominance compensation network to restore the vivid color appearance, because the loss of precision has diffused into all three channels of $H_{RGB}$.

Therefore, we propose to make implicit color space conversion in the feature domain. The key problem is precision gap between $Y$ channel and $U$, $V$ channels, we compute residuals $Res_U$ and $Res_V$ respectively under the guidance of $Y$ channel, which are added to $U$, $V$ channels to compensate the precision gap. Then we convert features from YUV to RGB by fusing different source channels. According to Eq. (6), features of $R$ channel come from $Y$ and $V$ channels, while features of $G$ channel come from $Y$, $U$, and $V$ channels.

To reconstruct temporally consistent HDR videos, we exploit the correlation between consecutive input frames by introducing a recurrent structure after the fusion of RGB features. It maintains a hidden state that is computed from current input as well as the encoded past states from the previous input. With recurrent structure, the temporal complementary and redundant information through time can be well exploited to alleviate flickering artifacts and reduce noise in HDR video reconstruction. Details of recurrent block are described in Sec. 3.3.

For a natural appearance of HDR results $H$, we apply Generative Adversarial Networks (GANs) [13] architecture to perform chrominance compensation. The network described above is viewed as a generator that learns to recover realistic color appearance in HDR images. We train a discriminator simultaneously that accepts the output of the generator and the corresponding real HDR images. It distinguishes the reality of color appearance, then propagates adversarial loss back to both the generator and discriminator.

### 3.3 Extension to HDR Video Reconstruction

We extend the preliminary NeurImg-HDR [14] to HDR video reconstruction by introducing the recurrent block.

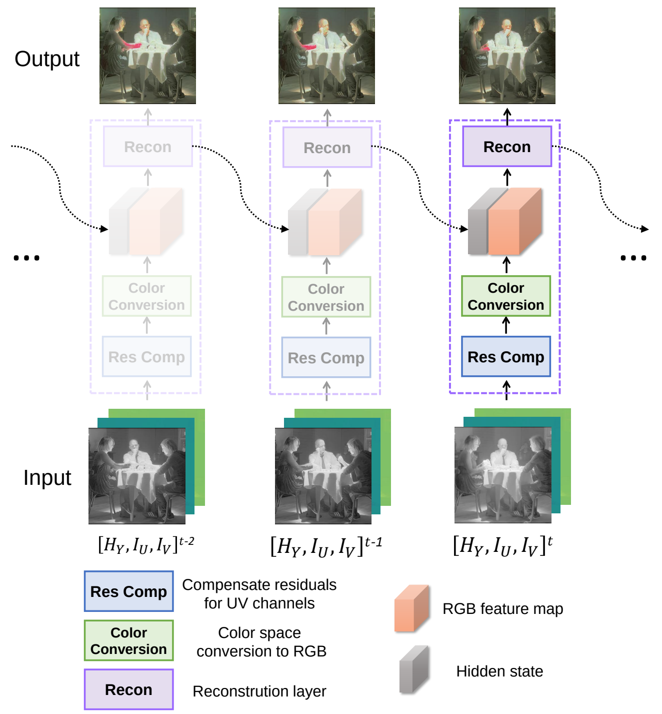


Figure. 6. The architecture of recurrent block in chrominance compensation network.

Since the proposed method merges two images captured by two different sensors simultaneously, it can be treated as operating in the "single-shot" mode, which is readily to be applied to HDR video capture and reconstruction. We split the video sequence into a series of frames and reconstruct the HDR image frame-wisely. However, naively applying the NeurImg-HDR pipeline [14] in a frame-by-frame manner leads to abrupt and incoherent changes from frame to frame. Because the differences between consecutive frames can be regarded as small disturbances like translations or noise added to the same image, which may cause the trained neural networks to output quite unstable results [53]. Such disturbance and instability inevitably introduce temporal inconsistencies when merging a sequence of frames to generate video. Therefore, dense motion estimation such as optical flow between frames is required to enforce temporal coherence [23], or alternatively, a temporal hidden state should be preserved to maintain the coherence between consecutive frames [32].

As illustrated in Fig. 6, we design a recurrent block with a hidden state in the chrominance compensation network, which takes the last state from the previous frame and fuses it with the RGB feature map to reconstruct the HDR result for current frame. During the training process, the hidden state is initialized to zero. The recurrent block maintains a temporal window to update the hidden state and accumulate gradients step by step. Due to the limit of GPU memory size, we set the window size to 10. For inference process, the hidden state is initialized at the beginning of a video and the size of temporal window is infinite.

### 3.4 Loss Functions

We first introduce two basic loss functions: pixel loss $\mathcal{L}_{pixel}$ and perceptual loss [16] $\mathcal{L}_{perc}$ that all the three sub-networks use. Pixel loss computes the $\ell_1$ norm distance between network prediction $\mathcal{F}(x)$ and ground truth $\hat{y}$:

$$\mathcal{L}_{pixel} = \left\|\mathcal{F}(x) - \hat{y}\right\|_1 . \tag{7}$$

The perceptual loss is defined based on the feature maps of images extracted by the VGG-16 network [45] pre-trained on ImageNet:

$$\begin{aligned}\mathcal{L}_{perc} = \sum_h \Big( &\left\|\phi_h(\mathcal{F}(x)) - \phi_h(\hat{y})\right\|_2^2 \\ &+ \left\|\boldsymbol{G}_h^{\phi}(\mathcal{F}(x)) - \boldsymbol{G}_h^{\phi}(\hat{y})\right\|_2^2 \Big) ,\end{aligned} \tag{8}$$

where $\phi_h$ denotes the feature map convoluted from $h$-th layer of the VGG-16 [45], $\boldsymbol{G}_h^{\phi}$ is the Gram matrix of feature maps $\phi_h$ of two input images. Both of the two parts are computed by $\ell_2$ norm. We use the layers 'relu1_2', 'relu2_2', 'relu3_3' and 'relu4_3' of VGG-16 network [45] in our experiments to compute perceptual loss.

**Loss functions of upsampling network and luminance fusion network.** The upsampling network learns to super-resolve intensity maps to the corresponding resolution of LDR images. We define the loss function of upsampling network as:

$$\mathcal{L}_U = \alpha_1 \mathcal{L}_{pixel} + \alpha_2 \mathcal{L}_{perc}, \tag{9}$$

where $\alpha_1$ and $\alpha_2$ are the weights for different parts of loss function. We set $\alpha_1 = 100.0$ and $\alpha_2 = 3.0$.

The luminance fusion network reconstructs images in the linear luminance domain, which covers a wide range of values. Directly calculating losses between the output image $H_Y$ and ground truth $\hat{H}_Y$ may cause the loss function to be dominated by large values (bright pixels) of $H_Y$, while the effect of small values (dark pixels) tends to be ignored. Therefore, it is reasonable to compute the loss function between $H_Y$ and $\hat{H}_Y$ after tone mapping. The range of pixel values are compressed by the following function proposed by [17] after normalized to $[0, 1]$:

$$\mathcal{T}(H_Y) = \frac{\log(1 + \mu H_Y)}{\log(1 + \mu)}, \tag{10}$$

where $\mathcal{T}(\cdot)$ is the tone mapping operator and $\mu$ (set to be 5000) denotes the amount of compression. The tone mapping operator is computationally effective and differentiable, thus easy for back-propagation.

The loss function of luminance fusion network is similar to that of the upsampling network, except for calculating the distance between $\mathcal{T}(H_Y)$ and $\mathcal{T}(\hat{H}_Y)$:

$$\mathcal{L}_L = \alpha_3 \mathcal{L}_{pixel} + \alpha_4 \mathcal{L}_{perc}. \tag{11}$$

We set $\alpha_3 = 100.0$ and $\alpha_4 = 3.0$ in $\mathcal{L}_L$.

**Loss functions of chrominance compensation network.** As for the chrominance compensation network, in addition to pixel loss and perceptual loss, we introduce adversarial loss from the discriminator. The losses of generator and discriminator are inherited from traditional GANs [13]:

$$\mathcal{L}_{gene} = \mathbb{E}_{H_{YUV}}[(\mathcal{D}(\mathcal{G}(H_{YUV})) - 1)^2], \tag{12}$$

$$\mathcal{L}_{disc} = \frac{1}{2}\left(\mathbb{E}_H[\mathcal{D}(H)-1)^2] + \mathbb{E}_{H_{YUV}}[(\mathcal{D}(\mathcal{G}(H_{YUV})))^2]\right), \tag{13}$$

where $\mathcal{G}$ and $\mathcal{D}$ are generator and discriminator, and $\mathbb{E}$ is the expectation function. We denote the input of chrominance compensation network $[H_Y, I_U, I_V]$ as $H_{YUV}$ and the final output HDR image as $H$. The total loss of chrominance compensation generator is:

$$\mathcal{L}_C = \alpha_5\mathcal{L}_{pixel} + \alpha_6\mathcal{L}_{perc} + \alpha_7\mathcal{L}_{gene}. \tag{14}$$

We set $\alpha_5 = 100.0$, $\alpha_6 = 3.0$, and $\alpha_7 = 10.0$. The weighting parameters $\alpha_i$ in loss functions balance the contributions of different parts of losses. Please refer to supplementary material for detailed analysis of how these weighting parameters are determined.

### 3.5 Dataset Preparation

Learning-based methods rely heavily on training data. However, there are no appropriate large-scale real HDR image datasets suitable for our purpose. Therefore, we collect HDR images from various image sources and video sources. Since the proposed network has two different types of images as input, we analyze the data formation process of different input and simulate each of them. For LDR images, we synthesize them from HDR images like taking photos with a virtual camera [9]. The formation process of LDR image $I$ from HDR image consists of 4 main steps: dynamic range clipping, noise simulation, non-linear mapping, and quantization. As for intensity maps, we simulate them in accordance with the data generation mechanism of two different types of neuromorphic cameras. Please refer to supplementary material for more details of data simulation.

### 3.6 Training Strategy

The proposed network is implemented by PyTorch, and we use ADAM optimizer [21] during the training process with a batch size of 2. We use instance normalization with the activation function of LeakyReLU in the luminance fusion network. The output of the network is activated by a Sigmoid function that maps pixel values to the range of $[0, 1]$. Both of the three networks are initialized with Xavier initialization [12]. During training, we apply phase-to-phase training for better learning efficiency, instead of learning all from scratch in an end-to-end manner. We train the upsampling network firstly with the input of low-resolution intensity maps $X$ and ground truth $X^{HR}$. Then we fix the upsampling network and train the luminance fusion network with the input of $I_Y$ and $X^{SR}$. Finally, we fix the previous two networks and train the chrominance compensation network with the input of $[H_Y, I_U, I_V]$. 600 epochs of training enables the networks to converge. The initial learning rate is $10^{-5}$, during the first 400 epochs it is fixed, in the next 200 epochs, it decays to 0 with a linear strategy.

## 4 Hybrid Camera and HES-HDR Dataset

### 4.1 System Setup

In order to demonstrate the effectiveness of the proposed method on real-world scenarios, we build a hybrid camera, which is composed of a conventional RGB camera and a neuromorphic camera (DAVIS346, or spike camera (Vidar) [15]) with the same lens. The prototype and specifications are illustrated in Fig. 7 and Table 1. There is a beam splitter (Thorlabs CCM1-BS013) with 50% splitting in front of the two sensors, which splits the incoming light and sends them to different sensors with the same view. We write a synchronization script to trigger two sensors simultaneously. Furthermore, the mobility of our hybrid system allows us to take photos both indoor and outdoor, which helps to validate that the proposed method is applicable to various scenarios.

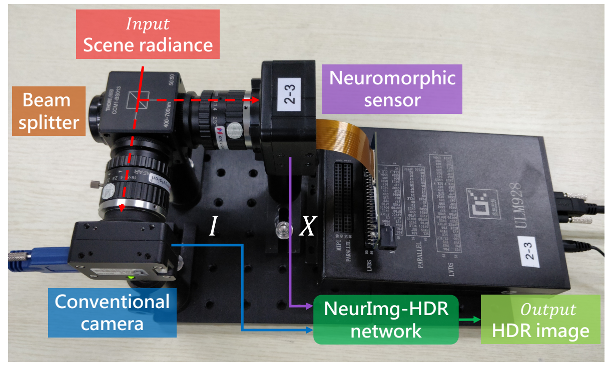


Figure. 7. The prototype of our hybrid camera, which is composed of a conventional RGB camera and a neuromorphic camera. Radiance information is recorded simultaneously by two sensors.

TABLE 1
The detailed specifications on spatial resolution, frame rate (FR) and dynamic range (DR) of our hybrid camera. In our implementation, the neuromorphic camera could be either an event camera (DAVIS346) or a spike camera (Vidar) [15].

| Camera Model | Conventional Camera | Neuromorphic Camera | |
|---|---|---|---|
| | FLIR Chameleon 3 | DAVIS346 | Vidar |
| Resolution | 2448×2048 | 346×260 | 400×250 |
| FR (fps) | 35 | up to $1M$ | $40K$ |
| DR (dB) | 60 | 120 | 100 |

### 4.2 Dataset Collection

We build a dataset named **H**ybrid **E**vent & **S**pike **HDR** (HES-HDR) dataset using the hybrid camera to evaluate the fusion of neuromorphic and RGB hybrid signals. We capture HDR images and videos for various scenarios and collect two types of hybrid signals (*e.g.*, event-RGB or spike-RGB) for each scene with spatial alignment and temporal synchronization. HES-HDR dataset includes both outdoor and indoor HDR scenarios. All the videos include global motion and/or local motion. In total, there are 20 video sequences, including 10 of event-RGB HDR videos and 10 spike-RGB HDR videos. Detailed introduction of the HES-HDR dataset can be found in the supplementary material

## 5 Experiments

### 5.1 Quantitative Evaluation using Synthetic Data

We compare two state-of-the-art deep learning based iTMO methods: Liu *et al.* [27] and Santos *et al.* [42]. The previous

TABLE 2
Quantitative evaluations of the proposed NeuImg-HDR+ and comparing methods. These scores are averaged across all images in the whole test dataset. ↑ (↓) represents the higher (lower) the better results. The best results are in red, and the second best results are in blue.

| Method | PSNR-pu↑ | SSIM-pu↑ | PSNR-t↑ | SSIM-t↑ | LPIPS-t↓ | HDR-VDP↑ | HDR-VQM↓ | TCM↑ |
|---|---|---|---|---|---|---|---|---|
| NeurImg-HDR+ | 26.31 | 0.754 | 24.01 | 0.905 | 0.199 | 9.215 | 0.288 | 0.778 |
| NeurImg-HDR [14] | 22.21 | 0.709 | 20.01 | 0.858 | 0.204 | 7.400 | 0.451 | 0.721 |
| Liu *et al.* [27] | 18.08 | 0.568 | 17.75 | 0.726 | 0.229 | 5.219 | 0.338 | 0.706 |
| Santos *et al.* [42] | 9.66 | 0.311 | 11.28 | 0.612 | 0.288 | 3.361 | 0.265 | 0.453 |
| LDR×2 [5] | 17.14 | 0.596 | 16.49 | 0.712 | 0.337 | 5.806 | 0.425 | 0.584 |

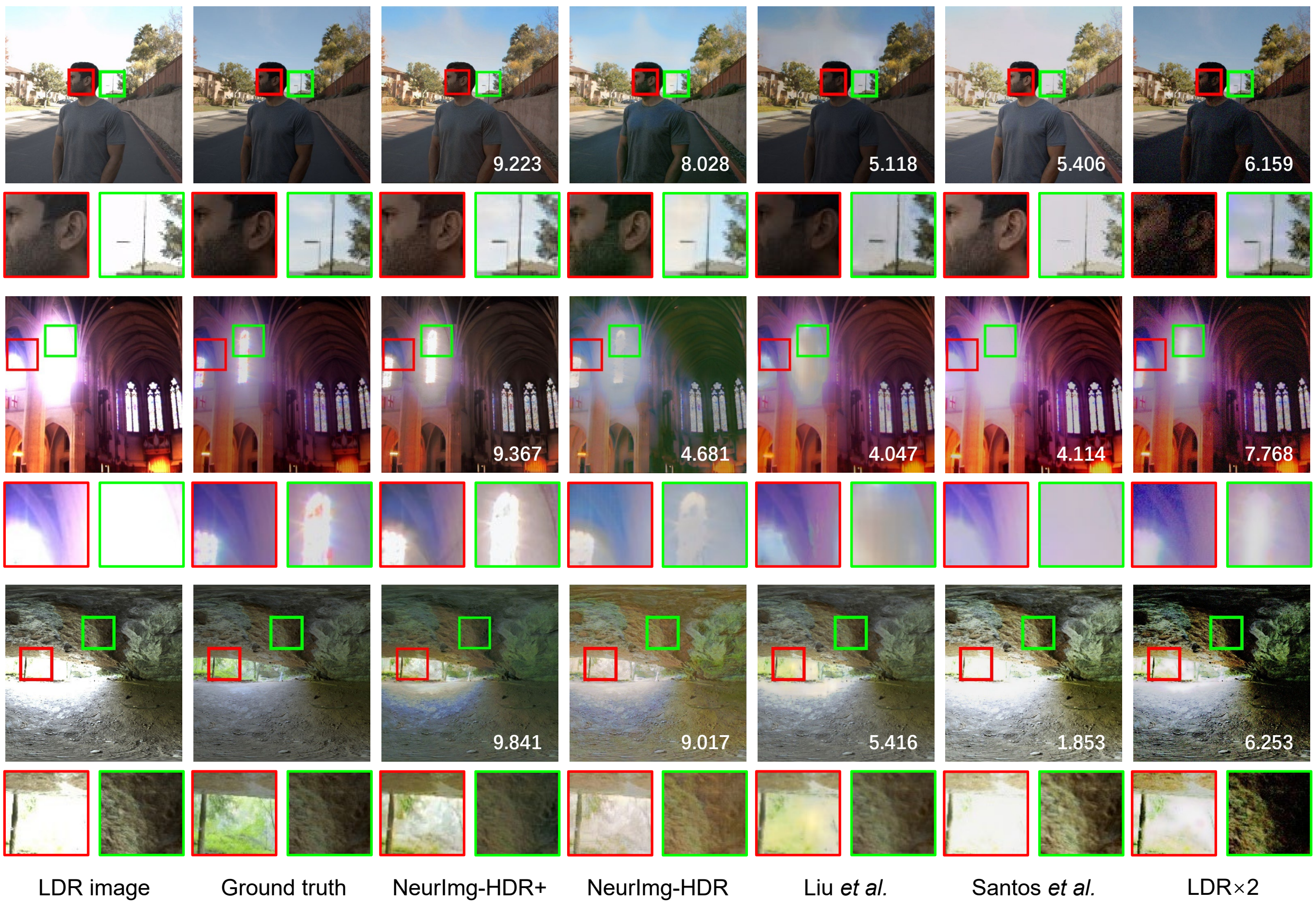


Figure. 8. Comparison between the proposed method and state-of-the-art deep learning based inverse tone mapping methods: Liu *et al*. [27] and Santos *et al*. [42]. We also compare with NeurImg-HDR [14] and a state-of-the-art approach [5] of merging two LDR images, denoted as LDR×2. The Q-Scores of HDR-VDP [33] are displayed in each image. Please zoom-in on the electronic versions for better details.

NeurImg-HDR [14] and a state-of-the-art method denoted as LDR×2 [5] that merges an over- and an under-exposed images are also included in comparison. The results are shown in Table 2 and Fig. 8. For LDR×2 method, we generate two LDR images with different exposure time from the HDR ground truth. We test different exposure ratios $\lambda$ for the whole test dataset, and find that $\lambda = 4$ achieves the highest performance. Therefore, we choose the optimal $\lambda$ as the comparison results in our experiments. Detailed analysis about the limitations of merging two LDR images can be found in Sec. 8 in the supplementary material. For the sake of fairness, we omit the comparison to merging three or more LDR images with different exposures.

Thanks to the extended dynamic range information provided by intensity maps, the proposed approach is able to recover rich texture details in the HDR results. For example, in the second row of Fig. 8, the outline of the intense light source (red inset) is clearly visible in our results, while this is not the case for other iTMO methods. Although merging two LDR images extends the dynamic range (more reliable than single-image solutions), it easily suffers from noise artifact due to the limited dynamic range covered by two LDR images, as shown in the man's face in the first case. It is hard to obtain both HDR and detailed scene radiance using merely two LDR images.

Besides visual comparison, we conduct quantitative evaluations using various metrics as shown in Table 2. In the linear domain, we conduct the widely adopted HDR-

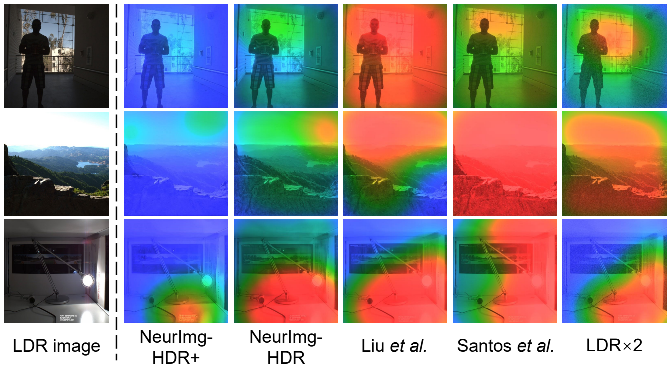


Figure. 9. Comparisons on quality maps calculated from HDR-VDP evaluation metrics [33]. Visual differences increase from blue to red in the quality maps.

VDP-3 [33] (range in $[0, 10]$) for images. For HDR videos, we use HDR-VQM [34] and TCM [49] metrics (both range in $[0, 1]$) to evaluate the HDR restoration quality and temporal consistency. In the perceptually uniform pixel domain [29], we compute peak signal-to-noise ratio (PSNR-pu), and structural similarity (SSIM-pu). Besides, we evaluate HDR images after tone mapping using Eq. (10) by conducting PSNR-t, SSIM-t, and learned perceptual image patch similarity (LPIPS-t) [50] metrics.

The HDR-VDP metrics [33] compute visual difference and predict the visibility and quality between the reconstructed and ground truth HDR images. It produces the quality map and Q-Score for each HDR image to indicate the quality of HDR reconstruction. Figure 9 shows the quality maps of different methods, which display the difference probability between a predicted HDR image and the ground truth. We set the peak luminance (in $cd/m^2$) to 200 and display contrast as $1000 : 1$ when conducting HDR-VDP [33] evaluation. Both visual comparisons and quantitative evaluation results show that the proposed approach achieves much higher quality in HDR image reconstruction compared to other state-of-the-art methods.

We test the proposed method and the comparing approaches on 13 different HDR video sequences with the number of frames varies from 151 to 834. Compared to iTMO methods [27], [42], videos from NeurImg-HDR+ recover much more details on both over-exposed and under-exposed regions. For LDR×2 [5], we set video frames with a short-long exposure mechanism following the frame generation process in their paper. However, videos reconstructed from LDR×2 [5] suffer from noise and "flickering" artifacts due to the exposure gap between consecutive frames. Please refer to the supplementary video for more results.

### 5.2 Results on Real-world Images and Videos

We capture photos and videos for both indoor and outdoor high dynamic range scenes to evaluate the effectiveness of the proposed method. HDR event streams are firstly converted to intensity maps using E2VID [41]. While for spikes, we apply a time window [15] to integrate spikes to get intensity maps. Figure 10 shows HDR results reconstructed on our HES-HDR dataset on both event camera (DAVIS346) and spike camera (Vidar). The input images are firstly fused in the luminance domain (denoted as $H_Y$ in Fig. 10) and then compensated by the chrominance information to get the final colorful HDR images. Results show that the proposed method can successfully fuse the input $I$ and $X$ to reconstruct high-quality HDR images. For example, the texture of dome building (the second case of Vidar) is over-exposed due to the strongly reflected sunlight, but the detailed texture could be captured by the neuromorphic cameras, and recovered in the fusion results using our method. We conduct the quantitative evaluation on real-world images by capturing multiple LDR images with exposure-bracketing. We merge LDR images as the ground truth and compute PSNR-t, SSIM-t, and LPIPS-t for tone mapped results, and HDR-VDP [33] in the linear domain, the results are shown in Table 3.

TABLE 3
Quantitative comparison on real-world data.

| Method | PSNR-t↑ | SSIM-t↑ | LPIPS-t↓ | HDR-VDP↑ |
|---|---|---|---|---|
| NeurImg-HDR+ | **23.25** | **0.952** | **0.164** | **7.847** |
| NeurImg-HDR [14] | 21.22 | 0.936 | 0.247 | 5.552 |

Thanks to the high temporal resolution property of neuromorphic cameras, we extend our model to high-frame-rate (HFR) video reconstruction. The misalignment on temporal domain can be alleviated by deformable convolution (DCN) [6], which introduces diverse offsets in multiple feature levels. DCN achieves implicit alignment and reduces warping errors effectively compared to explicit flow-based alignment [4]. Besides, DCN is a plug & play module without much modification to the original network architecture. Thus, we plug DCN in the luminance fusion network that aligns features from RGB frames to those from intensity maps before fusion, which achieves HFR HDR video reconstruction in the luminance domain. Please refer to the supplementary video for more HDR and HFR videos on our HES-HDR dataset.

### 5.3 High-resolution Reconstruction

The proposed model can handle higher spatial resolution (a typical DLSR or camera phone image with millions of pixels) once we upsample the low-resolution intensity to the corresponding resolution of LDR image. We trained upsampling network with different scaling factors $(2\times, 4\times, 8\times)$ to bridge the huge spatial resolution gap between intensity maps and LDR frames.

We show HDR results with different resolutions (denoted as $2\times$, $4\times$ and $8\times$) on both synthetic data and real-captured images in Fig. 11. The proposed method takes high-resolution LDR frames as input to achieve detailed textures in reconstruction. For example, the contour of the sculpture (green box in the top right case) and the edges of windows (red box in the bottom left case) are much clearer in $8\times$ results, which preserves the high-resolution details from LDR input. We can reconstruct up to $3200 \times 2000$ HDR results on Vidar-based and $2768 \times 2080$ on DAVIS-based hybrid camera. The spatial resolution is limited by the sensor size of neuromorphic cameras. If we use neuromorphic camera with a larger sensor size, such as Prophesee Gen 4 [38] with $1280 \times 720$ pixels, achieving even higher resolution could be possible.

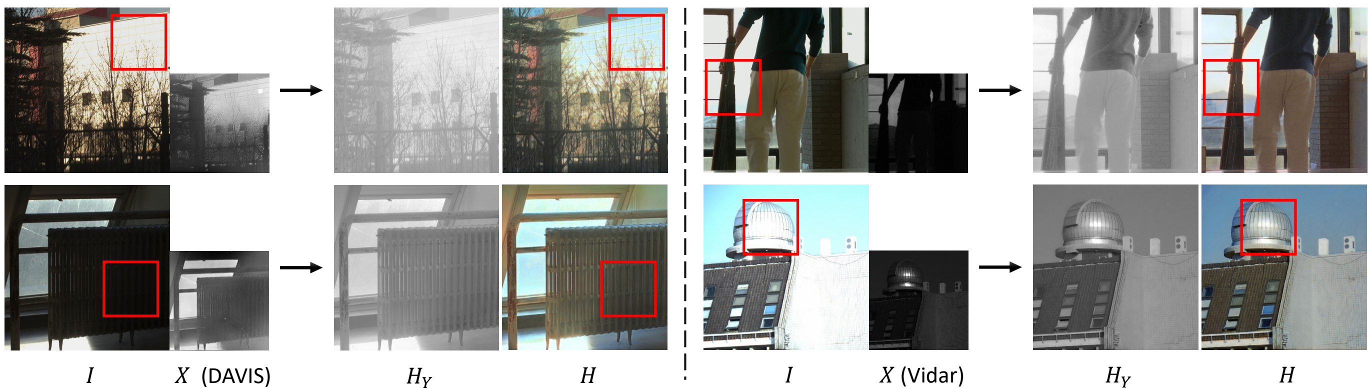


Figure. 10. Real data results reconstructed by NeurImg-HDR+. The LDR images are captured by conventional cameras and the intensity maps are acquired by DAVIS (the left three cases) and spike camera (the right three cases), respectively.

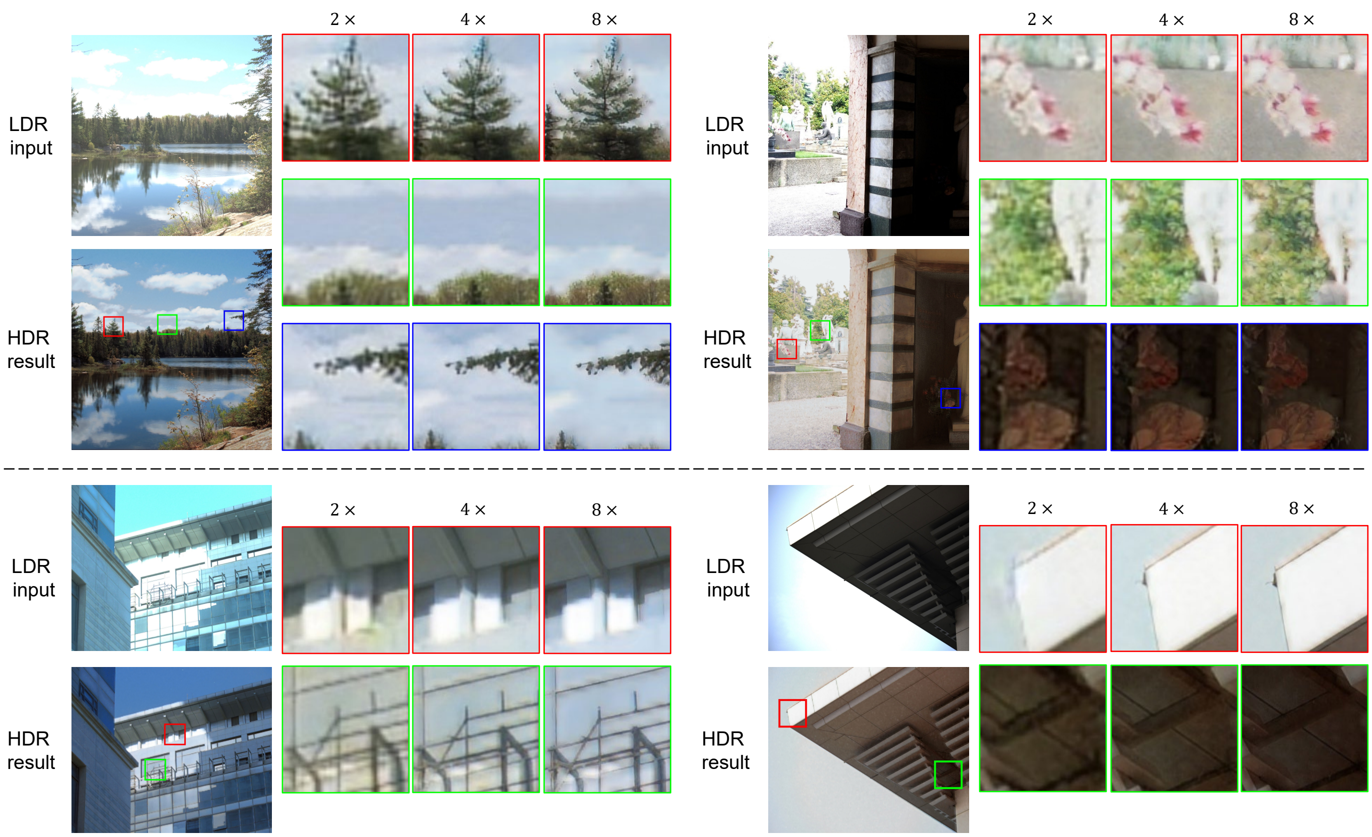


Figure. 11. High-resolution reconstruction in different scales. The top samples are synthetic images, and the bottom two samples are real data. The insets are the HDR results of $2\times$, $4\times$, and $8\times$ SR for intensity maps. Please zoom-in on the electronic version for better details.

However, when handling extremely large spatial resolution gap (*e.g.*, 8× SR), some blurry artifacts are unavoidable in completely saturated regions, such as the blur windows contour (red box in the bottom left case). Because these saturated regions in LDR images are filtered by the attention masks, and the HDR result can only rely on low-resolution intensity maps in these regions.

### 5.4 Ablation Studies

In this section, we conduct extensive ablation experiments to analyze the design of network structure and combination of loss functions by comparing with different variants. Quantitative comparison for different variants is shown in Table 4.

**Comparison with NeurImg-HDR.** The major differences in network architecture between NeurImg-HDR+ and NeurImg-HDR [14] are the encoder of $X^{SR}$ and chrominance compensation network. We propose an independent upsampling network to super-resolve the intensity map to different resolutions corresponding to $I$, instead of concatenating multi-scale feature maps in the decoder of luminance fusion network like [14]. We convert the color space from YUV to RGB in feature space in the chrominance compensation network to overcome the dynamic range gap between color channels. The discriminator added on the chrominance

TABLE 4
Quantitative comparison of different variants of the proposed method. The best results are in red, and the second best results are in blue.

| Method | PSNR-pu↑ | SSIM-pu↑ | PSNR-t↑ | SSIM-t↑ | LPIPS-t↓ | HDR-VDP↑ |
|---|---|---|---|---|---|---|
| NeurImg-HDR+ | 26.31 | 0.754 | 24.01 | 0.905 | 0.199 | 9.215 |
| NeurImg-HDR [14] | 22.21 | 0.709 | 20.01 | 0.858 | 0.204 | 7.400 |
| w/o attention masks | 24.15 | 0.710 | 22.48 | 0.891 | 0.204 | 8.493 |
| single encoder | 21.27 | 0.622 | 20.30 | 0.836 | 0.305 | 7.781 |
| end-to-end training | 24.95 | 0.742 | 22.83 | 0.893 | 0.206 | 8.890 |
| $\mathcal{L}_{pixel}$ loss | 22.86 | 0.603 | 20.77 | 0.848 | 0.422 | 8.295 |
| $\mathcal{L}_{pixel} + \mathcal{L}_{perc}$ loss | 26.09 | 0.748 | 23.46 | 0.901 | 0.205 | 9.208 |
| $\mathcal{L}_{\ell_2} + \mathcal{L}_{perc} + \mathcal{L}_{adv}$ loss | 26.27 | 0.747 | 23.57 | 0.900 | 0.205 | 9.196 |

compensation network provides the adversarial loss for compensating chrominance information, which makes HDR results more natural compared to NeurImg-HDR [14].

**Without attention masks.** We validate the effectiveness of the attention mask module by removing it and then compare the reconstruction results with the complete network. Without attention masks, it is difficult for the network to accurately distinguish the information to reserve or discard, hence leads to some artifacts and low-quality reconstruction. The over-exposed regions cannot fully take advantage of the HDR intensity map to recover structural details.

**Single encoder architecture.** We compare our network with a single encoder architecture, which removes the encoder of $X^{SR}$ in the luminance fusion network. This can be achieved by concatenating $X^{SR}$ and $I_Y$ at first, then sending the 2-channel tensor to a single encoder. In this case, two images from different domains are directly combined instead of fused at multi-scale feature space, which causes performance to drop a lot.

**End-to-end training.** The proposed network is trained in a phase-to-phase manner. If we train three sub-networks simultaneously in an end-to-end manner, they cannot be optimized for their own objectives effectively, which makes it difficult for the whole network to converge simultaneously. However, put the loss function variants aside, the variant with end-to-end training has relatively better performance than other variants. It is because there are no architecture or loss function changes. Different training mechanism has less impact on the final performance of the proposed network.

**Loss functions.** We investigate the effect of different terms in loss functions. The loss functions we used in the proposed NeurImg-HDR+ is $\mathcal{L}_{pixel} + \mathcal{L}_{perc} + \mathcal{L}_{adv}$. The variants are trained with only pixel loss (denoted as $\mathcal{L}_{pixel}$), removing adversarial loss (denoted as $\mathcal{L}_{pixel} + \mathcal{L}_{perc}$), and replacing pixel loss with $\ell_2$ norm (denoted as $\mathcal{L}_{\ell_2} + \mathcal{L}_{perc} + \mathcal{L}_{adv}$). Results show that removing adversarial loss has the minimum effect that achieves 3 runner-ups in Table 4.

**Without recurrent block.** To validate the effectiveness of recurrent block in maintaining temporal consistency of videos, we remove the recurrent block in chrominance compensation network, and test on 13 synthetic HDR videos. The variant without recurrent block achieves 0.326 in HDR-VQM metrics (lower is better) and 0.693 in TCM metrics (higher is better). Compared to the model with recurrent block, it is 11.7% and 12.3% worse in these two metrics, respectively. HDR videos reconstructed with recurrent block achieves better temporal consistency and less flickering artifacts.

## 6 Conclusion

We propose an HDR imaging method using the hybrid camera, which fuses the LDR frames and the intensity maps to reconstruct visually pleasing HDR videos. The preliminary NeurImg-HDR approach [14] has been improved in various aspects to achieve more natural color appearance, higher resolution reconstruction, and HDR video generation. Besides, we analyze the limitations of merging two LDR images and validate the superiority of the NeurImg fusion approach. Extensive experiments on synthetic data and the HES-HDR dataset captured by our hybrid camera demonstrate that the proposed method outperforms state-of-the-art comparing methods.

**Limitation and discussion.** We have tried to conduct frame interpolation and generate HFR videos in the luminance domain when capturing fast-moving scenes. It verifies that it is potentially possible to generate HFR HDR videos with the proposed method. However, for color restoration in HFR HDR videos, there is unsatisfactory color distortion due to the huge loss of chrominance information in both spatial and temporal domains. Since there are no HFR chrominance channels as references, the chrominance compensation results may degrade.

## References

[1] Francesco Banterle, Alessandro Artusi, Kurt Debattista, and Alan Chalmers. *Advanced high dynamic range imaging*. AK Peters/CRC Press, 2017. 1

[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. 1, 2

[3] Christian Brandli, Raphael Berner, Minhao Yang, Shih-Chii Liu, and Tobi Delbruck. A 240× 180 130 db 3 $\mu$s latency global shutter spatiotemporal vision sensor. *Journal of Solid-State Circuits*, 2014. 1, 3

[4] Kelvin CK Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Understanding deformable alignment in video super-resolution. In *Proc. of AAAI Conference on Artificial Intelligence*, 2021. 9

[5] 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 Internatoinal Conference on Computer Vision*, 2021. 2, 8, 9

[6] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen We. Deformable convolutional networks. In *Proc. of Internatoinal Conference on Computer Vision*, 2017. 9

[7] Paul E Debevec and Jitendra Malik. Recovering high dynamic range radiance maps from photographs. In *Proc. of ACM SIGGRAPH*, 1997. 1, 2, 3

[8] Peiqi Duan, Zihao Wang, Boxin Shi, Oliver Cossairt, Tiejun Huang, and Aggelos Katsaggelos. Guided event filtering: Synergy between intensity images and neuromorphic events for high performance imaging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 3

[9] 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, 4, 7

[10] Yuki Endo, Yoshihiro Kanamori, and Jun Mitani. Deep reverse tone mapping. *ACM Transactions on Graphics (Proc. of ACM SIGGRAPH Asia)*, 2017. 1, 2

[11] Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J Davison, Jörg Conradt, Kostas Daniilidis, et al. Event-based vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 3

[12] 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. 7

[13] 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. 5, 6

[14] 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, 4, 5, 6, 8, 9, 10, 11

[15] Tiejun Huang, Yajing Zheng, Zhaofei Yu, Rui Chen, Yuan Li, Ruiqin Xiong, Lei Ma, Junwei Zhao, Siwei Dong, Lin Zhu, et al. 1000× faster camera and machine vision with ordinary devices. *Engineering*, 2022. 1, 3, 7, 9

[16] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Proc. of European Conference on Computer Vision*, 2016. 6

[17] Nima Khademi Kalantari and Ravi Ramamoorthi. Deep high dynamic range imaging of dynamic scenes. *ACM Transactions on Graphics (Proc. of ACM SIGGRAPH)*, 2017. 2, 6

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

[19] Nima Khademi Kalantari, Eli Shechtman, Connelly Barnes, Soheil Darabi, Dan B Goldman, and Pradeep Sen. Patch-based high dynamic range video. *ACM Trans. Gr.*, 2013. 2

[20] Erum Arif Khan, Ahmet Oguz Akyuz, and Erik Reinhard. Ghost removal in high dynamic range images. In *Proc. of Internatoinal Conference on Computational Photography*, 2006. 2

[21] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 7

[22] 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

[23] Wei-Sheng Lai, Jia-Bin Huang, Oliver Wang, Eli Shechtman, Ersin Yumer, and Ming-Hsuan Yang. Learning blind video temporal consistency. In *Proc. of European Conference on Computer Vision*, 2018. 6

[24] 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

[25] 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*, 2020. 2

[26] 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

[27] 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. 1, 2, 7, 8, 9

[28] Steve Mann and Rosalind W. Picard. On being 'undigital' with digital cameras: Extending dynamic range by combining differently exposed pictures. In *Proceedings of IS&T*, 1995. 1, 2

[29] Rafal Mantiuk and Maryam Azimi. A novel perceptually uniform encoding for adapting existing quality metrics for HDR. In *Picture Coding Symposium*, 2021. 9

[30] Belen Masia, Sandra Agustin, Roland W. Fleming, Olga Sorkine, and Diego Gutierrez. Evaluation of reverse tone mapping through varying exposure conditions. *ACM Transactions on Graphics (Proc. of ACM SIGGRAPH Asia)*, 2009. 2

[31] Christopher A. Metzler, Hayato Ikoma, Yifan Peng, and Gordon Wetzstein. Deep optics for single-shot high-dynamic-range imaging. In *Proc. of Computer Vision and Pattern Recognition*, 2020. 2

[32] Sayed Mohammad Mostafaviisfahani, Yeongwoo Nam, Jonghyun Choi, and Kuk-Jin Yoon. E2SRI: Learning to super-resolve intensity images from events. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 6

[33] Manish Narwaria, Rafal Mantiuk, Mattheiu P. Da Silva, and Patrick Le Callet. HDR-VDP-2.2: A calibrated method for objective quality prediction of high-dynamic range and standard images. *Journal of Electronic Imaging*, 2015. 8, 9

[34] Manish Narwaria, Matthieu Perreira Da Silva, and Le Callet Patrick. HDR-VQM: An objective quality measure for high dynamic range video. *Signal Processing: Image Communication*, 2015. 9

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

[36] Tae-Hyun Oh, Joon-Young Lee, Yu-Wing Tai, and In So Kweon. Robust high dynamic range imaging by rank minimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014. 2

[37] 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. 3

[38] Etienne Perot, Pierre de Tournemire, Davide Nitti, Jonathan Masci, and Amos Sironi. Learning to detect objects with a 1 megapixel event camera. In *Proc. of International Conference on Neural Information Proceeding Systems*, 2020. 9

[39] Christoph Posch, Daniel Matolin, and Rainer Wohlgenannt. An asynchronous time-based image sensor. In *Proc. of International Symposium on Circuits and Systems*, 2008. 3

[40] K Ram Prabhakar, Susmit Agrawal, Durgesh Kumar Singh, Balraj Ashwath, and R Venkatesh Babu. Towards practical and efficient high-resolution hdr deghosting with CNN. In *Proc. of European Conference on Computer Vision*, 2020. 2

[41] 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. 3, 9

[42] Marcel Santana Santos, Ren Tsang, and Nima Khademi Kalantari. Single image HDR reconstruction using a CNN with masked features and perceptual loss. *ACM Transactions on Graphics (Proc. of ACM SIGGRAPH)*, 2020. 1, 2, 7, 8, 9

[43] Jo Schlemper, Ozan Oktay, Liang Chen, Jacqueline Matthew, Caroline Knight, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention-gated networks for improving ultrasound scan plane detection. *arXiv preprint arXiv:1804.05338*, 2018. 4

[44] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proc. of Computer Vision and Pattern Recognition*, 2016. 4

[45] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proc. of International Conference on Learning Representations*, 2015. 6

[46] 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. 2

[47] Lvdi Wang, Li-Yi Wei, Kun Zhou, Baining Guo, and Heung-Yeung Shum. High dynamic range image hallucination. In *Proc. of the Eurographics conference on Rendering Techniques*, 2007. 1

[48] Ziwei Wang, Yonhon Ng, Cedric Scheerlinck, and Robert Mahony. An asynchronous kalman filter for hybrid event cameras. In *Proc. of Internatoinal Conference on Computer Vision*, 2021. 3

[49] Chun-Han Yao, Chia-Yang Chang, and Shao-Yi Chien. Occlusion-aware video temporal consistency. In *Proc. of ACM International Conference on Multimedia*, 2017. 9

[50] 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. 9

[51] Hang Zhao, Boxin Shi, Christy Fernandez-Cull, Sai-Kit Yeung, and Ramesh Raskar. Unbounded high dynamic range photography using a modulo camera. In *Proc. of Internatoinal Conference on Computational Photography*, 2015. 2

[52] Jing Zhao, Ruiqin Xiong, Hangfan Liu, Jian Zhang, and Tiejun Huang. Spk2ImgNet: Learning to reconstruct dynamic scene from continuous spike stream. In *Proc. of Computer Vision and Pattern*

*Recognition*, 2021. 3

[53] Stephan Zheng, Yang Song, Thomas Leung, and Ian J. Goodfellow. Improving the robustness of deep neural networks via stability training. In *Proc. of Computer Vision and Pattern Recognition*, 2016. 6

[54] Chu Zhou, Hang Zhao, Jin Han, Chang Xu, Chao Xu, Tiejun Huang, and Boxin Shi. UnModNet: Learning to unwrap a modulo image for high dynamic range imaging. *Proc. of International Conference on Neural Information Proceeding Systems*, 2020. 2

[55] 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. 3

**Jin Han** is currently a PhD candidate at the Graduate School of Information Science and Technology, the University of Tokyo. He received the B.Sc. degree in computer science from Sichuan University in 2018, and the M.Sc. degree in machine intelligence from Peking University in 2021. His research interests lie in neuromorphic cameras, event-based vision, and image restoration. He has served as a reviewer for CVPR, ICCV, ECCV, IJCV, TMM, *etc*.

**Yixin Yang** is a Ph.D. student in the School of Computer Science, Peking University. Her research interests span event-based imaging and vision, hybrid-camera super-resolution and HDR reconstruction.

**Peiqi Duan** is a PhD student in the School of Computer Science, Peking University. His research interests span event-based imaging and vision, single image super resolution and HDR image reconstruction. He has served as a reviewer/program committee member for IJCV, CVPR, ICCV, ECCV, *etc*.

**Chu Zhou** is a PhD student in the School of Artificial Intelligence, Peking University. His research interests span event-based vision, polarization-based vision, and HDR imaging. He has served as a reviewer for CVPR, ICCV, ECCV, *etc*.

**Lei Ma** is an associate research professor from National Biomedical Imaging Center, Peking University. He is the department head of the Life Simulation Research Center at the Beijing Academy of Artificial Intelligence (BAAI). He obtained the BS degree from Zhejiang University, MS degree from Digital ART Laboratory of Shanghai Jiao Tong University and PhD degree from the State Key Laboratory of Computer Science, Chinese Academy of Sciences. During 2010-2012, He worked for Autodesk China Research and Development Center as a graphic engineer. His research interests include realistic image synthesis, virtual reality and brain-inspired artificial intelligence.

**Chao Xu** received the B.E. degree from Tsinghua University, Beijing, China, in 1988, the M.S. degree from the University of Science and Technology of China, Hefei, China, in 1991, and the Ph.D. degree from the Institute of Electronics, Chinese Academy of Sciences, Beijing, in 1997. From 1991 to 1994, he was an Assistant Professor with the University of Science and Technology of China. Since 1997, he has been with the School of Electronics Engineering and Computer Science (EECS), Peking University, Beijing, where he is currently a Professor. His research interests are in image and video coding, processing, and understanding. He has authored or coauthored more than publications and five patents in these fields.

**Tiejun Huang** received the B.Sc. and M.Sc. degrees in computer science from Wuhan University of Technology, China in 1992 and 1995, respectively, and the Ph.D. degree in pattern recognition and image analysis from Huazhong (Central China) University of Science and Technology in 1998. He is currently a professor with the School of Computer Science, Peking University, and the Director of the Beijing Academy for Artificial Intelligence. His research areas include visual information processing and neuromorphic computing. He is a Fellow of CAAI, CCF, CSIG and vice chair of the China National General Group on AI Standardization. He published 300+ peer-reviewed papers on leading journals and conferences, and also co-editor of 4 ISO/IEC standards, 5 National standards and 4 IEEE standards. He holds 100+ granted patents. Professor Huang received National Award for Science and Technology of China (Tier-2) for three times (2010, 2012 & 2017).

**Imari Sato** received a BS degree in policy management from Keio University in 1994. After studying at Robotics Institute of Carnegie Mellon University as a visiting scholar, she received MS and Ph.D. degrees in interdisciplinary information studies from the University of Tokyo in 2002 and 2005, respectively. In 2005, she joined the National Institute of Informatics, where she is currently a professor/director of the Digital Contents and Media Sciences Research Division. Concurrently, she serves as a professor at the University of Tokyo and a visiting professor at Tokyo Institute of Technology. Her primary research interests are physics-based vision, spectral analysis, image-based modeling, and medical image analysis. She has received various research awards, including The Young Scientists' Prize from The Commendation for Science and Technology by the Minister of Education, Culture, Sports, Science and Technology (2009), and Microsoft Research Japan New Faculty award (2011).

**Boxin Shi** Boxin Shi received the BE degree from the Beijing University of Posts and Telecommunications, the ME degree from Peking University, and the PhD degree from the University of Tokyo, in 2007, 2010, and 2013. He is currently a Boya Young Fellow Assistant Professor and Research Professor at Peking University, where he leads the Camera Intelligence Lab. Before joining PKU, he did research with MIT Media Lab, Singapore University of Technology and Design, Nanyang Technological University, National Institute of Advanced Industrial Science and Technology, from 2013 to 2017. His papers were awarded as Best Paper Runner-Up at ICCP 2015 and selected as Best Papers from ICCV 2015 for IJCV Special Issue. He has served as an editorial board member of IJCV and an area chair of CVPR/ICCV. He is a senior member of IEEE.

# Supplementary Material: Hybrid High Dynamic Range Imaging fusing Neuromorphic and Conventional Images

Jin Han, Yixin Yang, Peiqi Duan, Chu Zhou, Lei Ma, Chao Xu,
Tiejun Huang, *Senior Member, IEEE,* Imari Sato, and Boxin Shi, *Senior Member, IEEE*

## 7 Data Simulation Process

The simulation of different types of data is important to make the trained neural network generalize to real data. The conventional RGB cameras and neuromorphic cameras operate in quite different manners, and output data with different mechanisms and formats. For example, RGB cameras capture images or videos with shutter to control the incident light. While event cameras and spike cameras respond to scene brightness asynchronously in each pixel. Therefore, it is necessary to analyze and model the data generation process and noise pattern of RGB cameras and neuromorphic cameras when conducting accurate simulation. We build our training and testing dataset by collecting HDR images from various image sources [1], [2], [9], [12], [13], [30], [37] and video sources [3], [4], [11], [24], [35]. The data simulation process is described as follows.

### 7.1 LDR Image Simulation

For LDR images, we synthesize them from HDR images like taking photos with a virtual camera [8]. Given the radiance of a scene $E$ and exposure time $\Delta t$, the HDR image $H$ is formed by $H = E \times \Delta t$. Then the formation process of LDR image $I$ from HDR image consists of 4 main steps: dynamic range clipping, noise simulation, non-linear mapping, and quantization, which is denoted as:

$$I = \lfloor 255 \cdot f(\max(\min(H, 1), 0) + n) \rfloor, \tag{17}$$

where $f$ is camera response function and $n$ represents noise. Therefore, we simulate LDR images according to the formation pipeline denoted in Eq. (17). The irradiance values of HDR images are firstly rescaled to $[0, 1]$, then multiplied by random exposure time $\Delta t$. We clip pixel values larger than $1.0$ as the saturated regions. By modeling photon sensing with Poisson distribution and the remaining stationary disturbances with Gaussian distribution, we add Poisson-Gaussian noise [10] to generate noisy LDR images. In our simulation, darker regions in LDR images suffer from more severe noise, which is consistent with real images from conventional RGB cameras. Finally, we apply non-linear mapping with different camera response curves from the database of response functions (DoRF) [15] and quantize them as 8-bit LDR images.

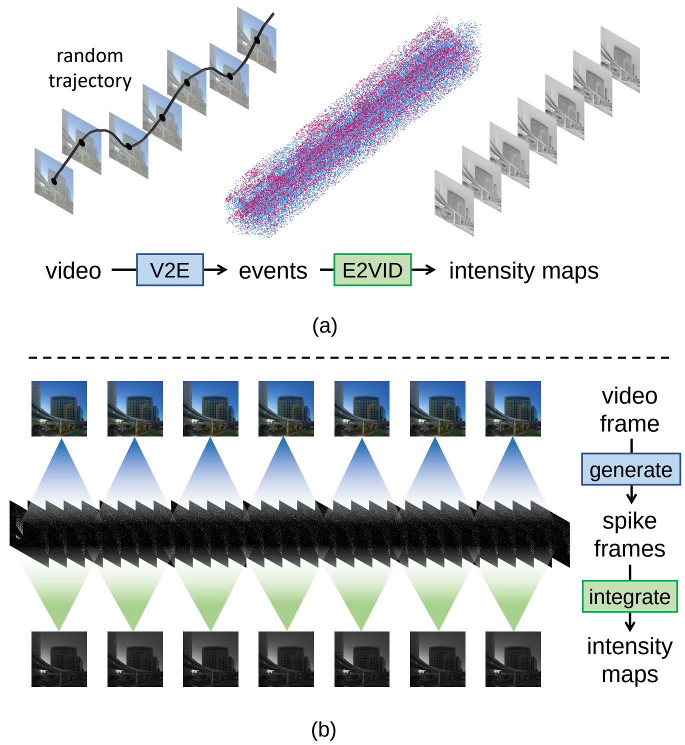


Figure. 12. The simulation process of event-based and spike-based intensity maps. (a) Event-based intensity maps simulation. (b) Spike-based intensity maps simulation.

### 7.2 Intensity Map Simulation

As for the intensity maps, we simulate them in accordance with the data generation mechanism of two different types of neuromorphic cameras.

**Event-based intensity map.** Event cameras detect the changes of brightness[1] and output a sequence of event

1. We use brightness as a perceived quantity, which refers to log intensity for scenes with uniform light.

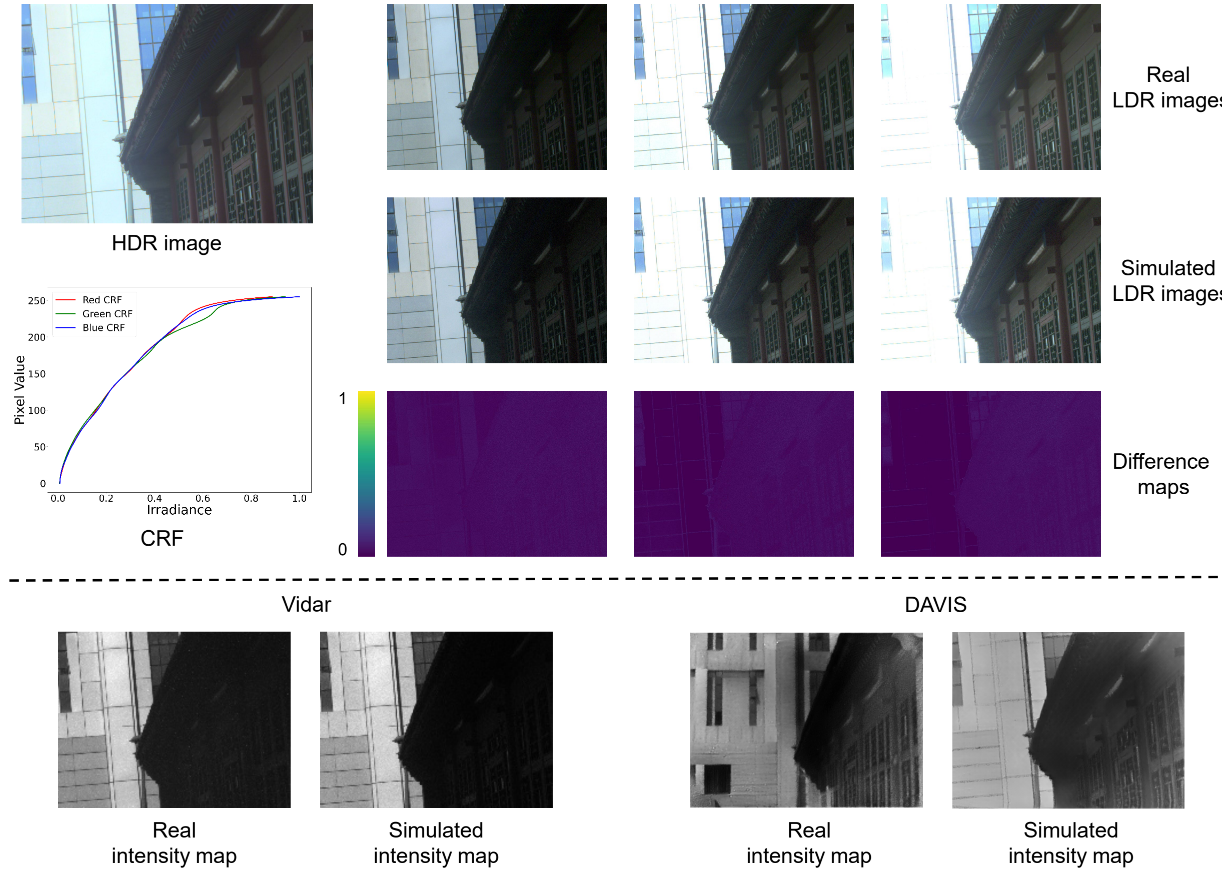


Figure. 13. Visual comparison of both LDR images (upper part) and intensity maps (bottom part) between synthetic data real data. The difference maps are computed from real LDR images subtract their corresponding synthetic LDR images. Since we have to switch the neuromorphic camera of hybrid camera, there exists misalignment in the intensity map from event camera (DAVIS [5]).

streams that contains timestamp, location, and polarity of brightness changes. Thanks to the HDR property of event sensors, the HDR radiance is recorded in a *differential* manner by event cameras. To simulate events, we generate a randomly moving trajectory for each HDR image and move it along the trajectory to get an HDR video. Then we use the event simulator (V2E) [7] to simulate event streams based on the movement between two consecutive frames. We set the threshold of event triggering to $0.18$ with a variance of $0.03$. The leak noise and temporal noise rates are set to $0.01$ and $0.001$, respectively. The parameters of V2E [7] are in accordance with real event cameras for more accurate simulation.

Intensity maps are then reconstructed from sparse event streams in an "integration" manner [23] or by a trained neural network [20], [31], [32], [36], [40]. Among those methods of reconstructing intensity maps from events, we choose the E2VID [32] network to transfer event streams into intensity maps. The process of events and intensity maps simulation is illustrated in Fig. 12 (a). Due to the limited resources of HDR videos, we use such a way to generate a large scale training event data.

**Spike-based intensity map.** Intensity maps can also be acquired from a spike camera (Vidar) [18]. Each pixel of the spike camera accumulates luminance independently, and outputs temporally asynchronous spikes. The accumulator at each pixel gathers luminance digitalized by the A/D converter. Once the accumulated intensity reaches a predefined threshold, a spike (indicated as a pixel value of $1$) is fired at this time stamp, then the corresponding accumulator is reset in which all the charges are drained. If there are no spike fired at this timestamp, we get $0$ for this pixel. Thus, spike cameras output spike frames with binary values in a high-temporal resolution ($40000$ spike frames per-second). We can easily find that the HDR scenes can be well recorded in an *integrated* manner by spike cameras due to the independent spiking mechanism. The bright regions will trigger dense spike streams because high luminance means a high frequency of spike firing, and vice versa. We first rescale pixel values of HDR frames to $[0, 1]$, then simulate spike frames for each HDR frame by regarding the pixel luminance values as spikes firing probabilities. Since Vidar suffers from dark current noise in low light intensity, we add fixed pattern noise [39] on each spike frame to achieve more realistic simulation.

To get the intensity map from spike frames, we apply a moving time window to integrate the spikes in a specific period, and the intensity map can be computed by counting these spikes pixel-wisely [18], as shown in Fig. 12 (b).

### 7.3 Similarity between Synthetic Data and Real Data

To demonstrate the effectiveness of our simulation, we show the similarity between our synthetic data and real-captured

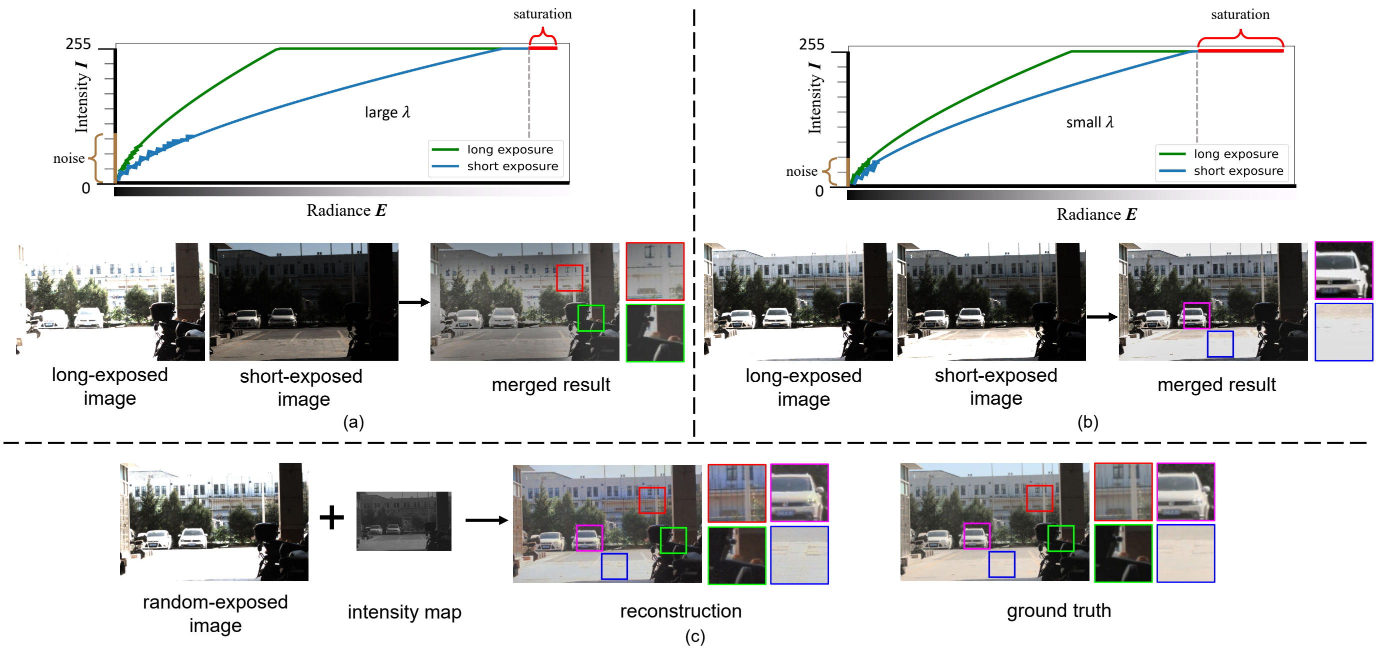


Figure. 14. Comparison between merging two LDR images with different exposure values and our approach of merging one LDR image and an intensity map using a real data example. (a) Merging two images with a large exposure ratio: The short-exposed image suffers from large noise in the low radiance area, while the long-exposed image is truncated in the region of sky and ground. The merged HDR image cannot preserve detailed information in both high radiance and low radiance regions. (b) Merging two images with a small exposure ratio: It reduces the noise, but loses information in the region around the reflection on the car and the ground because both LDR images are saturated in this area. (c) The fusion of RGB image and neuromorphic data achieves high-quality HDR reconstruction with broader dynamic range recovery and better details preservation by fusing HDR information encoded in intensity map and structural details in LDR image.

data in Fig. 13. We capture multiple LDR images with exposure bracketing and merge them to reconstruct the HDR image as ground truth. The camera response function (CRF) is estimated during the process. Then, we conduct our data simulation process to generate LDR images and intensity maps. For LDR images, we can easily get the exposure ratios by computing the linearized LDR images divided by HDR ground truth. We select three pairs of synthetic and real LDR images, and compute the difference maps between them. For intensity maps, we use both the spike camera (Vidar [18]) and the event camera (DAVIS346 [5]) to capture the same scenario. The intensity maps are reconstructed by integration of spike frames from Vidar [18], or by E2VID [32] from event streams. The results demonstrate our simulation is quite similar to real data on both LDR images and intensity maps.

## 8 Why not merge two LDR images?

Since we combine images from two different cameras, it is natural to consider why not replacing the neuromorphic camera with a much cheaper conventional camera and merging two LDR images to get an HDR result. In this section, we analyze the advantages of combining with an intensity map comparing with an additional LDR image.

For an extreme case, if we capture two images (for simplicity, we use LDR images that are captured using cameras with a linear CRF) with an exposure ratio of $256:1$, which means the saturation pixels in the short exposed image is set to be the darkest pixel in the long exposed image, covers the dynamic range up to 96 dB according to [29]. When the dynamic range of a scene is not very high, which can be well covered by two LDR images, merging these two images can achieve reasonably good results. However, two LDR images cannot cover very high dynamic range scenarios. In such a case, the advantage of NeuImg-HDR fusion naturally appears. An intensity map captured by a neuromorphic camera covers a much higher dynamic range (*e.g.*, 120 dB for DAVIS346) than any LDR image. However, images captured by a conventional camera suffer from noise or saturation if the exposure time is too short or too long. We analyze merging two LDR images with different exposure ratios $\lambda$, and demonstrate that the results from the NeurImg fusion method outperform that from merging two LDR images in very high dynamic range situation.

When merging two LDR images for HDR reconstruction, artifacts from noise and saturation are unavoidable. We provide such analysis using an example illustrated in Fig. 14. Firstly, we use our hybrid camera to capture a sequence of LDR images with different exposure time, while fixing other parameters like aperture and ISO. Then the only variable is shutter speed. So we can get the ground truth HDR image by merging these LDR images. Finally, we merge the selected two LDR images to acquire an HDR image using a state-of-the-art weighting and averaging method [26].

For case in Fig. 14 (a), we select LDR images with the shortest exposure time and longest exposure time to cover both very bright information (the outline of the distant building), and very dark details (the motorbike in the right side) to reconstruct a very high dynamic range scene. The image with long exposure (green line) has a large area of saturation while the image with a short exposure (blue line) is mainly influenced by noise. However, a too large exposure ratio brings the loss of detailed information, such as the artifacts on the car and ground. The merged result is mainly influenced by the too large exposure gap.

In contrast, if we try to preserve the detailed information accurately, it is inevitable to sacrifice the dynamic range due to the limit of conventional CMOS or CCD sensors, as shown in Fig. 14 (b). We choose LDR images with a smaller exposure ratio $\lambda$ to recover more accurate details. But the very high radiance area in the scene cannot be captured by either of the two differently exposed images. As a result, although the detailed information reconstructed by such an exposure ratio is less noisy than the case in Fig. 14 (a), it is impossible to recover the scene radiance out of dynamic range bound (*e.g.*, the reflection on the car and the ground). The merged result in this situation is dominant by saturation artifact in the over-exposed region. Since taking a good trade-off to suppress both noise and saturation artifacts by only merging two LDR images is practically difficult, existing exposure bracketing HDR approaches usually need more than three LDR images.

The proposed NeurImg fusion approach essentially differs from merging two LDR images. As shown in Fig. 14 (c), since the neuromorphic cameras capture intensity maps with a much higher dynamic range than any ordinary LDR image, we do not need to worry about how to balance the ratio of exposure time between two LDR images. The LDR image just needs to be exposed in an appropriate setting (neither too bright nor too dark) to the keep majority of chrominance information valid. Although intensity maps are noisy and low-resolution, the NeurImg fusion pipeline bridges domain gaps between the LDR images and the intensity maps as stated in Sec. 3 to realize HDR reconstruction. The zoom-in boxes show that the proposed method achieves much higher quality in both high (with little saturation, green inset) and low radiance (with little noise, red inset) regions.

TABLE 5
Quantitative comparison between the upsampling network and other pixel-interpolation methods on the performance of final results. The best values are in **bold**.

| | Method | PSNR-t↑ | SSIM-t↑ | LPIPS-t↓ |
|---|---|---|---|---|
| 2× SR | Upsampling net | **24.01** | **0.905** | **0.199** |
| | Bilinear | 23.55 | 0.903 | 0.200 |
| | Bicubic | 23.55 | 0.899 | 0.204 |
| 4× SR | Upsampling net | **22.65** | **0.893** | **0.281** |
| | Bilinear | 22.49 | 0.892 | 0.284 |
| | Bicubic | 21.37 | 0.870 | 0.310 |
| 8× SR | Upsampling net | **23.94** | **0.951** | **0.260** |
| | Bilinear | 23.86 | 0.951 | 0.265 |
| | Bicubic | 23.68 | 0.944 | 0.276 |

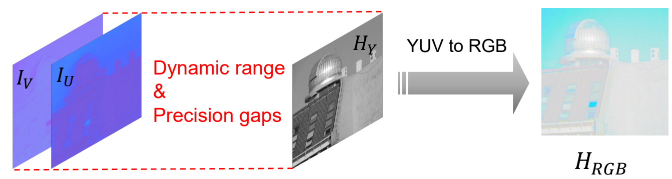


Figure. 15. A real example of explicit color conversion.

## 9 Discussion of Network Architecture

### 9.1 Effectiveness of Upsampling Network

We train the upsampling networks (2×, 4×, and 8×) to bridge the spatial resolution gap between intensity maps and LDR images. Compared to basic pixel interpolation methods like bilinear or bicubic interpolation, the trained upsampling network achieves better performance in final results. When using basic pixel interpolation methods, the noise in intensity maps will be preserved and enlarged in interpolated results. However, the upsampling network is trained with clean high-resolution intensity maps, which not only achieves super-resolution, but also suppresses noise in intensity maps. As shown in Table 5, the final results with our upsampling networks outperform other basic pixel interpolation methods.

### 9.2 Implicit *vs.* Explicit Color Space Conversion

In chrominance compensation network, we use implicit color space conversion from *YUV* to *RGB*. In the preliminary version of NeurImg-HDR [16], we used explicit color space conversion, which ignored the dynamic range gap and precision gap (*e.g.*, 8-bit unsigned integer data *vs.* 32-bit float data) between HDR luminance channel and LDR chrominance channels. If we simply concatenate the HDR luminance channel ($H_Y$) and LDR chrominance channels ($I_U, I_V$) as a 3-channel tensor $[H_Y, I_U, I_V]$, and explicitly transfer it to *RGB* color space, the converted $H_{RGB}$ loses precision in all three channels (*R*, *G*, and *B*), and tends to be colorless, especially after tone mapping, as shown in the $H_{RGB}$ in Fig. 15. In this case, it becomes more difficult for chrominance compensation network to restore the vivid color appearance, because the loss of precision has diffused into all three channels of $H_{RGB}$.

However, the implicit color space conversion considers the dynamic range gap and precision gap between luminance channel and chrominance channels by computing and compensating the residuals for *U*, *V* channels. Then the compensated chrominance channels $I_U, I_V$ have the same precision scale with HDR luminance channel $H_Y$. The $Y$, $U$, and $V$ channels are converted to $R$, $G$, and $B$ channels respectively in the feature levels. To properly assign weights for features from different color channels, we apply squeeze and excitation [17] operation in chrominance compensation network when converting to *RGB* color space. The final results demonstrate the advantages of implicit color space conversion over explicit one on both visual and quantitative evaluations.

### 9.3 Effectiveness of Recurrent Block

We use the recurrent block in chrominance compensation network to suppress the flickering artifacts when reconstructing HDR videos. Recurrent block has been proved to be an effective way in relieving flickering artifacts in previous works of video construction [19], [21], [32]. The recurrent block can be integrated to the network by plugging in a hidden state with chrominance compensation network as

TABLE 6
Quantitative comparison between the proposed NeurImg-HDR+ with recurrent block and DVP [25] on temporal consistency.

| | Recurrent block | DVP [25] |
|---|---|---|
| TCM↑ | **0.778** | 0.666 |

shown in Fig. 4 in the main manuscript. It doesn't increase the whole parameters and computation cost too much and achieves good performance in relieving flickering artifacts. In Table 6, we compare our recurrent-based network with deep video prior (DVP) [25], which regards the flickering artifacts as the noise in temporal domain, and use another network trained independently to suppress this kind of "noise". We evaluate the temporal consistency of test videos using temporal consistency metrics (TCM) metrics [38]. The results show that recurrent block outperforms DVP [25] in preserving temporal consistency when reconstructing HDR videos.

## 10 Analysis of Weighting Parameters in Loss Functions

In this section, we analyze how the weighting parameters $\alpha_i$ in loss functions for each sub-network are determined. There are three basic loss functions: pixel loss $\mathcal{L}_{pixel}$, perceptual loss [22] $\mathcal{L}_{perc}$, and adversarial loss [14] $\mathcal{L}_{adv}$, that we combine to optimize the network. Since the perceptual loss is the sum of $\ell_2$ norm of multiple layers from VGG-16 [34] network, the value of perceptual loss at the beginning of training is much larger than pixel loss, which is the $\ell_1$ norm distance between two images normalized in the range $[0, 1]$. Since the pixel loss basically minimizes the distance between the output and ground truth compared to perceptual loss, it is necessary to enlarge the weighting parameter of pixel loss to the same scale of perceptual loss to avoid that the total loss function is dominated by perceptual loss. While the adversarial loss for the generator is like an auxiliary loss to make the results more natural and vivid for human perception. So the scale of adversarial loss should be lower than the other two. The losses after multiplied by weighting parameters are plot in Fig. 16. If the weighting parameters are not suitably set, it will be difficult for the network to converge.

Besides the analysis above, we have conducted comprehensive ablation experiments to find an optimal combination of weighting parameters. As shown in Table 7, we firstly analyze how to balance the weights between $\mathcal{L}_{pixel}$ and $\mathcal{L}_{perc}$ in luminance fusion network[2], which is optimized by loss function $\mathcal{L}_L$ expressed in Eq. (11). Since the initial values of perceptual loss are much larger than pixel loss, we only change the weighting parameter of pixel loss to balance the weights between them. We find that the combination of weighting parameters as 100.0 for $\mathcal{L}_{pixel}$ and 3.0 for $\mathcal{L}_{perc}$ achieves the best results in comprehensive evaluations. Then for chrominance compensation network, which is optimized by loss function $\mathcal{L}_C$ expressed in Eq. (14), the adversarial loss is an extra auxiliary loss compared to $\mathcal{L}_L$.

2. Since the output of luminance fusion network are single-channel images, we can not compute HDR-VDP [28] metrics for them.

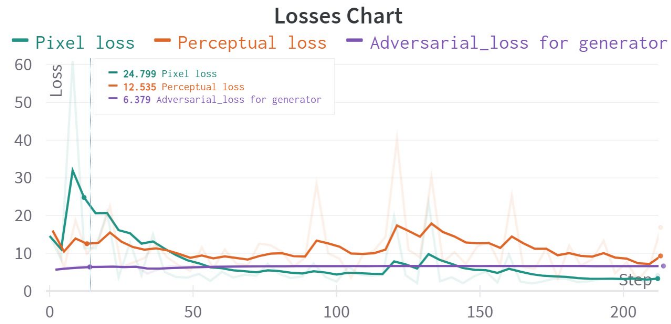


Figure. 16. The losses chart of the beginning 200 steps during training process. All the basic losses are multiplied by their weighting parameters, which are 100.0 for pixel loss, 3.0 for perceptual loss, and 10.0 for adversarial loss, respectively. The transparent curves behind are losses before smoothing.

TABLE 7
Quantitative comparison of variants with different weighting parameters. The best values are in red, and the second best values are in blue. The optimal variants are highlighted with gray.

| Loss Functions | Weighting parameters $\mathcal{L}_{pixel}$ | $\mathcal{L}_{perc}$ | $\mathcal{L}_{adv}$ | PSNR-t↑ | SSIM-t↑ | LPIPS-t↓ | HDR-VDP↑ |
|---|---|---|---|---|---|---|---|
| $\mathcal{L}_L$ | 1.0 | 3.0 | - | 18.77 | 0.844 | 0.160 | - |
| | 10.0 | 3.0 | - | 23.56 | 0.908 | 0.115 | - |
| | 100.0 | 3.0 | - | 24.76 | 0.932 | 0.136 | - |
| | 1000.0 | 3.0 | - | 25.99 | 0.921 | 0.188 | - |
| | 10000.0 | 3.0 | - | 24.28 | 0.911 | 0.232 | - |
| $\mathcal{L}_C$ | 100.0 | 3.0 | 1.0 | 23.71 | 0.904 | 0.204 | 9.210 |
| | 100.0 | 3.0 | 10.0 | 24.01 | 0.905 | 0.199 | 9.215 |
| | 100.0 | 3.0 | 100.0 | 23.76 | 0.904 | 0.202 | 9.206 |
| | 100.0 | 3.0 | 1000.0 | 23.66 | 0.902 | 0.201 | 9.183 |

We fix the weighting parameters of $\mathcal{L}_{pixel}$ and $\mathcal{L}_{perc}$ the same as $\mathcal{L}_L$ and change the weights of $\mathcal{L}_{adv}$. Results show that chrominance compensation network has the optimal performance when setting the parameter of $\mathcal{L}_{adv}$ to 10.0. Finally, the weighting parameters in loss functions are determined by theoretical analysis and ablation experiments.

## 11 Computational Cost

In this section, we analyze the number of parameters, the training time, and the inference speed of the proposed network. The number of parameters of upsampling networks, luminance fusion network, and chrominance compensation network are $2.00M$, $33.51M$, and $8.46M$, respectively. The total number of parameters of our network is $43.98M$. Since the proposed network has three sub-networks, and they are trained phase-to-phase, the total training time is the sum of three sub-networks, which is around 18 hours on an NVIDIA Titan RTX graphics card. For inference speed, we test our model on 70 HDR images, and compute the average inference speed. For a $512 \times 512$ image, our approach spends $101.55ms$ on an NVIDIA RTX 3080 Ti graphics card. Compared with preliminary NeurImg-HDR [16], which has $42.80M$ parameters with a inference speed of $69.89ms$ per image, the NeurImg-HDR+ has a comparable network size and spends more time on inference phase though, there is a huge improvement of the performance on restoring HDR images and videos.

## 12 HES-HDR DATASET

In this section, we describe the detailed information of the collected **H**ybrid **E**vent & **S**pike **HDR** (HES-HDR) dataset. We use the hybrid camera to capture various scenarios and build our dataset of RGB-neuromorphic video pairs. As shown in Table 8, in total, there are 20 video pairs, including 10 videos captured using the event camera (DAVIS346 [5]) and 10 videos captured using the spike camera (Vidar [18]). The HES-HDR dataset covers both indoor and outdoor HDR scenarios with camera motion or/and scene motion. We put a simple description to each video for easy reference. All the RGB frames are provided in $.jpg$ format. Event data are provided in stream-like $.txt$ format, and Spike data are provided in spike frame-like $.npz$ format.

## 13 GEOMETRIC CALIBRATION

In this section, we introduce the geometric calibration between two sensors of the hybrid camera. Since the two different views captured by the hybrid camera contain inevitable misalignment, we address this issue by conducting geometric calibration and cropping the center part from two views to extract the well-aligned regions as $I$ and $X$ for reconstruction. We consider homography and radial distortion between two camera views. Since event camera needs intensity changes to generate event signals, we choose to use a blinking checkerboard pattern displayed on a screen while keeping the hybrid camera system stationary. In order to extract the angular points from event data, we integrate the captured events over a time window (the time window should be no longer than the blinking period) to reconstruct the checkerboard image. As for the spike camera, the checkerboard pattern should be fixed without blinking and we just need to integrate a small period of spikes (around 300 spikes) data to reconstruct the intensity of the checkerboard.

It is easy for a conventional RGB camera to capture the stable checkerboard pattern. Then we convert it to grayscale. The angular points on the checkerboard are detected as the key points for calibration. The 2D-based calibration includes a homography transformation estimated based on the central key points and an anti-distortion transformation estimated based on all the key points. We crop the overlapped area of two images and force the height and width of $I$ to be even number multiples of those of $X$ for the purpose of following an upsampling operation by the proposed network.

## 14 ADDITIONAL HDR RESULTS

In addition to Fig. 8 in the main paper, we provide more comparisons on synthetic data between the proposed NeurImg-HDR+ and other methods in Fig. 17, including NeurImg-HDR [16], Liu *et al*. [27], Santos *et al*. [33], and LDR×2 [6]. We also show more results on real data in Fig. 18. More video results on synthetic data and real data are shown in the supplementary video.

## REFERENCES


[1] Funt et al. HDR dataset. https://www2.cs.sfu.ca/~colour/data/funt_hdr/. 1

[2] sIBL archive. http://www.hdrlabs.com/sibl/archive.html. 1

[3] A Banitalebi-Dehkordi, M Azimi, Y Dong, MT Pourazad, and P Nasiopoulos. Quality assessment of high dynamic range (HDR) video content using existing full-reference metrics. *ISO/IEC JTC1/SC29/WG11, France*, 2014. 1

[4] Amin Banitalebi-Dehkordi, Maryam Azimi, Mahsa T Pourazad, and Panos Nasiopoulos. Compression of high dynamic range video using the HEVC and h. 264/avc standards. In *International Conference on Heterogeneous Networking for Quality, Reliability, Security and Robustness*, 2014. 1

[5] Christian Brandli, Raphael Berner, Minhao Yang, Shih-Chii Liu, and Tobi Delbruck. A 240× 180 130 db 3 $\mu$s latency global shutter spatiotemporal vision sensor. *Journal of Solid-State Circuits*, 2014. 2, 3, 6

[6] 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 Internatoinal Conference on Computer Vision*, 2021. 6

[7] Tobi Delbruck, Yuhuang Hu, and Zhe He. V2E: From video frames to realistic DVS event camera streams. *arXiv preprint arXiv:2006.07722*, 2020. 2

[8] 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)*, Nov. 2017. 1

[9] Mark D Fairchild. The HDR photographic survey. In *Color and imaging conference*. Society for Imaging Science and Technology, 2007. 1

[10] Alessandro Foi, Mejdi Trimeche, Vladimir Katkovnik, and Karen Egiazarian. Practical poissonian-gaussian noise modeling and fitting for single-image raw-data. *IEEE Transactions on Image Processing*, 2008. 1

[11] 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 SPIE Electronic Imaging*, 2014. 1

[12] 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. *arXiv preprint arXiv:1704.00090*, 2017. 1

[13] Mathieu Garon, Kalyan Sunkavalli, Sunil Hadap, Nathan Carr, and Jean-François Lalonde. Fast spatially-varying indoor lighting estimation. In *Proc. of Computer Vision and Pattern Recognition*, 2019. 1

[14] 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. 5

[15] Michael D Grossberg and Shree K Nayar. What is the space of camera response functions? In *Proc. of Computer Vision and Pattern Recognition*, 2003. 1

[16] 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. 4, 5, 6

[17] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proc. of Computer Vision and Pattern Recognition*, 2018. 4

[18] Tiejun Huang, Yajing Zheng, Zhaofei Yu, Rui Chen, Yuan Li, Ruiqin Xiong, Lei Ma, Junwei Zhao, Siwei Dong, Lin Zhu, et al. 1000× faster camera and machine vision with ordinary devices. *Engineering*, 2022. 2, 3, 6

[19] Yan Huang, Wei Wang, and Liang Wang. Video super-resolution via bidirectional recurrent convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 4

[20] S. Mohammad Mostafavi I., Lin Wang, Yo-Sung Ho, and Kuk-Jin Yoon. 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

[21] Takashi Isobe, Xu Jia, Shuhang Gu, Songjiang Li, Shengjin Wang, and Qi Tian. Video super-resolution with recurrent structure-detail network. In *Proc. of European Conference on Computer Vision*, 2020. 4

[22] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Proc. of European Conference on Computer Vision*, 2016. 5

[23] H Kim, A Handa, R Benosman, SH Ieng, and AJ Davison. Simultaneous mosaicing and tracking with an event camera. In *Proc. of the British Machine Vision Conference*, 2014. 2
[24] Joel Kronander, Stefan Gustavson, Gerhard Bonnet, and Jonas Unger. Unified HDR reconstruction from raw CFA data. In *Proc. of Internatoinal Conference on Computational Photography*, 2013. 1
[25] Chenyang Lei, Yazhou Xing, and Qifeng Chen. Blind video temporal consistency via deep video prior. *Proc. of International Conference on Neural Information Proceeding Systems*, 2020. 5
[26] 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*, 2020. 3
[27] 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. 6
[28] Manish Narwaria, Rafal Mantiuk, Mattheiu P. Da Silva, and Patrick Le Callet. HDR-VDP-2.2: A calibrated method for objective quality prediction of high-dynamic range and standard images. *Journal of Electronic Imaging*, 2015. 5
[29] Shree K Nayar and Tomoo Mitsunaga. High dynamic range imaging: Spatially varying pixel exposures. In *Proc. of Computer Vision and Pattern Recognition*, 2000. 3
[30] Hiromi Nemoto, Pavel Korshunov, Philippe Hanhart, and Touradj Ebrahimi. Visual attention in LDR and HDR images. In *Proc. of International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, 2015. 1
[31] 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
[32] 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. 2, 3, 4
[33] Marcel Santana Santos, Ren Tsang, and Nima Khademi Kalantari. Single image HDR reconstruction using a CNN with masked features and perceptual loss. *ACM Transactions on Graphics (Proc. of ACM SIGGRAPH)*, 2020. 6
[34] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proc. of International Conference on Learning Representations*, 2015. 5
[35] Li Song, Yankai Liu, Xiaokang Yang, Guangtao Zhai, Rong Xie, and Wenjun Zhang. The SJTU HDR video sequence dataset. In *International Conference on Quality of Multimedia Experience*, 2016. 1
[36] Wenming Weng, Yueyi Zhang, and Zhiwei Xiong. Event-based video reconstruction using transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 2
[37] 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. 1
[38] Chun-Han Yao, Chia-Yang Chang, and Shao-Yi Chien. Occlusion-aware video temporal consistency. In *Proc. of ACM International Conference on Multimedia*, 2017. 5
[39] Lin Zhu, Jianing Li, Xiao Wang, Tiejun Huang, and Yonghong Tian. NeuSpike-Net: High speed video reconstruction via bio-inspired neuromorphic cameras. In *Proc. of Internatoinal Conference on Computer Vision*, 2021. 2
[40] 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

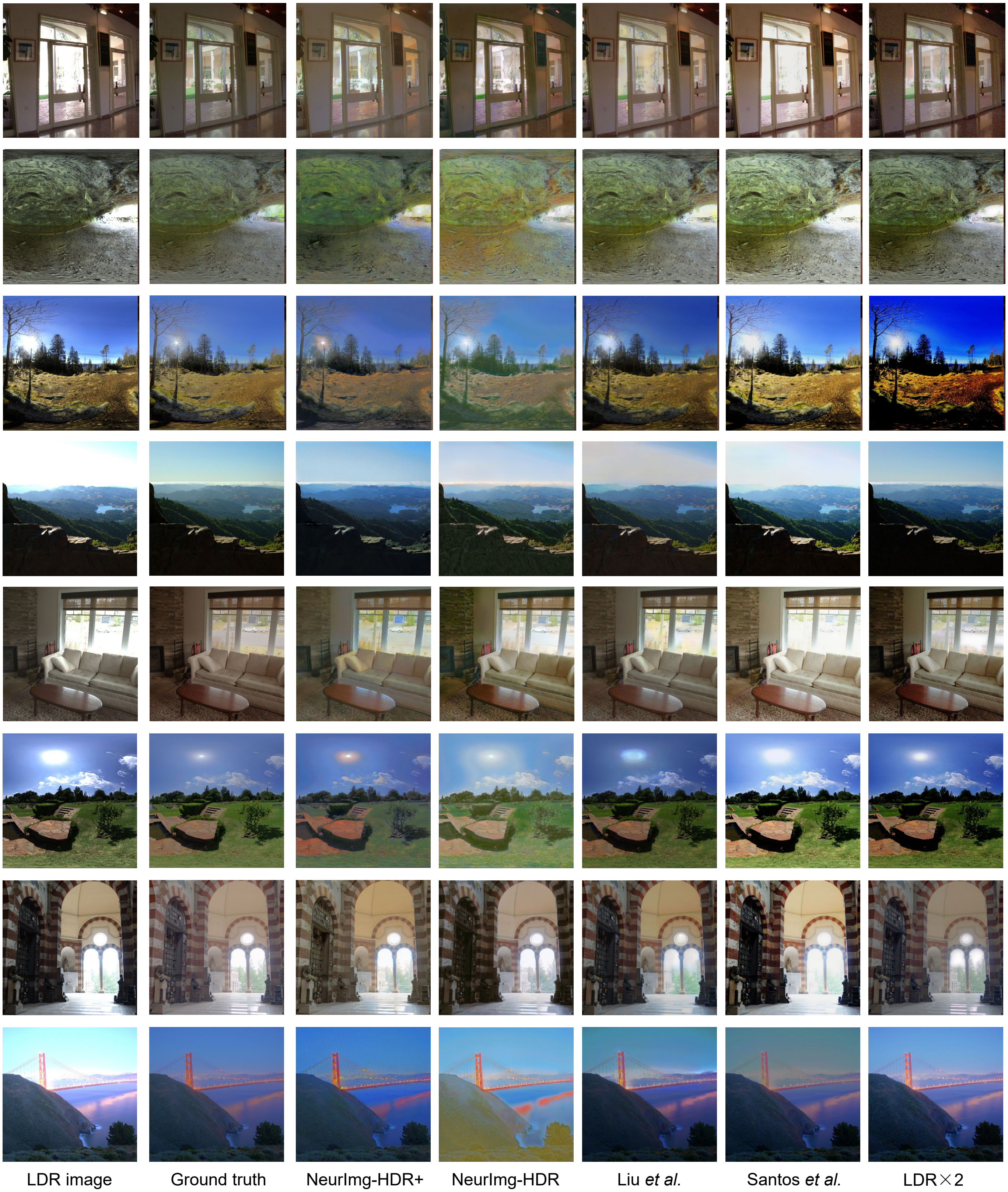


Figure. 17. More visual results on synthetic data.

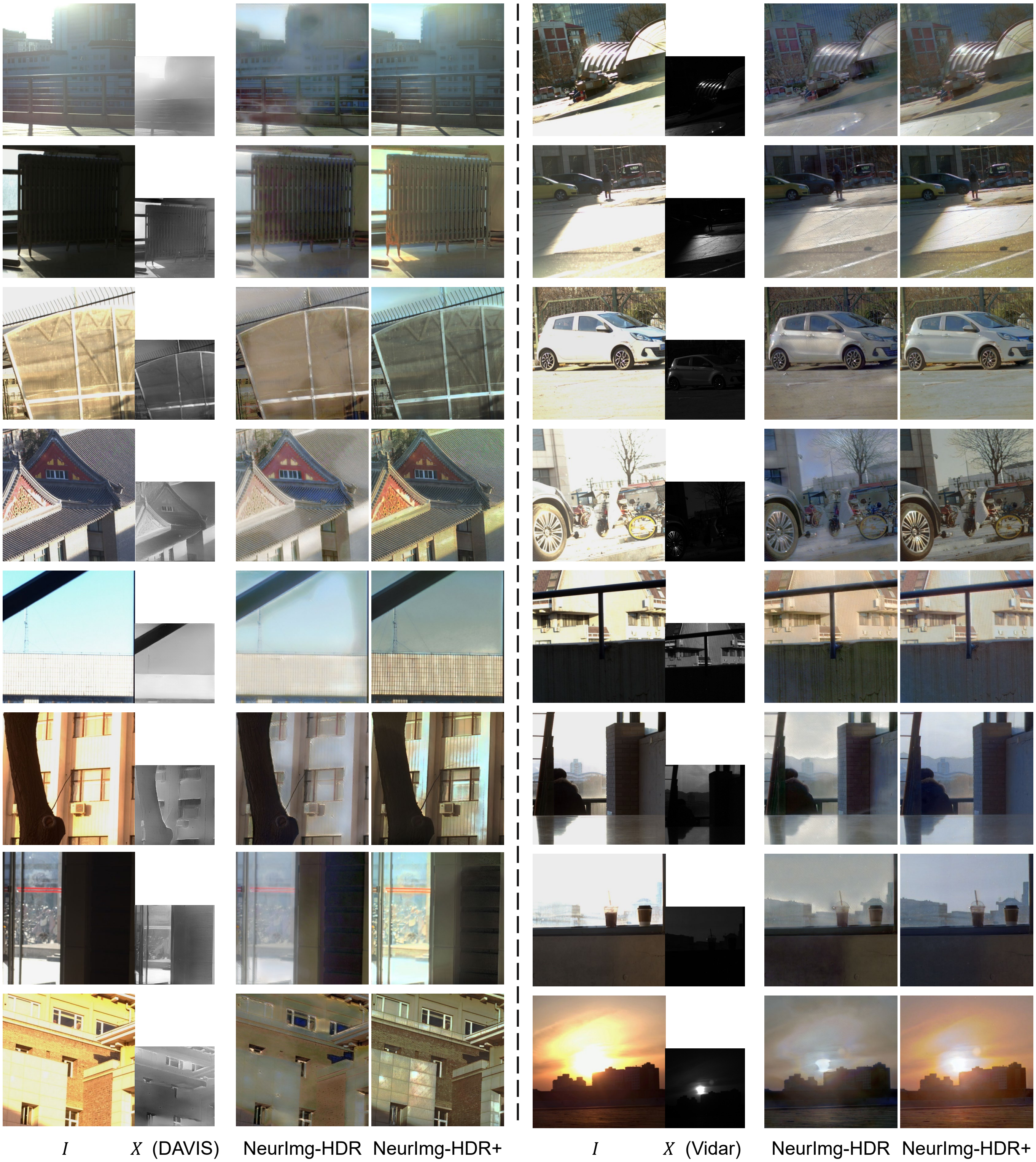


Figure. 18. More visual results on real data.

TABLE 8
Details of our HES-HDR dataset.

| **serial number** | **neuromorphic camera** | **# of frames** | **spatial resolution** | **indoor/ outdoor** | **camera motion** | **scene motion** | **description** |
|---|---|---|---|---|---|---|---|
| #event_01 | DAVIS346 | 164 | 260×346 | outdoor | ✓ | | the sun shining on the ground |
| #event_02 | DAVIS346 | 391 | 237×329 | indoor | ✓ | ✓ | capturing the outside scene through windows |
| #event_03 | DAVIS346 | 328 | 237×329 | indoor | ✓ | | capturing the windows from indoor |
| #event_04 | DAVIS346 | 389 | 237×329 | outdoor | ✓ | | the wall reflecting the sunlight |
| #event_05 | DAVIS346 | 378 | 237×329 | outdoor | ✓ | | static car and fence with camera motion |
| #event_06 | DAVIS346 | 396 | 237×329 | outdoor | ✓ | | capturing the buildings with camera motion |
| #event_07 | DAVIS346 | 212 | 237×329 | indoor | ✓ | | capturing the windows from indoor |
| #event_08 | DAVIS346 | 162 | 237×329 | indoor | ✓ | | capturing the outside scene through windows |
| #event_09 | DAVIS346 | 311 | 237×329 | outdoor | ✓ | ✓ | a building reflecting the sunlight |
| #event_10 | DAVIS346 | 350 | 237×329 | outdoor | ✓ | | windows of a building reflecting the sunlight |
| #spike_01 | Vidar | 398 | 250×400 | outdoor | ✓ | ✓ | a car passing away |
| #spike_02 | Vidar | 397 | 250×400 | outdoor | ✓ | ✓ | the sun shining on the ground |
| #spike_03 | Vidar | 387 | 250×400 | outdoor | ✓ | | a static car |
| #spike_04 | Vidar | 145 | 250×400 | outdoor | ✓ | | a static car |
| #spike_05 | Vidar | 387 | 250×400 | outdoor | ✓ | | a static car and bicycles |
| #spike_06 | Vidar | 135 | 250×400 | outdoor | ✓ | | a static car and fences |
| #spike_07 | Vidar | 397 | 250×400 | outdoor | ✓ | | the roof of a building |
| #spike_08 | Vidar | 212 | 250×400 | indoor | | ✓ | a passenger going down the stairs (short) |
| #spike_09 | Vidar | 396 | 250×400 | indoor | | ✓ | a passenger going down the stairs (long) |
| #spike_10 | Vidar | 397 | 250×400 | outdoor | ✓ | | capturing the sun directly |