Multi-Bracket High Dynamic Range Imaging with Event Cameras

Nico Messikommer*,[1] Stamatios Georgoulis*,[2] Daniel Gehrig[1] Stepan Tulyakov[2]
Julius Erbach[2] Alfredo Bochicchio[2] Yuanyou Li[2] Davide Scaramuzza[1]

[1]Dept. of Informatics, Univ. of Zurich and Dept. of Neuroinformatics, Univ. of Zurich and ETH Zurich
[2]Huawei Technologies, Zurich Research Center

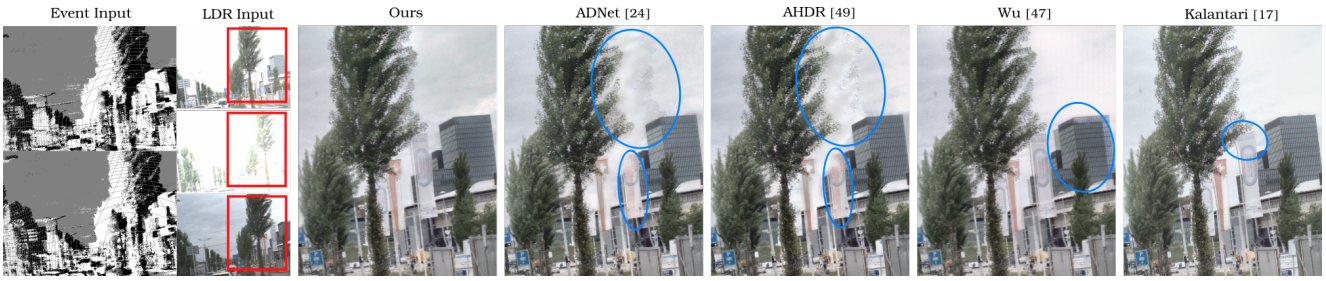

Figure 1. We propose to leverage the high dynamic range properties of an event camera to enhance a multi-bracket pipeline for HDR imaging. Events substantially improve the alignment of LDR images in scenes with both object and camera motion.

Abstract

Modern high dynamic range (HDR) imaging pipelines align and fuse multiple low dynamic range (LDR) images captured at different exposure times. While these methods work well in static scenes, dynamic scenes remain a challenge since the LDR images still suffer from saturation and noise. In such scenarios, event cameras would be a valid complement, thanks to their higher temporal resolution and dynamic range. In this paper, we propose the first multi-bracket HDR pipeline combining a standard camera with an event camera. Our results show better overall robustness when using events, with improvements in PSNR by up to 5dB on synthetic data and up to 0.7dB on real-world data. We also introduce a new dataset containing bracketed LDR images with aligned events and HDR ground truth.

Multimedia Material

Additional qualitative results can be viewed in this video: https://youtu.be/fw9-gNg6cM8

1. Introduction

Natural scenes have considerable variations in their illumination. On a sunny day, the same scene may depict a bright sky or sun as well as deep shadows with a brightness ratio of 1:10,000. The human eye is accustomed to perceiving such a *dynamic range* in natural scenes. Hence we expect the same from photos. However, conventional cameras have to set a global exposure time for the entire image and compress its full dynamic range into 10-14 bits. This is achieved through clipping, compression, and quantization of intensity values. As a result, captured images look less vivid and unimpressive. This problem is becoming more prominent as displays can natively support high dynamic range content.

Exposure bracketing [4] is a popular method for acquiring *high dynamic range (HDR)* photos without special hardware. The method operates by capturing several *low dynamic range (LDR)* photos of the same scene under different exposures, aligning them, and fusing them together. This method provides great results when there is no camera or scene motion. Unfortunately, in the age of handheld smartphone photography, this solution has practical limitations. In the presence of scene or camera motion, this method must deal with LDR image misalignments and degradations. Several works [18, 21, 24, 50, 52, 54] tried to tackle these problems, but the latter cannot be robustly solved using standard techniques, *e.g.* image-based alignment, because the bracketed LDR images violate the *brightness constancy assumption* [18, 45]. *Exposure compensation* on the bracketed LDR images is often used as a countermeasure, yet image saturation, noise, and motion blur still pose real challenges for these image-based HDR works.

By contrast, the human eye can reliably perceive a scene in a high dynamic range. Event cameras [1, 40] are novel *neuromorphic* vision sensors that attempt to mimic the high

*equal contribution

dynamic range and the high speed response of biological vision systems. Instead of measuring synchronously absolute intensity frames at fixed time intervals, event cameras only measure the *changes* in logarithmic intensity and do this independently for each pixel, resulting in an asynchronous stream of *events*. The resulting data have high temporal resolution, HDR and do not suffer from motion blur. Recent works have leveraged the outstanding properties of event cameras to generate high-speed video reconstructions with HDR properties from events [34, 46, 58, 62]. Nonetheless, event cameras only measure changes in brightness, and thus global image reconstruction from events is ill-posed. This places fundamental limits on these event-based HDR methods, which are further aggravated by persisting technical limitations of the current event sensor technology, *i.e.*, low spatial resolution, and lack of events in low contrast regions.

To bypass these limitations, hybrid HDR works [14, 49] proposed to combine a single LDR image captured by a frame camera with events from an auxiliary event camera, thus leveraging the advantages of both. The authors proposed different ways to enhance the luminance of the LDR image using the added information from events, but they still rely on the chrominance of the LDR image for color. Therefore, these methods have to hallucinate color in the areas where the LDR image is saturated. Moreover, since low contrast parts of the scene do not trigger events, these methods suffer from poor details and are highly dependent on motion since static scenes do not generate events.

In this paper, we propose to marry image-based exposure bracketing with event-based vision to get the best of both worlds. By doing so, we can enhance key parts of the HDR imaging pipeline, like the alignment of bracketed LDR images and their fusion to HDR at the feature level. Our main contributions can be summarized as:

1. We introduce EHDR, the first method that combines bracketed LDR images and synchronized events for HDR imaging. In doing so, EHDR is more robust than image-based HDR works to LDR image saturation, noise and motion blur, and, unlike event-based and hybrid HDR works, can faithfully reproduce color and low contrast details regardless of scene motion.

2. We propose a deformable feature alignment module that leverages motion information from both images and events to guide the learning of kernel offsets and modulation masks, which, in turn, leads to significant performance improvements in PSNR and SSIM on both synthetic and real data.

3. To evaluate the proposed method and facilitate future research on the topic, we collect the first HDR dataset consisting of sequences with ground truth HDR, bracketed LDR images, and synchronized and aligned event data, called HDR-ERGB.

2. Related Work

Current HDR solutions can be classified into: single-exposure and multi-exposure methods that only rely on standard image sensors; event-based methods that only use neuromorphic sensors; and hybrid methods that utilize both. For a more detailed overview, we refer to [47]

Single-exposure methods infer an HDR image [5,23,27, 36,37] or a stack of differently exposed LDR images [6,20] from a single LDR image, and can therefore be applied to legacy LDR content. They operate by inverting *non-linear mapping*, *quantization* and *saturation clipping* applied during image acquisition; thus, they are also called *inverse tone mapping (iTM)* methods. Deep learning iTM approaches [5, 6, 23, 27, 37] recently achieved good results in recovering saturated details by utilizing large image context; however, they still essentially hallucinate details in saturated regions from the surrounding non-saturated context, and, thus, are not suitable for commercial applications.

Other works [2, 12, 17, 56] expose hidden details in dark areas by tone mapping; however, they are also not able to recover details that are not present in the original image.

Multi-exposure HDR methods acquire multiple LDR images under different exposures and fuse them into an HDR image. The LDR images with different exposures can be acquired at once using a *special camera system*, *e.g.* system with two image sensors and common lens with beam splitter [42] or two separate cameras [32, 43]. However, such systems have higher power consumption and memory requirements. There exist more exotic systems too, such as systems with pixel-wise exposure control [39], out-of-range intensity warping [59], or gradient encoding [44], but the latter are expensive and inaccessible to regular users.

Alternatively, LDR images can be acquired by standard camera hardware using *exposure bracketing* as in [3, 18, 24, 25, 28, 30, 33, 38, 50, 52, 53, 55] by taking several photos of the same scene under different exposures. However, this technique has a longer capturing time, which, in turn, can result in misaligned and blurry LDR images for longer exposures due to camera motion or non-static scenes. The latter can be avoided by synthetic exposure, *i.e.* capturing multiple underexposed images with a high gain and synthesizing long exposed images [15, 22]. Still, synthetic exposure provides limited dynamic range improvement and still suffers from the images' misalignment problem. Note that, a naive fusion of the misaligned LDR images is prone to ghosting artifacts in the HDR image.

Multi-exposure alignment (see survey [45]). A simple solution to the alignment problem is to *reject moving object pixels* [29, 52] or use robust exposure fusion [15, 22]. However, these methods often fail to identify moving objects and are unable to reconstruct them in HDR. Another solution is to explicitly *estimate and compensate motion* between LDR images [9,13]. These methods register all expo-

sures to a reference LDR image by estimating global transformations (e.g., homography) or optical flow. Since different exposures have intensity differences, motion estimation is performed using robust similarity measures [15], or features [9], or is performed after exposure compensation [18]. Still, it is hard to directly estimate motion from bracketed LDR images because long exposures suffer from saturation and motion blur, while short exposures from image noise. Also, image-based motion estimation methods suffer from ambiguity when the motion is large.

Neuromorphic sensor. All HDR approaches described so far rely on frame-based cameras that capture the entire image at once. However, the dynamic range of frame-based cameras is limited as they need to encode all scene intensities into a fixed number of bits (e.g., 8 bits). While frame-based cameras are the sensor of choice for computer vision applications, new biologically-inspired *neuromorphic sensors* [8] are gaining popularity. Instead of measuring the intensity of every pixel, these sensors report asynchronous events whenever the log-intensity change at an individual pixel reaches a certain threshold, called *contrast threshold*. Since neuromorphic sensors do not encode absolute intensity levels, they do not saturate in extreme lighting conditions, such as bright daylight and night, and they currently have $>$ 120 dB dynamic range versus 50 dB - 80 dB for a frame camera. Because of these properties, event cameras have recently become popular for HDR reconstruction.

Event-based methods [34, 46, 58, 62] reconstruct intensity frames from events using deep learning. These methods do not explicitly target HDR imaging, but they rather reconstruct HDR-like images as a consequence of using events. Their reconstruction quality is typically low in part because current event cameras have low spatial resolution ($\leq$ 1 MP) and do not provide events in low contrast regions. These problems may be alleviated in the future through the advancement of the current sensor technology. Yet, the resolution of event cameras will always lag behind frame-based cameras due to the higher complexity of their pixel circuitry. Still, there are more fundamental limitations. First, the reconstruction of an image from events is an ill-posed problem due to the lack of absolute intensity information and varying contrast thresholds. Therefore, event-based methods often produce images with incorrect global contrast. Second, event cameras are unable to capture details below the contrast threshold, which becomes evident in low light as noted in [58]. Therefore, the results of purely event-based methods typically lack details. Finally, the quality of event-based image reconstruction is motion-dependent, usually performing poorly on scenes with little motion.

Hybrid single-exposure HDR methods [14, 49] combine an LDR image captured by a high-resolution frame camera with events acquired by an auxiliary event camera. These methods propose different ways to enhance the luminance of the LDR image using the added information from events. Yet, they still rely on the chrominance of the LDR image when it comes to color, as most event cameras currently do not provide color information. Inevitably, in saturated or underexposed regions, these approaches ought to hallucinate color, and sometimes even structure, from the non-saturated nearby context, which is unreliable when the saturated regions are large. Essentially, current single-exposure hybrid works suffer the same drawbacks as iTM methods, that is, they hallucinate results, yet they produce more educated guesses guided by the added event information. In contrast, we propose the *first multi-exposure hybrid method* that eliminates any hallucination component and instead relies on actual measurements generated by combining exposure bracketing with event-based vision.

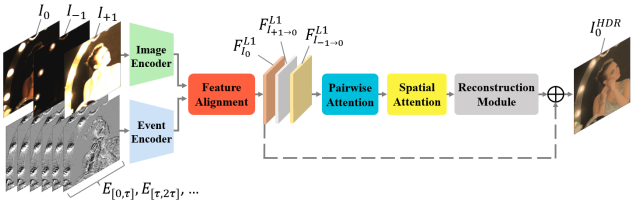

Figure 2. Overview of our proposed multi-bracket HDR pipeline. The events and frames are used to generate aligned features, which are processed by a pairwise and a spatial attention module. Finally, the reconstruction module outputs the HDR prediction.

3. Method

3.1. Method Overview

Let us assume the following setup. An image sensor captures a burst of bracketed LDR images (I_0, I_{-1}, I_{+1}) corresponding to mid, short and long exposure times, respectively. The images are captured by a handheld camera and have significant camera and scene motion between them. In parallel, an event sensor records a stream of asynchronous events $E = ((t_0, x_0, y_0, p_0), (t_1, x_1, y_1, p_1), \dots)$, with t denoting the time that an event was triggered, (x, y) the spatial position of the event, and p its polarity (positive or negative). We assume that the event data and the images are temporally synchronized and spatially aligned and have the same resolution, as though as they come from one sensor (*e.g.* a hybrid sensor). Our goal is to reconstruct the HDR image at mid exposure time I_0^{HDR}.

Fig. 2 gives an overview of our method, called EHDR. As a pre-processing step, the stream of raw events E is split into fixed-duration chunks ($E_{[0,\tau]}$, $E_{[\tau,2\tau]}$, ...), with each chunk containing all events within a time window τ. The events within each chunk are represented as a *voxel grid* [11, 61] with 5 equally-sized temporal bins. The bracketed LDR images and chunked events are passed to an *Image Encoder* and *Event Encoder*, respectively. The resulting multi-scale feature representations of each non-

reference image are aligned to the reference image I_0 by a *Feature Alignment* module using the chunked event features. In a next step, the aligned feature representations ($F^{L1}_{I_{-1\to 0}}$, $F^{L1}_{I_{+1\to 0}}$) and the reference features ($F^{L1}_{I_0}$) are concatenated and inputted to a *Pairwise Attention* module that fuses them into a single feature representation that is consequently passed to a *Spatial Attention* module that aims to recover fine details by deep spatial feature transform. Finally, the attended feature representation is decoded by a *Reconstruction* module that produces the HDR image prediction $\hat{I}_0^{HDR}$. Below, we explain the individual modules of EHDR in more detail.

3.2. Encoder & Reconstruction Modules

Our system consists of two encoders, *i.e.*, *Image Encoder* and *Event Encoder*, that follow the same multi-scale architecture. In particular, each encoder uses 5 residual blocks [16], each having a (Conv2d, ReLU, Conv2d) design with a residual connection. Unless stated otherwise, all convolutions use 3x3 kernels with 64 channels, padding 1, stride 1, and no BatchNorm layers. The 5 residual blocks are followed by 2 down-sampling blocks, each with a (Conv2d, LeakyReLU, Conv2d, LeakyReLU) design with the first Conv2d having a stride of 2, essentially generating a pyramid of feature representations across 3 scales (L1: 1, L2: 1/2, L3: 1/4). The multi-scale architecture allows for coarse to fine feature alignment inside the *Feature Alignment* module, such that larger motions can be compensated too. Note that, the input to the *Event Encoder* module are chunked events in voxel grid format with 5 channels, while the *Image Encoder* module expects images with 6 channels, as in prior HDR works [18, 19, 52]. In particular, the non-linear LDR image I_j is concatenated with its exposure compensated linearized version $\tilde{I}_j = f^{-1}(I_j)/t_j$, where f is the camera response function, simplified to a gamma curve with $\gamma = 2.2$ in our case[1], and t_j is the exposure time of image j. This exposure compensation procedure approximates brightness constancy among the LDR images required for alignment purposes.

The Reconstruction module consists of 10 residual blocks that follow the exact same design as the encoder ones. Note that, a skip connection from the feature representation of the reference LDR image $F^{L1}_{I_0}$ is added to the output of this module before decoding it to the HDR image $\hat{I}_0^{HDR}$ using a Conv2d layer (64 to 3 channels).

3.3. Feature Alignment Module

Exposure bracketing in handheld photography of dynamic scenes may lead to image misalignments, which need to be resolved in order to avoid ghosting artifacts in the HDR image. Existing HDR works estimate some form of 'motion', be it global or local, between the LDR images after exposure compensation [18, 21, 24, 50, 52, 54]. The latter is implicitly required for motion estimation which is dependent on brightness constancy in the LDR images. However, even exposure compensation can not guarantee brightness constancy, as the LDR images also suffer from saturation, noise, or motion blur. This renders motion estimation from bracketed LDR images an ill-posed problem. Events do not suffer from saturation and motion blur, and carry fine-grained information about motion between bracketed LDR images due to their high temporal resolution. As a result, they are a natural fit for image alignment purposes in dynamic scenes, relaxing the strong assumptions of brightness constancy and motion linearity in current HDR works. However, events are sparse by nature and absent in low contrast regions due to the limited contrast sensitivity of current event cameras, rendering them incomplete for motion estimation in every image patch if used on their own. To leverage this complementarity between images and events and get the best of both worlds, in this paper, we propose to combine these two sources of motion information for LDR image alignment in HDR photography.

To this end, we design a *Feature Alignment* module that leverages information from both images and events to align the non-reference images at the feature level. Our starting

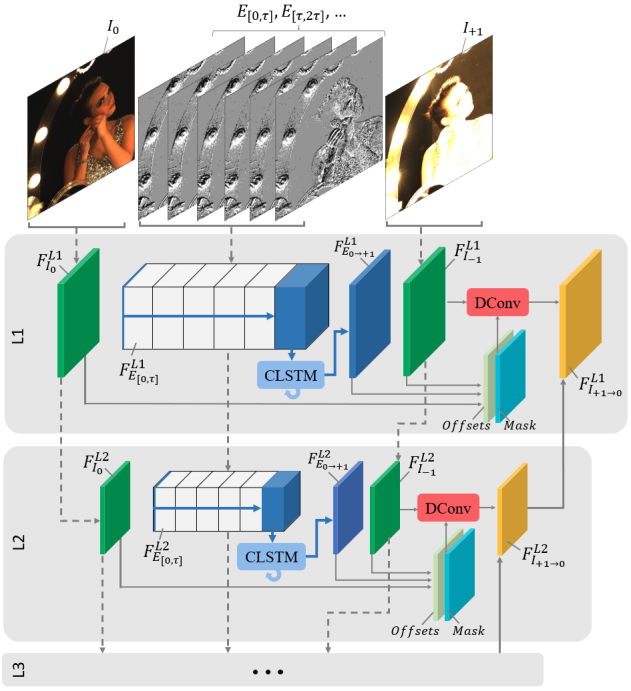

Figure 3. Overview of our alignment module. The event features (blue) are processed by a ConvLSTM (CSTM) and afterwards combined with image features (green) to compute the offsets and masks for the deformable convolution (DConv). The output of the deformable convolutions are the aligned image features (yellow).

[1]We assume that a proper image linearization has been performed in advance using the camera response function (CRF) computed from camera calibration techniques. Hence, gamma can replace CRF in this case.

point is the Pyramid, Cascading, and Deformable (PCD) module introduced in EDVR [48], used in HDR too [24]. The PCD module uses a pyramid of feature representations to estimate kernel offsets and modulation masks for modulated deformable convolutions [60] at each pyramidal level separately, which are then used to gradually 'align' the feature representations of non-reference images to the reference image. We build upon this core idea of deformable feature alignment, but introduce the following modifications:

(1) The kernel offsets and modulation masks at level $L_l, l \in \{1, 2, 3\}$ are jointly computed from image ($F^{L_l}_{I_0}$, $F^{L_l}_{I_{+1}}$) and chunked event ($F^{L_l}_{E_{[0,\tau]}}$, $F^{L_l}_{E_{[\tau,2\tau]}}$, ..., $F^{L_l}_{E_{[-\tau,+1]}}$) features. Fig. 3 illustrates this procedure. Although image features can be directly used for the computation, chunked event features first need to be integrated across the entire time window ($\{0, \tau, ..., +1\}$). We achieve this by introducing ConvLSTM modules [51], one at each level L_l, that output the integrated event features $F^{L_l}_{E_{0\rightarrow+1}}$. An added benefit of the ConvLSTM modules is that they can 'compensate' the camera motion in-between chunked events, which inevitably causes events to appear in different locations than the one they were originally triggered. The integrated event features $F^{L_l}_{E_{0\rightarrow+1}}$ are concatenated with the image features ($F^{L_l}_{I_0}$, $F^{L_l}_{I_{+1}}$) and passed through a (3 x (Conv2d, LeakyReLU), Conv2d) block that returns the offsets and masks of each level l. Note that, we described the procedure for time window ($\{0, \tau, ..., +1\}$), but the exact same holds for time window ($\{0, \tau, ..., -1\}$)

(2) The kernel offsets for level l are learned as a residual to the kernel offsets estimated at level $l - 1$. By doing so, we encourage a coarse-to-fine learning of motion that takes into account estimates from the low-resolution feature representations with larger receptive field. Thus, allowing for compensation of larger motions.

3.4. Attention Modules

After the *Feature Alignment* module, the resulting feature representations ($F^{L1}_{I_0}$, $F^{L1}_{I_{-1\rightarrow0}}$, $F^{L1}_{I_{+1\rightarrow0}}$) are passed through a pair of consecutive attention modules.

First, *Pairwise Attention* aims to fuse information from the different LDR images. To design it, we draw inspiration from the HDR generation procedure of ground truth data. In case of no camera or scene motion, an HDR image can be approximated via simple weighted averaging of the exposure compensated LDR images [18] Assuming that motion has been compensated in the previous module, we extend this concept to the feature level. In particular, we use attention blocks with a (Conv2d, LeakyReLU, Conv2d, Sigmoid) design each, that are applied pairwise between ($F^{L1}_{I_0}$, $F^{L1}_{I_{-1\rightarrow0}}$, $F^{L1}_{I_{+1\rightarrow0}}$) and $F^{L1}_{I_0}$, and provide per-pixel and per-channel blending weights. The latter are used as guides for weighted averaging in the feature level, resulting in a single merged feature representation.

Next, *Spatial Attention* aims to recover fine details from the merged feature representation. For this, we directly utilize the multi-scale spatial attention from the TSA module in EDVR [48].

3.5. HDR-ERGB Dataset

As there is no publicly available HDR dataset which features RGB images and synchronized event data, we captured a new dataset named HDR Events and RGB (HDR-ERGB) dataset. The hybrid imaging system used to capture our dataset combines a high-resolution RGB camera synchronized with a high-resolution event sensor. The RGB camera is a FLIR Blackfly S with a resolution of 4000×3000 and global shutter. The event camera is a Prophesee Gen4 with a resolution of 1280×720. The two sensors share a similar FOV and are mounted in a beam splitter setup, containing a mirror which splits the incoming light to the event and frame camera, and ensures alignment between events and high-resolution frames.

To record bracketed LDR images with events and HDR ground truth, we divide the recording procedure into two steps. In step 1, we acquire HDR ground truth by mounting the imaging system on a tripod and recording a set of 9 bracketed LDR images (0, ± 1, ± 2, ± 4 fstops) with no camera or scene motion. In step 2, we acquire the bracketed LDR images (0, ± 2 fstops) and synchronized events with camera and/or scene motion. To simulate dynamic scenes, we follow a procedure similar to [18], and capture camera motion by simply moving the tripod and scene motion by asking people to move. That is, the persons stand still during HDR ground truth recording (step 1), and receive an audio signal to start moving after the reference bracketed LDR image (0 fstops) is taken (step 2). Note that, a stream of events from the event camera is synchronously recorded with the bracketed LDR images at the beginning of step 2. Following this procedure, the reference bracketed LDR image of the motion affected recording (step 2) is aligned with the ground truth HDR recording (step 1). Stereo alignment between events and bracketed LDR images is performed by camera calibration and rectification, which results in events and images with a resolution of 960×688. In total, we have collected 53 dynamic scenes (scene motion, with or without camera motion) and 12 static scenes (only camera motion). For more details, visit the supplementary materials.

4. Results

4.1. Experimental Settings

HDR dataset with synthetic events. We use the HDM-HDR-2014 dataset [7] that contains real-world HDR video sequences, which, in turn, can be used to synthesize events. To generate bracketed LDR images (0±3 f-stops) with re-

Table 1. Quantitative comparison with state-of-the-art multi-bracket HDR approaches on the synthetic HDM-HDR-2014.

Method	±3 f-stops			
	PSNR-$\mu\uparrow$	SSIM-$\mu\uparrow$	LPIPS$\downarrow$	HDR-VDP2$\uparrow$
Kalantari [18]	39.53	98.21	0.0310	45.33
AHDR [52]	39.70	98.49	0.0230	47.52
ADNet [24]	40.14	98.79	0.0222	47.37
Ours w/o events	40.42	98.67	0.0211	47.38
Ours	**45.86**	**98.88**	**0.0161**	**53.21**

alistic noise characteristics from the HDR video sequences, we follow the exact same procedure as in [19], with the exception of not applying tone perpetuation. To simulate the high-speed nature of event cameras, we use a frame skip of 2 between bracketed LDR images. Synchronized events are generated using the VID2E simulator [10]. Following [41], we set the contrast threshold in VID2E to match the event rate per frame of HDR-ERGB, in order to ensure realistic event data rates. The HDM-HDR-2014 dataset was also used in the NTIRE 2021 Multi-Frame HDR Challenge [31]. However, we can not use the challenge dataset, as we lack access to the test set images required to synthesize events.

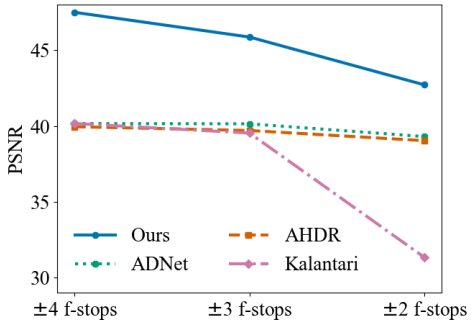

Figure 4. Qualitative experiments showing the effect of changing the f-stop values.

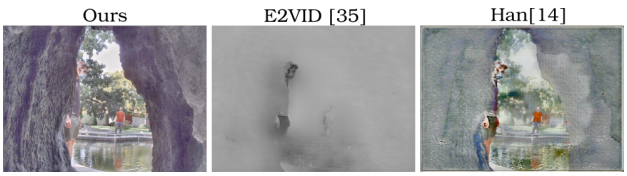

Figure 5. The comparison with existing event-based HDR method.

Baselines. As we propose the first multi-bracket HDR method with events, we are restricted to comparing against image-based multi-bracket HDR works. From the latter, we select the winner of the NTIRE 2021 Multi-Frame HDR Challenge [31] and the current state-of-the art method called ADNet [24], the HDR method of Kalantari et el. [18], and AHDR [52]. Additionally, we include the method of Wu et al. [50] for comparison on HDR-ERGB only, since we could not train it with reasonable performance on the synthetic dataset. Existing event-based methods [14, 35] naturally perform worse, see Fig. 5, since they tackle a harder task, where only events and a single LDR image are given as input. Thus, we did not include them in our comparisons

Training details. For supervision, we use a combination of L1 and LPIPS [57] losses, with weights 1, on the μ-law HDR images. μ-law is a compression introduced in [18], defined as $T = \log(1 + \mu H)/\log(1 + \mu)$ with $\mu = 5000$, that simulates a differentiable tonemapping operation. The following random augmentations are applied: scale, crop (256x256 patches), rotate (90 degrees), flip (horizontal and vertical), and color channel swap. We a use batch size of 4, Adam optimizer with an initial learning rate of 1×10^{-4} that decreases by a factor of 2 every 15 epochs (300 epochs on HDR-ERGB), and train for 60 epochs (1500 epochs on HDR-ERGB). To guarantee a fair comparison, we trained all baselines with the same settings from scratch on both HDM-HDR-2014 and HDR-ERGB.

Testing details. In HDM-HDR-2014, we test on 5 sequences (bistro_03, carousel_fireworks_09, fireplace_02, fishing_closeshot, poker_fullshot). We evaluate at the original image resolution with a cropped border of 10 pixels, *i.e.* 1900×1060, which is applied to remove the black pixels at the border. In HDR-ERGB, we test on 9 challenging sequences containing real noise for both images and events. Since we recorded ground truth HDR with up to ±4 f-stops but bracketed LDR with 0±2 f-stops, we can evaluate all methods on ±2 f-stop or ±4 f-stop range. The latter can test the hallucination capabilities of all methods outside the recorded dynamic range.

Metrics. We report results for the LPIPS, PSNR, SSIM and HDR-VDP-2 [26] metrics. As commonly done, we compute all metrics on the tonemapped images using μ-law (-μ) except HDR-VDP-2, which is calculated using linear HDR images and default parameters (PPD: 52.72).

4.2. Comparison on HDM-HDR-2014

We first evaluate all methods on the HDM-HDR-2014 dataset containing synthesized events. The results in Table 1 verify the substantial benefit of including events in a multi-bracket HDR approach. Our method significantly outperforms the state-of-the-art baselines on all the evaluated metrics. To showcase the advantages of our method, an example containing a fast-moving object overexposed in the reference frame is shown in Fig. 7. Using the high-speed and high dynamic range events, our method can take the necessary information from the short exposure and reconstruct the glass without major artifacts. The image-based baselines fail at properly aligning the moving object to its saturated counterpart in the reference frame.

Synthetic evaluation enables us to test all methods under

Table 2. Quantitative comparison with state-of-the-art multi-bracket HDR approaches on our recorded HDR-ERGB.

Method	±4 f-stops GT				±2 f-stops GT			
	PSNR-μ↑	SSIM-μ↑	LPIPS↓	HDR-VDP2↑	PSNR-μ↑	SSIM-μ↑	LPIPS↓	HDR-VDP2↑
Kalantari [18]	31.15	84.75	0.130	48.14	31.94	85.01	0.124	49.39
AHDR [52]	31.98	86.79	0.132	50.97	32.53	87.20	0.127	53.25
ADNet [24]	31.91	86.87	0.131	49.90	32.59	87.21	0.127	**53.54**
Wu [50]	32.16	86.47	0.132	50.18	32.73	86.85	0.127	51.64
Ours	**32.84**	**87.84**	**0.112**	**51.86**	**33.32**	**88.22**	**0.108**	53.32

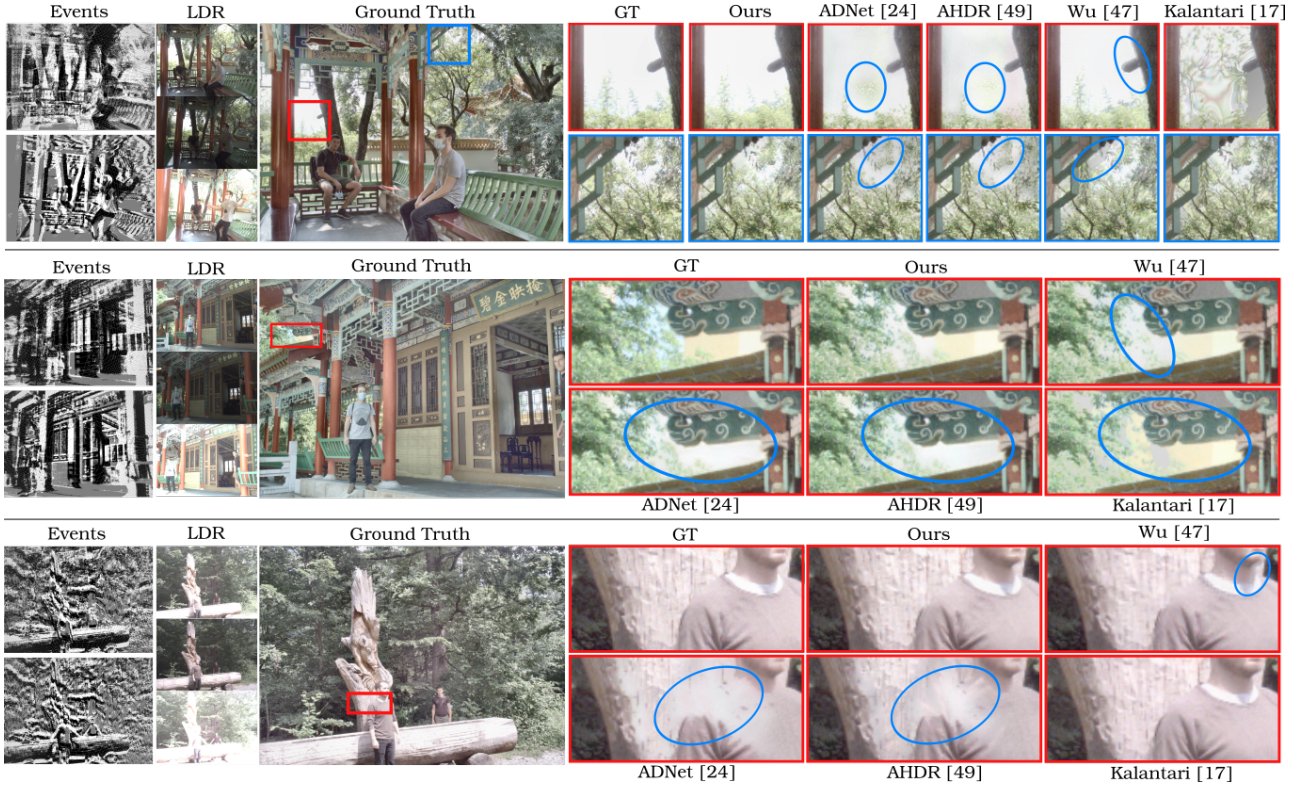

Figure 6. Our method can reliably align the LDR images without generating artifacts and can reconstruct thin structures e.g. leaves on a tree. In comparison, the baselines suffer from misalignment artifacts and have difficulties reconstructing thin details. Failure cases are highlighted with a blue circle.

different f-stops for the LDR brackets, which effectively increases or decreases the dynamic range recorded in the input LDRs. The plot in Fig. 4 illustrates the achieved PSNR score of the methods for the different f-stops. It can be observed that our method shows higher performance with only $\pm$ 2 f-stops compared to all baselines at the larger range of $\pm$ 4 f-stops. This shows the potential of our method to reduce the acquisition time for multi-bracket HDR, minimizing the risk for LDR image misalignments. We refer to Table 1 in the supplement for the numerical results of this comparison.

Table 3. Network ablation study on HDM-HDR-2014.

Method	Ours	w/o feat. align.	w/o temp. att.	w/o spat. att.
PSNR-μ↑	**45.86**	40.37	45.50	45.32

4.3. Comparison on HDR-ERGB

To test under more realistic conditions, we evaluate all methods on our challenging HDR-ERGB Dataset. We report the results for ground truth construction with different dynamic ranges in Table 2. Our method outperforms the state-of-the-art baselines in all metrics, except for HDR-VDP2 on the $\pm$2 f-stop ground truth. In general, the comparison on the $\pm$2 f-stop ground truth is more challenging for our method since the network needs to decide which events should be discarded as events contain a higher dynamic range than the ground truth image. Compared to the synthetic HDM-HDR-2014 dataset, the events in our HDR-ERGB dataset contain real sensor noise, which can explain the lower performance improvement of EHDR on real data.

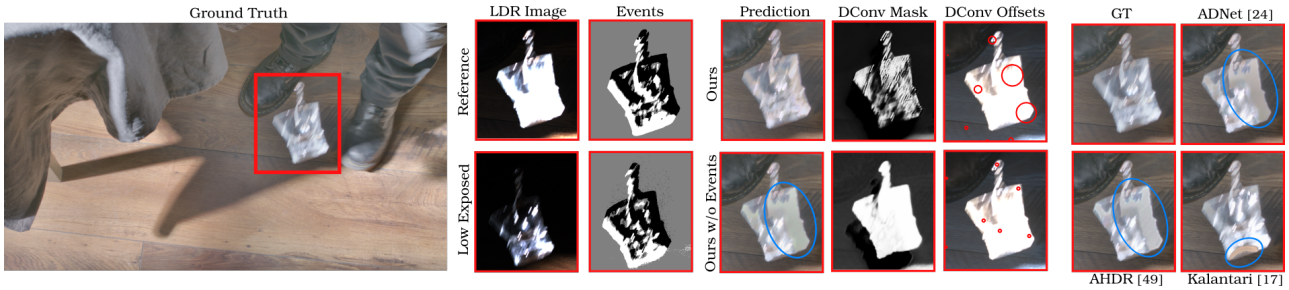

Figure 7. Qualitative example of the HDM-HDR 2014 dataset showing a falling glass. It can be observed that events provide more reliable motion information under fast motion than the different exposed LDR images. The comparison to the image-based version of our method shows a better HDR reconstruction due to a more detaild deformable modulation mask (DConv Mask) and larger deformable offsets (DConv Offsets). Finally, our method constructs a more accurate HDR image than the evaluated state-of-the-art baselines.

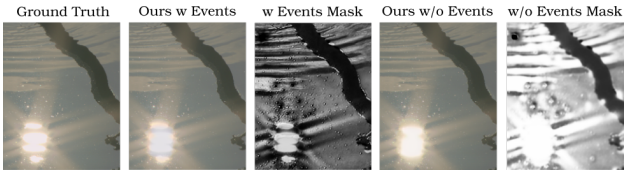

Figure 8. The impact of the high-dynamic-range in events.

The qualitative results on HDR-ERGB validate the advantages of EHDR. Since we use events and images, our method achieves a better LDR alignment compared to pure image-based alignment methods [18, 50], which exhibit ghosting artifacts due to misalignment, see Fig. 6 top. Moreover, the image-based flow can fail in textureless regions, which leads to severe artifacts in the HDR prediction for [18]. HDR methods relying on deep alignment [24, 52] suffer from artifacts in the same textureless regions as well, see Fig. 6 top. Additionally, they generate ghosting artifacts for large motions, observable in Fig. 1. Overall, our method achieves a robust alignment and is able to construct thin structures like leaves, shown in Fig. 6 top and middle.

4.4. Ablation Studies

To verify the effect of events in our pipeline, we evaluate EHDR with and without event data input, which results in a pure image-based HDR method. The image-based version of our architecture does not use the events in the feature alignment module and instead only uses the image features. By including events, we see a performance boost of 5.4 dB, confirming the benefits of events.

This improvement can also be observed in the reconstructed HDR images, which contain more details than the image-only method, especially in objects, which are oversaturated in all the LDR brackets. In Fig.. 8, the shape of the sun reflecting in water can only be inferred properly when modulation masks have access to events (via the proposed feature alignment module). This shows that the high-dynamic-range of events is leveraged by the modulation masks of deformable convolutions to guide HDR generation in extreme conditions.

To provide more insights on how events affect our method, we visualize the kernel offsets (DConv Offsets) (Fig. 7) computed for the deformable convolutions. By comparing the offsets between image and the combination of image and events, we see an enlarged receptive field visualized by the red circle on the glass falling down. Thus, it can be concluded that the events improve the alignment by providing more accurate motion information. Furthermore, we visualize the modulated mask for the deformable convolutions (DConv Mask). The mask predicted with events exhibits thin structure details, whereas the image-based uses a uniform weighting on the moving object.

Finally, we ablate the introduced network components by removing them from the architecture. The results in Tab. 3 show that each component improves the performance whereby the feature alignment has the largest impact.

5. Conclusion

We presented the first approach for multi-bracket HDR imaging with events. EHDR fuses motion information from images and events to enhance key parts of the HDR pipeline. As verified by our experiments, events significantly increase the performance on the real and synthetic data, confirming the robustness of our approach against misalignments. Our approach also requires less f-stops to achieve the same performance as image-based alternatives. Finally, we recorded the first dataset that contains bracketed LDR images and synchronized events with HDR ground truth.

6. Acknowledgement

This work was supported by Huawei Zurich Research Center; NCCR Robotics, a National Centre of Competence in Research, funded by the Swiss National Science Foundation (grant number 51NF40_185543); the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement No. 864042).

References

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

[2] Jianrui Cai, Shuhang Gu, and Lei Zhang. Learning a deep single image contrast enhancer from multi-exposure images. *Transactions on Image Processing*, 2018. 2

[3] Sungil Choi, Jaehoon Cho, Wonil Song, Jihwan Choe, Jisung Yoo, and Kwanghoon Sohn. Pyramid inter-attention for high dynamic range imaging. *Sensors*, 20(18), 2020. 2

[4] Paul E. Debevec and Jitendra Malik. Recovering high dynamic range radiance maps from photographs. In *SIGGRAPH*, 1997. 1

[5] Gabriel Eilertsen, Joel Kronander, Gyorgy Denes, Rafał K Mantiuk, and Jonas Unger. Hdr image reconstruction from a single exposure using deep cnns. *TOG*, 2017. 2

[6] Yuki Endo, Yoshihiro Kanamori, and Jun Mitani. Deep reverse tone mapping. *SIGGRAPH ASIA*, 2017. 2

[7] 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. SPIE*, volume 9023, pages 9023 – 9023 – 10, 2014. 5

[8] Guillermo Gallego, Tobi Delbruck, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew Davison, Jörg Conradt, Kostas Daniilidis, and Davide Scaramuzza. Event-based vision: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2020. 3

[9] Orazio Gallo, Alejandro Troccoli, Jun Hu, Kari Pulli, and Jan Kautz. Locally non-rigid registration for mobile hdr photography. In *WCVPR*, 2015. 2, 3

[10] Daniel Gehrig, Mathias Gehrig, Javier Hidalgo-Carrió, and Davide Scaramuzza. Video to Events: Recycling video datasets for event cameras. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2020. 6

[11] Daniel Gehrig, Antonio Loquercio, Konstantinos G. Derpanis, and Davide Scaramuzza. End-to-end learning of representations for asynchronous event-based data. In *Int. Conf. Comput. Vis. (ICCV)*, 2019. 3

[12] Chunle Guo, Chongyi Li, Jichang Guo, Chen Change Loy, Junhui Hou, Sam Kwong, and Runmin Cong. Zero-reference deep curve estimation for low-light image enhancement. In *CVPR*, 2020. 2

[13] Yoav HaCohen, Eli Shechtman, Dan B Goldman, and Dani Lischinski. Non-rigid dense correspondence with applications for image enhancement. *TOG*, 2011. 2

[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 *CVPR*, 2020. 2, 3, 6

[15] Samuel W Hasinoff, Dillon Sharlet, Ryan Geiss, Andrew Adams, Jonathan T Barron, Florian Kainz, Jiawen Chen, and Marc Levoy. Burst photography for high dynamic range and low-light imaging on mobile cameras. *TOG*, 2016. 2, 3

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 770–778, 2016. 4

[17] Yifan Jiang, Xinyu Gong, Ding Liu, Yu Cheng, Chen Fang, Xiaohui Shen, Jianchao Yang, Pan Zhou, and Zhangyang Wang. Enlightengan: Deep light enhancement without paired supervision. *CoRR*, 2019. 2

[18] Nima Khademi Kalantari and Ravi Ramamoorthi. Deep high dynamic range imaging of dynamic scenes. *SIGGRAPH*, 2017. 1, 2, 3, 4, 5, 6, 7, 8

[19] Nima Khademi Kalantari and Ravi Ramamoorthi. Deep hdr video from sequences with alternating exposures. In *Computer Graphics Forum*, volume 38, pages 193–205. Wiley Online Library, 2019. 4, 6

[20] Siyeong Lee, Gwon Hwan An, and Suk-Ju Kang. Deep recursive hdri: Inverse tone mapping using generative adversarial networks. In *ECCV*, 2018. 2

[21] Sang-Hoon Lee, Haesoo Chung, and Nam Ik Cho. Exposure-structure blending network for high dynamic range imaging of dynamic scenes. *IEEE Access*, 8:117428–117438, 2020. 1, 4

[22] Orly Liba, Kiran Murthy, Yun-Ta Tsai, Tim Brooks, Tianfan Xue, Nikhil Karnad, Qiurui He, Jonathan T Barron, Dillon Sharlet, Ryan Geiss, et al. Handheld mobile photography in very low light. *TOG*, 2019. 2

[23] Yu-Lun Liu, Wei-Sheng Lai, Yu-Sheng Chen, Yi-Lung Kao, Ming-Hsuan Yang, Yung-Yu Chuang, and Jia-Bin Huang. Single-image hdr reconstruction by learning to reverse the camera pipeline. In *CVPR*, 2020. 2

[24] Zhen Liu, Wenjie Lin, Xinpeng Li, Qing Rao, Ting Jiang, Mingyan Han, Haoqiang Fan, Jian Sun, and Shuaicheng Liu. Adnet: Attention-guided deformable convolutional network for high dynamic range imaging. In *CVPRW*, pages 463–470, 2021. 1, 2, 4, 5, 6, 7, 8

[25] Kede Ma, Zhengfang Duanmu, Hanwei Zhu, Yuming Fang, and Zhou Wang. Deep guided learning for fast multi-exposure image fusion. *IEEE Transactions on Image Processing*, 2019. 2

[26] Rafał Mantiuk, Kil Joong Kim, Allan G Rempel, and Wolfgang Heidrich. Hdr-vdp-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions. *TOG*, 30(4):1–14, 2011. 6

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

[28] Yuzhen Niu, Jianbin Wu, Wenxi Liu, Wenzhong Guo, and Rynson WH Lau. Hdr-gan: Hdr image reconstruction from multi-exposed ldr images with large motions. *IEEE Transactions on Image Processing*, 30:3885–3896, 2021. 2

[29] Tae-Hyun Oh, Joon-Young Lee, Yu-Wing Tai, and In So Kweon. Robust high dynamic range imaging by rank minimization. *PAMI*, 2014. 2

[30] Yafei Ou, Prasoon Ambalathankandy, Masayuki Ikebe, Shinya Takamaeda, Masato Motomura, and Tetsuya Asai. Real-time tone mapping: A state of the art report. *CoRR*, 2020. 2

[31] Eduardo Perez-Pellitero, Sibi Catley-Chandar, Ales Leonardis, and Radu Timofte. Ntire 2021 challenge on high dynamic range imaging: Dataset, methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 691–700, June 2021. 6

[32] K Ram Prabhakar, V Sai Srikar, and R Venkatesh Babu. Deepfuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs. In *ICCV*, 2017. 2

[33] Zhiyuan Pu, Peiyao Guo, M. Salman Asif, and Zhan Ma. Robust high dynamic range (hdr) imaging with complex motion and parallax. In *Computer Vision – ACCV 2020*, pages 134–149, Cham, 2021. 2

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

[35] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. High speed and high dynamic range video with an event camera. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2019. 6

[36] Allan G Rempel, Matthew Trentacoste, Helge Seetzen, H David Young, Wolfgang Heidrich, Lorne Whitehead, and Greg Ward. Ldr2hdr: on-the-fly reverse tone mapping of legacy video and photographs. *TOG*, 2007. 2

[37] Marcel Santana Santos, Tsang Ing Ren, and Nima Khademi Kalantari. Single image hdr reconstruction using a cnn with masked features and perceptual loss. *TOG*, 2020. 2

[38] Pradeep Sen, Nima Khademi Kalantari, Maziar Yaesoubi, Soheil Darabi, Dan B Goldman, and Eli Shechtman. Robust patch-based hdr reconstruction of dynamic scenes. *TOG*, 2012. 2

[39] Ana Serrano, Felix Heide, Diego Gutierrez, Gordon Wetzstein, and Belen Masia. Convolutional sparse coding for high dynamic range imaging. In *Computer Graphics Forum*, 2016. 2

[40] Bongki Son, Yunjae Suh, Sungho Kim, Heejae Jung, Jun-Seok Kim, Changwoo Shin, Keunju Park, Kyoobin Lee, Jinman Park, Jooyeon Woo, et al. A 640×480 dynamic vision sensor with a 9μm pixel and 300meps address-event representation. In *ISSCC*, 2017. 1

[41] Timo Stoffregen, Cedric Scheerlinck, Davide Scaramuzza, Tom Drummond, Nick Barnes, Lindsay Kleeman, and Robert Mahony. Reducing the sim-to-real gap for event cameras. *Eur. Conf. Comput. Vis. (ECCV)*, Aug. 2020. 6

[42] Michael D Tocci, Chris Kiser, Nora Tocci, and Pradeep Sen. A versatile hdr video production system. *TOG*, 2011. 2

[43] Marc Comino Trinidad, Ricardo Martin Brualla, Florian Kainz, and Janne Kontkanen. Multi-view image fusion. In *ICCV*, 2019. 2

[44] Jack Tumblin, Amit Agrawal, and Ramesh Raskar. Why i want a gradient camera. In *CVPR*, 2005. 2

[45] Okan Tarhan Tursun, Ahmet Oğuz Akyüz, Aykut Erdem, and Erkut Erdem. The state of the art in hdr deghosting: A survey and evaluation. In *Computer Graphics Forum*. Wiley Online Library, 2015. 1, 2

[46] Lin Wang, S. Mohammad Mostafavi I., 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 *CVPR*, 2019. 2, 3

[47] Lin Wang and Kuk-Jin Yoon. Deep learning for hdr imaging: State-of-the-art and future trends. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2021. 2

[48] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 5

[49] Ziwei Wang, Yonhon Ng, Cedric Scheerlinck, and Robert Mahony. An asynchronous kalman filter for hybrid event cameras. *CoRR*, 2020. 2, 3

[50] Shangzhe Wu, Jiarui Xu, Yu-Wing Tai, and Chi-Keung Tang. Deep high dynamic range imaging with large foreground motions. In *ECCV*, 2018. 1, 2, 4, 6, 7, 8

[51] SHI Xingjian, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *NeurIPS*, pages 802–810, 2015. 5

[52] Qingsen Yan, Dong Gong, Qinfeng Shi, Anton van den Hengel, Chunhua Shen, Ian Reid, and Yanning Zhang. Attention-guided network for ghost-free high dynamic range imaging. In *CVPR*, 2019. 1, 2, 4, 6, 7, 8

[53] Qingsen Yan, Bo Wang, Lei Zhang, Jingyu Zhang, Zheng You, Qinfeng Shi, and Yanning Zhang. Towards accurate hdr imaging with learning generator constraints. *Neurocomputing*, 428:79–91, 2021. 2

[54] Q. Yan, L. Zhang, Y. Liu, Y. Zhu, J. Sun, Q. Shi, and Y. Zhang. Deep hdr imaging via a non-local network. *IEEE Transactions on Image Processing*, 29:4308–4322, 2020. 1, 4

[55] Qingsen Yan, Yu Zhu, and Yanning Zhang. Robust artifact-free high dynamic range imaging of dynamic scenes. *Multimedia Tools and Applications*, 2019. 2

[56] Wenhan Yang, Shiqi Wang, Yuming Fang, Yue Wang, and Jiaying Liu. From fidelity to perceptual quality: A semi-supervised approach for low-light image enhancement. In *CVPR*, 2020. 2

[57] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2018. 6

[58] Song Zhang, Yu Zhang, Zhe Jiang, Dongqing Zou, Jimmy Ren, and Bin Zhou. Learning to see in the dark with events. In *ECCV*, 2020. 2, 3

[59] Hang Zhao, Boxin Shi, Christy Fernandez-Cull, Sai-Kit Yeung, and Ramesh Raskar. Unbounded high dynamic range photography using a modulo camera. In *ICCP*, 2015. 2

[60] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *CVPR*, pages 9308–9316, 2019. 5

[61] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Unsupervised event-based optical flow using motion compensation. In *ECCV*, pages 0–0, 2018. 3

[62] Yunhao Zou, Yinqiang Zheng, Tsuyoshi Takatani, and Ying Fu. Learning to reconstruct high speed and high dynamic range videos from events. In *CVPR*, pages 2024–2033, 2021. 2, 3