Multi-Bracket High Dynamic Range Imaging with Event Cameras

Nico Messikommer*,1 Stamatios Georgoulis*,2 Daniel Gehrig1 Stepan Tulyakov2 Julius Erbach2 Alfredo Bochicchio2 Yuanyou Li2 Davide Scaramuzza1

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

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 [21, 24, 27, 57, 59, 61] 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 [21, 52]. 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, 47] are novel
neuromorphic vision sensors that attempt to mimic the high
dynamic range and the high speed response of biological vi-
sion 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 res-
olution, 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 [41, 53, 65, 69]. 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 meth-
ods, 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 [17, 56]
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 pro-
posed 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
therapy 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 im-
gages and their fusion to HDR at the feature level. Our main
contributions can be summarized as:

1. We introduce EH HDR, the first method that combines
bracketed LDR images and synchronized events for
HDR imaging. In doing so, EH HDR is more robust
than image-based HDR works to LDR image satura-
tion, 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 signifi-
cant 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, bracket-
eted 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 [54]

Single-exposure methods infer an HDR image [6, 26, 30,
43, 44] or a stack of differently exposed LDR images [7, 23]
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 ap-
plied during image acquisition; thus, they are also called
inverse tone mapping (iTM) methods. Deep learning iTM
approaches [6, 7, 26, 30, 44] recently achieved good results
in recovering saturated details by utilizing large image con-
text; however, they still essentially hallucinate details in sat-
urated regions from the surrounding non-saturated context,
and, thus, are not suitable for commercial applications.

Other works [2, 15, 20, 63] 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. sys-
tem with two image sensors and common lens with beam
splitter [49] or two separate cameras [38, 50]. However,
such systems have higher power consumption and memory
requirements. There exist more exotic systems too, such
as systems with pixel-wise exposure control [46], out-of-
range intensity warping [66], or gradient encoding [51], 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, 21, 27,
28, 33, 36, 39, 45, 57, 59, 60, 62] 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 ex-
posures 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 synthe-
sizing long exposed images [18, 25]. Still, synthetic ex-
posure 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 [52]). A simple
solution to the alignment problem is to reject moving ob-
ject pixels [34, 59] or use robust exposure fusion [18, 25].
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 \([12, 16]\). These methods register all exposures 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 \([18]\), or features \([12]\), or is performed after exposure compensation \([21]\). 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** \([11]\) 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 \(> 120\, \text{dB}\) dynamic range versus \(50\, \text{dB} - 80\, \text{dB}\) for a frame camera. Because of these properties, event cameras have recently become popular for HDR reconstruction.

**Event-based methods** \([41, 53, 65, 69]\) 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 (\(\leq 1\, \text{MP}\)) 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 \([65]\). 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 \([17, 56]\) 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.

### 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 \(E = \{(t_0, x_0, y_0, p_0), (t_1, x_1, y_1, p_1), \ldots\}\), 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 \((E_{[i, z]}, E_{[z+1, z]}, \ldots)\), with each chunk containing all events within a time window \(\tau\). The events within each chunk are represented as a **voxel grid** \([14, 68]\) 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 \((F_{I_0}^{L_1}, F_{I_{-1}^{-z_0}}^{L_1}, F_{I_{+1}^{z_0}}^{L_1})\) and the reference features \((F_{I_0}^{L_1})\) 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 [19], 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 [21, 22, 59]. In particular, the non-linear LDR image $I_j$ is concatenated with its exposure compensated linearized version $I_j^L = f^{-1}(I_j)/t_j$, 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 $F_0^{L1}$ 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).

\(^1\)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.

### 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 [21,24,27,57,59,61]. 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 point is the Pyramid, Cascading, and Deformable (PCD) module introduced in EDVR [55], used in HDR too [27]. The PCD module uses a pyramid of feature representations to estimate kernel offsets and modulation masks for modulated deformable convolutions [67] \(^2\) at each pyramidal level separately, which are then used to gradually ‘align’ the feature representations of non-reference images to the reference ones.

\(^2\)The modulated deformable convolution is computed as $y(\rho) = \sum_{k=1}^{K} w_k \cdot x(p + p_k + \Delta p_k) \cdot \Delta m_k$ with $\Delta p_k \in \mathbb{R}$ and $\Delta m_k \in [0, 1]$ being the learnable kernel offsets and modulation masks respectively.
grated event features than the one they were originally triggered. The inte-

which inevitably causes events to appear in different loca-

pensate’ the camera motion in-between chunked events,

benefit of the ConvLSTM modules is that they can ‘com-

hers for time window (τ, . . . , τ) and passed through a (Conv2d,

LeakyReLU), Conv2d) block that returns the offsets and

masks of each level L

l

L

features first need to be integrated across the en-

tire time window (τ, . . . , τ). Although im-

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).

ence image. We build upon this core idea of deformable fea-

1) The kernel offsets and modulation masks at level

L

l

l ∈ \{1, 2, 3\} are jointly computed from image (\(F_{0,l}^{L} \), \(F_{0,l+1}^{L}\)) and chunked event (\(F_{E_{0,\tau}}^{L} \), \(F_{E_{\tau,2\tau}}^{L} \), . . . , \(F_{E_{\tau-\tau,1}}^{L}\)) 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 [38], one at each level \(L_l\), that output the integrated event features \(F_{E_{0,\tau,l+1}}^{L}\). 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 \(F_{E_{0,\tau,l+1}}^{L}\) are concatenated with the image features (\(F_{0,l}^{L_l} \), \(F_{l+1}^{L_l}\)) 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 fea-

ture representations (\(F_{0,l}^{L_1}, F_{l+1}^{L_1}, F_{l+2}^{L_1}\)) 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 [21]. 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 (\(F_{0,l}^{L_1}, F_{l+1}^{L_1}, F_{l+2}^{L_1}\)) and \(F_{l+1}^{L_1}\), 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 uti-

lize the multi-scale spatial attention from the TSA module

in EDVR [55].

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 mount-

ing 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 [21], and capture camera motion by simply moving the tripod and scene motion by

\[ \hat{I}_j = I_j^p / \gamma \]

Here, \(\alpha_j(\gamma)\) represent per-pixel blending weights.

For the case of 3 bracketed LDR images (\(I_j, j \in \{0, -1, +1\}\)), the HDR image can be computed as

\[ I_{HDR}(\gamma) = \frac{\sum_j \alpha_j(\gamma) I_j(\gamma)}{\sum_j \alpha_j(\gamma)}, \]

where \(\gamma = \{0, 1, 2\}\).
Table 1. Quantitative comparison with state-of-the-art multi-bracket HDR approaches on the synthetic HDM-HDR-2014.

<table>
<thead>
<tr>
<th>Method</th>
<th>PSNR-$\mu$↑</th>
<th>SSIM-$\mu$↑</th>
<th>LPIPS↓</th>
<th>HDR-VDP2↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kalantari [21]</td>
<td>39.53</td>
<td>98.21</td>
<td>0.0310</td>
<td>45.33</td>
</tr>
<tr>
<td>AHDR [59]</td>
<td>39.70</td>
<td>98.49</td>
<td>0.0230</td>
<td>47.52</td>
</tr>
<tr>
<td>ADNet [27]</td>
<td>40.14</td>
<td>98.79</td>
<td>0.0222</td>
<td>47.37</td>
</tr>
<tr>
<td>Ours w/o events</td>
<td>40.42</td>
<td>98.67</td>
<td>0.0211</td>
<td>47.38</td>
</tr>
<tr>
<td>Ours</td>
<td><strong>45.86</strong></td>
<td><strong>98.88</strong></td>
<td><strong>0.0161</strong></td>
<td><strong>53.21</strong></td>
</tr>
</tbody>
</table>

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 f-stops) 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 [8] 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 realistic noise characteristics from the HDR video sequences, we follow the exact same procedure as in [22], 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 [13]. Following [48], 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 [37]. However, we can not use the challenge dataset, as we lack access to the test set images required to synthesize events.

**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 [37] and the current state-of-the-art method called ADNet [27], the HDR method of Kalantari et al. [21], and AHDR [59]. Additionally, we include the method of Wu et al. [57] for comparison on HDR-ERGB only, since we could not train it with reasonable performance on the synthetic dataset. Existing event-based methods [17, 42] 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 [64] losses, with weights 1, on the $\mu$-law HDR images. $\mu$-law is a compression introduced in [21], 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 use a batch size of 4, Adam optimizer with an initial learning rate of $1 \times 10^{-4}$ 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_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...
Table 2. Quantitative comparison with state-of-the-art multi-bracket HDR approaches on our recorded HDR-ERGB.

<table>
<thead>
<tr>
<th>Method</th>
<th>±4 f-stops GT</th>
<th>±2 f-stops GT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PSNR-µ↑</td>
<td>SSIM-µ↑</td>
</tr>
<tr>
<td>Kalantari [21]</td>
<td>31.15</td>
<td>84.75</td>
</tr>
<tr>
<td>AHDR [59]</td>
<td>31.98</td>
<td>86.79</td>
</tr>
<tr>
<td>ADNet [27]</td>
<td>31.91</td>
<td>86.87</td>
</tr>
<tr>
<td>Wu [57]</td>
<td>32.16</td>
<td>86.47</td>
</tr>
<tr>
<td>Ours</td>
<td><strong>32.84</strong></td>
<td><strong>87.84</strong></td>
</tr>
</tbody>
</table>

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.

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

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

<table>
<thead>
<tr>
<th>Method</th>
<th>Ours</th>
<th>w/o feat. align.</th>
<th>w/o temp. att.</th>
<th>w/o spat. att.</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSNR-µ↑</td>
<td><strong>45.86</strong></td>
<td>40.37</td>
<td>45.50</td>
<td>45.32</td>
</tr>
</tbody>
</table>
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 detailed 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. Observed that our method shows higher performance with only ±2 f-stops compared to all baselines at the larger range of ±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.

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 ±2 f-stop ground truth. In general, the comparison on the ±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.

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 [21, 57], 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 [21]. HDR methods relying on deep alignment [27, 59] 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 over-saturated in all the LDR brackets. In Fig., 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).

Supplementary: Multi-Bracket High Dynamic Range Imaging with Event Cameras

7. f-stops Experiments on HDM-HDR-2014

Since we can synthetically create input LDR brackets with different exposure times on HDM-HDR-2014, we also evaluate the influence of different dynamic ranges in the input on the performance of all evaluated baselines. Note that, we keep the same exposure value for the mid-exposed LDR image, and only modify the $\pm f$-stops range for the short- and long-exposed LDR images. Fig. 4 in the main manuscript gives a concise overview of the Table 4, which reports in addition to PSNR also the SSIM and LPIPS [64] metric. It can be observed that our approach trained with only $\pm 2 f$-stops range is outperforming the baselines trained with $\pm 4 f$-stops range based on the SSIM and LPIPS metric as well.

8. Limitations

As can be seen in Fig. 9, our recorded HDR-ERGB dataset contains on purpose challenging samples with very large non-uniform motion. In these samples, our proposed method still faces some challenges in aligning correctly the moving parts with the reference frame, partially due to saturation in the moving parts. However, our methods still achieves the best results compared qualitatively to all of the tested state-of-the-art methods.

9. Licenses for Code and Dataset

As stated on the project page of HDM-HDR-2014 at https://www.hdm-stuttgart.de/vmlab/hdm-hdr-2014/#FTPdownload, the "Academic and educational use of the HdM-HDR-2014 data set is free". In our code framework, we used several code packages which are either under MIT or BSD-2-Clause License. We refer to the submitted code for the corresponding source URL.

10. HDR-ERGB Dataset

In this section, we describe in more detail how the sequences of our HDR-ERGB dataset were recorded using a beam splitter setup containing an event and RGB camera, see Fig 11. We show several examples of our dataset in Fig. 10.

10.1. Synchronization

In our beam splitter setup, the event and RGB cameras are hardware synchronized. Specifically, the RGB camera sends a trigger signal at the start and end of each exposure to our event camera, where those are recorded as external trigger events. These trigger events contain a precise timestamp for the clock time of the event camera, that allows to temporally synchronize the images with the event stream.

10.2. Calibration

To calibrate the setup, we use the E2Calib toolbox [32]. We initially record a checkerboard pattern with both cameras. The toolbox then uses E2VID [40, 42] to reconstruct intensity images from the asynchronous events, which are temporally synchronized with the standard frames. The calibration tool Kalibr [9, 10, 31, 35] is then used to obtain the intrinsics and extrinsics of both cameras. In a final step, the cameras are rectified using the camera parameters of the event camera. In practice, the event and RGB camera still have a small parallax ($< 1$ mm) in the z-direction of the corresponding camera coordinate system. However, since our HDR-ERGB dataset does not contain scenes recorded at a close distance to the camera, the stereo rectification provides pixel-accurate alignment.

10.3. Ground Truth HDR Acquisition

As described in the paper, we record bracketed LDR images and events in two steps. In the first step, we record a set of 9 bracketed LDR images on a steady tripod, which are then used to create the HDR ground truth. In step two, events and three LDR brackets are recorded containing camera and scene motion in form of moving persons. This way, the first LDR frame of the dynamic sequence is aligned with the ground truth HDR frame. In order to achieve this, a one second pause was introduced in between the first and second LDR image to allow the photographer
Table 4. Quantitative comparison with state-of-the-art multi-bracket HDR approaches on the synthetic HDM-HDR-2014. The input LDR images are simulated with different f-stops to evaluate the performance with changing dynamic range in the input. Our methods outperform all of the baselines, even showing a better performance compared to the baselines provided with more dynamic range information in the input.

<table>
<thead>
<tr>
<th>Method</th>
<th>±2 f-stops</th>
<th>±4 f-stops</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PSNR↑</td>
<td>SSIM↑</td>
</tr>
<tr>
<td>Kalantari [21]</td>
<td>31.34</td>
<td>97.41</td>
</tr>
<tr>
<td>AHDR [59]</td>
<td>39.03</td>
<td>98.76</td>
</tr>
<tr>
<td>ADNet [27]</td>
<td>39.30</td>
<td>98.82</td>
</tr>
<tr>
<td>Ours</td>
<td>42.71</td>
<td>99.02</td>
</tr>
</tbody>
</table>

Figure 9. Challenging example of our HDR-ERGB dataset. All of the tested methods struggle to correctly align the motion information from the LDR brackets to the reference frame. Nevertheless, our methods generate quantitatively the HDR image closest to the ground truth.

and people in the shot to react. To construct the HDR ground truth from the $N = 9$ static LDR images, a simple triangle weighting scheme was adapted from [5] and [21] to merge the images. The LDR images were first linearized using the inverse camera response function and divided by the normalized exposure to map them into the HDR domain. $H_i = \frac{I_i}{t_i}$ with $t_i$ being the exposure time of the image divided by the shortest exposure time in the sequence. In our notation the exposures are sorted from the shortest exposure time with index $i = 0$ to the longest exposure time with index $i = N$. The images $H_i$ are then averaged with a weighted average scheme inspired by [5] and [21].

$$H_{gt} = \frac{\sum_{i=0}^{N} \alpha_i(p) \ast H_i(p)}{\sum_{i=0}^{N} \alpha_i(p)}$$ (1)

The weights $\alpha_i$ are obtained from the triangle functions depicted in Fig.12 evaluated on the $i$-indexed LDR image.
Figure 10. Our recorded dataset includes samples containing only camera motion (top row), only scene motion (second top row) and combined camera and scene motion (third and last row).

<table>
<thead>
<tr>
<th>Events</th>
<th>Mid LDR</th>
<th>Low LDR</th>
<th>High LDR</th>
<th>HDR</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Image" /></td>
<td><img src="image2.png" alt="Image" /></td>
<td><img src="image3.png" alt="Image" /></td>
<td><img src="image4.png" alt="Image" /></td>
<td><img src="image5.png" alt="Image" /></td>
</tr>
<tr>
<td><img src="image6.png" alt="Image" /></td>
<td><img src="image7.png" alt="Image" /></td>
<td><img src="image8.png" alt="Image" /></td>
<td><img src="image9.png" alt="Image" /></td>
<td><img src="image10.png" alt="Image" /></td>
</tr>
</tbody>
</table>

Figure 11. The beamsplitter setup used to record our new HDR-ERGB dataset. It combines an event and RGB camera by projecting the scene via a beamplitter mirror to both cameras.

![Diagram](image11.png)

RGB Camera

Beam Splitter

Event Camera

Figure 12. The triangle functions used as weights $\alpha$ to generate the ground truth HDR image.

![Plot](image12.png)
References


[17] 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


[21] Nima Khademi Kalantari and Ravi Ramamoorthi. Deep high dynamic range imaging of dynamic scenes. SIGGRAPH, 2017. 1, 2, 3, 4, 5, 6, 7, 8, 10


from low dynamic range content. In Computer Graphics Forum, 2018. 2


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


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


[56] Ziwei Wang, Yonhon Ng, Cedric Scheerlinck, and Robert Mahony. An asynchronous kalman filter for hybrid event cameras. In ICCV, pages 448–457, October 2021. 2, 3


[59] 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


[65] 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


