HDR Reconstruction from Bracketed Exposures and Events

Richard Shaw
richard.shaw@huawei.com
Sibi Catley-Chandar
sibi.catley.chandar@huawei.com
Aleš Leonardis
ales.leonardis@huawei.com
Eduardo Pérez-Pellitero
e.perez.pellitero@huawei.com

Huawei Noah's Ark Lab
London, UK

Abstract

Reconstruction of high-quality HDR images is at the core of modern computational photography. Significant progress has been made with multi-frame HDR reconstruction methods, producing high-resolution, rich and accurate colour reconstructions with high-frequency details. However, they are still prone to fail in dynamic or largely over-exposed scenes, where frame misalignment often results in visible ghosting artifacts. Recent approaches attempt to alleviate this by utilizing an event-based camera (EBC), which measures only binary changes of illuminations. Despite their desirable high temporal resolution and dynamic range characteristics, such approaches have not outperformed traditional multi-frame reconstruction methods, mainly due to the lack of colour information and low-resolution sensors. In this paper, we propose to leverage both bracketed LDR images and simultaneously captured events to obtain the best of both worlds: high-quality RGB information from bracketed LDRs and complementary high frequency and dynamic range information from events. We present a multi-modal end-to-end learning-based HDR imaging system that fuses bracketed images and event modalities in the feature domain using attention and multi-scale spatial alignment modules. We propose a novel event-to-image feature distillation module that learns to translate event features into the image-feature space with self-supervision. Our framework exploits the higher temporal resolution of events by sub-sampling the event streams using a sliding window, enriching our combined feature representation. Our proposed approach surpasses state-of-the-art (SoTA) multi-frame HDR reconstruction methods using synthetic and real events, with a 2dB and 1dB improvement in PSNR-L and PSNR-μ on the HdM HDR dataset, respectively.

1 Introduction

High dynamic range (HDR) imaging techniques extend the luminance range capturable beyond conventional or low dynamic range (LDR) cameras. The vision and graphics communities have developed numerous HDR strategies over recent years, summarized by [36, 39].

Figure 1: Our learning-based method produces high-quality HDR images, leveraging bracketed LDRs and events. The two modalities provide complementary information; events: high frequency and dynamic range, LDRs: colour and fine detail. Left: input LDR sequence with different exposures (short, medium, long) and event stream, where the colour denotes event polarity (red positive, blue negative). Right: input LDR reference and our HDR result.

Traditionally, HDR methods involve capturing multiple LDRs with varying exposure values (bracketed exposures) and merging them with different weights to reconstruct an HDR image [6]. The seminal work of [18] extends multi-frame fusion to dynamic scenes by curating paired dynamic input LDRs with static ground truth HDR labels, and training end-to-end deep neural networks. This paradigm has been very successful over recent years and now defines the state-of-the-art in HDR reconstruction [21, 24]. However, these methods have limitations intrinsic to the camera sensor and bracketing strategy: frames generally need to be aligned to a reference frame due to sequential capturing. Finding frame correspondences is challenging and can be affected by e.g. non-rigid motion or disoclussions, resulting in motion-related artifacts. Furthermore, the dynamic range per LDR is limited; thus, each exposure bracket will inevitably miss some parts of the scene (i.e. under- or over-exposed). This is especially problematic for the reference frame, as alignment to areas that suffer information loss is not well defined in terms of photometric loss. Other approaches reconstruct from only a single LDR [19]; an ill-posed problem where texture details in poorly-exposed regions are hallucinated from neighbouring areas or priors learned through neural networks [8]. Multi-frame methods, however, continue to out-perform single image approaches [27].

Event-based cameras (EBC) have recently garnered significant attention from researchers due to their unique properties distinct from conventional frame-based cameras. EBCs are novel bio-inspired sensors presenting a paradigm shift in how visual information is acquired; while a standard camera captures intensity images at a fixed frame rate, EBCs detect changes in per-pixel log intensity $L = \log(I)$ (brightness) asynchronously. An event $E_i = (x_i, y_i, t_i, p_i)$ is triggered at pixel (x_i, y_i) at time t_i when the brightness increment since the last event at that pixel, i.e. $\Delta L(x_i, y_i, t_i) = L(x_i, y_i, t_i) - L(x_i, y_i, t_i - \Delta t_i)$ exceeds a contrast threshold $\pm C$, i.e. $\Delta L(x_i, y_i, t_i) = p_i C$, where $C > 0$ and polarity $p_i \in \{+1, -1\}$ is the sign of the brightness change [10]. This enables very high temporal resolution capture (in the order of μs), with high dynamic range (140dB vs 60dB) and low power consumption, making them appealing for HDR applications [10]. Despite these advantages, EBCs generally have low spatial resolution and typically only record grayscale information and thus have so far struggled to produce high-resolution, colour-accurate and artifact-free image reconstructions.

In this work, we address these limitations by exploiting the strengths of each modality. As Fig. 1 shows, we propose a multi-modal HDR imaging method combining bracketed ex-

posures from a frame-based camera and high temporal resolution and dynamic range events from an EBC. The main contributions of this paper are: **(1)** A multi-modal architecture that combines bracketed LDRs and events trained in an end-to-end manner. **(2)** An event-to-image distillation module that transforms event features into the image feature space without needing an intermediate intensity image, trained with self-supervision from corresponding LDR features. **(3)** An event window sampling mechanism that leverages their high temporal resolution by extracting subsets of events and spatially aligning them in feature space.

2 Related Work

HDR from Bracketed LDRs: Bracketed HDR methods capture differently exposed LDRs and merge them in a weighted fashion [6]. In dynamic scenes, the LDR images must be aligned to the reference, often using optical flow, and then processed with a reconstruction network. [18] pioneered this two-stage process, using classical optical flow [20] to align low- and high-exposure images to the medium frame and a CNN to merge and correct alignment errors. [44] used a similar approach, but instead of optical flow, compute a homography for background alignment, relying on the CNN to correct foreground motion implicitly. [45] followed this pipeline but introduced attention to suppress undesired information (misaligned and badly-exposed regions) from the LDRs before merging. [26] and [29] replaced classical optical flow with learning-based approaches, e.g. FlowNet [7], and [30] save computation by computing flow at low resolution and upscaling. Despite performance improvements, these methods still suffer ghosting, particularly for fast-moving objects and saturated regions. Recent developments include [24] with a GAN-based approach and [31] using a weakly supervised training strategy. Most relevant to ours, [21] introduced deformable convolution alignment, which we adopt in our work, and won the NTIRE'21 HDR challenge [27].

Intensity Reconstruction from Events: [3] was one of the first to explore intensity image reconstruction from events; however, it was restricted to known camera motion: a rotating event camera for panoramic imaging. [2] advanced to generic camera motion, estimating joint intensity and optical flow with cost function minimization. In the seminal E2Vid, [33] was among the first to employ a learning-based approach, utilizing a recurrent neural network (RNN) for video reconstruction. Other approaches include [48] using an RNN and [15] employing a conditional GAN. Extensions to [33] include: reducing network parameters [35], improving generalization [38], and enhancing temporal consistency using optical flow [47]. Although image reconstruction from events has progressed considerably, results are typically low resolution, grayscale, and exhibit artifacts. Moreover, events primarily reflect edge information, and these approaches hallucinate details in textureless regions. [16, 41, 43] proposed reconstructing high-resolution HDR images from low-resolution events, and [42] tried to learn more robust event representations by jointly learning HDR images with downstream tasks such as segmentation and depth estimation via knowledge distillation [40]. However, these methods still suffer in dealing with event sparsity and fail to produce detailed colour reconstructions compared to image-based HDR methods.

Event-guided HDR Reconstruction: Rather than reconstructing intensity images solely from events, previous approaches have used events to guide the LDR to HDR mapping. Notably, [13] proposed a multi-modal system and learning framework using a single LDR and intensity map generated by events. However, the method has two main drawbacks: 1) the event intensity map is generated using off-the-shelf network E2Vid [34] and thus, they do not optimize end-to-end, resulting in a point of model failure, and 2) using a single LDR limits

their ability to handle scenes with extreme brightness ranges. These two aspects hinder the algorithm's performance, which suffers from colour artifacts in over-exposed regions, and its quantitative performance falls short when compared to SoTA multi-frame HDR methods that only use a conventional camera. We solve these issues using an end-to-end HDR network that learns from LDRs and events jointly. Specifically, we enhance bracketed multi-frame HDR reconstruction with multiple event streams, enabling the processing of more complex scenes. Moreover, we leverage information between the events and LDRs using knowledge distillation. Our unified HDR framework directly fuses images and events in the feature domain without relying on the intermediate step of event intensity image generation.

3 Proposed Method

Given a sequence of n LDR images with different exposure values $\{I_1, I_2, ..., I_n\}$ captured at timestamps $\{t_1, t_2, ..., t_n\}$ and a stream of input events $\{E_i\}_{t_0}^{t_n}$ our aim is to reconstruct a single HDR image H aligned to the reference frame I_{ref} at timestamp t_{ref}. In our implementation, we use three input LDR images corresponding to short, medium and long exposures, specifying the middle frame as the reference $I_{ref} = I_2$. To generate inputs to our model, we follow [18, 44, 45] forming a linearized image L_i for each I_i as follows: $L_i = I_i^{\gamma} / T_i, i = 1, 2, 3$, where T_i is the exposure time of image I_i. Setting $\gamma = 2.2$ approximates the inverse gamma correction while dividing by the exposure time adjusts the images to have consistent brightness.

Events and LDRs are acquired simultaneously but at different frequencies, i.e. LDRs are acquired at low frequency $\{t_1, t_2, t_3\}$ and events at high frequency $E_i \in [t_0, t_3]$ where t_0 is the beginning of the event stream before the first LDR is acquired. Events provide additional information in-between the low-frequency LDRs and parts of the event stream correspond to the acquisition of a different LDR image, therefore we partition the input event stream into three chunks corresponding to the LDR timestamps: $\{E_1, E_2, E_3\} = \{E_{t_0 \to t_1}, E_{t_1 \to t_2}, E_{t_2 \to t_3}\}$. Therefore, our proposed network g can be defined as $\hat{H} = g(\{I_i\}, \{L_i\}, \{E_i\}; \theta)$, where $\hat{H}$ denotes the reconstructed HDR image and θ the network parameters.

Following [21], instead of concatenating the inputs and processing jointly, we use a multi-branch pipeline where each input modality is processed separately before fusion. Specifically, for LDR images $\{I_i\}$, we learn attention feature maps with a spatial attention module $\mathcal{A}$ to suppress misaligned and badly-exposed regions. Gamma-corrected linear images $\{L_i\}$ are processed using a pyramidal, cascading and deformable (PCD) alignment module $\mathcal{P}^L$ to handle scene or camera motion. We extend the method for events $\{E_i\}$, using a separate PCD module $\mathcal{P}^E$ to spatially align event features to the reference timestamp. Finally, a feature distillation module $\mathcal{D}$ transforms intermediately sampled events $\{E_j\}$ into the image feature space. Therefore, more accurately our end-to-end network g can be described as:

$$\hat{H} = g\left(\mathcal{A}(I_i), \mathcal{P}^L(L_i), \mathcal{P}^E(E_i), \mathcal{D}(E_j); \theta\right). \tag{1}$$

3.1 Network Architecture

In this section, we provide an overview of our approach (shown in Fig. 2). Our architecture is composed of five components: 1) an LDR spatial attention module $\mathcal{A}$, 2) a linear image alignment module $\mathcal{P}^L$, 3) an event alignment module $\mathcal{P}^E$, 4) an event-to-image feature distillation and alignment network $\mathcal{D}$, and 5) a fusion and HDR reconstruction network.

1) LDR Attention Module: Following [45], a spatial attention module $\mathcal{A}$ learns attention maps from the three input LDR images. Given LDRs $\{I_1, I_2, I_3\}$ we extract features using

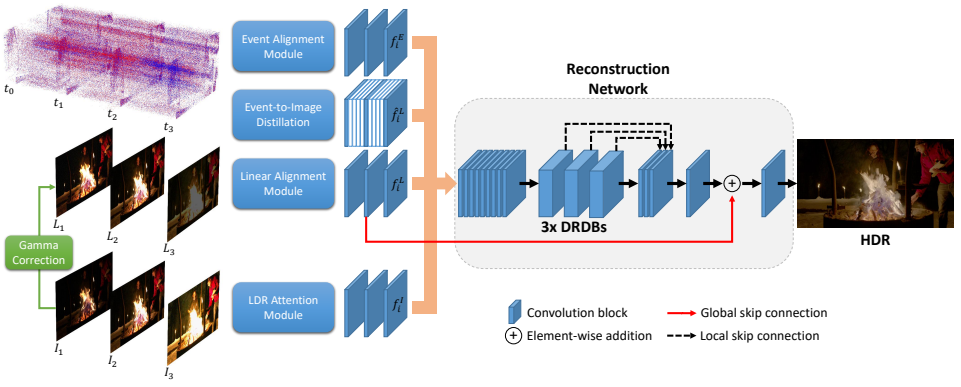

Figure 2: Model architecture. LDRs pass through an attention module, while gamma-corrected linear images and events are spatially aligned using pyramidal deformable convolution (PCD) alignment modules. Events are temporally sub-sampled and translated to pseudo-image features using distillation. Finally, the reconstruction network, comprising residual dense blocks, fuses the aforementioned input branches producing an HDR image.

a single convolutional layer. For each non-reference LDR image ($I_i \neq I_2$), we concatenate the features with the reference features f^I_{ref} as input to the attention module, comprising two convolutional layers and a sigmoid function, generating an attention map in the range 0-1. Element-wise multiplication of LDR features with the corresponding attention map generates spatially attenuated features for each LDR image: $f^I_i = \mathcal{A}\left(I_i, I_{ref}\right), i = 1,3$.

2) Linear Alignment Module: Following [21], we use a PCD alignment module $\mathcal{P}^L$ to align gamma-corrected linear images $\{L_i\}$ at the feature level. As shown by [4], alignment at the feature level is typically better than at the image level, and we do this with deformable convolution [5]. We extract multi-scale feature pyramids using strided convolutions for each $\{L_i\}$ and perform PCD alignment to the reference features f^L_{ref} at each scale: $f^L_i = \mathcal{P}^L\left(L_i, L_{ref}\right), i = 1,3$.

3) Event Alignment Module: For the event modality we employ a separate PCD alignment module $\mathcal{P}^E$ to perform spatial alignment in the event feature domain. [43] demonstrated that alignment of events using deformable convolutions is effective for the task of intensity image reconstruction. In detail, we align the partitioned input event streams $E_1 = \{E_{t_0 \to t_1}\}$ and $E_3 = \{E_{t_2 \to t_3}\}$ to the events accumulated during the capture of the reference frame corresponding to the reference timestamp $E_{ref} = \{E_{t_1 \to t_2}\}$: $f^E_i = \mathcal{P}^E\left(E_i, E_{ref}\right), i = 1,3$.

4) Event-to-Image Distillation: To leverage complementary information between events and images, we introduce a novel feature distillation module $\mathcal{D}$ that learns to transform events into the image feature domain since our end goal is to predict an HDR image. Events accumulated during different parts of the event stream correspond to the acquisition of a different LDR image, i.e. three partitions of the event stream $\{E_1, E_2, E_3\} = \{E_{t_0 \to t_1}, E_{t_1 \to t_2}, E_{t_2 \to t_3}\}$ correspond to gamma-corrected linear images $\{L_1, L_2, L_3\}$ at times $\{t_1, t_2, t_3\}$. Therefore, to translate events into image features, we apply a self-supervising ℓ_2 loss between the extracted event features f^E_i and the corresponding linear image features f^L_i at each timestamp i:

$$\mathcal{L}_{\mathcal{D}} = \sum_{s=1}^{S} \sum_{i=1}^{3} \left(f^E_{i,s} - \text{sg}\left(f^L_{i,s}\right)\right)^2, \tag{2}$$

where sg(·) indicates a stop-gradient, i.e. the learnt linear image features f_i^L are treated as self-supervising labels. This encourages domain transfer of event features into image features. We apply this loss at each scale $s \in S$ of the feature pyramid. The event-to-image distillation network exploits the higher temporal resolution of events by sub-sampling the input event stream using a sliding window approach. Sampling chunks of events in-between the LDR keyframes and passing these through the learnt distillation network enables us to predict intermediate pseudo-image features $\hat{f}_i^L$ (event features transformed to image domain), thereby enriching our combined feature representation. This process is shown in Fig. 3.

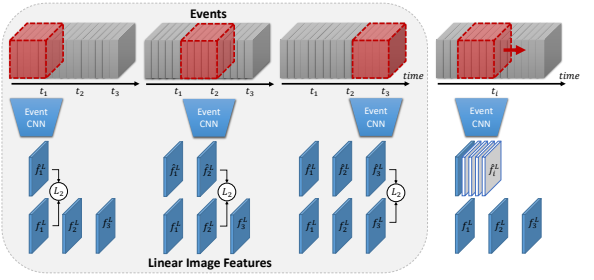

Figure 3: Event-to-image feature distillation transforms events into image domain features. Left: a self-supervision training strategy employs a loss on the corresponding LDR features. Right: intermediate features are generated by sub-sampling events with a sliding window.

5) HDR Reconstruction Network: The feature maps from the aforementioned input branches (LDRs after attention, aligned linear images, aligned events, and event-to-image features) are concatenated as input to the HDR fusion network. The fusion network follows good practices common in recent literature, e.g. [21, 45], consisting of three dilated residual dense blocks (DRDBs), with dilated convolutions [46] to increase the receptive field, and global and local skip connections. As shown by the ablation results in section 4, we find that adding these submodules leads to improved HDR quality with less ghosting and better detail recovery.

3.1.1 Loss Function

Following previous work [21, 45] we use the μ-law to map from the linear HDR image to the tonemapped image $\mathcal{T}(H) = \log(1+\mu H)/\log(1+\mu)$, where H is the linear HDR image, $\mathcal{T}(H)$ is the tonemapped image and $\mu = 5000$. We then estimate the ℓ_1-norm between the predicted and ground truth HDR images as follows: $\mathcal{L}_{HDR} = ||\mathcal{T}(\hat{H}) - \mathcal{T}(H)||_1$. In addition to the tonemapped reconstruction loss, we use a self-supervising loss given by Eq. (2) on the features generated from our event-to-image feature distillation module. Therefore, the total loss is the sum of the HDR reconstruction loss and distillation losses: $\mathcal{L}_{total} = \mathcal{L}_{HDR} + \mathcal{L}_{\mathcal{D}}$.

3.2 Training Details

Training Data: We model bracketed exposure LDRs from ground truth HDR video frames and generate synthetic event data using ESIM: Event Camera Simulator [32]. We obtain ground truth HDR frames from the HdM HDR dataset [3], which contains sequences with varied scenes, lighting and motion. Following [28], using four scenes for validation/testing and 25 scenes for training, ensuring no scene overlap between training and testing/validation splits, we obtain 1500, 60 and 201 samples for training, validation and testing, respectively.

Image Formation Model: To generate LDRs $\{I_i\}$ we use the pixel measurement model [14]: $I_i = \min(\Phi T/g + I_0 + n, I_{max})$, where Φ is scene brightness, T exposure time, g sensor gain, I_0 offset current, n sensor noise and I_{max} saturation point. We approximate Φ by the ground truth HDR image and generate LDRs by modifying T for any three consecutive frames.
Noise Model: We include a noise signal n whose variance comes from three sources: photon noise, read noise, and ADC gain and quantization: $\text{Var}(n) = \Phi/g^2 + \sigma^2_{read}/g^2 + \sigma^2_{ADC}$ [27].
Event Generation Model: To generate high-frequency events, we follow Vid2e [11] by temporally upsampling the HdM HDR video sequences using Super SloMo [17]. Whereas prior works use tonemapped LDR video as input, to retain high dynamic range, we first μ-tonemap the ground truth HDR frames to the nonlinear domain, as the pre-trained Super SloMo network was trained on tonemapped LDR videos. The resulting interpolated frames are converted back to the linear domain with the reverse tonemap transformation. ESIM then processes the upsampled frames with a contrast ratio of $C = 0.5$ generating binary event streams $\{E_i\} = \{(x_i, y_i, t_i, p_i)\}$. Examples of synthetic event data are shown in Fig. 4.
Event Representation: To process events with a CNN, we discretize the time axis into B bins. Following [47], we extend the effective temporal resolution beyond B by weighted accumulation of events. Given N events $\{E_i\}_{i=0,\dots,N-1} = \{(x_i, y_i, t_i, p_i)\}_{i=0,\dots,N-1}$, we scale the timestamp range $\Delta t = t_{N-1} - t_0$ to $[0, B-1]$; each event distributing a polarity p_i to the two closest spatio-temporal voxels: $E(x_i, y_i, t_i) = \sum_i p_i \max(0, 1 - |t_i - t_i^*|)$, where $t_i^* = \frac{B-1}{\Delta t}(t_i - t_0)$ is the normalized timestamp. We use $B = 5$ as is typically used [33]. Note, increasing B has limited influence on reconstruction at the expense of increased computational cost.
Implementation Details: Each module consists of 3×3 convolutions, extracting 64 feature channels, and the network ends with a ReLU predicting an unbounded linear HDR image. In training, we randomly sample crops of $256 \times 256 \times 3$ from the LDR images and corresponding crops of $256 \times 256 \times 5$ from the event voxel grids. Augmentations consist of horizontal/vertical flipping and 90° rotations. We train each model for 2000 epochs with batch size 16, a learning rate of 10^{-4} and Adam optimizer. We use a stepped learning rate schedule decaying by factor 10 every 500 epochs. Future work will aim to provide implementation under popular learning frameworks *e.g.* [1, 25].

4 Results

This section presents results on synthetic and real event data, using PSNR in the linear and tonemapped domains (PSNR-L and μ) and HDR-VDP-2 (HV2) [22]. We compare 8 other methods: three non-learning bracketed LDR methods Debevec [6], Mertens [23], Sen [37], two SoTA learning-based bracketed LDR methods AHDR [45], ADNet [21], event-only method E2Vid [33], and two Neuromorphic HDR [13] models (Neuro^1 and Neuro^2) which combine a single LDR with an event intensity map. Neuro^1 follows the training procedure of [13] generating a synthetic event intensity map from the Poisson reconstruction of the ground truth HDR image gradients. Neuro^2 is the same model with the intensity map generated by E2Vid. These models are the best performing and most relevant models to our method. Note that E2Vid is disadvantaged as it only uses event data and cannot reconstruct colour information reliably. Thus, we compute metrics for this model in the grayscale domain.
Synthetic Events Dataset: Quantitative results using synthetic event data generated from the HdM HDR video test set are displayed in Table 1 (left). Our method performs significantly better in all metrics over the other methods, with an over 2dB increase in PSNR-L and almost 1dB increase in PSNR-μ over the closest performing model ADNet. Note that E2Vid

Table 1: Quantitative results on the HDM test set with synthetic events (left) and on DSEC test set with real events (right). Best performer on each dataset denoted in bold.

		HDM dataset			DSEC dataset		
Method	Inputs	PSNR-L	PSNR-μ	HV2	PSNR-L	PSNR-μ	HV2
Debevec [8]	3 LDRs	26.52	15.83	40.87	10.19	13.20	68.10
Mertens [23]	3 LDRs	31.71	19.73	39.92	13.89	17.43	68.53
Sen [37]	3 LDRs	32.44	28.32	37.05	15.53	25.02	70.17
AHDR [15]	3 LDRs	37.79	36.59	49.11	36.52	34.64	74.47
ADNet [21]	3 LDRs	39.18	36.89	50.06	37.17	35.12	75.14
E2Vid [33]	Events	22.44	14.68	38.90	12.04	10.84	64.29
Neuro^1 [13]	LDR + Events	32.34	30.97	48.63	28.26	31.62	72.76
Neuro^2 [13]	LDR + Events	27.45	25.98	36.72	22.58	24.76	68.92
Ours	3 LDRs + Events	**41.81**	**37.84**	**55.79**	**38.13**	**35.83**	**76.65**

performs particularly poorly here because of its RNN architecture which is unstable for short sequences, and the fact that many frames in the HdM dataset have relatively little motion. The network falls into a failure case with a sparse event signal, unable to reconstruct the image properly. In contrast, our method can rely on the other input modality of the bracketed LDRs to maintain good HDR reconstruction performance even with a lack of events.

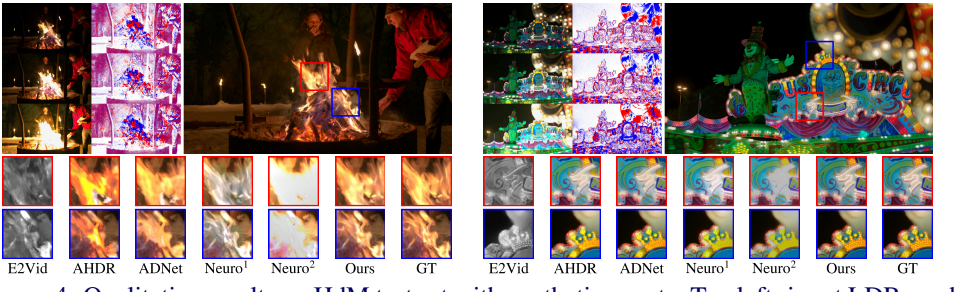

Figure 4: Qualitative results on HdM test set with synthetic events. Top left: input LDRs and events, top right: predicted HDR image using our method, bottom row: comparison crops.

Real Events Dataset: As there are no publicly available real-world datasets with both HDR images and events, we use the DSEC dataset [12] to experiment with real events. In DSEC, a car-mounted rig captures LDR video and hardware synchronized events. Due to FoV, optical center and resolution differences, events are spatially aligned to the video frames using calibration. We choose well-exposed frames as pseudo ground truth HDR (no clipped shadows/highlights) and degrade them with the exposure model, noise, clipping and quantization to generate short, medium and long exposure LDRs as discussed in section 3.2. Events are extracted from the event stream corresponding to the timestamps of the selected frames.

Quantitative results are shown in Table 1 (right). Similar to the synthetic dataset, our approach outperforms all other methods in each metric, with gaps of roughly 1dB and 0.7dB (PSNR-L and μ) to the second best-performer ADNet. Qualitative results are shown in Fig. 5. We outline in the crops noticeably better texture reconstruction in fast-moving objects (note, in this driving dataset, most motion occurs towards the edges of the frame, with the middle being relatively static), such as the tunnel surface and the lane edge. High-contrast structures, e.g. the traffic signal, are better recovered, and even in challenging regions of over-exposure object and scene structure is better preserved (e.g. rock to sky transition).

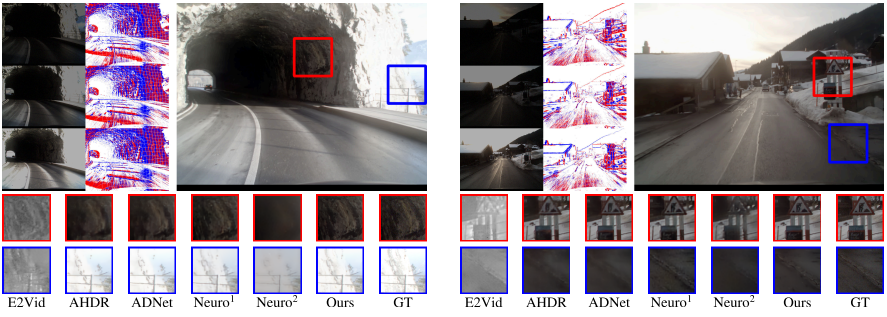

Figure 5: Qualitative results on the DSEC test set with real events. Top left: input LDRs and events, top right: predicted HDR image using our method, bottom row: comparison crops.

Table 2: Ablation studies for different model components on the HdM test set. See Sec. 4.1.

	Imbalanced Params				Balanced Params			
Method	Params(M)	PSNR-L	PSNR-μ	HV2	Params(M)	PSNR-L	PSNR-μ	HV2
Images-only	2.81	39.18	36.89	50.06	6.14	37.75	36.33	48.92
+ Event alignment	4.38	40.97	37.35	55.14	6.27	39.55	36.62	52.15
+ Event sub-sampling	4.67	41.32	37.60	54.85	6.74	40.39	36.98	53.33
+ Event-to-image distill.	6.13	**41.81**	**37.84**	**55.79**	6.13	**41.81**	**37.84**	**55.79**

4.1 Ablation Studies

We perform ablation studies examining the performance of each component of our model. Quantitative results for each ablation on the HdM test set are shown in Table 2 Left. We train each model equally in the four following scenarios. First, *Images-only* uses only bracketed LDRs as input, equivalent to ADNet. Second, *Images + Event alignment* is bracketed LDRs and the event alignment module discussed in section 3.1 (3). Third, *Images + Event sub-sampling* is bracketed LDRs and sub-sampling of the event stream with alignment to the reference without feature distillation. And fourth, *Images + Event-to-image distillation* is our complete model with sub-sampled events passed through the distillation module discussed in section 3.1 (4). The results show that each model configuration increases PSNR, thus validating our model design. The best result comes from the *Images + Event-to-image distillation* model, demonstrating the advantage of using feature distillation from events to images, rather than aligning the sub-sampled events in the event-feature domain.

Note that deleting modules for each ablation modifies the architecture, reducing the effective number of parameters, which may account for the changes in performance. We address this by increasing the number of features in the remaining modules to ensure the number of parameters in the network remains roughly constant for each configuration (Table 2 Right). Adding extra parameters leads to over-fitting; the Images-only model with balanced parameters only scores PSNR-L/μ 37.75/36.33dB. Even with balanced parameters, the performance improves with each module addition, and our complete configuration still performs the best.

Furthermore, we observe that the additional input information of events does not necessarily lead to performance improvement. To support this, we ran the following experiment: concatenating events with the input LDRs and feeding into the baseline ADNet. This leads to a performance decrease, scoring only PSNR-L/μ of 37.91/36.01dB because the input modalities occupy different domains. Thus, our specific architecture design is essential to leverage this additional information appropriately and achieve the reported performance gains.

5 Conclusions

We have presented a learning-based method leveraging events and bracketed LDRs to improve HDR reconstruction, with benefits over event- or image-only methods. In static scenes where event-only methods fail (no event signal), our method resorts back to the input LDRs and is equivalent to SoTA image-based method ADNet. In dynamic scenes, where image-based methods struggle, our method leverages high-frequency events to better align and reconstruct details. Our method is better at handling dynamic highlights, e.g. flashing lights and fast-moving textures, which image-based methods struggle to align due to over-exposure. Aligning events and images in feature space enriches our feature representation and leads to better reconstructions, and the event-to-image domain distillation allows the system to find an optimal feature space for both events and images without an event-based intensity guide image. We validated our method on both synthetic and real events, and conducted ablation studies supporting our contributions. Our method obtains significant improvements over other SoTA algorithms in all the measured metrics and noticeably improved visual results.

References

[1] MindSpore. https://www.mindspore.cn/.

[2] Patrick Bardow, Andrew J. Davison, and Stefan Leutenegger. Simultaneous Optical Flow and Intensity Estimation from an Event Camera. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 884–892, 2016.

[3] Ahmed Nabil Belbachir, Stephan Schraml, Manfred Mayerhofer, and Michael Hofstätter. A Novel HDR Depth Camera for Real-Time 3D 360° Panoramic Vision. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 425–432, 2014.

[4] Kelvin C. K. Chan, Xintao Wang, K. Yu, Chao Dong, and Chen Change Loy. Understanding Deformable Alignment in Video Super-Resolution. In *AAAI*, 2021.

[5] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, et al. Deformable Convolutional Networks. In *IEEE International Conference on Computer Vision (ICCV)*, pages 764–773, 2017.

[6] Paul E. Debevec and Jitendra Malik. Recovering High Dynamic Range Radiance Maps from Photographs. In *SIGGRAPH '08*, 2008.

[7] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Häusser, Caner Hazirbas, et al. FlowNet: Learning Optical Flow with Convolutional Networks. *IEEE International Conference on Computer Vision (ICCV)*, pages 2758–2766, 2015.

[8] Gabriel Eilertsen, Joel Kronander, Gyorgy Denes, Rafał K. Mantiuk, and Jonas Unger. HDR Image Reconstruction from a Single Exposure Using Deep CNNs. *ACM Transactions on Graphics (TOG)*, 36:1 – 15, 2017.

[9] Jan Froehlich, Stefan Grandinetti, Bernd Eberhardt, Simon Walter, Andreas Schilling, et al. Creating Cinematic Wide Gamut HDR-video for the Evaluation of Tone Mapping Operators and HDR-displays. In *Digital Photography X*, volume 9023, pages 279 – 288, 2014.

[10] Guillermo Gallego, Tobi Delbruck, G. Orchard, Chiara Bartolozzi, Brian Taba, et al. Event-based Vision: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44:154–180, 2022.

[11] Daniel Gehrig, Mathias Gehrig, Javier Hidalgo-Carrió, and Davide Scaramuzza. Video to Events: Recycling Video Datasets for Event Cameras. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[12] Mathias Gehrig, Willem Aarents, Daniel Gehrig, and Davide Scaramuzza. DSEC: A Stereo Event Camera Dataset for Driving Scenarios. *IEEE Robotics and Automation Letters*, 6:4947–4954, 2021.

[13] Jin Han, Chu Zhou, Peiqi Duan, Yehui Tang, Chang Xu, et al. Neuromorphic Camera Guided High Dynamic Range Imaging. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1727–1736, 2020.

[14] Samuel W. Hasinoff, Frédo Durand, and William T. Freeman. Noise-optimal Capture for High Dynamic Range Photography. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 553–560, 2010.

[15] Sayed Mohammad Mostafavi Isfahani, Lin Wang, Yo-Sung Ho, and Kuk jin Yoon. Event-based High Dynamic Range Image and Very High Frame Rate Video Generation Using Conditional Generative Adversarial Networks. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10073–10082, 2019.

[16] Sayed Mohammad Mostafavi Isfahani, Jonghyun Choi, and Kuk-Jin Yoon. Learning to Super Resolve Intensity Images From Events. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2765–2773, 2020.

[17] Huaizu Jiang, Deqing Sun, V. Jampani, Ming-Hsuan Yang, Erik G. Learned-Miller, et al. Super SloMo: High Quality Estimation of Multiple Intermediate Frames for Video Interpolation. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9000–9008, 2018.

[18] Nima Khademi Kalantari and Ravi Ramamoorthi. Deep High Dynamic Range Imaging of Dynamic Scenes. *ACM Transactions on Graphics (TOG)*, 36:1 – 12, 2017.

[19] Zeeshan Khan, Mukul Khanna, and Shanmuganathan Raman. FHDR: HDR Image Reconstruction from a Single LDR Image Using Feedback Network. *GlobalSIP*, pages 1–5, 2019.

[20] Ce Liu. Exploring New Representations and Applications for Motion Analysis. *PhD Thesis*, 2009.

[21] Zhen Liu, Wenjie Lin, Xinpeng Li, Qing Rao, Tingting Jiang, et al. ADNet: Attention-guided Deformable Convolutional Network for High Dynamic Range Imaging. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2021.

[22] Rafat K. 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. In *SIGGRAPH*, 2011.

[23] Tom Mertens, Jan Kautz, and Frank Van Reeth. Exposure Fusion. *15th Pacific Conference on Computer Graphics and Applications (PG'07)*, pages 382–390, 2007.

[24] Yuzhen Niu, Jianbin Wu, Wenxi Liu, Wenzhong Guo, and Rynson W. H. Lau. HDR-GAN: HDR Image Reconstruction From Multi-Exposed LDR Images With Large Motions. *IEEE Transactions on Image Processing*, 30:3885–3896, 2021.

[25] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. 2019.

[26] Feiyue Peng, Maojun Zhang, Shiming Lai, Hanlin Tan, and Shen Yan. Deep HDR Reconstruction of Dynamic Scenes. *IEEE 3rd International Conference on Image, Vision and Computing (ICIVC)*, pages 347–351, 2018.

[27] Eduardo Pérez-Pellitero, Sibi Catley-Chandar, Aleš Leonardis, and Radu Timofte. NTIRE 2021 Challenge on High Dynamic Range Imaging: Dataset, Methods and Results. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 691–700, 2021.

[28] Eduardo Pérez-Pellitero, Sibi Catley-Chandar, Richard Shaw, Aleš Leonardis, Radu Timofte, et al. NTIRE 2022 Challenge on High Dynamic Range Imaging: Methods and Results. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1008–1022, 2022.

[29] K. Prabhakar, Rajat Arora, Adhitya Swaminathan, Kunal Pratap Singh, et al. A Fast, Scalable, and Reliable Deghosting Method for Extreme Exposure Fusion. *IEEE International Conference on Computational Photography (ICCP)*, pages 1–8, 2019.

[30] K. Prabhakar, Susmit Agrawal, Durgesh Kumar Singh, Balraj Ashwath, and R. Venkatesh Babu. Towards Practical and Efficient High-Resolution HDR Deghosting with CNN. In *European Conference on Computer Vision (ECCV)*, 2020.

[31] K. Prabhakar, G. A. Senthil, Susmit Agrawal, R. Venkatesh Babu, Rama Krishna, et al. Labeled from Unlabeled: Exploiting Unlabeled Data for Few-shot Deep HDR Deghosting. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4873–4883, 2021.

[32] Henri Rebecq, Daniel Gehrig, and Davide Scaramuzza. ESIM: An Open Event Camera Simulator. *Conference on Robotics Learning (CoRL)*, October 2018.

[33] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. Events-To-Video: Bringing Modern Computer Vision to Event Cameras. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3852–3861, 2019.

[34] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. High Speed and High Dynamic Range Video with an Event Camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43:1964–1980, 2021.

[35] Cedric Scheerlinck, Henri Rebecq, Daniel Gehrig, Nick Barnes, Robert E. Mahony, et al. Fast Image Reconstruction with an Event Camera. *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 156–163, 2020.

[36] Pradeep Sen and Cecilia Aguerrebere. Practical High Dynamic Range Imaging of Everyday Scenes: Photographing the world as we see it with our own eyes. *IEEE Signal Processing Magazine*, 33:36–44, 2016.

[37] Pradeep Sen, Nima Khademi Kalantari, Maziar Yaesoubi, Soheil Darabi, Dan B. Goldman, et al. Robust Patch-based HDR Reconstruction of Dynamic Scenes. *ACM Transactions on Graphics (TOG)*, 31:1 – 11, 2012.

[38] Timo Stoffregen, Cedric Scheerlinck, Davide Scaramuzza, Tom Drummond, Nick Barnes, et al. Reducing the Sim-to-Real Gap for Event Cameras. In *European Conference on Computer Vision (ECCV)*, 2020.

[39] 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*, 2021.

[40] Lin Wang and Kuk-Jin Yoon. Knowledge Distillation and Student-Teacher Learning for Visual Intelligence: A Review and New Outlooks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[41] Lin Wang, Tae-Kyun Kim, and Kuk-Jin Yoon. EventSR: From Asynchronous Events to Image Reconstruction, Restoration, and Super-Resolution via End-to-End Adversarial Learning. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8312–8322, 2020.

[42] Lin Wang, Yujeong Chae, and Kuk-Jin Yoon. Dual Transfer Learning for Event-based End-task Prediction via Pluggable Event to Image Translation. *ArXiv*, abs/2109.01801, 2021.

[43] Lin Wang, Tae-Kyun Kim, and Kuk-Jin Yoon. Joint Framework for Single Image Reconstruction and Super-Resolution with an Event Camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[44] Shangzhe Wu, Jiarui Xu, Yu-Wing Tai, and Chi-Keung Tang. Deep High Dynamic Range Imaging with Large Foreground Motions. In *European Conference on Computer Vision (ECCV)*, 2018.

[45] Qingsen Yan, Dong Gong, Qinfeng Shi, Anton van den Hengel, Chunhua Shen, et al. Attention-Guided Network for Ghost-Free High Dynamic Range Imaging. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1751–1760, 2019.

[46] Fisher Yu, Vladlen Koltun, and Thomas A. Funkhouser. Dilated Residual Networks. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 636–644, 2017.

[47] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Unsupervised Event-based Optical Flow Using Motion Compensation. In *European Conference on Computer Vision (ECCV) Workshops*, 2018.

[48] Yunhao Zou, Yinqiang Zheng, Tsuyoshi Takatani, and Ying Fu. Learning to Reconstruct High Speed and High Dynamic Range Videos from Events. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2024–2033, 2021.