Event-based Image Enhancement Under High Dynamic Range Scenarios

Jingchong Weng, Boyang Li, and Kai Huang

Sun Yat-sen University, Guangzhou, China
wengjch3@mail2.sysu.edu.cn
{liby83, huangk36}@mail.sysu.edu.cn

Abstract. Event cameras, as bio-inspired vision sensors with a high dynamic range, are capable of addressing the problems of local overexposure or underexposure that conventional frame-based cameras encounter in scenarios with high dynamic range or fluctuating lighting conditions. Due to the modality gap between the two types of cameras, simple direct fusion is not feasible. Additionally, the ghosting artifacts caused by the deviation in the camera positions and frame-rates also affects the quality of final fused image. To solve the problems, this paper proposes a joint framework that combines locally poor-exposed frames with event streams captured by the event camera to enhance the images with detailed textures in high dynamic range scenarios. Specifically, a lightweight multi-scale receptive field block is employed for rapid modality conversion from event streams to frames. Besides, a dual-branch fusion module is proposed to align features and remove ghosting artifacts. Experimental results demonstrate that the proposed method effectively mitigates information loss in both highly bright and dark regions of images across a range of extreme lighting conditions, generating the both realistic and natural images.

Keywords: Image enhancement · Event camera · High dynamic range.

1 Introduction

The dynamic range of illumination in a natural scene varies greatly. Due to the constraints inherent in conventional frame-based cameras, the dynamic range captured in a single image is significantly narrower compared to that of natural scenes [28], which leads to the instances of local overexposure or underexposure. This challenge not only severely impacts image quality but also complicates downstream tasks, *e.g.*, object detection, tracking and visual SLAM.

Currently, numerous studies focus on broadening the dynamic range of conventional cameras using the images with different exposure times [37]. The existing methods are categorized into two ways: High Dynamic Range (HDR) image reconstruction and multi-exposure image fusion (MEF). The former merges snapshots taken at different exposure times using the Camera Response Function (CRF) to reconstruct HDR images [2,41], displayed on the ordinary display

device through tone mapping (TM). The latter approach directly employs MEF technology to fuse input images with different exposure times into an image with rich information and vivid colors [32, 40], which do not consider camera curve calibration, HDR reconstruction, and tone mapping. However, in dynamic scenes, the ego-motion of cameras and moving objects within the scene cause misalignment in multi-exposure image sequences, leading to blurred and ghosting artifacts in the fusion results. Additionally, the high frame rate imaging requirements in high-speed scenes constrain the camera exposure time, resulting in the loss of scene information in dark areas.

Unlike conventional cameras that capture images at fixed frame rates, event cameras, inspired by biological vision systems [4], measure changes in light intensity and generate asynchronous events to capture scene information. Event cameras have advantages including high dynamic range (140 dB), high temporal resolution (1 µs), and low power consumption. However, event streams represent a fundamentally different format from conventional images, rendering existing computer vision algorithms inapplicable. As a result, the common approach is to reconstruct images from event streams. Some methods [25, 27, 35] have explored various strategies to enhance image reconstruction quality in different scenarios. Nevertheless, these event-based methods lack sufficient image information to compute the absolute intensity of each pixel. Furthermore, while maintaining image reconstruction quality, the models should be as lightweight as possible to preserve the low latency of event cameras, which has not been fully explored.

Event-based cameras, with their high dynamic range capabilities, are beneficial to capture contour details in high dynamic range and motion scenarios, overcoming the overexposure and underexposure challenges faced by conventional cameras. Yet, conventional cameras still offer richer detail in well-exposed areas. Therefore, we exploit to leverage event streams to supplement the scene information of a single overexposure or underexposure image for high-quality enhancing in high dynamic range scenes. Since event cameras do not capture color information, our approach focuses on enhancing grayscale frames.

To achieve our objective, two challenges require to be addressed: (i) The modality conversion from the event streams to images is challenging. We aim to maintain the reconstructed image quality from the event streams while optimizing the model's real-time performance to handle high-speed dynamic scenes. (ii) Effectively aligning and fusing scene information from frames and the event streams is another challenge. There is a frame-rate difference and a spatial deviation between the two types of cameras, which necessitates the design of a robust alignment strategy to avoid ghosting artifacts.

In this paper, in response to the degradation in high dynamic range scenes, we propose the an event-based image enhancement framework, which takes the event stream and locally overexposure or underexposure frames as inputs and outputs high-quality images with detailed textures. Fig. 1 shows the visual results of our proposed framework. Specifically, an event encoder is proposed based on receptive field blocks (RFB) to extract multi-scale features from the event streams, enabling fast and efficient modality conversion. To suppress ghosting

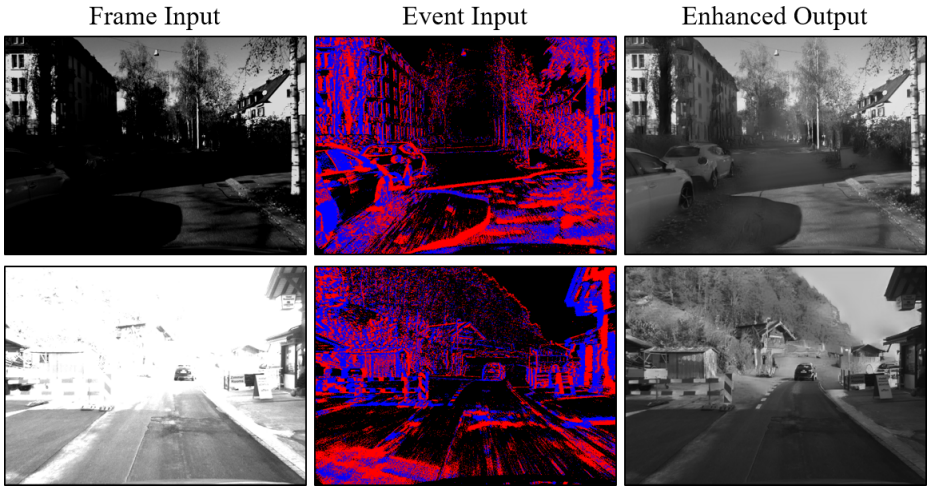

Fig. 1: The visual results of our event-based image enhancement method

artifacts during fusion, a dual-branch fusion module (DFM) is designed based on deformable convolutional block and spatial attention block to achieve feature alignment and fusion from event streams to frame. To sum up, the contributions of this paper are as follows:

- An event-based image enhancement framework is proposed to solve the visual degradation problem in high dynamic range scenes.
- To solve the problem arising from the modality conversion of event streams into images, receptive field blocks are employed to facilitate swift and efficient transformation of event streams into image representations.
- To suppress ghosting artifacts, a dual-branch feature alignment module is designed to achieve feature alignment from event streams to images.
- Extensive experiments on an outdoor driving scene dataset demonstrate the effectiveness of the proposed framework in high dynamic range scenarios.

The remainder of the paper is organized as follows. Section 2 reviews the related work. Section 3 introduces the methodology. Section 4 presents the experiments. Section 5 makes a conclusion.

2 Related Work

2.1 Frame-based imaging in high dynamic range scenes

Compared to HDR image reconstruction, multi-exposure fusion methods are simpler, more cost-effective, and efficient alternative for achieving high dynamic range imaging tasks [37]. To counter the challenges in dynamic scenes, recent approaches leverage motion detection and image registration to reduce motion

artifacts [19,30]. Specifically, learning-based methods are utilized for optical flow estimation and deep image fusion [22, 23]. Additionally, attention mechanisms are integrated into the system to bolster the robustness of optical flow-based registration [8, 38]. Although numerous studies have investigated deep learning for exposure correction in images, recovering details in overexposed or underexposed areas without additional cues is still a significant challenge [1, 6]. As a result, a model that better captures the complex relationships between different exposure levels and the original scene is required in the task.

2.2 Event-based image reconstruction

As the representation of event streams is different from conventional images, related work on event-based image reconstruction is discussed in this section. Due to the high dynamic range of event cameras, reconstructing images from pure events holds tremendous potential for perception in dynamic scenes. Early works [5, 14] achieved image reconstruction by moving the event-camera in a static scene. To address the movement restriction, Bardow *et al.* [3] and Munda *et al.* [21] proposed a dynamic energy minimization framework to reconstruct intensity images in motion scenes. Recently, Wang *et al.* [33] and Mostafavi *et al.* [20] used generative adversarial networks (GANs) to reconstruct images from events. Rebecq *et al.* [25] and Scheerlinck *et al.* [27] achieved stable reconstruction results through a recurrent neural network. Weng *et al.* [35] presented a hybrid CNN-Transformer network to further promote the reconstruction quality. However, these event-based methods lack sufficient events information to compute the absolute intensity of each pixel in the complicated scenarios.

2.3 Combination of event and frame-based camera

Conventional cameras are constrained by the limited dynamic range while event-cameras capture the visual information across a much broader dynamic range with limited texture features. By integrating event-cameras with frame-based cameras, the strengths of both technologies allow for the retention of texture details while also capturing the full range of light levels, thus improving the overall visual information. Wang *et al.* [31] proposed a sparse learning and integration method aimed at enhancing image clarity through deblurring and achieving super-resolution. Han *et al.* [10] suggested reconstructing intensity maps from the combination of events and low dynamic range (LDR) frames. Jiang *et al.* [12] proposed a event-guided low-illumination image enhancement method. Yang *et al.* [42] proposed a reconstruction framework with events and LDR videos.

3 Methodology

This section introduces our proposed event-based image enhancement method under high dynamic range scenes. We first introduce the event representation method. Then, we describe the specific architecture of our approach and the loss functions used for training.

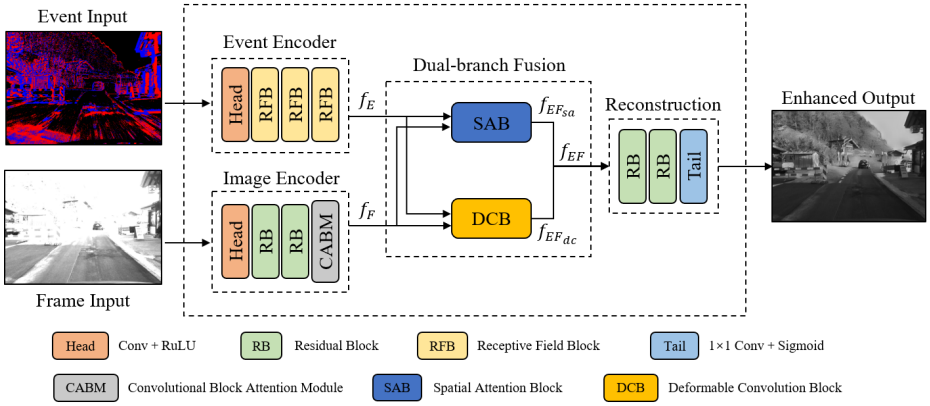

Fig. 2: Overview of our proposed image enhancement framework. The event streams and poorly exposed frames are used as input to obtain the event and frame features by event encoder and image encoder respectively. Next, the final enhanced output is fused and reconstructed by the dual-branch fusion and reconstruction module.

3.1 Event Representation

The event streams obtained from an event-camera is represented by $\varepsilon_n = \{e_k\}_{k=1}^{N}$, where N is the number of events. Each $e_i \in \varepsilon_n$ can be represented by a tuple (x_i, y_i, t_i, p_i), where x and y represent the spatial position of the event, t represents the timestamp of the event, and $p = \pm 1$ represents the polarity of the event. In order to input the event stream ε_n into frame-based reconstruction, we need to convert it into a fixed-size tensor representation [43]. The event stream is transformed into an event voxel grid E using time bilinear interpolation. E is divided into B bins using the following equation:

$$E_{j\in[0,B-1]} = \sum_{i=0}^{N} p_i max\left(0, 1 - \left|j - \frac{t_i - t_0}{\Delta t}(B-1)\right|\right) \tag{1}$$

where $\Delta t = t_{N-1} - t_0$ is the time span of the N events (t_{N-1} is the end time and t_0 is the start time), and this method evenly distributes the entire stream ε_n in $[t_0, t_{N-1}]$ into B contiguous bins. In this paper, we set $B = 5$ for all experiments. Eventually, we convert the event stream to an event tensor $E \in \mathbb{R}^{5\times W\times H}$, where W and H are the spatial sizes of the frame.

3.2 Network Architecture

As illustrated in Fig. 2, our proposed model consists of an event encoder, an image encoder, a dual-branch fusion module, and a reconstruction module. Instead of adopting a U-Net [26] style structure, which downsamples the input first before upsampling, our model operates at full resolution to preserve the informative details. To address the challenges brought about by the different modalities of event

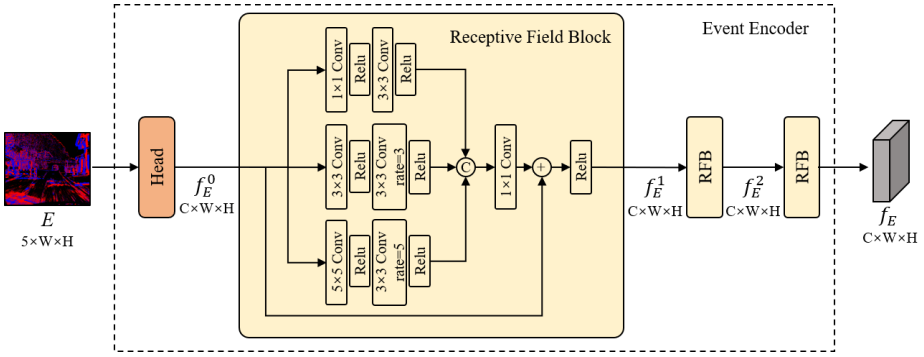

Fig. 3: The detailed structure of event encoder based on receptive field blocks.

streams and frame in image enhancement tasks, namely, image frames are only adjusted in the original domain, while event streams need to be reconstructed across domains, we design the event encoder and image encoder with different architectures. A dual-branch fusion module is leveraged to achieve multi-modal feature fusion. Finally, the enhanced clear images could be obtained from the fused feature using the reconstruction module. We introduce the structures of event encoder, image encoder and reconstruction module in detail respectively, and the most critical dual-branch fusion module will be explained separately in the Sec. 3.3.

Event Encoder. We propose an event encoder based on the Receptive Field Block (RFB) [17], which can capture a broader range of contextual information by expanding the receptive field. As shown in Fig. 3, the RFB module contains three branches, each branch has convolutions with a different kernel size and different dilation rate. The features from each branch are aggregated through a concatenation and 1×1 convolution, effectively extracts and integrates multi-scale features from event data. Given the event tensor $E \in \mathbb{R}^{5\times W\times H}$, the head which consists of a convolution with ReLU is employed to obtain the initial event feature $f_E^0 \in \mathbb{R}^{C\times W\times H}$. In our work, we set C as 16. Then f_E^0 is passed to the stack of 3 RFBs for layer-by-layer feature extraction and fusion, generating the event feature $f_E \in \mathbb{R}^{C\times W\times H}$.

Image Encoder. As shown in Fig. 2, given the frame $F \in \mathbb{R}^{1\times W\times H}$, the head is employed for obtain the initial image feature $f_F^0 \in \mathbb{R}^{C\times W\times H}$. Then we stack 2 residual blocks [11] as a lightweight encoder. In order to adjust the global exposure of the frame, a convolutional block attention module (CBAM) [36] is used to perform channel and spatial attention. Through the above steps, we obtain the frame feature $f_F \in \mathbb{R}^{C\times W\times H}$.

Reconstruction Module. In reconstruction stage, we decode the fused feature f_{EF} obtained from the dual-branch fusion module (DFM) to reconstruct the enhanced image. We stack two residual modules to fine-tune the f_{EF}. The enhanced image $F_{out} \in \mathbb{R}^{1\times W\times H}$ is finally reconstructed using a tail consisting of convolution and sigmoid.

3.3 Dual-branch Fusion Module

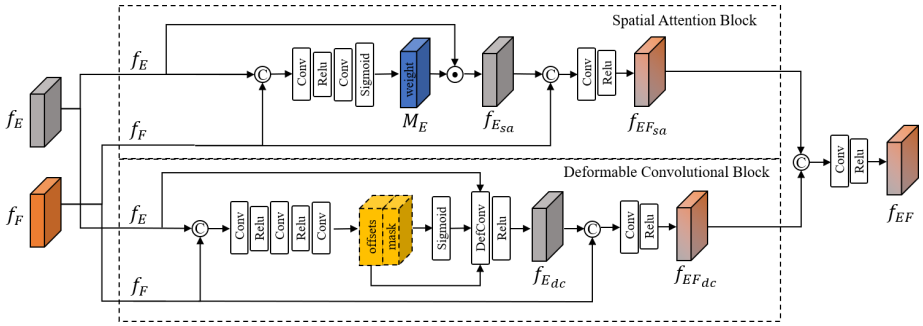

Fig. 4: Dual-branch Fusion Module

The dual-branch fusion module (DFM) is proposed to effectively integrate the features of the event and frame, as shown in Fig. 4. The DFM comprises two routes, each taking the event features f_E and the frame features f_F as input. One branch adopts a Spatial Attention Block [39] to produce the aggregated feature $f_{EF_{sa}}$. A second branch adopts a Deformable Convolutional Block [18] to produce the aligned feature $f_{EF_{dc}}$. Then both $f_{EF_{sa}}$ and $f_{EF_{dc}}$ are concatenated and fused through a convolution and a relu, generating the fused feature f_{EF}.

Spatial Attention Block. As the Spatial Attention Block [39] allows the network to extract features of particular areas of the inputs, we use it to adaptively supplement poorly exposed regions of conventional frames with events without the need for hand-crafted masks [10, 42]. As depicted in Fig. 4, the event attention map M_E is obtained from $[f_E, f_F]$ after 2 convolutions follows by a relu and a sigmoid respectively. The weighted event feature $f_{E_{sa}}$ is computed by performing point-wise multiplication on f_E and M_E. Finally we obtain the aggregated feature $f_{EF_{sa}}$ by fusing $[f_{E_{sa}}, f_F]$ through a convolution and a relu.

Deformable Convolutional Block. Deformable convolutions have been used for image alignment in feature space, aiding video restoration [34] and HDR reconstruction [18] tasks. Inspired by this, we design a lightweight deformable convolutional block to adaptively align cross-modal features for conventional frames and events. As depicted in Fig. 4, the offsets and the mask are predicted according to f_E and f_F with a stack of convolutions and relu activations. With offsets, mask and f_E, the aligned event feature $f_{E_{dc}}$ of the conventional frame can be computed by the deformable convolution. Similar to Spatial Attention Block, we fuse $f_{E_{dc}}$ and f_F to obtain the aligned feature $f_{EF_{dc}}$.

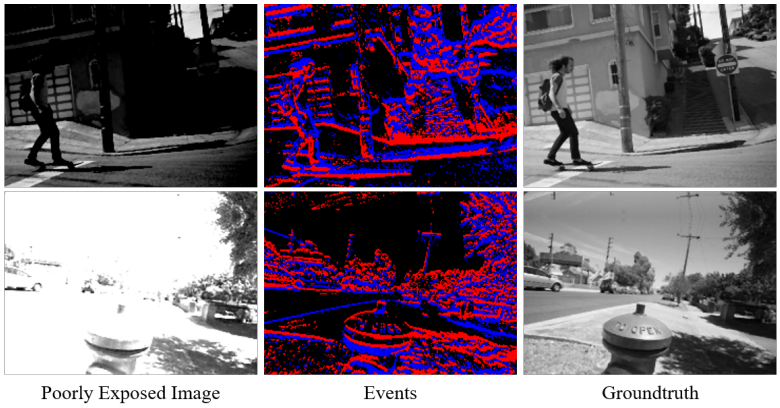

Fig. 5: Example samples of our training dataset.

3.4 Loss Functions

We train our network by minimizing the loss function which use a combination of mean square error (MSE) loss and perceptual similarity (LPIPS) loss [13], with equal weights. Our total loss $\mathcal{L}$ can be written as:

$$\mathcal{L} = \mathcal{L}_{mse} + \mathcal{L}_{perc} \tag{2}$$

The MSE loss $\mathcal{L}_{mse}$ is used to evaluate the pixel-level error between the enhanced image I_{out} and groundtruth image I_{gt}, which is defined as:

$$\mathcal{L}_{\text{mse}} = \|I_{gt} - I_{out}\|_2^2. \tag{3}$$

The LPIPS loss $\mathcal{L}_{perc}$ is introduced to measure the feature similarity between images, which is defined based on the feature maps extracted by the VGG16 [29] model pre-trained on ImageNet [7]:

$$\mathcal{L}_{perc} = \sum_l \left(||\phi^l(I_{gt}) - \phi^l(I_{out})||_2^2 \right) \tag{4}$$

where ϕ^l represents the l-th layer of the feature extraction network.

4 Experiments

4.1 Experimental setup

Training dataset. For the training of our model, a large amount of triplets (events, poorly exposed image, groudtruch image) are required. However, there are no readily available datasets containing these triplets. To synthesize such a dataset, We first leverage the training data obtained from an event-based image reconstruction method E2VID [25], which yields a substantial number of

corresponding event streams and images with a resolution of 240 × 180. The event streams are generated by the event simulator ESIM [24], while MS-COCO images [16] are mapped onto a 3D plane and random 6-DOF camera movements are utilized to trigger events. From these data, we select those images rich in scene texture and luminance information to form our groundtruth image, which represent well-exposed images in high dynamic range scenes. We adjust the exposure by modifying the pixel values of the groundtruth image to generate locally underexposed and overexposed images. Additionally, to mimic the imaging perspective deviation caused by positional discrepancies between event cameras and traditional cameras in real-world scenarios, we perform minor affine transformations on the event streams for data augmentation. In total, there are 47852 frames for training, and 2523 frames for validation. Some examples are shown in Fig. 5.

Testing dataset. We evaluate our model on a publicly released stereo event camera dataset in driving scenarios: DESC [9]. In DESC, events and images are from different cameras with different resolutions. As the calibration parameters are provided, we warp the images to the event locations with a resolution of 640 × 480. Since DSEC covers a large variety of illumination conditions, we selected well-exposed image sequences with rich scene information as groundtruth. Similar to the training set, we adjust the exposure of the groundtruth images to synthesize locally underexposed or overexposed images for testing.

Evaluation metrics. For quantitative evaluation, we consider three widely-used evaluation metrics: (i) Peak Signal-to-Noise Ratio (PSNR): PSNR quantifies the fidelity between the enhanced image and the ground truth. A higher PSNR value suggests less distortion, indicating superior enhancement performance.(ii) Structural Similarity Index (SSIM): SSIM assesses the resemblance between two images based on three components: brightness, contrast, and structure. A higher SSIM value denotes better enhancement quality. (iii) Perceptual Similarity (LPIPS): LPIPS evaluates the similarity of features between images rather than comparing the images directly, as described in Eq. 4. A lower LPIPS value signifies greater similarity.

Implementation details. Our network is implemented using the Pytorch framework. Adam [15] is utilized as the optimizer with the initial learning rate 0.0002, which is decayed by a factor of 0.9 every 20 epochs. Our model is trained for 100 epochs with batch size of 16 on 1 NVIDIA Tesla V100 GPU.

4.2 Comparisons with State-of-the-art Methods

We compare the proposed method to five state-of-the-art imaging methods, including an frame-based exposure correction method: MSEC [1]; two pure event-based image reconstruction method: FireNet [27] and E2VID [25]; the grayscale variant of two event guided HDR reconstruction method: NeurImg-HDR [10] and HDRev [42].

Quantitative evaluation. We quantitatively compare all methods on DSEC dataset and summarize the results in Tab. 1. Our method outperforms existing methods in terms of all metrics. Regarding PSNR, SSIM, and LPIPS, we improve

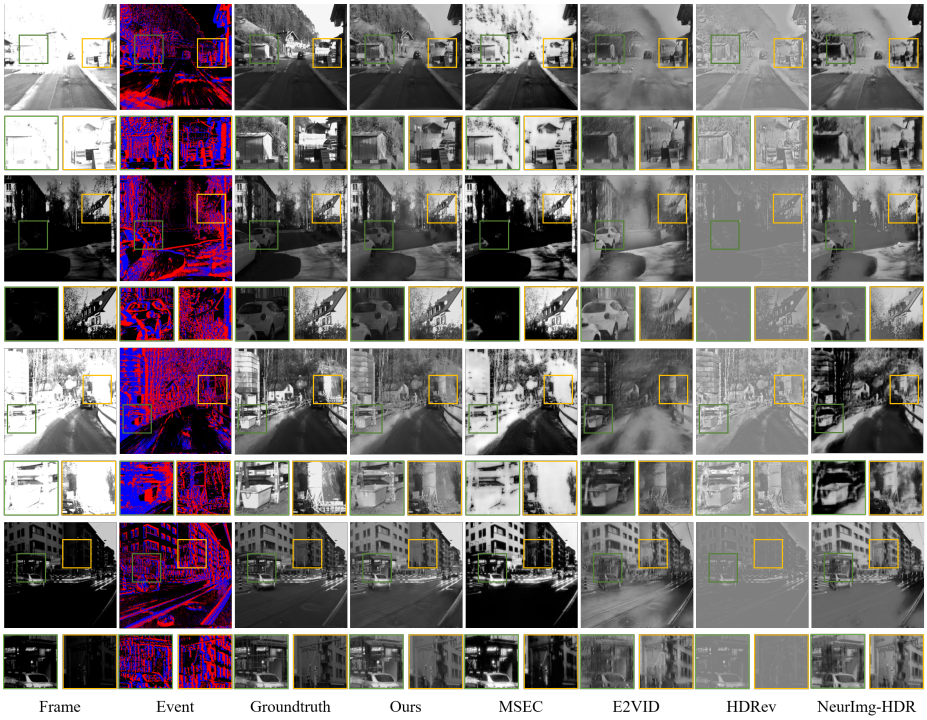

Fig. 6: Visual quality comparisons on DSEC dataset.

6.007 dB, 0.287, and 0.118 respectively, compared to the second-best methods. Event-based methods typically underperform frame-supported methods on all metrics, due to the absence of image information.

Qualitative comparisons. We also conduct extensive qualitative comparisons for all methods, as shown in Fig. 6. In over/under-exposed regions of the input frames, most details are significantly lost. Frame-based MSEC [1] can perform exposure correction to enhance image contrast, but can not restore the lost scene details. Due to the sparsity of events, event-based E2VID [25] results in severe artifacts and fails to recover sufficient scene information. NeurImg-HDR [10] fuses the reconstructed image of E2VID [25] with the frame. However, this post-fusion method is seriously affected by the quality of the event-based reconstructed image. HDRev [42] employs a fusion strategy using a shared representation space, which results in overall image distortion. In comparison, Our proposed method not only corrects image exposure and recovers the lost details, but also effectively predicts the brightness information of overexposed or underexposed regions.

Furthermore, in Fig. 7 we show the image enhancement result of our method in an extremely overexposed scene of a car exiting a tunnel. The enhanced image restored the cliff curve outside the tunnel, which is important for improving driving safety.

Table 1: Quantitative comparison on DSEC dataset. ↑ (↓) means higher (lower) is better. Best in bold, the second best with underline.

Methods	PSNR ↑	SSIM ↑	LPIPS ↓
MSEC [1]	<u>14.125</u>	0.475	0.436
FireNet [27]	9.372	0.314	0.513
E2VID [25]	11.746	0.393	0.483
HDRev [42]	9.630	<u>0.484</u>	<u>0.339</u>
NeurImg-HDR [10]	12.384	0.414	0.422
Ours	**20.132**	**0.771**	**0.211**

Fig. 7: An exiting tunnel scenario in DSEC dataset. The scene information outside the tunnel has been lost in the image frame.

4.3 Ablation study

Influence of Event Encoder. We evaluated the impact of our event encoder by replacing the feature extractor RFBs with residual blocks. As shown in Tab. 2, when the RFB is removed, PSNR and SSIM drop significantly by 3.187 dB and 0.071 respectively. In order to further verify the effectiveness of RFB. We connect the event encoder and the tail block to build the purely event-based reconstruction variant RFB-Net. As shown in Tab. 3, compare with FireNet [27] and E2VID [25], RFB-Net achieves a balance between imaging quality and computational performance. Especially for LPIPS, RFB-Net outperforms the other two methods with only 51k parameters and 15.6 flops.

Influence of Image Encoder. We evaluated the impact of our image encoder by directly removing the CBAM module. As shown in Tab. 2, We can see that the PSNR and SSIM decreases sharply by 3.610 dB and 0.076 respectively, which shows the effectiveness of the global attention adjustment for frame features in our image enhancement task.

Influence of Dual-branch Fusion Module. We built three models to evaluate the impact of the dual-branch fusion module (DFM). (i) W/o DFM. We replace the DFM with a direct concatenation and convolution layer. (ii) W/o DCB. We remove the deformable convolutional block in DFM. (iii) W/o SAB. We remove the spatial attention block in DFM. Tab. 2 shows that the PSNR values of these three models decreased by 3.795, 0.411 and 0.213 respectively. This result shows the effectiveness of our dual-branch fusion module. Moreover, we visualize the behavior of DFM in Fig. 8. In particular, we illustrate 2 pooly exposed frames, the respective event features f_E from event encoder, the weighted event

Table 2: Ablation results

Methods	PSNR ↑	SSIM ↑	LPIPS ↓
W/o RFB	17.450	0.712	0.241
W/o CABM	17.027	0.707	0.241
W/o DFM	16.842	0.689	0.253
W/o DCB	20.226	0.774	0.217
W/o SAB	20.450	0.768	0.218
Ours	**20.637**	**0.783**	**0.204**

Table 3: Quantitative comparison of the event-based variant RFB-Net with two other methods, in terms of imaging quality and computational performance.

Methods	PSNR ↑	SSIM ↑	LPIPS ↓	Parameters(k)	Flops(G)
FireNet [27]	9.372	0.314	0.513	38	12.6
E2VID [25]	**11.746**	**0.393**	0.483	10700	147.2
RFB-Net	10.589	0.363	**0.462**	51	15.6

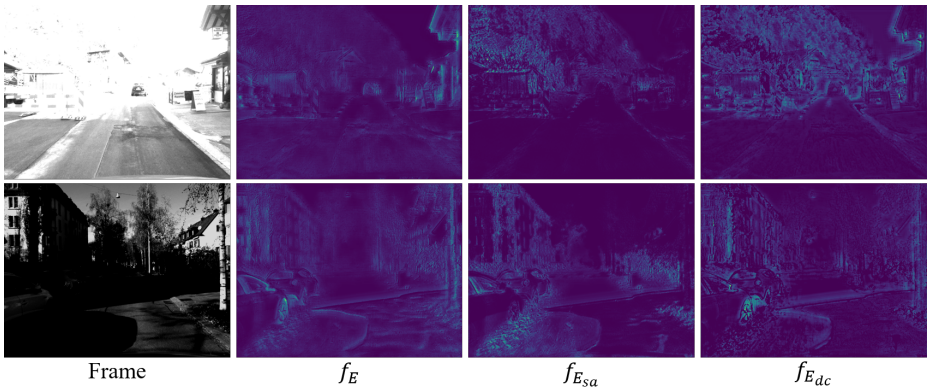

Fig. 8: Qualitative representation of the behavior from dual-branch fusion module (DFM). f_E is the event feature from event encoder. $f_{E_{sa}}$ is the weighted event feature from spatial attention block (SAB). $f_{E_{dc}}$ is the aligned event feature from deformable convolutional block (DCB).

features $f_{E_{sa}}$ from spatial attention block and the aligned event features $f_{E_{dc}}$ from deformable convolutional block. Only the first channel is visualized for all features. Fig. 8 effectively illustrates how the spatial attention block uses the event features to restore scene details in overexposed and underexposed region of the frames. And the deformable convolutional block effectively enhances the event features and aligns them to the frame features.

5 Conclusion

In this paper, we propose an event-based image enhancement framework to solve the visual degradation problem in high dynamic range scenes. We architect corresponding encoders for the two distinct modalities of input. We design a dual-branch fusion strategy to achieve efficient cross-modal alignment and fusion imaging between events and images. Extensive experiments demonstrate the effectiveness of the proposed framework in high dynamic range scenarios.

Acknowledgments. This work was supported by Guangxi Key R & D Program (No. GuikeAB24010324) and the Basic Research Operating Expenses of Universities-Young Teachers Cultivation Programs (Grant No. 24qnpy143).

References

1. Afifi, M., Derpanis, K.G., Ommer, B., Brown, M.S.: Learning multi-scale photo exposure correction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9157–9167 (2021)
2. Banterle, F., Artusi, A., Debattista, K., Chalmers, A.: Advanced high dynamic range imaging. AK Peters/CRC Press (2017)
3. Bardow, P., Davison, A.J., Leutenegger, S.: Simultaneous optical flow and intensity estimation from an event camera. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 884–892 (2016)
4. Brandli, C., Berner, R., Yang, M., Liu, S.C., Delbruck, T.: A 240× 180 130 db 3 μs latency global shutter spatiotemporal vision sensor. IEEE Journal of Solid-State Circuits **49**(10), 2333–2341 (2014)
5. Cook, M., Gugelmann, L., Jug, F., Krautz, C., Steger, A.: Interacting maps for fast visual interpretation. In: The 2011 International Joint Conference on Neural Networks. pp. 770–776. IEEE (2011)
6. Cui, Z., Li, K., Gu, L., Su, S., Gao, P., Jiang, Z., Qiao, Y., Harada, T.: You only need 90k parameters to adapt light: a light weight transformer for image enhancement and exposure correction. arXiv preprint arXiv:2205.14871 (2022)
7. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
8. Deng, Y., Liu, Q., Ikenaga, T.: Multi-scale contextual attention based hdr reconstruction of dynamic scenes. In: Twelfth International Conference on Digital Image Processing (ICDIP 2020). vol. 11519, pp. 413–419. SPIE (2020)
9. Gehrig, M., Aarents, W., Gehrig, D., Scaramuzza, D.: Dsec: A stereo event camera dataset for driving scenarios. IEEE Robotics and Automation Letters **6**(3), 4947–4954 (2021)
10. Han, J., Zhou, C., Duan, P., Tang, Y., Xu, C., Xu, C., Huang, T., Shi, B.: Neuromorphic camera guided high dynamic range imaging. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1730–1739 (2020)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
12. Jiang, Y., Wang, Y., Li, S., Zhang, Y., Zhao, M., Gao, Y.: Event-based low-illumination image enhancement. IEEE Transactions on Multimedia (2023)

13. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14. pp. 694–711. Springer (2016)
14. Kim, H., Handa, A., Benosman, R., Ieng, S.H., Davison, A.J.: Simultaneous mosaicing and tracking with an event camera. J. Solid State Circ **43**, 566–576 (2008)
15. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
16. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014)
17. Liu, S., Huang, D., et al.: Receptive field block net for accurate and fast object detection. In: Proceedings of the European conference on computer vision (ECCV). pp. 385–400 (2018)
18. Liu, Z., Lin, W., Li, X., Rao, Q., Jiang, T., Han, M., Fan, H., Sun, J., Liu, S.: Adnet: Attention-guided deformable convolutional network for high dynamic range imaging. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 463–470 (2021)
19. Ma, K., Li, H., Yong, H., Wang, Z., Meng, D., Zhang, L.: Robust multi-exposure image fusion: a structural patch decomposition approach. IEEE Transactions on Image Processing **26**(5), 2519–2532 (2017)
20. Mostafavi, M., Wang, L., Yoon, K.J.: Learning to reconstruct hdr images from events, with applications to depth and flow prediction. International Journal of Computer Vision **129**, 900–920 (2021)
21. Munda, G., Reinbacher, C., Pock, T.: Real-time intensity-image reconstruction for event cameras using manifold regularisation. International Journal of Computer Vision **126**(12), 1381–1393 (2018)
22. Peng, F., Zhang, M., Lai, S., Tan, H., Yan, S.: Deep hdr reconstruction of dynamic scenes. In: 2018 IEEE 3rd International Conference on Image, Vision and Computing (ICIVC). pp. 347–351. IEEE (2018)
23. Prabhakar, K.R., Arora, R., Swaminathan, A., Singh, K.P., Babu, R.V.: A fast, scalable, and reliable deghosting method for extreme exposure fusion. In: 2019 IEEE International Conference on Computational Photography (ICCP). pp. 1–8. IEEE (2019)
24. Rebecq, H., Gehrig, D., Scaramuzza, D.: Esim: an open event camera simulator. In: Conference on robot learning. pp. 969–982. PMLR (2018)
25. Rebecq, H., Ranftl, R., Koltun, V., Scaramuzza, D.: Events-to-video: Bringing modern computer vision to event cameras. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3857–3866 (2019)
26. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. pp. 234–241. Springer (2015)
27. Scheerlinck, C., Rebecq, H., Gehrig, D., Barnes, N., Mahony, R., Scaramuzza, D.: Fast image reconstruction with an event camera. In: WACV. pp. 156–163 (2020)
28. Shen, R., Cheng, I., Basu, A.: Qoe-based multi-exposure fusion in hierarchical multivariate gaussian crf. IEEE Transactions on Image Processing **22**(6), 2469–2478 (2012)
29. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)

30. Ulucan, O., Ulucan, D., Turkan, M.: Ghosting-free multi-exposure image fusion for static and dynamic scenes. Signal Processing **202**, 108774 (2023)
31. Wang, B., He, J., Yu, L., Xia, G.S., Yang, W.: Event enhanced high-quality image recovery. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16. pp. 155–171. Springer (2020)
32. Wang, J., Li, X., Liu, H.: Exposure fusion using a relative generative adversarial network. IEICE TRANSACTIONS on Information and Systems **104**(7), 1017–1027 (2021)
33. Wang, L., Kim, T.K., Yoon, K.J.: Eventsr: From asynchronous events to image reconstruction, restoration, and super-resolution via end-to-end adversarial learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8315–8325 (2020)
34. Wang, X., Chan, K.C., Yu, K., Dong, C., Change Loy, C.: Edvr: Video restoration with enhanced deformable convolutional networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. pp. 0–0 (2019)
35. Weng, W., Zhang, Y., Xiong, Z.: Event-based video reconstruction using transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2563–2572 (2021)
36. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV). pp. 3–19 (2018)
37. Xu, F., Liu, J., Song, Y., Sun, H., Wang, X.: Multi-exposure image fusion techniques: A comprehensive review. Remote Sensing **14**(3), 771 (2022)
38. Yan, Q., Gong, D., Shi, J.Q., Van Den Hengel, A., Shen, C., Reid, I., Zhang, Y.: Dual-attention-guided network for ghost-free high dynamic range imaging. International Journal of Computer Vision pp. 1–19 (2022)
39. Yan, Q., Gong, D., Shi, Q., Hengel, A.v.d., Shen, C., Reid, I., Zhang, Y.: Attention-guided network for ghost-free high dynamic range imaging. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1751–1760 (2019)
40. Yan, Q., Gong, D., Zhang, P., Shi, Q., Sun, J., Reid, I., Zhang, Y.: Multi-scale dense networks for deep high dynamic range imaging. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 41–50. IEEE (2019)
41. Yan, Q., Zhang, L., Liu, Y., Zhu, Y., Sun, J., Shi, Q., Zhang, Y.: Deep hdr imaging via a non-local network. IEEE Transactions on Image Processing **29**, 4308–4322 (2020)
42. Yang, Y., Han, J., Liang, J., Sato, I., Shi, B.: Learning event guided high dynamic range video reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13924–13934 (2023)
43. Zhu, A.Z., Yuan, L., Chaney, K., Daniilidis, K.: Unsupervised event-based learning of optical flow, depth, and egomotion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 989–997 (2019)