Towards Robust Event-guided Low-Light Image Enhancement: A Large-Scale Real-World Event-Image Dataset and Novel Approach

Guoqiang Liang[1] Kanghao Chen[1] Hangyu Li[1] Yunfan Lu[1] Lin Wang[1,2*]
[1]AI Thrust, HKUST(GZ) [2]Dept. of Computer Science and Engineering, HKUST
{gliang041,kchen879,hli886,ylu066}@connect.hkust-gz.edu.cn, linwang@ust.hk

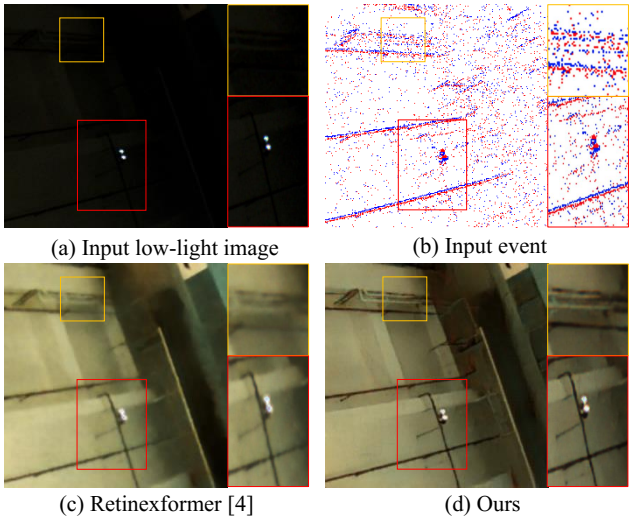
Project Page: https://vlislab22.github.io/eg-lowlight/

Abstract

*Event camera has recently received much attention for low-light image enhancement (LIE) thanks to their distinct advantages, such as high dynamic range. However, current research is prohibitively restricted by the lack of large-scale, real-world, and spatial-temporally aligned event-image datasets. To this end, we propose a real-world (indoor and outdoor) dataset comprising over **30K** pairs of images and events under both low and normal illumination conditions. To achieve this, we utilize a robotic arm that traces a consistent **non-linear** trajectory to curate the dataset with spatial alignment precision under **0.03mm**. We then introduce a matching alignment strategy, rendering 90% of our dataset with errors less than **0.01s**. Based on the dataset, we propose a novel event-guided LIE approach, called **EvLight**, towards robust performance in real-world low-light scenes. Specifically, we first design the multi-scale holistic fusion branch to extract holistic structural and textural information from both events and images. To ensure robustness against variations in the regional illumination and noise, we then introduce a Signal-to-Noise-Ratio (SNR)-guided regional feature selection to selectively fuse features of images from regions with high SNR and enhance those with low SNR by extracting regional structure information from events. Extensive experiments on our dataset and the synthetic SDSD dataset demonstrate our EvLight significantly surpasses the frame-based methods,* e.g., *[4] by **1.14** dB and **2.62** dB, respectively.*

1. Introduction

Images captured under sub-optimal lighting conditions often exhibit various types of degradation such as poor visibility, noise, and inaccurate color [23]. For this reason, low-light image enhancement (LIE) serves as an essential task in ameliorating low-light image quality. LIE is crucial for downstream tasks, *e.g.*, face detection [27, 50] and nighttime semantic segmentation [29]. Recently, with the emergence of deep learning, abundant frame-based methods have been proposed, ranging from enhancing contrast [54], removing noise [47] to correcting color [38]. Although the performance has been remarkably boosted, these methods often suffer from unbalanced exposure and color distortion when the visual details, *e.g.*, edges, provided by frame-based cameras are less distinctive, as shown in Fig. 1 (c).

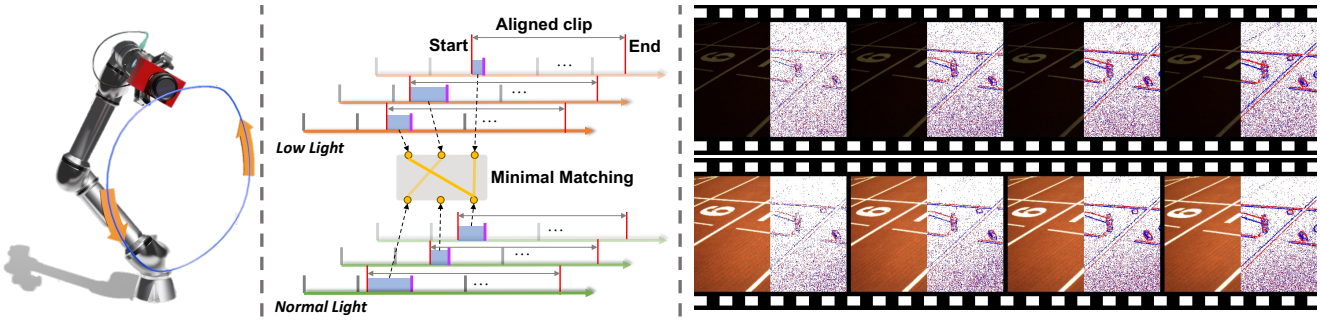


(a) Input low-light image (b) Input event
(c) Retinexformer [4] (d) Ours

Figure 1. A challenging example of our dataset containing an extremely low-light image (a) and sparse events (b). Compared with the result from a SOTA frame-based method Retinexformer [4] (c), our EvLight (d) not only recovers the structure details (*e.g.*, the pipe on the ceiling) but also avoids over-enhancement and saturation in the bright regions (*e.g.*, the lights).

Event cameras are bio-inspired sensors that generate event streams with high dynamic range (HDR), high temporal resolution, *etc*. [33, 55]. However, few research efforts have been made in combining both frame-based and event

*Corresponding author



(a) Data Capture (b) Temporal Alignment (c) Examples of SDE Dataset

Figure 2. (a) An illustration of collecting spatially-aligned image-event dataset by mounting a DAVIS 346 event camera on the robotic arm and recording the sequences with the same trajectory receptively. (b) An overview of our matching alignment strategy. (c) An example of our dataset with images and paired events captured in low-light (with an ND8 filter) and normal-light conditions.

cameras to address the LIE task [18, 24, 25, 52] to date. A hurdle is the prohibitive lack of large-scale real-world datasets with spatial-temporally aligned images and events. For example, [52] proposes an unsupervised framework without the need for paired event-image data, and [24, 25] leverage the synthetic datasets for training. Nonetheless, these methods are less competent for applications in real-world low-light scenarios. LIE dataset [18] is a real-world event-image dataset with paired low-light/normal-light sequences, obtained by simply adjusting indoor lamplight (artificial light fluctuations) and outdoor exposure time while maintaining a fixed camera position. Thus, similar to the previous frame-based dataset SMID [6], this dataset is only limited to static scenes.

In this paper, we propose a large-scale real-world dataset, named **SDE** dataset – containing over 30K pairs of spatio-temporally aligned images and events (see examples in Fig. 2 (c)) – captured under both low-light and normal-light conditions (Sec. 3). To construct such a dataset, the inherent difficulty stems from the complexities involved in ensuring precise spatial and temporal alignment between paired low-light and normal-light sequences, especially for dynamic scenes in nonlinear motion. To achieve this, we design a robotic alignment system to guarantee spatial alignment, where a DAVIS346 event camera [35] is mounted on a Universal UR5 robotic arm, see Fig. 2 (a). Our system shows a remarkable spatial accuracy with an error margin of merely 0.03mm, a significant improvement over the frame-based dataset, SDSD [39] with the error of 1mm. Moreover, unlike the setup of uniform linear motion in SDSD and the static scene in the LIE dataset [18], our system embraces non-linear motions with complex trajectories. This significantly enhances the diversity of our dataset for real-world scenarios. As for temporal alignment, a direct way to obtain aligned sequences is to clip them according to the specific motion start and end timestamps. However, even with the same camera and robot setting, the intervals (blue regions in Fig. 2 (b)) between motion start timestamps (left red line) and the timestamps of the initial frame (magenta line) in each clipped sequence are different, resulting in random temporal errors. To this end, we propose a novel matching alignment strategy to reduce the temporal discrepancies.

Buttressed by the dataset, we propose an event-guided LIE approach, called **EvLight**, towards the robust performance in real-world low-light scenes. The underlying premise is that – while low-light images deliver crucial color contents and events offer essential edge details – both modalities may be corrupted by different kinds of noise, yielding different noise distributions. Therefore, directly fusing the features of both modalities, as commonly done in [18], may also aggravate the noise in different regions of the two inputs, as shown in the blue box area in Fig. 5 (g).

To tackle these problems, our key idea is to fuse event and image features holistically, followed by a selective region-wise manner to extract the textural and structural information with the guidance of Signal-to-Noise-Ratio (SNR) prior information. To ensure robustness against variations in the regional illumination and noise, we further introduce an SNR-guided feature selection to extract features of images from regions with high SNR and those of events from regions with low SNR. This preserves the regional textural and structural information (Sec. 4.2). Then, we design an attention-based holistic fusion branch to coarsely extract holistic structural and textural information from both events and images (Sec. 4.3). Finally, a fusion block with channel attention is employed to fuse the holistic feature with the regional feature of images and events.

We conduct extensive experiments by comparing with the frame-based *e.g.*, [4] and event-guided *e.g.*, [25] methods on our real-world dataset and SDSD dataset (frame-based dataset) [39] with events generated from the event simulator [15]. The experiments show that our EvLight works decently for enhancing diverse underexposed images under extremely low-light conditions, as depicted in Fig. 1.

2. Related Work

LIE Datasets. The performance of learning-based methods heavily relies on the quality of the training datasets [10]

Dataset	Release	Dynamic Scene	With Ground Truth	Numbers
DVS-Dark [52]	✗	✓	✗	17,765
LIE [18]	✗	✗	✓	2,231
EvLowLight [24]	✗	✓	✗	—
Ours	✓	✓	✓	31,477

Table 1. A summary of existing real-world image-event datasets. Note that images in DVS-Dark are gray-scale.

for either images [3, 5, 7] or videos [6, 10, 17, 22, 39, 40]. For example, SDSD [39] obtains a pair of videos under various light conditions from a scene by mounting the camera on a mechatronic system. In this paper, we mainly focus on the event-image datasets. A summary of existing image-event datasets for low-light enhancement is shown in Tab. 1. EvLowLight [24] only includes low-light images/events without corresponding normal-light images/events as ground truth, while DVS-Dark [52] provides unpaired low-light/normal-light images/events. LIE [18] is a real-world image-event dataset, captured by adjusting the camera's light intake in a static scene, wherein events are triggered by the light changes (indoor) and exposure times (outdoor). In contrast, we present a real-world dataset with over $30K$ spatially and temporally aligned image-event pairs (both indoor and outdoor), using a robotic alignment system, considering the non-linear motion.

Frame-based LIE. Frame-based methods for low-light image enhancement can be divided into non-learning-based methods [1, 11, 12, 28, 46] and learning-based methods [4, 7, 9, 38, 41, 44, 45, 48, 49, 53, 54]. Non-learning-based methods typically rely on handcrafted features, such as histogram equalization [1, 28] and the Retinex theory [11, 12, 46]. Nonetheless, these methods lead to the absence of adaptivity and efficiency [44]. With the development of deep learning, an increasing number of learning-based methods have emerged, which can be bifurcated as Retinex-based methods [4, 7, 9, 44, 53, 54] and non-Retinex-based methods [38, 41, 45, 48, 49]. Specially, SNR-Aware [48] collectively exploits Signal-to-Noise-Ratio-aware transformers and convolutional models to dynamically enhance pixels with spatial-varying operations. However, these frame-based approaches often result in blurry outcomes and low Structural Similarity (SSIM) due to the buried edge in low-light images.

Event-based LIE. Event cameras enjoy HDR and provide rich edge information even under low-light scenes [55]. Zhang *et al*. [52] focuses on reconstructing grayscale images from low-light events but faces challenges in preserving original details using only brightness changes from events. Recently, some researchers have utilized events as guidance for low-light image enhancement [18, 19], low-light video enhancement [24, 25], and deblurring for low-light images [56]. ELIE [18] utilizes a residual fusion module to blend event and image for low light enhancement. Liu *et al.* [25] address artifacts in prior low-light video enhancement methods by synthesizing events from adjacent images for intensity and motion information, and propose a fusion transform module to fuse these event features with image features. EvLowLight [24] establishes temporal coherence by jointly estimating motion from both events and frames while ensuring the spatial coherence between events and frames with different spatial resolutions. However, these methods directly fuse features extracted from events and images without considering the discrepancy of the noise at the different local regions in events and images.

3. Our SDE Dataset

Capturing paired dynamic sequences from real-world scenes presents a formidable challenge, primarily attributed to the complexity involved in ensuring spatial and temporal alignment under varying illumination conditions. The first line of approaches employs a stereo camera system to simultaneously record the identical scenes, using non-linear transformations and cropping like DPED [16]. However, it struggles with SIFT keypoint computation and matching [26] in the low light. This hinders the identification of overlapped video segments. The second line of approaches [17, 22] constructs an optical system incorporating a beam splitter, allowing two cameras to capture a unified view. Nonetheless, achieving impeccable alignment with such systems remains challenging, resulting in spatial misalignments, as mentioned in [22, 24, 31]. The third line of approaches, *e.g*., SDSD [39] proposes a mechatronic system mounting the camera on an electric slide rail to capture low-light/normal-light videos separately (two rounds). However, SDSD is constrained by the limited *linear* motion of the electric slide rail. Differently, we design a robotic alignment system, equipped with an event camera to capture paired RGB images and events, under both low-light and normal-light conditions. Our system features the non-linear motions with complex trajectories.

1) Data Capture with Spatial Alignment. To ensure the spatial alignment of paired sequences, a robotic arm (Universal UR5), exhibiting a minimal repeated error of 0.03mm, is equipped to capture sequences following an identical trajectory. We set the robotic system with a predefined trajectory and a DAVIS 346 event camera with fixed parameters, *e.g*. exposure time. Firstly, paired image and event sequences are acquired under normal lighting conditions. Subsequently, an ND8 filter is applied to the camera lens, which facilitates the capture of low-light sequences while maintaining consistent camera parameters, such as exposure time and frame intervals.

2) Temporal Alignment of Low-light/Normal-light sequences. The alignment of SDSD [39] dataset involves a manual selection of the initial and final frames of each paired video, based on the motion states depicted in the videos, leading to inevitable bias. To mitigate this problem,

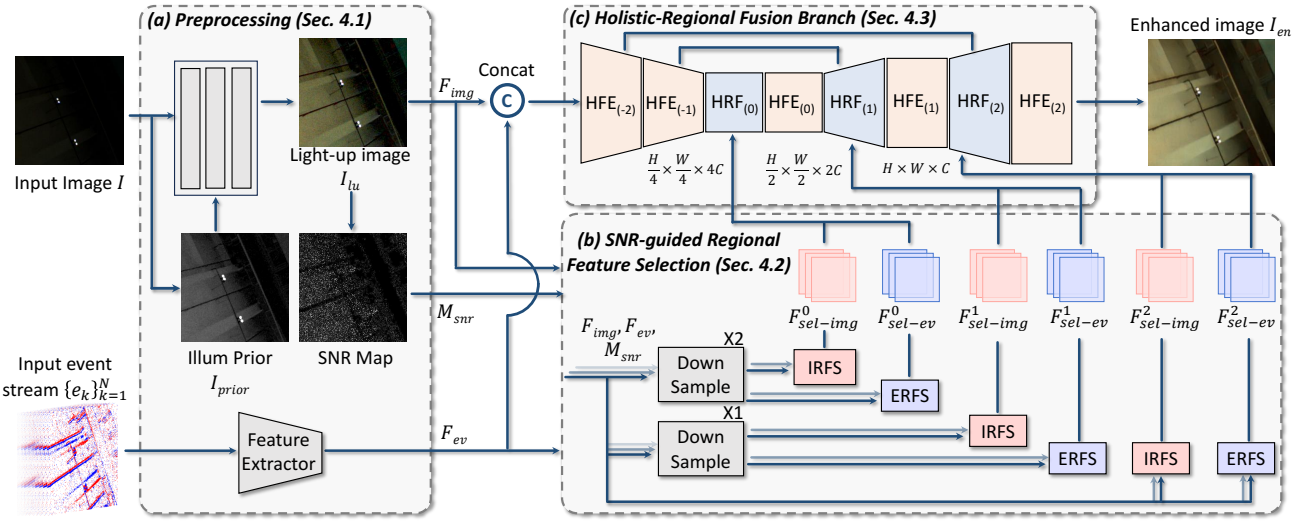


Figure 3. **An overview of our framework**. Our method consists of three parts, **(a)** Preprocessing (Sec. 4.1), **(b)** SNR-guided Regional Feature Selection (Sec. 4.2), and **(c)** Holistic-Regional Fusion Branch (Sec. 4.3). Specifically, SNR-guided Regional Feature Selection consists of two parts: Image-Regional Feature Selection (IRFS) and Event-Regional Feature Selection (ERFS). Additionally, Holistic-Regional Fusion Branch encompasses Holistic Feature Extraction (HFE) and Holistic-Regional Feature Fusion (HRF).

initial temporal alignment is performed by trimming the sequences based on the start and end timestamps of a predefined trajectory. However, even with consistent settings for exposure time and frame intervals, there exists a variable time interval between the start timestamp of the trajectory and the first frame timestamp captured post-initiation of the trajectory in each sequence. The bias causes the misalignment between each low-light image and its normal-light image pair, particularly in complex motion paths.

To achieve further alignment, we introduce a matching alignment strategy, wherein sequences from each scene are captured multiple times to minimize the alignment error to the largest extent, as shown in Fig. 2 (b). Practically, we capture 6 paired event-image sequences per scene —three in low-light and three in normal-light conditionals. These 6 sequences are trimmed to the predefined trajectory's start and end timestamps, ensuring uniform content across all videos. Subsequently, the time intervals between the trajectory's start timestamps and the initial frame timestamps of each trimmed sequence are calculated. As shown in Fig. 2 (b), the time intervals (blue regions) of 6 sequences are different, and we match the low-light sequence with the normal-light sequence, which has the minimal absolute errors of their time intervals; thus, we can reduce the misalignment caused by the random time interval. With the matching alignment strategy, we achieve a remarkable precision, with 90% of the datasets, exhibiting temporal alignment errors below 0.01s, and maximum errors of 0.013s and 0.027s for our indoor and outdoor datasets, respectively.

4. The Proposed EvLight Framework

Based on our SDE dataset, we further propose a novel event-guided LIE framework, called **EvLight**, as depicted in Fig. 3. Our goal is to selectively fuse the features of the image and events to achieve robust performance for event-guided LIE. EvLight takes the low-light image $\mathbf{I}$ and paired event stream $\{\mathbf{e}_k\}_{k=1}^N$ with N events as inputs and outputs an ehnanced image $\mathbf{I}_{en}$. Our pipeline consists of three components: **1)** Preprocessing, **2)** SNR-guided Regional Feature Selection, and **3)** Holistic-Regional Fusion Branch.

Event Representation. Given an event stream $\{\mathbf{e}_k\}_{k=1}^N$, we follow [30] to obtain the event voxel grid $\mathbf{E}$ by assigning the polarity of each event to the two closest voxels. The bin size is set to 32 in all the experiments.

4.1. Preprocessing

Initial Light-up. As demonstrated in recent frame-based LIE methods [4, 41, 49], coarsely enhancing the low-light image benefits the image restoration process and further boosts the performance. For simplicity, we follow Retinexformer [4] for the initial enhancement. As shown in the Fig. 3, we estimate the initial light-up image $\mathbf{I}_{lu}$ as:

$$\mathbf{I}_{lu} = \mathbf{I} \odot \mathbf{L}, \mathbf{L} = \mathcal{F}(\mathbf{I}, \mathbf{I}_{prior}), \tag{1}$$

where $\mathbf{I}_{prior} = max_c(\mathbf{I})$ denotes the illumination prior map, with max_c denoting the operation that computes the max values for each pixel across channels. $\mathcal{F}$ outputs the estimated illumination map $\mathbf{L}$, which is then applied to the input image $\mathbf{I}$ through a pixel-wise dot product.

The SNR Map. Following the previous approaches [2, 8, 48], we estimate the SNR map based on the initial light-up image $\mathbf{I}_{lu}$ and make it an effective prior for the SNR-guided regional feature selection in Sec. 4.2. Given the initial light-up image $\mathbf{I}_{lu}$, we first convert it into grayscale one $\mathbf{I}_g$, *i.e.*, $\mathbf{I}_g \in \mathbb{R}^{H\times W}$, followed by computing the SNR map $\mathbf{M}_{snr} = \tilde{\mathbf{I}}_g / abs(\mathbf{I}_g - \tilde{\mathbf{I}}_g)$, where $\tilde{\mathbf{I}}_g$ is the denoised coun-

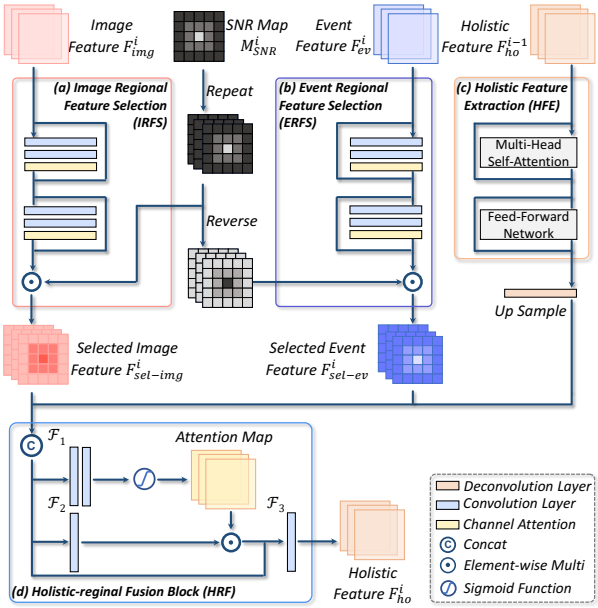


Figure 4. Details of each block in SNR-guided Regional Feature Selection and Holistic-Regional Fusion Branch's decoder.

terpart of $\mathbf{I}_g$. In practice, similar to SNR-Net [48], the denoised counterpart is calculated with the mean filter.

Feature Extraction. Image feature $\mathbf{F}_{img}$ of light-up image $\mathbf{I}_{lu}$ and event feature $\mathbf{F}_{ev}$ of the event voxel grid $\mathbf{E}$ are initially extracted with $conv3 \times 3$.

4.2. SNR-guided Regional Feature Selection

In this section, we aim to *selectively extract the regional features from either images or events*. We design an image-regional feature selection (IRFS) block to select image feature with higher SNR values, thereby obtaining image-regional feature, less affected by noise. However, SNR map assigns low SNR values to not only high-noise regions but also edge-rich regions. Consequently, solely extracting features from regions with high SNR values can inadvertently suppress edge-rich regions. To address this, we introduce an event-regional feature selection (ERFS) block for enhancing edges in areas with poor visibility and high noise.

As shown in Fig. 3, inputs of this module include the image feature $\mathbf{F}_{img}$, the event feature $\mathbf{F}_{ev}$, and the SNR map $\mathbf{M}_{snr}$. Firstly, the image feature $\mathbf{F}_{img}$ and event feature $\mathbf{F}_{ev}$ are down-sampled with $conv4 \times 4$ layers with the stride of 2 and SNR map $\mathbf{M}_{snr}$ undergoes an averaging pooling with the kernel size of 2. These donwsampling operations are represented as '*Down Sample*' in Fig. 3 and we attain different scale image feature $\mathbf{F}^i_{img} \in \mathbb{R}^{\frac{H}{2^{2-i}} \times \frac{W}{2^{2-i}} \times 2^{2-i}C}$, event feature $\mathbf{F}^i_{ev} \in \mathbb{R}^{\frac{H}{2^{2-i}} \times \frac{W}{2^{2-i}} \times 2^{2-i}C}$, and SNR map $\mathbf{M}^i_{snr} \in \mathbb{R}^{\frac{H}{2^{2-i}} \times \frac{W}{2^{2-i}}}$ where $i = 0, 1, 2$. Then, the image feature $\mathbf{F}^i_{img}$ and event feature $\mathbf{F}^i_{ev}$ are selected with the guidance of SNR map $\mathbf{M}^i_{snr}$ in IRFS block, and ERFS block. These two blocks then output the selected image features $\mathbf{F}^i_{sel-img}$ and event features $\mathbf{F}^i_{sel-ev}$, respectively. We now describe the details of these two blocks.

Image-Regional Feature Selection (IRFS) Block. As depicted in Fig. 4 (a), for an image feature $\mathbf{F}^i_{img}$, we initially process it through two residual blocks [13] to extract regional information and yield the output $\hat{\mathbf{F}}^i_{img}$. Each block comprises two $conv3 \times 3$ layers and an efficient channel attention layer [37]. The SNR map $\mathbf{M}^i_{snr}$ is then expanded along the channel to align with the image feature's channel dimensions. Then, we normalize it and make it within the range of $[0, 1]$. We then apply a predefined threshold on the SNR map to attain $\hat{\mathbf{M}}^i_{snr}$. To emphasize regions with higher SNR values and attain the selected image feature $\mathbf{F}^i_{sel-img}$, we perform an element-wise multiplication $\odot$ between the extended SNR map and the image feature $\hat{\mathbf{F}}^i_{img}$, formulated as:

$$\mathbf{F}^i_{sel-img} = \hat{\mathbf{M}}^i_{snr} \odot \hat{\mathbf{F}}^i_{img}. \tag{2}$$

Event-Regional Feature Selection (ERFS) Block. Edge-rich regions in the initial light-up image, particularly those underexposed, exhibit low SNR values. Additionally, we observe that events in high SNR regions (*e.g.*, well-illuminated smooth planes) are predominantly leak noise and shot noise events. Consequently, we design the ERFS block that utilizes the inverse of the SNR map to selectively enhance edges in low-visibility, high-noise areas, and to suppress noise events in sufficiently illuminated regions. The initial processing in this block follows a similar architecture to that used for the IRFS block, with $\mathbf{F}^i_{ev}$ as the input and $\hat{\mathbf{F}}^i_{ev}$ as the output. Given the SNR map $\hat{\mathbf{M}}^i_{snr}$, we obtain the reserve of SNR map $\bar{\mathbf{M}}^i_{snr}$ by $\mathbb{1}$ - $\hat{\mathbf{M}}^i_{snr}$. To obtain the selected event-regional feature $\mathbf{F}^i_{sel-ev}$, the element-wise multiplication product $\odot$ between the reserve of SNR map and the event feature is carried out, which is formulated as:

$$\mathbf{F}^i_{sel-ev} = \bar{\mathbf{M}}^i_{snr} \odot \hat{\mathbf{F}}^i_{ev}. \tag{3}$$

4.3. Holistic-Regional Fusion Branch

In this section, we aim to *extract the holistic features from both the event features and image features, so as to build up long-range channel-wise dependencies between them.* Besides, the holistic features are enhanced with the selected image-regional and event-regional features in the holistic-region feature fusion process.

Fig. 3 (c) depicts our holistic-regional fusion branch, which employs a UNet-like architecture [32] with the skip connections. This branch takes the concatenated feature of image $\mathbf{F}_{img}$ and event $\mathbf{F}_{ev}$ from the preprocessing stage (Sec. 4.1) as the input and the enhanced image $\mathbf{I}_{en}$ as the output. In the contracting path, there are 2 layers and the output of each layer is $\mathbf{F}^{i+1}_{ho} \in \mathbb{R}^{\frac{H}{2^{2-|i+1|}} \times \frac{W}{2^{2-|i+1|}} \times 2^{2-|i+1|}C}$ where $i = -2, -1$. In the i-th layer, the holistic feature $\mathbf{F}^i_{ho}$ first undergoes the holistic feature extraction (HFE) block. Then with a strided

Input	Method	SDE-in			SDE-out			SDSD-in			SDSD-out		
		PSNR↑	PSNR*↑	SSIM↑	PSNR↑	PSNR*↑	SSIM↑	PSNR↑	PSNR*↑	SSIM↑	PSNR↑	PSNR*↑	SSIM↑
Event Only	E2VID+ (ECCV'20) [34]	15.19	15.92	0.5891	15.01	16.02	0.5765	13.48	13.67	0.6494	16.58	17.27	0.6036
Image Only	SNR-Net (CVPR'22) [48]	20.05	21.89	0.6302	22.18	22.93	0.6611	24.74	25.30	0.8301	24.82	26.44	0.7401
	Uformer (CVPR'22) [43]	21.09	22.75	0.7524	22.32	23.57	0.7469	24.03	25.59	0.8999	24.08	25.89	0.8184
	LLFlow-L-SKF (CVPR'23) [45]	20.92	22.22	0.6610	21.68	23.41	0.6467	23.39	24.13	0.8180	20.39	24.73	0.6338
	Retinexformer (ICCV'23) [4]	21.30	23.78	0.6920	22.92	23.71	0.6834	25.90	25.97	0.8515	26.08	28.48	0.8150
Image+Event	ELIE (TMM'23) [18]	19.98	21.44	0.6168	20.69	23.12	0.6533	27.46	28.30	0.8793	23.29	28.26	0.7423
	eSL-Net (ECCV'20) [36]	21.25	23.19	0.7277	22.42	24.39	0.7187	24.99	25.72	0.8786	24.49	26.36	0.8031
	Liu *et al.* (AAAI'23) [25]	21.79	23.88	0.7051	22.35	23.89	0.6895	27.58	28.43	0.8879	23.51	27.63	0.7263
	Ours	**22.44**	**24.81**	**0.7697**	**23.21**	**25.60**	**0.7505**	**28.52**	**29.73**	**0.9125**	**26.67**	**30.30**	**0.8356**

Table 2. **Comparisons on our SDE dataset and SDSD [39] dataset**. The highest result is highlighted in **bold** while the second highest result is highlighted in underline. Since E2VID+ [34] can only reconstruct grayscale images, its metrics are calculated in grayscale.

$conv4 \times 4$ down-sampling operation, the holistic feature $\mathbf{F}_{ho}^{i+1}$ is obtained. In the expansive path, the output of each layer is $\mathbf{F}_{ho}^{i}$ where $i = 0, 1, 2$. As shown in Fig. 4, the holistic feature $\mathbf{F}_{ho}^{i-1}$ is processed with the HFE block and $\hat{\mathbf{F}}_{ho}^{i-1}$ is produced. Then, the holistic feature $\hat{\mathbf{F}}_{ho}^{i-1}$ is up-sampled with a strided $deconv2 \times 2$ and it is fused with the selected regional image $\mathbf{F}_{sel-img}^{i}$ and event features $\mathbf{F}_{sel-ev}^{i}$ in the holistic-regional fusion (HRF) block.

Holistic Feature Extraction (HFE) Block. As shown in Fig. 4 (c), holistic feature extraction is mainly composed of a multi-head self-attention module and a feed-forward network. Given a holistic feature $\mathbf{F}_{ho}^{i-1}$, the feature can be processed as:

$$\begin{aligned} \hat{\mathbf{F}}_{mid}^{i-1} &= \text{Attention}(\mathbf{F}_{ho}^{i-1}) + \mathbf{F}_{ho}^{i-1}, \\ \hat{\mathbf{F}}_{ho}^{i-1} &= \text{FFN}(\text{LN}(\hat{\mathbf{F}}_{mid}^{i-1})) + \hat{\mathbf{F}}_{mid}^{i-1}, \end{aligned} \tag{4}$$

where $\hat{\mathbf{F}}_{mid}^{i-1}$ is the middle output, LN is the layer normalization, FFN represents the feed-forward network, and Attention signifies the channel-wise self-attention, analogous to the multi-head attention mechanism employed in [51].

Holistic-Regional Fusion (HRF) Block. This block first concatenates the selected image features $\mathbf{F}_{sel-img}^{i}$, selected event features $\mathbf{F}_{sel-ev}^{i}$, and up-sampled holistic features $\hat{\mathbf{F}}_{ho}^{i-1}$. This concatenated feature $\mathbf{F}_{cat}^{i}$ is then passed through $conv3\times3$ layers to generate a spatial attention map. Sequentially, the element-wise multiplication is performed between the attention map and the concatenated features, which can be denoted as:

$$\mathbf{F}_{ho}^{i} = \mathcal{F}_3(\sigma(\mathcal{F}_1(\mathbf{F}_{cat}^{i})) \odot \mathcal{F}_2(\mathbf{F}_{cat}^{i}) + \mathbf{F}_{cat}^{i}), \tag{5}$$

where $\mathcal{F}_i$ is the convolution operation indicated in Fig. 4 (d). σ and $\odot$ denote the Sigmoid function and the element-wise production, respectively.

Optimization. The loss function $\mathcal{L}$ utilized for training is articulated as: $\mathcal{L} = \sqrt{||\mathbf{I}_{en} - \mathbf{I}_{gt}||^2 + \epsilon^2} + \lambda||\Phi(\mathbf{I}_{en}) - \Phi(\mathbf{I}_{gt})||_1$, where λ is a hyper-parameter, ϵ is set to 10^{-4}, $\mathbf{I}_{en}$ and $\mathbf{I}_{gt}$ denote the enhanced and ground truth images, and Φ represents feature extraction using the Alex network [21].

5. Experiments

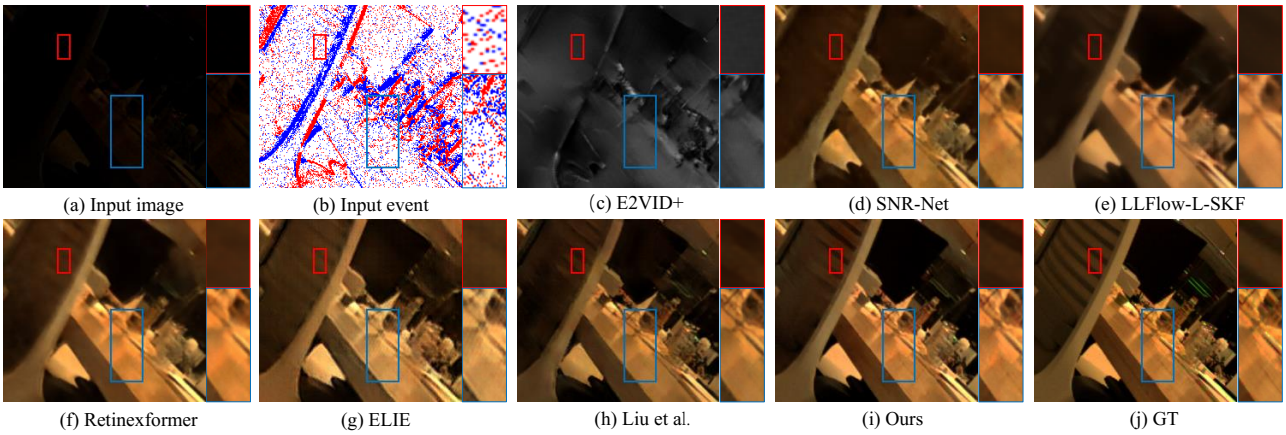
Implementation Details: We employ the Adam optimizer [20] for all experiments, with learning rates of $1e-4$ and $2e-4$ for SDE and SDSD datasets, respectively. Our framework is trained for 80 epochs with a batch size of 8 using an NVIDIA A30 GPU. We apply random cropping, horizontal flipping, and rotation for data augmentation. The cropping size is 256×256, and the rotation angles include 90, 180, and 270 degrees.

Evaluation Metrics: We use the peak-signal-to-noise ratio (PSNR) [14] and SSIM [42] for evaluation. Following the finetuning of the overall brightness of predicted results in previous methods [45, 53], we introduce the PSNR* as the metric to assess image restoration effectiveness beyond light fitting. The calculation of PSNR* is formulated as:

$$\begin{aligned} \text{PSNR*} &= \text{PSNR}(\mathbf{I}_{en} \times \mathbf{R}_{gt-en}, \mathbf{I}_{gt}), \\ \mathbf{R}_{gt-en} &= \text{Mean}(\text{Gray}(\mathbf{I}_{gt}))/\text{Mean}(\text{Gray}(\mathbf{I}_{en})), \end{aligned} \tag{6}$$

where $\mathbf{I}_{en}$, $\mathbf{I}_{gt}$, Gray, Mean, and PSNR represent the enhanced image, the ground-truth image, the operation of converting RGB images to grayscale ones, the operation of getting mean value, and the operation of calculating PSNR value, respectively.

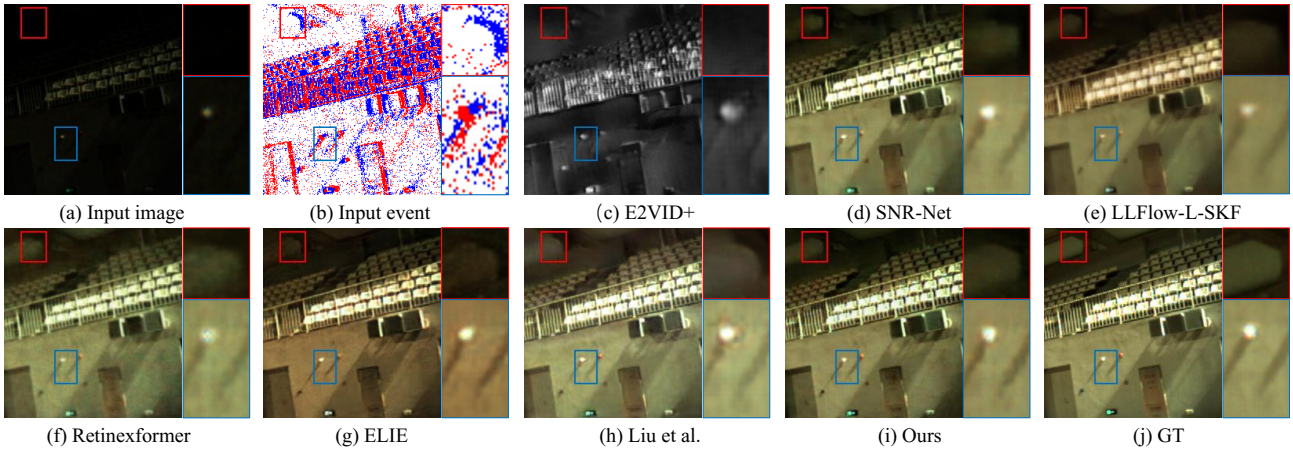
Datasets: **1) SED dataset** contains 91 image+event paired sequences (43 indoor and 48 outdoor sequences) captured with a DAVIS346 event camera [33] which outputs RGB images and events with the resolution of 346×260. For all collected sequences, 76 sequences are selected for training, and 15 sequences are for testing. **2) SDSD dataset [39]** provides paired low-light/normal-light videos with 1920×1080 resolution containing static and dynamic versions. We choose the dynamic version for simulating events and employ the same dataset split scheme as in SDSD [39]: 125 paired sequences for training and 25 paired sequences for testing. We first downsample the original videos to the same resolution (346×260) of the DAVIS346 event camera. Then, we input the resized images to the event simulator v2e [15] to synthesize event streams with noise under the default noisy model.



(a) Input image (b) Input event (c) E2VID+ (d) SNR-Net (e) LLFlow-L-SKF

(f) Retinexformer (g) ELIE (h) Liu et al. (i) Ours (j) GT

Figure 5. Qualitative results on our SDE-in dataset.



(a) Input image (b) Input event (c) E2VID+ (d) SNR-Net (e) LLFlow-L-SKF

(f) Retinexformer (g) ELIE (h) Liu et al. (i) Ours (j) GT

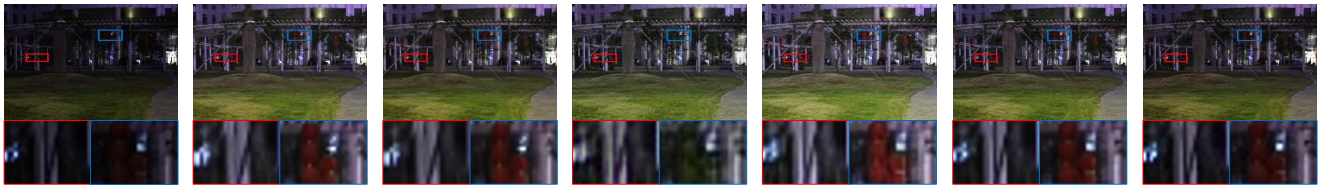
Figure 6. Qualitative results on our SDE-out dataset.

5.1. Comparison and Evaluation

We compare our method with recent methods with three different settings: **(I)** the experiment with events as input, including E2VID+ [34]. **(II)** the experiment with a RGB image as input, including SNR-Net [48], Uformer [43], LLFlow-L-SKF [45], and Retinexformer [4]. **(III)** the experiment with a RGB image and paired events as inputs, including ELIE [18], eSL-Net [36], and Liu *et al*. [25]. We reproduced ELIE [18] and Liu *et al*. [25] according to the descriptions in the papers, while the others are retrained with the released code. We replace the event synthesis module in [25] by inputting events captured with the event camera or generated from the event simulator [15].

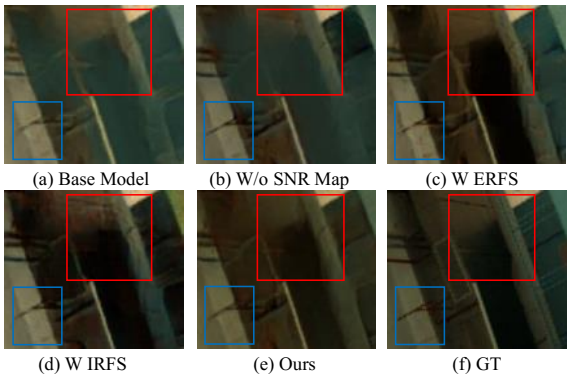
Comparison on our SDE Dataset: Quantitative results in Tab. 2 showcase our method's superior performance on the SDE dataset, outperforming baselines with higher PSNR by 0.65 dB for SDE-in and 0.29 dB for SDE-out. To assess image restoration effectiveness beyond light fitting, we computed PSNR* and our method also notably surpasses SOTA techniques, achieving higher PSNR* by 0.93 dB for SDE-in and 1.21 dB for SDE-out. This marks a significant validation of our approach for low-light image enhancement. Qualitatively, as depicted in Fig. 5 and Fig. 6 for indoor and outdoor scenes respectively, our method effectively reconstructs clear edges in dark areas (*e.g*., the red box areas in Fig. 5 and Fig. 6), surpassing frame-based methods like Retinexformer [4] and event-guided approaches such as Liu *et al*. [25]. Moreover, our method demonstrates less color distortion and noise on challenging regions (e.g., the wall in Fig. 6) than LLFlow-L-SKF [45] and ELIE [18], and Retinexformer [4], underscoring our method's robustness.

Comparison on the SDSD Dataset: To evaluate our method's generalization, we conducted comparisons on the SDSD dataset [39], with quantitative outcomes detailed in Tab. 2. Our method outperforms baselines significantly in PSNR, PSNR*, and SSIM, leading by more than 0.94 dB for SDSD-in and 0.59 dB for SDSD-out. Although ELIE and Liu *et al*. [25] surpass frame-based methods in SDSD-in dataset, they suffer from the overfitting in SDSD-out dataset which is demonstrated by the substantial disparity between PSNR and PSNR*. Qualitatively, as shown in Fig. 7, our method effectively restores underexposed images to more detailed structures, as highlighted in the red box area. Moreover, ELIE [18] tends to produce color distortions, as visible in the blue box area of Fig. 7 (d).



(a) Input image (b) SNR-Net (c) Retinexformer (d) ELIE (e) Liu et al. (f) Ours (g) GT

Figure 7. Qualitative results on SDSD dataset [39].



(a) Base Model (b) W/o SNR Map (c) W ERFS

(d) W IRFS (e) Ours (f) GT

Figure 8. Visualization of ablation results.

	Regional Feature Selection	SNR-guided	PSNR	SSIM
1	✗	✗	21.58	0.7001
2	✓	✗	21.86	0.7490
3	✓	✓	**22.44**	**0.7697**

Table 3. Ablation of SNR-guided regional feature selection.

	IRFS	ERFS	PSNR	SSIM
1	✗	✗	21.58	0.7001
2	✓	✗	21.92	0.7108
3	✗	✓	22.18	0.7525
4	✓	✓	**22.44**	**0.7697**

Table 4. Impact of each module of SNR-guided regional feature selection.

5.2. Ablation Studies and Analysis

We conduct ablation studies on SDE-in dataset to assess the effectiveness of each module of our method. The basic implementation, without SNR-guided regional feature selection as described in Sec. 4.2, is called the *Base* model.

Impact of Events: To reveal the impact of events, we conduct experiments on the *Base* model. The variant excluding events attains a PSNR of 21.35 dB and an SSIM of 0.6985, whereas adding events results in a 0.23 dB improvement in PSNR and a 0.002 improvement in SSIM. However, the *Base* model cannot fully explore the potential of events demonstrated by the limited improvement in SSIM.

Impact of SNR-guided regional feature selection: To verify it, we conduct an ablation study in Tab. 3. We replace the SNR map with an all-ones matrix and remove the whole selection module (the *Base* model). Compared with the *Base* model (1st row), regional feature selection with an all-ones matrix (2nd row) and SNR-guided regional feature selection (3rd row) yield 0.28 dB and 0.86 dB increase in PSNR, respectively, demonstrating the necessity of regional features and the SNR map. Although regional feature selection with an all-ones matrix and *Base* model both have color distortion (*e.g.*, the red box in Fig. 8 (a), (b)), (b) has better structure details than (a).

Impact of IRFS and ERFS: To verify them, we conduct an ablation study in Tab. 4. Compared with the *Base* model (1st row), image-regional feature selection (IRFS, 2nd row), event-regional feature selection (ERFS, 3rd row), and the combination of them (4th row) yields the 0.34 dB, 0.60 dB, and 0.86 dB increase in PSNR, respectively, demonstrating the necessity of the IRFS and ERFS block. As shown in Fig. 8, IRFS (d) or ERFS (c) can reduce the color distortion that appears in the *Base* model (a). With both IRFS and ERFS blocks, our results deliver the best visual quality (*e.g.*, red box and blue box in Fig. 8).

Generalization Ability: To assess the generalization capability of our EvLight, we carry out an experiment on the CED [33] and MVSEC [57] with the model trained on our SDE dataset. Moreover, we use the model, trained on the synthetic events from the SDSD dataset [39] to evaluate the generalization capacity on real events of our SDE dataset. *Detailed visual results are available in Suppl. Mat.*

6. Conclusion

This paper presented a large-scale real-world event-image dataset, called SDE, curated via a non-linear robotic path for high-fidelity spatial and temporal alignment, encompassing low and normal illumination conditions. Based on the real-world dataset, we designed a framework, EvLight, towards robust event-guided low-light image enhancement, which adaptively fuse the event and image features in a holistic and region-wised manner resulting in robust and superior performance. **Limitations and Future Work:** Due to inherent limitations of DAVIS346 event cameras, RGB images in our SDE dataset may exhibit partial chromatic aberrations and the moiré pattern. In the future, we will improve our hardware system to enable synchronous triggering of robots and event cameras, thereby significantly reducing labor costs associated with repetitive collection.

Acknowledgment. This paper is supported by the National Natural Science Foundation of China (NSF) under Grant No. NSFC22FYT45 and the Guangzhou City, University and Enterprise Joint Fund under Grant No.SL2022A03J01278.

References

[1] Tarik Arici, Salih Dikbas, and Yucel Altunbasak. A histogram modification framework and its application for image contrast enhancement. *IEEE Transactions on image processing*, 18(9):1921–1935, 2009. 3

[2] Antoni Buades, Bartomeu Coll, and J-M Morel. A non-local algorithm for image denoising. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, pages 60–65. Ieee, 2005. 4

[3] Vladimir Bychkovsky, Sylvain Paris, Eric Chan, and Frédo Durand. Learning photographic global tonal adjustment with a database of input/output image pairs. In *CVPR 2011*, pages 97–104. IEEE, 2011. 3

[4] Yuanhao Cai, Hao Bian, Jing Lin, Haoqian Wang, Radu Timofte, and Yulun Zhang. Retinexformer: One-stage retinex-based transformer for low-light image enhancement. *arXiv preprint arXiv:2303.06705*, 2023. 1, 2, 3, 4, 6, 7

[5] Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. Learning to see in the dark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3291–3300, 2018. 3

[6] Chen Chen, Qifeng Chen, Minh N Do, and Vladlen Koltun. Seeing motion in the dark. In *Proceedings of the IEEE/CVF International conference on computer vision*, pages 3185–3194, 2019. 2, 3

[7] Wenhan Yang Jiaying Liu Chen Wei, Wenjing Wang. Deep retinex decomposition for low-light enhancement. In *British Machine Vision Conference*, 2018. 3

[8] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. Image denoising with block-matching and 3d filtering. In *Image processing: algorithms and systems, neural networks, and machine learning*, pages 354–365. SPIE, 2006. 4

[9] Huiyuan Fu, Wenkai Zheng, Xiangyu Meng, Xin Wang, Chuanming Wang, and Huadong Ma. You do not need additional priors or regularizers in retinex-based low-light image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18125–18134, 2023. 3

[10] Huiyuan Fu, Wenkai Zheng, Xicong Wang, Jiaxuan Wang, Heng Zhang, and Huadong Ma. Dancing in the dark: A benchmark towards general low-light video enhancement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12877–12886, 2023. 2, 3

[11] Xueyang Fu, Delu Zeng, Yue Huang, Xiao-Ping Zhang, and Xinghao Ding. A weighted variational model for simultaneous reflectance and illumination estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2782–2790, 2016. 3

[12] Xiaojie Guo, Yu Li, and Haibin Ling. Lime: Low-light image enhancement via illumination map estimation. *IEEE Transactions on image processing*, 26(2):982–993, 2016. 3

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[14] Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th international conference on pattern recognition*, pages 2366–2369. IEEE, 2010. 6

[15] Yuhuang Hu, Shih-Chii Liu, and Tobi Delbruck. v2e: From video frames to realistic dvs events. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1312–1321, 2021. 2, 6, 7

[16] Andrey Ignatov, Nikolay Kobyshev, Radu Timofte, Kenneth Vanhoey, and Luc Van Gool. Dslr-quality photos on mobile devices with deep convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 3277–3285, 2017. 3

[17] Haiyang Jiang and Yinqiang Zheng. Learning to see moving objects in the dark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7324–7333, 2019. 3

[18] Yu Jiang, Yuehang Wang, Siqi Li, Yongji Zhang, Minghao Zhao, and Yue Gao. Event-based low-illumination image enhancement. *IEEE Transactions on Multimedia*, 2023. 2, 3, 6, 7

[19] Haiyan Jin, Qiaobin Wang, Haonan Su, and Zhaolin Xiao. Event-guided low light image enhancement via a dual branch gan. *Journal of Visual Communication and Image Representation*, 95:103887, 2023. 3

[20] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. 6

[22] Sohyun Lee, Jaesung Rim, Boseung Jeong, Geonu Kim, Byungju Woo, Haechan Lee, Sunghyun Cho, and Suha Kwak. Human pose estimation in extremely low-light conditions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 704–714, 2023. 3

[23] Chongyi Li, Chunle Guo, Linghao Han, Jun Jiang, Ming-Ming Cheng, Jinwei Gu, and Chen Change Loy. Low-light image and video enhancement using deep learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(12):9396–9416, 2021. 1

[24] Jinxiu Liang, Yixin Yang, Boyu Li, Peiqi Duan, Yong Xu, and Boxin Shi. Coherent event guided low-light video enhancement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10615–10625, 2023. 2, 3

[25] Lin Liu, Junfeng An, Jianzhuang Liu, Shanxin Yuan, Xiangyu Chen, Wengang Zhou, Houqiang Li, Yan Feng Wang, and Qi Tian. Low-light video enhancement with synthetic event guidance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1692–1700, 2023. 2, 3, 6, 7

[26] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60:91–110, 2004. 3

[27] Long Ma, Tengyu Ma, Risheng Liu, Xin Fan, and Zhongxuan Luo. Toward fast, flexible, and robust low-light image

enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5637–5646, 2022. 1

[28] Keita Nakai, Yoshikatsu Hoshi, and Akira Taguchi. Color image contrast enhacement method based on differential intensity/saturation gray-levels histograms. In *2013 International Symposium on Intelligent Signal Processing and Communication Systems*, pages 445–449. IEEE, 2013. 3

[29] Jingyi Pan, Sihang Li, Yucheng Chen, Jinjing Zhu, and Lin Wang. Towards dynamic and small objects refinement for unsupervised domain adaptative nighttime semantic segmentation. *arXiv preprint arXiv:2310.04747*, 2023. 1

[30] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. Events-to-video: Bringing modern computer vision to event cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3857–3866, 2019. 4

[31] Jaesung Rim, Haeyun Lee, Jucheol Won, and Sunghyun Cho. Real-world blur dataset for learning and benchmarking deblurring algorithms. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*, pages 184–201. Springer, 2020. 3

[32] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 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*, pages 234–241. Springer, 2015. 5

[33] Cedric Scheerlinck, Henri Rebecq, Timo Stoffregen, Nick Barnes, Robert Mahony, and Davide Scaramuzza. Ced: Color event camera dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 1, 6, 8

[34] Timo Stoffregen, Cedric Scheerlinck, Davide Scaramuzza, Tom Drummond, Nick Barnes, Lindsay Kleeman, and Robert Mahony. Reducing the sim-to-real gap for event cameras. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII 16*, pages 534–549. Springer, 2020. 6, 7

[35] Gemma Taverni, Diederik Paul Moeys, Chenghan Li, Celso Cavaco, Vasyl Motsnyi, David San Segundo Bello, and Tobi Delbruck. Front and back illuminated dynamic and active pixel vision sensors comparison. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 65(5):677–681, 2018. 2

[36] Bishan Wang, Jingwei He, Lei Yu, Gui-Song Xia, and Wen Yang. Event enhanced high-quality image recovery. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, pages 155–171. Springer, 2020. 6, 7

[37] Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. Eca-net: Efficient channel attention for deep convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11534–11542, 2020. 5

[38] Ruixing Wang, Qing Zhang, Chi-Wing Fu, Xiaoyong Shen, Wei-Shi Zheng, and Jiaya Jia. Underexposed photo enhancement using deep illumination estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6849–6857, 2019. 1, 3

[39] Ruixing Wang, Xiaogang Xu, Chi-Wing Fu, Jiangbo Lu, Bei Yu, and Jiaya Jia. Seeing dynamic scene in the dark: A high-quality video dataset with mechatronic alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9700–9709, 2021. 2, 3, 6, 7, 8

[40] Wei Wang, Xin Chen, Cheng Yang, Xiang Li, Xuemei Hu, and Tao Yue. Enhancing low light videos by exploring high sensitivity camera noise. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4111–4119, 2019. 3

[41] Yinglong Wang, Zhen Liu, Jianzhuang Liu, Songcen Xu, and Shuaicheng Liu. Low-light image enhancement with illumination-aware gamma correction and complete image modelling network. *arXiv preprint arXiv:2308.08220*, 2023. 3, 4

[42] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 6

[43] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17683–17693, 2022. 6, 7

[44] Wenhui Wu, Jian Weng, Pingping Zhang, Xu Wang, Wenhan Yang, and Jianmin Jiang. Uretinex-net: Retinex-based deep unfolding network for low-light image enhancement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5901–5910, 2022. 3

[45] Yuhui Wu, Chen Pan, Guoqing Wang, Yang Yang, Jiwei Wei, Chongyi Li, and Heng Tao Shen. Learning semantic-aware knowledge guidance for low-light image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1662–1671, 2023. 3, 6, 7

[46] Jun Xu, Yingkun Hou, Dongwei Ren, Li Liu, Fan Zhu, Mengyang Yu, Haoqian Wang, and Ling Shao. Star: A structure and texture aware retinex model. *IEEE Transactions on Image Processing*, 29:5022–5037, 2020. 3

[47] Ke Xu, Xin Yang, Baocai Yin, and Rynson WH Lau. Learning to restore low-light images via decomposition-and-enhancement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2281–2290, 2020. 1

[48] Xiaogang Xu, Ruixing Wang, Chi-Wing Fu, and Jiaya Jia. Snr-aware low-light image enhancement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17714–17724, 2022. 3, 4, 5, 6, 7

[49] Xiaogang Xu, Ruixing Wang, and Jiangbo Lu. Low-light image enhancement via structure modeling and guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9893–9903, 2023. 3, 4

[50] Jun Yu, Xinlong Hao, and Peng He. Single-stage face detection under extremely low-light conditions. In *Proceedings*

of the IEEE/CVF International Conference on Computer Vision, pages 3523–3532, 2021. 1

[51] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5728–5739, 2022. 6

[52] Song Zhang, Yu Zhang, Zhe Jiang, Dongqing Zou, Jimmy Ren, and Bin Zhou. Learning to see in the dark with events. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*, pages 666–682. Springer, 2020. 2, 3

[53] Yonghua Zhang, Jiawan Zhang, and Xiaojie Guo. Kindling the darkness: A practical low-light image enhancer. In *Proceedings of the 27th ACM international conference on multimedia*, pages 1632–1640, 2019. 3, 6

[54] Yonghua Zhang, Xiaojie Guo, Jiayi Ma, Wei Liu, and Jiawan Zhang. Beyond brightening low-light images. *International Journal of Computer Vision*, 129:1013–1037, 2021. 1, 3

[55] Xu Zheng, Yexin Liu, Yunfan Lu, Tongyan Hua, Tianbo Pan, Weiming Zhang, Dacheng Tao, and Lin Wang. Deep learning for event-based vision: A comprehensive survey and benchmarks. *arXiv preprint arXiv:2302.08890*, 2023. 1, 3

[56] Chu Zhou, Minggui Teng, Jin Han, Chao Xu, and Boxin Shi. Delieve-net: Deblurring low-light images with light streaks and local events. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1155–1164, 2021. 3

[57] Alex Zihao Zhu, Dinesh Thakur, Tolga Özaslan, Bernd Pfrommer, Vijay Kumar, and Kostas Daniilidis. The multivehicle stereo event camera dataset: An event camera dataset for 3d perception. *IEEE Robotics and Automation Letters*, 3 (3):2032–2039, 2018. 8