Coherent Event Guided Low-Light Video Enhancement

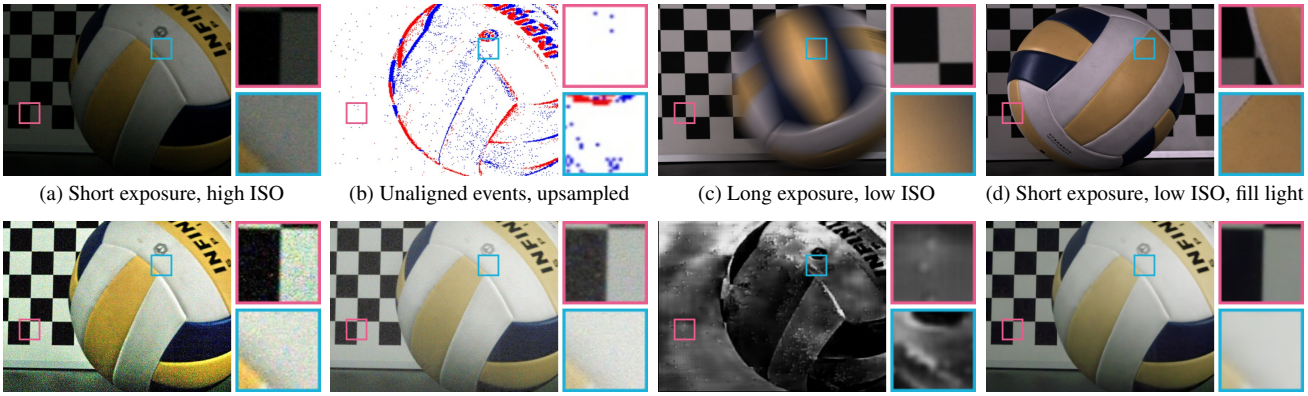
Jinxiu Liang[1,2] Yixin Yang[1,2] Boyu Li[1,2] Peiqi Duan[1,2] Yong Xu[3] Boxin Shi*[1,2]

[1] National Key Laboratory for Multimedia Information Processing, School of Computer Science, Peking University

[2] National Engineering Research Center of Visual Technology, School of Computer Science, Peking University

[3] School of Computer Science and Engineering, South China University of Technology

{cssherryliang, yangyixin93, liboyu, duanqi0001, shiboxin}@pku.edu.cn yxu@scut.edu.cn



(a) Short exposure, high ISO (b) Unaligned events, upsampled (c) Long exposure, low ISO (d) Short exposure, low ISO, fill light

(e) Result of [12] from only (a) (f) Result of [63] from only (a) (g) Result of [64] from only (b) (h) Our result from (a) and (b)

Figure 1. Comparison among different camera settings for capturing fast-moving scenes and different methods (conventional image enhancement [12], deep-learning-based video enhancement [63], event-to-video reconstruction [64], and the proposed event guided video enhancement) for low-light dynamic imaging. (a) and (b) are the inputs of the proposed method. Please refer to our project page[1] for the corresponding animations.

Abstract

With frame-based cameras, capturing fast-moving scenes without suffering from blur often comes at the cost of low SNR and low contrast. Worse still, the photometric constancy that enhancement techniques heavily relied on is fragile for frames with short exposure. Event cameras can record brightness changes at an extremely high temporal resolution. For low-light videos, event data are not only suitable to help capture temporal correspondences but also provide alternative observations in the form of intensity ratios between consecutive frames and exposure-invariant information. Motivated by this, we propose a low-light video enhancement method with hybrid inputs of events and frames. Specifically, a neural network is trained to establish spatiotemporal coherence between visual signals with different modalities and resolutions by constructing correlation volume across space and time. Experimental results on synthetic and real data demonstrate the superiority of the proposed method compared to the state-of-the-art methods.

*Corresponding author

[1] https://sherrycattt.github.io/EvLowLight

1. Introduction

Capturing fast-moving scenes without introducing blurry artifacts (Figure 1 (c)) is challenging, especially in an environment with insufficient illumination. A fast shutter speed helps to freeze motion (Figure 1 (a)), but it also causes excessive noise and low contrast. A popular choice in professional photography, such as sports recording and filmmaking, is to place large fill lights to allow sufficient illumination (Figure 1 (d)), however, they are limited in their portability and power requirements in uncontrolled environments.

Video enhancement aims at improving the degraded quality, in which the key is to exploit the temporal coherence [27, 59]. Its performance heavily depends on the quality of image-based optical flow estimation that assumes spatiotemporally small translation and brightness constancy; however, it becomes fragile for low-light frames whose features such as edges are less distinctive and are contaminated by noise (Figure 1 (e)). Despite recent advances in learning-based low-light video enhancement methods [63, 49, 20, 30, 5], it remains challenging to improve the quality of frames capturing fast-moving scenes where the pixel displacements are large (Figure 1 (f)).

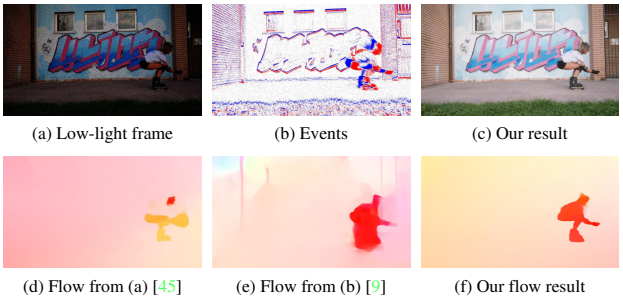
Event cameras asynchronously record logarithmic brightness changes with a high dynamic range, low latency, and low power cost [39], which raise promising directions for low-light imaging with events [64, 67]. Their unique advantage of high temporal resolution in the order of microseconds benefits motion estimation (Figure 2), providing reliable temporal coherence between frames in fast-moving scenes for video enhancement. They are free of the notion of exposure time and thus do not suffer from the well-known trade-off between strong blur using long exposure and low SNR using short exposure, which remains hard to be handled by existing deblurring or multi-exposure fusion methods. In this paper, we propose to utilize the high temporal resolution and high dynamic range information from events to guide low-light video enhancement. Specifically, motion is jointly estimated from events and frames for capturing *temporal coherence* as guidance to warp and integrate multimodal observations according to *the same scene points* for noise reduction.

However, it is non-trivial due to three types of misalignment: *i) Modality.* Events *asynchronously* record brightness changes, while frames *synchronously* record absolute brightness. They are inherently different modalities, whose gap is further increased by noise from their mechanisms in low-light conditions, making their translation [64] or fusion [37, 36] hard (Figure 1 (g)). *ii) Sensor.* Hybrid sensors with precisely aligned events and frames have low spatial resolutions (*e.g.*, 346 $\times$ 260 [14]). Although hybrid camera systems [48, 47] have high-resolution frames (*e.g.*, 2448 $\times$ 2048 [67]), online registration for each scene (with different depths and arbitrary lighting) that is important in obtaining stable results becomes fragile in low-light conditions since features for matching are too weak to be precisely extracted (Figure 1 (a) *vs.* (b)). *iii) Temporal resolution.* Pixels corresponding to the same scene points in events and frames are recorded in different temporal resolutions, between which the established correspondences should be robust to inaccuracy in motion estimation caused by limited photons.

To overcome the above challenges, the paper proposes to establish the spatiotemporal coherence between events and frames by following strategies:

- A *multimodal coherence modeling* module that establishes multiscale all-pair coherence between events and frames in the feature space to compensate for modality misalignment in sensing ability and the sensor-level misalignment under low-light conditions.
- A *temporal coherence propagation* module that samples features of consecutive events and frames corresponding to the same scene point for realizing coherence-aware aggregation in the local displacement field to improve the SNR of the reconstructed video.

The modules newly designed above extract *complementary information from events and frames*, enable a *misalignment-robust hybrid-imaging setting*, and integrate information across time for *effective denoising*. They compose our contribution to *the first event guided low-light video enhancement method* that captures high-quality videos in fast-moving scenes with short exposure. The superior performances against state-of-the-art methods for synthetic and real data make it potentially useful as an alternative to fill lights for low-light photography.



(a) Low-light frame (b) Events (c) Our result

(d) Flow from (a) [45] (e) Flow from (b) [9] (f) Our flow result

Figure 2. Comparison of optical flow estimation. Optical flow (d) and (e) are estimated from (a) low-light frames and (b) events by using the state-of-the-art methods [45] and [9], respectively, while (f) is estimated by the proposed joint optical flow estimation module from both (a) and (b).

2. Related Work

Low-light image enhancement. The most straightforward method to enhance low-light videos is to apply the low-light image enhancement method in a frame-by-frame manner. In recent years, deep-learning-based methods have achieved impressive results [29, 6]. Lore *et al*. [29] trained a stacked denoising auto-encoder to fit the desired intensity mapping. Chen *et al*. [6] used a U-Net to learn the mapping, with a focus on raw input instead of RGB. Cai *et al*. [2] and Xu *et al*. [57] proposed learning the mapping from low-light image to low/high-frequency parts of the normal-light reference sequentially. Xu *et al*. [58] introduced a transformer for SNR-aware low-light image enhancement. Some works attempt to relax the prerequisite on paired training images for deep-learning-based low-light image enhancement [21, 28]. Several methods [42, 52, 66, 50, 60, 54, 31] applied the Retinex decomposition model to incorporate more priors to alleviate ill-posedness.

Low-light video enhancement. Low-light video enhancement is a more difficult problem than image tasks. Chen *et al*. [5] proposed a paired low-light video dataset by capturing short- and long-exposure video pairs for static scenes. Lv *et al*. [30] replaced the commonly used 2D convolution layers for images with 3D layers for videos. Wang *et al*. [49] proposed a high-quality low-light video dataset in which the paired low-light and normal-light videos are obtained by mechatronic alignment of the cameras. The above methods focus on dynamic scenes where motion is caused by camera movement. Zhang *et al*. [63] focused on enhancing low-light videos where motion is caused by moving objects. Due to the difficulty in extracting distinct features for motion es-

timation, exploration of temporal redundancy in low-light videos for better noise reduction is still limited.

Event guided video enhancement. Event cameras have distinctive advantages in sensing scenes in high dynamic range with high temporal resolution and low power consumption [8]. Some works proposed directly reconstructing video from events by optimization [33], supervised [39, 40, 53, 43] and unsupervised [38, 64] learning. However, it is highly ill-posed to estimate the absolute intensity values in video from only brightness changes recorded in events. A more promising solution is to integrate events as brightness increments into the absolute intensity values of its neighboring frame by event double integral model [37, 36] and its deep-learning-based extensions for deblurring [22, 23, 41, 44, 56, 46], frame interpolation [61], and both [26, 65]. These works are designed for hybrid imaging sensors with events and frames aligned pixelwise, which have low spatial resolution partially due to consideration of data transmission efficiency [14]. Recently, hybrid camera systems with an event camera and a frame camera have been adopted for signal processing [51, 7, 1], video frame interpolation [48, 17, 47], high dynamic range imaging [15, 13], rolling shutter correction [69], pose estimation [62], and deblurring low-light images [67]. Currently, the question of how to utilize events for both noise reduction and exposure compensation for video enhancement remains unsolved.

3. Method

3.1. Physical Formulation of Events and Frames

The relation between a low-light frame L and its ground truth I is formulated as:

$$L = f_{\text{CRF}}(f_{\text{CRF}}^{-1}(I) \odot P) + N, \tag{1}$$

where f_{CRF} denotes the underlying camera response function that maps scene radiance in the linear domain into intensity in the nonlinear domain for better human perception, $\odot$ denotes the Hadamard multiplication, N is the zero mean noise, and P is the exposure parameter determined by camera setting, *e.g.*, exposure time $\Delta \cdot I$.

An event $e = (t, p, \sigma)$ at the pixel $p = (p_x, p_y)^\top$ and time t is triggered whenever the logarithmic change of irradiance R exceeds a pre-defined threshold θ (> 0),

$$\|\log R_t^{(p)} - \log R_{t-\delta t}^{(p)}\| \geq \theta, \tag{2}$$

where R_t denotes the instantaneous intensity at time t, and the polarity $\sigma \in \{-1, +1\}$ indicates {negative, positive} brightness changes. Similar to previous works, we use the representation of the 3D voxel grid for events[70]. By discretizing duration $\Delta t = t_{K-1} - t_0$ spanned by K events into B temporal bins, each event $e_k = (t_k, p_k, \sigma_k)$ distributes its polarity σ_k to the two closest voxels as follows [39]:

$$E_t^{(p)} = \sum_{p_k = p} \sigma_k \max(0, 1 - |t - \tilde{t}_k|), \tag{3}$$

where $\tilde{t}_k := \frac{B-1}{\Delta t}(t_k - t_0)$ is the normalized timestamp.

Given two consecutive frames L_{t_i}, L_{t_j}, the events $E_{t_i \to t_j}$ triggered within the period from t_i to t_j can be integrated as intensity increments, whose relationship is:

$$L_{t_j}^{(p)} = L_{t_i}^{(p)} \exp(\theta \int E_{t_i \to t_j}^{(p)} dt). \tag{4}$$

3.2. Method Overview

In this paper, our objective is to recover a frame I_t with reduced noise and increased contrast from neighboring low-light frames $\{L_i\}_{i=t-N}^{t+N}$ and their corresponding events $E_{t-N \to t+N}$ in the form of voxel grids in Eq. (3) with B bins, where N is the temporal radius. Following [48, 47], we consider a hybrid camera system in which the two cameras are not precisely aligned. Additionally, we consider a situation where the event camera has a *lower spatial resolution* compared to the frame camera. Without losing generality, we describe how to generate a high-quality image I_t from multimodal observations of frames L_t, L_{t+1} and the corresponding events $E_{t \to t+1}$.

The overall method is shown in Figure 3. First, two modal-specific feature encoders $\mathcal{F}_{\text{enc}}^L$ and $\mathcal{F}_{\text{enc}}^E$ extract features $\phi_t^L, \phi_{t \to t+1}^E$ from frames L_t and events $E_{t \to t+1}$, respectively. The multimodal coherence C^{modal} is estimated for global and local alignments. The global coherence to coarsely align $\phi_{t \to t+1}^E$ into ϕ_t^L is fit to compensate for the resolution gap and possible sensor misalignment. Pixelwise correspondences are modeled to fuse complementary information from optical flows $S_{t \to t+1}^E, S_{t \to t+1}^L$ estimated from events and frames, respectively.

With the temporal coherence C^{temp} estimated from consecutive frames, events and frames according to the same scene points across time are warped into the middle frame based on jointly estimated optical flows $S_{t \to t+1}$. They are then integrated into latent frames according to Eq. (4) implicitly and propagated temporally for noise reduction. The noise-reduced features ϕ_t^I are then used to reconstruct a denoised frame $\widetilde{L}_t$ by a decoder $\mathcal{F}_{\text{dec}}$. Finally, the frame I_t can be generated from an exposure parameter map P_t predicted by network $\mathcal{F}_{\text{exp}}$ by inverting Eq. (1):

$$I_t = f_{\text{CRF}}(f_{\text{CRF}}^{-1}(\widetilde{L}_t) \odot P_t^{-1}). \tag{5}$$

3.3. Multimodal Coherence Modeling

Multimodal coherence estimation. We propose to establish coherence between L_t and $E_{t \to t+1}$ across all the spatial support within the same duration, whose key component is a learned 4D correlation volume that models all-pair correspondence between the events and frames. First, feature maps $\varphi_t^L, \varphi_{t \to t+1}^E$ with different spatial resolutions for capturing visual similarity between events $E_{t \to t+1}$ and frames L_t are extracted by the feature extractors of the unimodal optical flow estimators $\mathcal{F}_{\text{flow}}^E$ and $\mathcal{F}_{\text{flow}}^L$. Then, the correlation

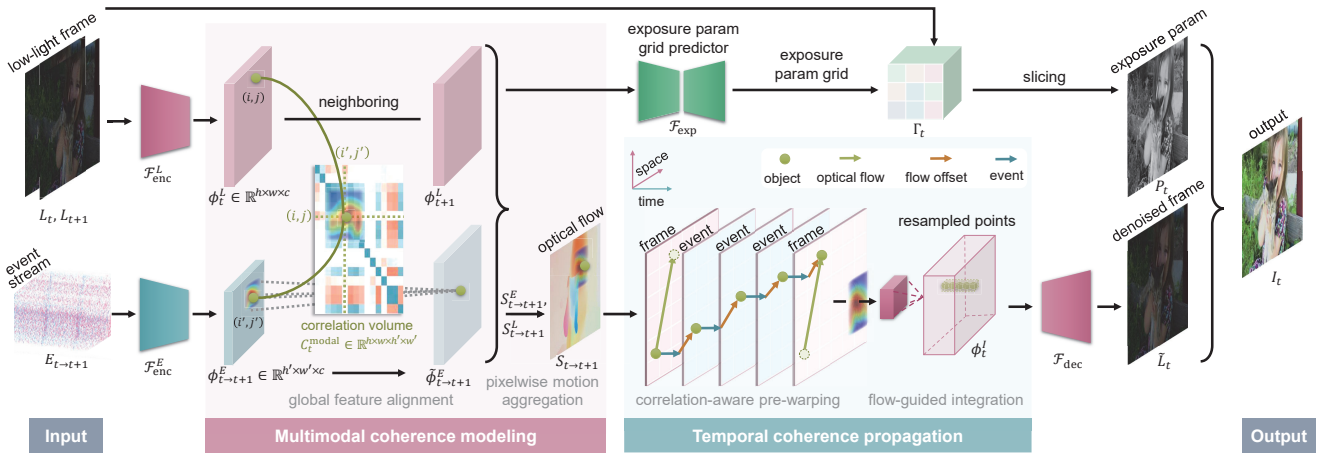


Figure 3. An overview of the proposed method. All-pair correlation volumes between each pixel of events and frames are computed from their features by using the proposed *multimodal coherence modeling* module firstly, which enables the event features to be aligned and the optical flow to be jointly estimated. In the subsequent module of *temporal coherence propagation*, observations corresponding to the same scene point are sampled and propagated across time to estimate the underlying clean frame. Parallelly, exposure parameters are extracted from both events and frames to produce a high-quality frame.

volume can be obtained by:

$$C_t^{\text{modal}}(L_t, E_{t\to t+1})_{pq} = \exp((\varphi_t^L)_p^\top(\varphi_{t\to t+1}^E)_q), \quad (6)$$

where p and q are pixel indices of the features extracted from events and frames, as shown in the red box in Figure 3. The exponent of the inner product between them is computed to rectify the magnitude of strong correlations and suppress the weak ones caused by misaligned pixels or modality discrepancy. To further enlarge the perceptive field, the volume is average-pooled along the spatial dimension of events by a factor of 2 for 3 times. Then new volume $C_{t,s}^{\text{modal}} \in \mathbb{R}^{h\times w\times(h'/2^s)\times(w'/2^s)}$ on a coarser scale s is obtained. Such a design helps to model correspondence across different scales while maintaining fine-grained image details in high resolution. The volumes in four scales are used in subsequent stages for correlation-aware aggregation.

This design is inspired by the recent work in dense correspondence matching [45] from frames of the same modality and resolution, which is different from our setting of cross-modality and cross-resolution. Comparatively, modality misalignment is handled by concatenation-based feature fusion [47] and channel-wise cross-attention [44], which ignore the spatially-unaligned and localized affinity between events and frames that benefits denoising.

Global feature alignment. We consider a situation where there is possible misalignment between the sensor of events and frames. Global misalignment caused by sensor misalignment or camera movements can be effectively regularized by a projection matrix with a few parameters. Considering robustness to error and noise in low-light conditions, we perform the global alignment in the feature space with low resolution. Specifically, the 3×3 projection matrix M_t is parameterized by a $2\times 2\times 2$ displacement cube D_t containing vectors of the four corner points of an image, which can be conveniently transformed into the projection matrix as in [25, 3]. The displacement cube D_t is initialized from an identical transform and then iteratively updated by module $\mathcal{F}_{\text{global}}$ from the correlation feature indexed from C_t^{modal}, which is downsampled into a fixed small spatial size of $(32\times 32)\times(32\times 32)$ for regularization and robustness. After the projection matrix M_t is obtained, the event features $\phi_{t\to t+1}^E$ are aligned by:

$$\widetilde{\phi}_{t\to t+1}^E = \mathcal{P}(\phi_{t\to t+1}^E, M_t), \quad (7)$$

where $\mathcal{P}$ denotes the projection transform.

Given global feature alignment for modeling the sensor misalignment and camera movement between events and frames, we additionally use pixelwise motion integration for the remaining correspondences caused by object movement within exposure time or patch recurrence.

Pixelwise motion aggregation. Event cameras have overwhelming advantages for motion estimation, especially for large displacements and occlusions. However, they usually have low spatial resolution and are only triggered in regions with "moving edges" [32], lacking information in the low-textured regions. Fortunately, although the features are weaker, fine-grained appearance with high resolution remains in low-light frames at the boundary timestamps, which can complement the motion information in events. Therefore, we propose to estimate the flow in the duration between t and $t+1$ jointly from events and boundary frames by aggregating optical flows $S_{t\to t+1}^E, S_{t\to t+1}^L$ estimated from features of frames $\varphi_t^L, \varphi_{t+1}^L$ and events $\varphi_{t\to t+1}^E$. Then, $S_{t\to t+1}^E$ on the coarser scale is projected to coordinates of the frames according to C_t^{modal} and aggregated into $S_{t\to t+1}^L$.

For computational *efficiency*, we consider only the mostly related local regions $\mathcal{N}(p')$ with radius r centered in p' of $S_{t\to t+1}^E$ for each pixel p in the sensor coordinates of frames, rather than all position correlations in self-attention [62] for

jointly estimating depth from events and frames:

$$\mathcal{N}(p')_r = \{p' + \delta p | \delta p \in \mathbb{Z}^2, \|\delta p\|_\infty \leq r\}, \quad (8)$$

The indices of this local neighborhood are used to obtain the correlation scores from the correlation volumes C_t^{modal}, which are used as the weights to aggregate complementary information from events to frames as the residual estimation:

$$\widetilde{S}^E_{t\to t+1}(p) = \frac{1}{Z} \sum_{q\in\mathcal{N}(p)_r} S^E_{t\to t+1}(q)(C_t^{\text{modal}})_{pq}, \quad (9)$$

where Z is a normalizing factor. It is updated by $\mathcal{F}^L_{\text{flow}}$ to obtain the fused flow $S_{t\to t+1}$ for exploiting temporal coherence subsequently. Such a correlation-aware scheme allows a misalignment-robust motion estimation.

3.4. Temporal Coherence Propagation

We follow the basic assumption of denoising that the noise has zero mean, which is expected to be filtered into a clean one even by a simple integrating operation. Recall that events record the logarithmic brightness changes of the scene, which can provide an alternative observation of intensity values for those scene points. However, the pixels from different timestamps corresponding to the same scene point have been shifted due to camera motion or moving objects in the scene in both the captured events and frames. Thus, for each pixel in the target frame to be denoised, we need to resample the points corresponding to the same scene point from neighboring events and frames according to the optical flow estimated in the previous stage.

The resampling process is illustrated in the blue box of Figure 3. Specifically, a pixel p in the target frame L_t appears at the pixel $p + \Delta_{t+1}(p)$ in the frame L_{t+1} with $\Delta_{t+1}(p) = S_{t\to t+1}(p)$. However, for asynchronous events, their duration of motion spanning t to $t+1$ is usually not constant, which is recorded with a temporal resolution of the order of microseconds. To fill the temporal gap between events and frames for more accurate motion estimation, we decompose the optical flow of events into two parts: one with constant velocity, which can be subsampled from $S_{t\to t+1}$; one with varying velocity across time, which can be learned as a flow offset field.

Correlation-aware pre-warping. The features of events and frames according to the same scene points across time are warped into the middle frame based on the estimated optical flows $S_{t\to t+1}$ firstly. Specifically, the features of supporting frames L_{t+1} are pre-warped toward the features of the target frame L_t using the estimated optical flow $S_{t\to t+1}$ in duration t and $t+1$ as:

$$\widetilde{\phi}^L_t = \mathcal{W}(\phi^L_{t+1}, S_{t\to t+1}). \quad (10)$$

For events, we estimate the motion at intermediate B timestamps between the boundary frames (+2). The event features $\widetilde{\phi}^E_{t\to t+1}$ are warped towards target frame feature ϕ^L_t according to the part with constant speed spanned in the space-time neighborhood as:

$$\widetilde{\phi}^E_t = \sum_{\tau\in[1,B]} \mathcal{W}(\widetilde{\phi}^E_{t\to t+1}(\tau), \frac{\tau}{B+2} S_{t\to t+1}), \quad (11)$$

where $\widetilde{\phi}^E_{t\to t+1}(\tau)$ denotes the event features in the τth slice. $\mathcal{W}$ denotes a correlation-aware variant of the back-warping operation as:

$$\mathcal{W}(\phi_{t+1}, S_{t\to t+1})_p = \frac{1}{Z} \sum_{q\in\mathcal{N}(p)_r} \phi_{t+1}(q)(C_t^{\text{temp}})_{pq}, \quad (12)$$

where Z is for normalizion, and the temporal coherence C^{temp} are estimated from consecutive frames:

$$C_t^{\text{temp}}(L_t, L_{t+1})_{pq} = \exp((\varphi^L_t)^\top_p (\varphi^L_{t+1})_q). \quad (13)$$

Flow-guided integration. To compensate for the temporal misalignment between events and frames, we propose to predict the flow offsets $O_{t\to t+1}$ from concatenation of $\{\phi^L_t, \widetilde{\phi}^L_t, \widetilde{\phi}^E_t\}$ and $S_{t\to t+1}$ along the channel dimension:

$$\Delta_{t\to t+1} = \mathcal{F}_{\text{offset}}(\phi^L_t, \widetilde{\phi}^L_t, \widetilde{\phi}^E_t). \quad (14)$$

They are used as the residue to the flow, which jointly composes the offset fields for the deformable convolution layers for implicitly integrating events and frames into latent frames, and propagating temporal coherence across timestamps t for noise reduction as:

$$\phi^I_t = \mathcal{F}_{\text{prop}}(\phi^I_{t+1}, \widetilde{\phi}^E_{t\to t+1}; \Delta_{t\to t+1} + S_{t\to t+1}). \quad (15)$$

The features ϕ^I_t are then decoded into denoised frames.

3.5. Exposure Parameter Estimation

For simplicity, we use the gamma curve with parameter $1/2.2$ to approximate the camera response function [11], and the network to predict the parameters is chosen as a very lightweight one [10]. As shown on the top of the blur box in Figure 3, the exposure parameter grid in a low-dimension space-range bilateral space is extracted from features of events and frames on a pixelwise (spatial variant) basis:

$$\Gamma_t = \mathcal{F}_{\text{exp}}(\phi^L_t, \widetilde{\phi}^E_{t\to t+1}), \quad (16)$$

Then the exposure parameters P_t are sampled to the same resolution of the input frame from the grid Γ_t under the guidance of the spatial index as well as the input low-light frame L_t:

$$P_t = \mathcal{S}(\Gamma_t, L_t), \quad (17)$$

where $\mathcal{S}$ denotes the slicing operation [10]. Then it is applied to the estimated clean frame $\widetilde{L}_t$ according to Eq. (5).

3.6. Training Details

Loss function. To train the whole framework, we use a combination of ℓ_1 loss and gradient loss:

$$\mathcal{L} = \mathcal{L}_{\text{rec}} + \lambda_{\text{grad}}\mathcal{L}_{\text{grad}} + \lambda_{\text{exposure}}\mathcal{L}_{\text{exposure}}, \quad (18)$$

where $\lambda_{\text{grad}} = 10$ and $\lambda_{\text{exposure}} = 100$ are hyper-parameters to balance the contributions of different terms. The first two terms are used to regularize the fidelity between the

predicted normal-light frames I_t and the ground truth $\widetilde{I}_t$ in both intensity and gradient domain:

$$\mathcal{L}_{\text{rec}} = \sum_{i=t-N}^{t+N} \|\widetilde{I}_t - I_t\|_1, \tag{19}$$

and

$$\mathcal{L}_{\text{grad}} = \sum_{i=t-N}^{t+N} \|\nabla(\widetilde{I}_t) - \nabla(I_t)\|_1. \tag{20}$$

The third term is used to regularize the exposure parameters to correctly enhance a blurred version of low-light frames into its normal-light counterpart:

$$\mathcal{L}_{\text{exposure}} = \sum_{i=t-N}^{t+N} \|f_{\text{CRF}}(f_{\text{CRF}}^{-1}(\mathcal{G}(L_t)); P_t)) - \mathcal{G}(\widetilde{I}_t)\|_1. \tag{21}$$

where $\mathcal{G}$ is a Gaussian blurring operation with variance 2 to improve the robustness to noise of the exposure parameter.

Data preparation. The training dataset is constructed following the protocol proposed in Zhang *et al*. [63], which contains 107 pairs of synthetic normal-light and low-light video (6208 frames). Their work focuses only on taking low-light videos as input. The final dataset contains 87 videos randomly split for training and 20 for testing. Following [63], we synthesize low-light frames L_t without noise from normal-light ones I_t using gamma correction and linear scaling with the same parameter setting:

$$L_t(p) = \beta \times (\alpha \times I_t(p))^{\gamma}, \tag{22}$$

where α, β, γ are sampled from a uniform distribution $\mathcal{U}(0.9, 1), \mathcal{U}(0.5, 1), \mathcal{U}(2, 3.5)$, respectively. To meet our requirement of hybrid inputs of events and low-light videos, we further synthetic the events using the video to event simulator v2e [19]. The spatial resolution of all frames is 854×480, while it is 427×240 for events to simulate the discrepancy in spatial resolution between the two modalities. The events and frames of a hybrid camera system are hard to be perfectly aligned in practice. To take this into consideration, we apply random perspective transforms between them as [25].

Noise simulation. Noise in low-light conditions is the key factor that we take care of. For events, insufficient illumination brings distinctive degradation such as limited bandwidth, more leaky events, and shot noise, which we simulate following [67]. For frames, the commonly used noise model Gaussian-Poisson distribution can be modeled by a signal-dependent Gaussian distribution [18]:

$$L(p) = \mathcal{N}(I(p), \sigma_r^2 + \sigma_s I(p)), \tag{23}$$

where $\mathcal{N}$ denotes the Gaussian distribution and the noise parameters σ_r and σ_s are both sampled from uniform distribution $\mathcal{U}(0.01, 0.04)$ in [63].

To enable the proposed method to generalize in complex real-world scenarios, we propose to use a more practical degradation process for low-light frames. First, when the intensity is small, the Poisson distribution has very different characteristics from a signal-dependent Gaussian distribution. In our experiments, the shot noise is sampled from the Poisson distribution with a noise scale sampled from $\mathcal{U}(0.05, 2.5)$ instead of a Gaussian approximated one. Second, JPEG compression often occurs in digital images. The caused artifacts become more significant in low-light conditions where the features are weak. We include JPEG compression in the simulation process with a quality factor sampled from $\mathcal{U}(50, 95)$.

Other implementation details. We apply random cropping, horizontal flipping, and rotation for data augmentation. The cropping size is 128×128, and the rotation angles include 90, 180, and 270 degrees. The learning rate is set to 1×10^{-4}, and the model is trained by Adam Optimizer [24] with default parameters for 50 epochs on a single NVIDIA TITAN RTX GPU. Following the common practice in flow-based video restoration [16, 35, 4, 68], the optical flow estimators $\mathcal{F}_{\text{flow}}^{E}$ and $\mathcal{F}_{\text{flow}}^{L}$ are initialized from pre-trained models for events [9] and frames [45], respectively.

Table 1. Quantitative comparison on synthetic low-light video enhancement dataset proposed in [63]. ↑ (↓) means higher (lower) is better.

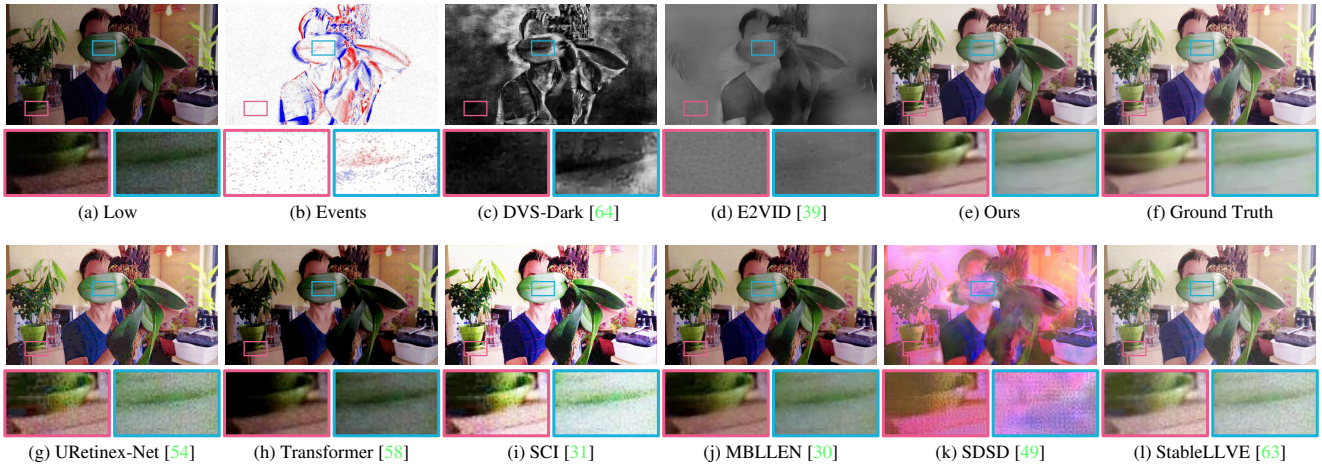
Method		PSNR↑	SSIM↑	LPIPS↓
Pure event	DVS-Dark [64]	9.84	0.2617	0.6205
	E2VID [39]	11.63	0.3208	0.6154
Image-based	LIME [12]	15.08	0.3402	0.5327
	SCI [31]	16.90	0.4187	0.5081
	Transformer [58]	15.58	0.6146	0.3965
	URetinex-Net [54]	20.20	0.5157	0.4797
Video-based	MBLLEN [30]	16.72	0.6051	0.5493
	SDSD [49]	12.35	0.2604	0.7187
	StableLLVE [63]	21.79	0.7176	0.3856
Hybrid	Ours	**22.81**	**0.8180**	**0.2747**

4. Experiments

In this section, we evaluate the effectiveness of the proposed method using real and synthetic data. Ablation studies are conducted to verify the effectiveness of the proposed modules. Throughout all experiments, we adopt the commonly used metrics PSNR, SSIM, and LPIPS to evaluate the performance of different methods quantitatively.

4.1. Comparison with State-of-the-Art Methods

We compare the proposed method with nine state-of-the-art methods, including four image-based methods LIME [12], SCI [31], Transformer [58], URetinex-Net [54]; three video-based methods MBLLEN [30], StableLLVe [63], and SDSD [49]; and two event-based restoration methods DVS-Dark [64] and E2VID [39]. All results for comparison are produced from their official codes with recommended hyperparameters. Note that LIME [12] is the state-of-the-art conventional method, while the others are all learning-



(a) Low (b) Events (c) DVS-Dark [64] (d) E2VID [39] (e) Ours (f) Ground Truth

(g) URetinex-Net [54] (h) Transformer [58] (i) SCI [31] (j) MBLLEN [30] (k) SDSD [49] (l) StableLLVE [63]

Figure 4. Visual quality comparison of enhancement results on synthetic data.

based. DVS-Dark [64] and E2VID [39] can only reconstruct grayscale frames; we convert the color frames into grayscale ones for numerical comparison.

To quantitatively evaluate the effectiveness of the proposed method, we conduct comparisons using synthetic data proposed in [63] and the one synthesized by the proposed noise simulation process for handling complex real-world degradation. Quantitative results in the low-light video dataset proposed in [63] are reported in Table 1, demonstrating the proposed method's superior performance. Thanks to the introduced visual modality of event streams for effective motion estimation and denoising, the proposed method outperforms the others in terms of all metrics by a large margin. To enable a noise-robust low-light video enhancement with a more generalizable ability, we propose a noise simulation process that better characterizes the real noise and compression artifacts. The comparison results on the synthetic dataset processed by the proposed noise simulation are reported in Table 2. The superior performance further demonstrates the effectiveness of the proposed method. The corresponding visual comparison results are shown in Figure 4. Directly utilizing the high dynamic and high temporal resolution property of events to reconstruct video is quite ill-posed, which is evidenced in the results of DVS-Dark [64] and E2VID [39]. Moreover, DVS-Dark [64] is an adversarial learning method trained from unpaired data, which could cause model collapse. Unlike the other video enhancement methods, temporal redundancy in low-light videos is hard to be solely exploited due to the weakened features and significant noise. It can be seen that none of the frame-based methods can well suppress the severe noise. SDSD [49] produces inferior results, which might be due to the use of an unsuitable upsampling layer [34] and batch normalization [55], which have been found to produce artifacts in low-level tasks. In comparison, MBLLEN [30], StableLLVe [63], and the proposed method successfully recover global illumination. However, the noise is significant in the results of MBLLEN [30] and StableLLVe [63]. In comparison, our method can suppress noise well and recover pleasing illumination. It validates the effectiveness of introducing event cameras for better motion estimation and alternative intensity observations.

Table 2. Quantitative comparison on our synthetic data with more severe noise. ↑ (↓) means higher (lower) is better.

Method		PSNR↑	SSIM↑	LPIPS↓
Pure event	DVS-Dark [64]	9.81	0.2654	0.5748
	E2VID [39]	16.20	0.5729	0.5077
Image-based	LIME [12]	15.66	0.4274	0.4817
	SCI [31]	15.96	0.5166	0.4667
	Transformer [58]	15.81	0.5847	0.4087
	URetinex-Net [54]	20.87	0.5898	0.4499
Video-based	MBLLEN [30]	17.77	0.5896	0.3823
	SDSD [49]	12.60	0.2822	0.7225
	StableLLVE [63]	19.37	0.6992	0.3857
Hybrid	Ours	**23.98**	**0.8369**	**0.2794**

The existing real-world dataset proposed in [20] is unsuitable for our task because it only takes events as input rather than events and frames, and their released events are stacked using a method different from ours. To evaluate the proposed method for real-world scenarios, we build a hybrid camera system consisting of an industrial camera (FLIR Chameleon 3 Color, with resolution 1920×1280 at 20 fps) and an event camera (DAVIS346, with resolution 346×260) via a beam splitter (Thorlabs CCM1-BS013) mounted in front of the two cameras with 50% optical splitting. Visual comparisons are shown in Figure 5. It shows that the proposed method can suppress noise as well as compensate for the exposure for high-quality output. The event-based video reconstruction methods cannot recover high-quality texture due to the trigger mechanism in event cameras. The result shows that while all other methods produce obvious artifacts or events

(a) Low (b) Events (c) DVS-Dark [64] (d) E2VID [39] (e) Ours (f) LIME [12]

(g) URetinex-Net [54] (h) Transformer [58] (i) SCI [31] (j) MBLLEN [30] (k) SDSD [49] (l) StableLLVE [63]

Figure 5. Visual quality comparison of enhancement results on real-captured data.

Table 3. Ablation results. ↑ (↓) means higher (lower) is better.

Methods	PSNR↑	SSIM↑	LPIPS↓
w/o events	22.86	0.8147	0.3021
w/o global feature alignment	23.41	0.8241	0.2974
w/o pixelwise motion aggregation	22.89	0.8248	0.2976
w/o temporal coherence propagation	23.67	0.8231	0.2967
w/o exposure parameter estimation	19.83	0.7944	0.3248
w/o noise simulation	22.15	0.7290	0.3592
Ours	**23.98**	**0.8369**	**0.2794**

that magnify the noise, the proposed method can produce smooth results with well-suppressed noise.

4.2. Ablation Study

To validate the effectiveness of each component, we compare the proposed method with its five variants. Qualitative results are shown in Table 3. Without the optical flows extracted from events, the results drastically decrease (the first row). This is because the features of low-light frames are weak, so extracting the optical flow solely from the frames with low contrast and low SNR is unreliable. The two visual modalities with totally different sensors and resolutions are hard to be precisely aligned. A good spatial alignment module to compensate for misalignment in low-light conditions, a good spatial alignment module is essential to utilize the high dynamic range and high temporal resolution property of events (the second row). Without the pixelwise motion aggregation module to jointly extract motion information from events and frames, performance drops (the third row). Figure 2 also shows its effectiveness in joint optical flow estimation. For temporal coherence propagation, we propose a flow-guided integration module tailored to fill the gap in temporal resolution between events and frames and propagate temporal redundancy information across frames (the fourth row). The exposure parameter estimation module is crucial for enhancing low-light frames to achieve high contrast and visually appealing results (the fifth row). The proposed noise simulation process is effective for robustness to noise, as shown in the sixth row.

5. Conclusion

In this paper, we present a deep learning framework for low-light video enhancement from hybrid inputs of low-light video and the corresponding events. Thanks to the proposed multimodal coherence module to compensate for the sensor misalignment between events and low-light frames and the temporal coherence propagation module to utilize temporal redundancy for improving SNR and contrast in low-light videos, our method can successfully suppress noise under challenging conditions of low-light videography.

Limitations. In this paper, we focus on utilizing high temporal resolution and high dynamic range in event cameras for noise suppression and exposure compensation in low-light videos. However, we cannot handle the situation where the visual signal is too weak to trigger enough events, for example, less than 0.5 lux. Moreover, we do not consider the color distortion that might occur in low-light conditions.

Acknowledgement

This work was supported by the National Key R&D Program of China (Grant No. 2021ZD0109800) and the National Natural Science Foundation of China (Grant No. 62088102, 62136001, 62072188). Jinxiu Liang was also supported by China Postdoctoral Science Foundation (Grant No. 2022M720236).

References

[1] Srutarshi Banerjee, Henry H. Chopp, Jianping Zhang, Zihao W. Wang, Peng Kang, Oliver Cossairt, and Aggelos Katsaggelos. A Joint Intensity-Neuromorphic Event Imaging System With Bandwidth-Limited Communication Channel. *IEEE Transactions on Neural Networks and Learning Systems*, 2022. 3

[2] Jianrui Cai, Shuhang Gu, and Lei Zhang. Learning a Deep Single Image Contrast Enhancer from Multi-Exposure Images. *IEEE Transactions on Image Processing*, 27(4), Apr. 2018. 2

[3] Si-Yuan Cao, Jianxin Hu, Zehua Sheng, and Hui-Liang Shen. Iterative Deep Homography Estimation. In *Proc. of Computer Vision and Pattern Recognition*, 2022. 4

[4] Kelvin C. K. Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. BasicVSR: The Search for Essential Components in Video Super-Resolution and Beyond. In *Proc. of Computer Vision and Pattern Recognition*, 2021. 6

[5] Chen Chen, Qifeng Chen, Minh N. Do, and Vladlen Koltun. Seeing Motion in the Dark. In *Proc. of International Conference on Computer Vision*, 2019. 1, 2

[6] Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. Learning to See in the Dark. In *Proc. of Computer Vision and Pattern Recognition*, 2018. 2

[7] Peiqi Duan, Zihao Wang, Boxin Shi, Oliver Cossairt, Tiejun Huang, and Aggelos Katsaggelos. Guided Event Filtering: Synergy between Intensity Images and Neuromorphic Events for High Performance Imaging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 3

[8] Guillermo Gallego, Tobi Delbruck, Garrick Michael Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew Davison, Jorg Conradt, Kostas Daniilidis, and Davide Scaramuzza. Event-based Vision: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 3

[9] Mathias Gehrig, Mario Millhäusler, Daniel Gehrig, and Davide Scaramuzza. E-RAFT: Dense Optical Flow from Event Cameras. In *Proc. 3DV*, 2021. 2, 6

[10] Michaël Gharbi, Jiawen Chen, Jonathan T. Barron, Samuel W. Hasinoff, and Frédo Durand. Deep Bilateral Learning for Real-time Image Enhancement. *ACM Transactions on Graphics (Proc. of ACM SIGGRAPH)*, 36(4):118:1–118:12, July 2017. 5

[11] M.D. Grossberg and S.K. Nayar. What is the space of camera response functions? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2003. 5

[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, Feb. 2017. 1, 6, 7

[13] Jin Han, Yixin Yang, Peiqi Duan, Chu Zhou, Lei Ma, Chao Xu, Tiejun Huang, Imari Sato, and Boxin Shi. Hybrid High Dynamic Range Imaging fusing Neuromorphic and Conventional Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 3

[14] Jin Han, Yixin Yang, Chu Zhou, Chao Xu, and Boxin Shi. EvIntSR-Net: Event Guided Multiple Latent Frames Reconstruction and Super-Resolution. In *Proc. of International Conference on Computer Vision*, 2021. 2, 3

[15] Jin Han, Chu Zhou, Peiqi Duan, Yehui Tang, Chang Xu, Chao Xu, Tiejun Huang, and Boxin Shi. Neuromorphic Camera Guided High Dynamic Range Imaging. In *Proc. of Computer Vision and Pattern Recognition*, 2020. 3

[16] Muhammad Haris, Greg Shakhnarovich, and Norimichi Ukita. Recurrent Back-Projection Network for Video Super-Resolution. In *Proc. of Computer Vision and Pattern Recognition*, 2019. 6

[17] Weihua He, Kaichao You, Zhendong Qiao, Xu Jia, Ziyang Zhang, Wenhui Wang, Huchuan Lu, Yaoyuan Wang, and Jianxing Liao. TimeReplayer: Unlocking the Potential of Event Cameras for Video Interpolation. In *Proc. of Computer Vision and Pattern Recognition*, 2022. 3

[18] G. E. Healey and R. Kondepudy. Radiometric CCD camera calibration and noise estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(3):267–276, Mar. 1994. 6

[19] Yuhuang Hu, Shih-Chii Liu, and Tobi Delbruck. V2e: From Video Frames to Realistic DVS Events. In *Proc. of Computer Vision and Pattern Recognition Workshops*, 2021. 6

[20] Haiyang Jiang and Yinqiang Zheng. Learning to See Moving Objects in the Dark. In *Proc. of International Conference on Computer Vision*, 2019. 1, 7

[21] Yifan Jiang, Xinyu Gong, Ding Liu, Yu Cheng, Chen Fang, Xiaohui Shen, Jianchao Yang, Pan Zhou, and Zhangyang Wang. EnlightenGAN: Deep Light Enhancement without Paired Supervision. *IEEE Transactions on Image Processing*, 30:2340–2349, 2021. 2

[22] Zhe Jiang, Yu Zhang, Dongqing Zou, Jimmy Ren, Jiancheng Lv, and Yebin Liu. Learning Event-Based Motion Deblurring. In *Proc. of Computer Vision and Pattern Recognition*, 2020. 3

[23] Taewoo Kim, Jeongmin Lee, Lin Wang, and Kuk-Jin Yoon. Event-guided Deblurring of Unknown Exposure Time Videos. In *Proc. of European Conference on Computer Vision*, 2022. 3

[24] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *Proc. ICLR*, 2015. 6

[25] Hoang Le, Feng Liu, Shu Zhang, and Aseem Agarwala. Deep Homography Estimation for Dynamic Scenes. In *Proc. of Computer Vision and Pattern Recognition*, 2020. 4, 6

[26] Songnan Lin, Jiawei Zhang, Jinshan Pan, Zhe Jiang, Dongqing Zou, Yongtian Wang, Jing Chen, and Jimmy Ren. Learning Event-Driven Video Deblurring and Interpolation. In *Proc. of European Conference on Computer Vision*, 2020. 3

[27] Ce Liu and William T. Freeman. A High-Quality Video Denoising Algorithm Based on Reliable Motion Estimation. In *Proc. of European Conference on Computer Vision*, 2010. 1

[28] Risheng Liu, Long Ma, Jiaao Zhang, Xin Fan, and Zhongxuan Luo. Retinex-Inspired Unrolling with Cooperative Prior Architecture Search for Low-Light Image Enhancement. In *Proc. of Computer Vision and Pattern Recognition*, 2021. 2

[29] Kin Gwn Lore, Adedotun Akintayo, and Soumik Sarkar. LLNet: A deep autoencoder approach to natural low-light image enhancement. *Pattern Recognit.*, 61:650–662, Jan. 2017. 2

[30] Feifan Lv, Feng Lu, Jianhua Wu, and Chongsoon Lim. MBLLEN: Low-Light Image/Video Enhancement using CNNs. In *Proc. of British Machine Vision Conference*, 2018. 1, 2, 6, 7

[31] Long Ma, Tengyu Ma, Risheng Liu, Xin Fan, and Zhongxuan Luo. Toward Fast, Flexible, and Robust Low-Light Image Enhancement. In *Proc. of Computer Vision and Pattern Recognition*, 2022. 2, 6, 7

[32] Anton Mitrokhin, Zhiyuan Hua, Cornelia Fermuller, and Yiannis Aloimonos. Learning Visual Motion Segmentation using Event Surfaces. In *Proc. of Computer Vision and Pattern Recognition*, 2020. 4

[33] Gottfried Munda, Christian Reinbacher, and Thomas Pock. Real-Time Intensity-Image Reconstruction for Event Cameras Using Manifold Regularisation. *International Journal of Computer Vision*, 126(12):1381–1393, 2018. 3

[34] Augustus Odena, Vincent Dumoulin, and Chris Olah. Deconvolution and Checkerboard Artifacts. *Distill*, 1(10):e3, Oct. 2016. 7

[35] Jinshan Pan, Haoran Bai, and Jinhui Tang. Cascaded Deep Video Deblurring Using Temporal Sharpness Prior. In *Proc. of Computer Vision and Pattern Recognition*, 2020. 6

[36] Liyuan Pan, Richard Hartley, Cedric Scheerlinck, Miaomiao Liu, Xin Yu, and Yuchao Dai. High Frame Rate Video Reconstruction based on an Event Camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 2, 3

[37] Liyuan Pan, Cedric Scheerlinck, Xin Yu, Richard Hartley, Miaomiao Liu, and Yuchao Dai. Bringing a Blurry Frame Alive at High Frame-Rate with an Event Camera. In *Proc. of Computer Vision and Pattern Recognition*, 2019. 2, 3

[38] Federico Paredes-Valles and Guido C. H. E. de Croon. Back to Event Basics: Self-Supervised Learning of Image Reconstruction for Event Cameras via Photometric Constancy. In *Proc. of Computer Vision and Pattern Recognition*, 2021. 3

[39] Henri Rebecq, Rene Ranftl, Vladlen Koltun, and Davide Scaramuzza. Events-to-Video: Bringing Modern Computer Vision to Event Cameras. In *Proc. of Computer Vision and Pattern Recognition*, 2019. 2, 3, 6, 7

[40] 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(6):1964–1980, June 2021. 3

[41] Wei Shang, Dongwei Ren, Dongqing Zou, Jimmy S. Ren, Ping Luo, and Wangmeng Zuo. Bringing Events Into Video Deblurring With Non-Consecutively Blurry Frames. In *Proc. of International Conference on Computer Vision*, 2021. 3

[42] Liang Shen, Zihan Yue, Fan Feng, Quan Chen, Shihao Liu, and Jie Ma. MSR-net:Low-light Image Enhancement using Deep Convolutional Network. *arXiv:1711.02488 [cs]*, Nov. 2017. 2

[43] 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 *Proc. of European Conference on Computer Vision*, 2020. 3

[44] Lei Sun, Christos Sakaridis, Jingyun Liang, Qi Jiang, Kailun Yang, Peng Sun, Yaozu Ye, Kaiwei Wang, and Luc Van Gool. Event-Based Fusion for Motion Deblurring with Cross-modal Attention. In *Proc. of European Conference on Computer Vision*, 2022. 3, 4

[45] Zachary Teed and Jia Deng. RAFT: Recurrent All-Pairs Field Transforms for Optical Flow. In *Proc. of European Conference on Computer Vision*, 2020. 2, 4, 6

[46] Minggui Teng, Chu Zhou, Hanyue Lou, and Boxin Shi. NEST: Neural Event Stack for Event-Based Image Enhancement. In *Proc. of European Conference on Computer Vision*, 2022. 3

[47] Stepan Tulyakov, Alfredo Bochicchio, Daniel Gehrig, Stamatios Georgoulis, Yuanyou Li, and Davide Scaramuzza. Time lens++: Event-based frame interpolation with parametric nonlinear flow and multi-scale fusion. In *Proc. of Computer Vision and Pattern Recognition*, 2022. 2, 3, 4

[48] Stepan Tulyakov, Daniel Gehrig, Stamatios Georgoulis, Julius Erbach, Mathias Gehrig, Yuanyou Li, and Davide Scaramuzza. Time Lens: Event-based video frame interpolation. In *Proc. of Computer Vision and Pattern Recognition*, 2021. 2, 3

[49] 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 *Proc. of International Conference on Computer Vision*, 2021. 1, 2, 6, 7

[50] Ruixing Wang, Qing Zhang, Chi-Wing Fu, Xiaoyong Shen, Wei-Shi Zheng, and Jiaya Jia. Underexposed Photo Enhancement using Deep Illumination Estimation. In *Proc. of Computer Vision and Pattern Recognition*, 2019. 2

[51] Zihao W. Wang, Peiqi Duan, Oliver Cossairt, Aggelos Katsaggelos, Tiejun Huang, and Boxin Shi. Joint Filtering of Intensity Images and Neuromorphic Events for High-Resolution Noise-Robust Imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 3

[52] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep Retinex Decomposition for Low-Light Enhancement. In *Proc. of British Machine Vision Conference*, 2018. 2

[53] Wenming Weng, Yueyi Zhang, and Zhiwei Xiong. Event-Based Video Reconstruction Using Transformer. In *Proc. of International Conference on Computer Vision*, 2021. 3

[54] 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 *Proc. of Computer Vision and Pattern Recognition*, 2022. 2, 6, 7

[55] Sitao Xiang and Hao Li. On the Effects of Batch and Weight Normalization in Generative Adversarial Networks. *arXiv:1704.03971 [cs]*, (arXiv:1704.03971), Dec. 2017. 7

[56] Fang Xu, Lei Yu, Bishan Wang, Wen Yang, Gui-Song Xia, Xu Jia, Zhendong Qiao, and Jianzhuang Liu. Motion Deblurring With Real Events. In *Proc. of International Conference on Computer Vision*, 2021. 3

[57] Ke Xu, Xin Yang, Baocai Yin, and Rynson W. H. Lau. Learning to Restore Low-Light Images via Decomposition-and-Enhancement. In *Proc. of Computer Vision and Pattern Recognition*, 2020. 2

[58] Xiaogang Xu, Ruixing Wang, Chi-Wing Fu, and Jiaya Jia. SNR-Aware Low-Light Image Enhancement. In *Proc. of Computer Vision and Pattern Recognition*, 2022. 2, 6, 7

[59] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T. Freeman. Video Enhancement with Task-Oriented Flow. *International Journal of Computer Vision*, 127(8):1106–1125, Aug. 2019. 1

[60] Wenhan Yang, Shiqi Wang, Yuming Fang, Yue Wang, and Jiaying Liu. From Fidelity to Perceptual Quality: A Semi-Supervised Approach for Low-Light Image Enhancement. In *Proc. of Computer Vision and Pattern Recognition*, 2020. 2

[61] Zhiyang Yu, Yu Zhang, Deyuan Liu, Dongqing Zou, Xijun Chen, Yebin Liu, and Jimmy S. Ren. Training Weakly Supervised Video Frame Interpolation With Events. In *Proc. of International Conference on Computer Vision*, 2021. 3

[62] Dehao Zhang, Qiankun Ding, Peiqi Duan, Chu Zhou, and Boxin Shi. Data Association Between Event Streams and Intensity Frames Under Diverse Baselines. In *Proc. of European Conference on Computer Vision*, 2022. 3, 4

[63] Fan Zhang, Yu Li, Shaodi You, and Ying Fu. Learning Temporal Consistency for Low Light Video Enhancement From Single Images. In *Proc. of Computer Vision and Pattern Recognition*, 2021. 1, 2, 6, 7

[64] Song Zhang, Yu Zhang, Zhe Jiang, Dongqing Zou, Jimmy Ren, and Bin Zhou. Learning to See in the Dark with Events. In *Proc. of European Conference on Computer Vision*, 2020. 1, 2, 3, 6, 7

[65] Xiang Zhang and Lei Yu. Unifying Motion Deblurring and Frame Interpolation with Events. In *Proc. of Computer Vision and Pattern Recognition*, 2022. 3

[66] Yonghua Zhang, Jiawan Zhang, and Xiaojie Guo. Kindling the Darkness: A Practical Low-light Image Enhancer. In *Proc. ACM MM*, 2019. 2

[67] Chu Zhou, Minggui Teng, Jin Han, Jinxiu Liang, Chao Xu, Gang Cao, and Boxin Shi. Deblurring Low-Light Images with Events. *International Journal of Computer Vision*, Feb. 2023. 2, 3, 6

[68] Kun Zhou, Wenbo Li, Liying Lu, Xiaoguang Han, and Jiangbo Lu. Revisiting Temporal Alignment for Video Restoration. In *Proc. of Computer Vision and Pattern Recognition*, 2022. 6

[69] Xinyu Zhou, Peiqi Duan, Yi Ma, and Boxin Shi. Evunroll: Neuromorphic events based rolling shutter image correction. In *Proc. of Computer Vision and Pattern Recognition*, 2022. 3

[70] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Unsupervised Event-Based Learning of Optical Flow, Depth, and Egomotion. In *Proc. of Computer Vision and Pattern Recognition*, 2019. 3