Low-Light Video Enhancement with Synthetic Event Guidance

Lin Liu[1], Junfeng An[2] *, Jianzhuang Liu[4], Shanxin Yuan[3†], Xiangyu Chen[6, 8],
Wengang Zhou[1], Houqiang Li[1], Yan Feng Wang[7], Qi Tian[5]

[1]CAS Key Laboratory of Technology in GIPAS, EEIS Department, University of Science and Technology of China
[2]Independent Researcher
[3]Queen Mary University of London
[4]Huawei Noah's Ark Lab
[5]Huawei Cloud BU
[6]University of Macau
[7]Cooperative medianet innovation center of Shanghai Jiao Tong University
[8]Shenzhen Institute of Advanced Technology (SIAT)

Abstract

Low-light video enhancement (LLVE) is an important yet challenging task with many applications such as photographing and autonomous driving. Unlike single image low-light enhancement, most LLVE methods utilize temporal information from adjacent frames to restore the color and remove the noise of the target frame. However, these algorithms, based on the framework of multi-frame alignment and enhancement, may produce multi-frame fusion artifacts when encountering extreme low light or fast motion. In this paper, inspired by the low latency and high dynamic range of events, we use synthetic events from multiple frames to guide the enhancement and restoration of low-light videos. Our method contains three stages: 1) event synthesis and enhancement, 2) event and image fusion, and 3) low-light enhancement. In this framework, we design two novel modules (event-image fusion transform and event-guided dual branch) for the second and third stages, respectively. Extensive experiments show that our method outperforms existing low-light video or single image enhancement approaches on both synthetic and real LLVE datasets. Our code will be available at https://gitee.com/mindspore/models/tree/master/research/cv/LLVE-SEG.

1 Introduction

The image quality in low-light or under-exposure conditions is often unsatisfactory, so image/video enhancement in low light has been an active research topic in computer vision (Wang et al. 2021; Jiang et al. 2022b). However, it is challenging due to strong noise, detail loss, non-uniform exposure, *etc*. These problems become even more serious in videos taken from dynamic scenes. In this paper, in contrast to low-light, we loosely call bright-light and day-light images/videos/events *normal-light* images/videos/events.

Most of fully-supervised deep-learning based low-light video enhancement (LLVE) methods (Lv et al. 2018; Jiang

(a) Traditional frames-based LLVE method

(b) Our event-image-fusion LLVE method

(c) Visual comparison with a recent LLVE method (best viewed on screen)

Figure 1: Comparison between (a) Traditional frames-based LLVE method and (b) Our event-image fusion LLVE method. (c) Visual comparison between a recent LLVE method SDSDNet (Wang et al. 2021) and ours. Our method can better remove noise, maintain details and avoid misaligned artifacts of multi-frame fusion.

and Zheng 2019; Wang et al. 2021) or video reconstruction methods (Xue et al. 2019; Wang et al. 2019b; Isobe et al. 2020; Dai et al. 2022) are based on the multi-frame alignment-and-enhancement framework. This pipeline firstly utilizes some techniques, *e.g.*, 3D convolution (Lv et al. 2018; Jiang and Zheng 2019), deformable convolution (Isobe et al. 2020; Dai et al. 2022), or flow-based alignment (Xue et al. 2019), to align the temporal information from adjacent frames to the reference frame, and

---

*Part of the work was done during an internship in Huawei Noah's Ark Lab.

†Corresponding author

then uses an enhancement network for noise removal and illumination correction (see Fig. 1(a)). However, when facing extreme low light or significant motion in videos, these algorithms may produce multi-frame fusion artifacts in the predicted images (see the results of SDSDNet in Fig. 1(c)).

Existing methods face some potential difficulties. First, sensor noise is not negligible in low signal-to-noise low-light scenarios. This noise hinders the network from learning the alignment of temporal features. Second, the interference between strong noise and image details causes the enhancement network to remove some image details inevitably.

In this paper, inspired by the low latency and high dynamic range of events, we use synthetic events to guide the enhancement and restoration of low-light videos. In general, events are captured by event cameras (e.g., DAVIS240C (Mueggler et al. 2017)), which contain sparse and asynchronous intensity changes of the scene, instead of the color and intensity as in normal images. Recently, Gehrig *et al.* (Gehrig et al. 2020) present a method that can convert videos recorded with conventional cameras into synthetic realistic events. The pioneering work applying event information to low-light enhancement is SIDE (Zhang et al. 2020). However, their work studies the transformation from real events to single images, and due to the difficulty of collecting real event-image pairs, it is an unpaired learning method. Differently, we focus on the fusion between synthetic enhanced events and video frames, where the synthetic events and frames are paired.

Unlike the conventional two-stage video reconstruction/enhancement pipeline, we propose a three-stage LLVE pipeline with the guidance of synthetic events: 1) event synthesis and restoration; 2) event-image fusion; 3) low-light enhancement (see Fig. 1(b)). The first stage is event-to-event, aiming to obtain normal-light events from low-light events. Due to the sparsity of events and the interpolation of fast motion by an event synthesis algorithm, our network can restore normal-light events well and solve the problems of noise, color shift, and detail loss based on the enhanced events (see Fig. 2).

Different from most event-image fusion methods (Pini, Borghi, and Vezzani 2018; Pini et al. 2019; Han et al. 2020; Paikin et al. 2021; He et al. 2022), which combine event and image features by simply concatenating them, we design a novel Event and Image Fusion Transform (EIFT) module for better fusing the sparse event features and image features in the second stage of our pipeline. And in the last stage, we design an Event-Guided Dual Branch (EGDB) module for low-light enhancement. Using the mask generated by the enhanced events, the image can be divided into areas with little change and areas with sharp change during a period of time. Moreover, the proposed global branch and local branch in EGDB can deal with these two kinds of areas, respectively. A transformer network is used in the global branch to capture global illumination information.

In summary, we make the following contributions:

- A novel three-stage pipeline with the guidance of synthetic events for low-light video enhancement is proposed. It consists of event synthesis and restoration, event image fusion, and low-light enhancement.
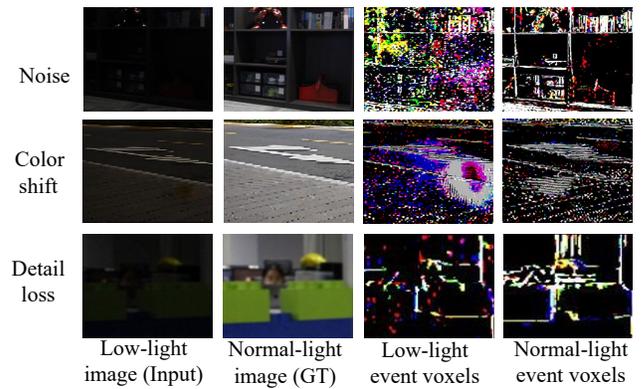
Figure 2: Challenges of low-light image/video enhancement. When comparing the events of normal-light images and low-light images, we find that low-light enhancement faces challenges of strong noise, color shift, and detail loss. These problems are hard to notice in normal-light images, but are apparent in the event voxels.

- We design a novel Event and Image Fusion Transform (EIFT) module for event-image fusion and an Event-Guided Dual Branch (EGDB) module for low-light enhancement.
- Extensive experiments on both synthetic and real LLVE datasets show that our method outperforms both the state-of-the-art (SOTA) low-light video and single image enhancement approaches.

2 Related Work

CNNs and Transformers are playing critical roles in image/video restoration/reconstruction (Liu et al. 2020a; Zheng et al. 2020, 2021; Liu et al. 2020b,c; Zheng et al. 2022; Dai et al. 2021; Li et al. 2022; Zhao et al. 2021b; Ershov et al. 2022; Wang et al. 2023). This section discusses the most related methods, including low-light video enhancement, event-image fusion, and low-level vision Transformers.

Low-Light Video Enhancement (LLVE). This line of work focuses on how to utilize temporal information from adjacent frames. They can be divided into paired learning and unpaired learning. In the former, most methods use the alignment-and-enhance framework for LLVE. Lv *et al.* (Lv et al. 2018) and Chen *et al.* (Chen et al. 2019) use 3D convolution and their networks can combine temporal information from image sequences. Wang *et al.* use a mechatronic system to collect a dataset named SDSD, where the normal-light and low-light image pairs are well aligned (Wang et al. 2021). They also use deformable convolutions in the multi-frame alignment stage. Some approaches try to solve the LLVE problem by combining an image-based model with temporal consistency loss (Chen et al. 2019; Zhang et al. 2020). However, most of the temporal consistency losses in LLVE (Zhang et al. 2021a) or other video restoration tasks (Yang et al. 2018; Zhao et al. 2021a) are optical-flow based, which may be limited by the error in optical flow esti-

mation. As for unpaired learning methods, to solve the lack of data problem, Triantafyllidou *et al.* (Triantafyllidou et al. 2020) convert normal-light videos to low-light videos using two CycleGAN (Zhu et al. 2017) networks. In this framework, normal-light videos are first converted into long exposure videos and then into low-light videos.

Event-Based Image/Video Reconstruction. The event camera can handle high-speed motion and high-dynamic scenes because of its dynamic vision sensor. Therefore, events can be applied to image or video reconstruction tasks including video deblurring (Jiang et al. 2020), super-resolution (Han et al. 2021; Jing et al. 2021), joint filtering (Wang et al. 2020), and tone mapping (Simon Chane et al. 2016; Han et al. 2020). Ours is the first to explore the use of synthetic events in low-light video enhancement.

Event-Image Fusion. Event features and image features are of different modals, with different characteristics. Event features reflect motion changes, so most values in the scene at a certain moment are zero. Image features in low light contain both strong noise and scene structure. Most works simply concatenate or multiply event features and image features (Pini, Borghi, and Vezzani 2018; Pini et al. 2019; Han et al. 2020; Paikin et al. 2021; He et al. 2022; Tomy et al. 2022). Han *et al.* (Han et al. 2020) fuse the low-dynamic-range image features and the event features for high-dynamic-range image reconstruction. He *et al.* (He et al. 2022) concatenate the optical flow, event, and image together for video interpolation. Tomy *et al.* (Tomy et al. 2022) multiply the multi-scale image and event features for robust object detection. Very recently, Cho *et al.* (Cho and Yoon 2022) design an event image fusion module for depth estimation. The differences between these methods and our work are: 1) we deal with the LLVE task; 2) they mainly use image features to densify sparse event features, while in our work, event and image features enhance each other.

Low-Level Vision Transformers. In recent years, Transformers have been applied to many fields (Chen et al. 2022a; Liu et al. 2022c) including low-level vision, and they can be divided into multi-task and task-specific Transformers. For multi-task Transformers, Chen *et al.* (Chen et al. 2021) and Liu *et al.* (Liu et al. 2022a) propose IPT and TAPE, respectively, which are pre-trained on several low-level vision tasks. Task-specific Transformers also obtain state-of-the-art performances in many applications including super-resolution (Yang et al. 2020a; Chen et al. 2022b), deraining (Xiao et al. 2022; Liu et al. 2022b; Jiang et al. 2022a), inpainting (Zeng, Fu, and Chao 2020), and HDR (Chen et al. 2023). In our paper, we use Transformer in the enhancement stage to deal with the areas in the videos without fast motion.

3 Method

In this section, we first describe our overall pipeline and then present the three stages of our method, including event synthesis and restoration, event-image fusion, and low-light enhancement.

As shown in Fig. 3, firstly, the original event voxels are obtained by synthesizing light sequences (Sec. 3.1). And the restored event voxels are generated from a U-Net (Sec. 3.1). The image and the restored event voxels are encoded to deep
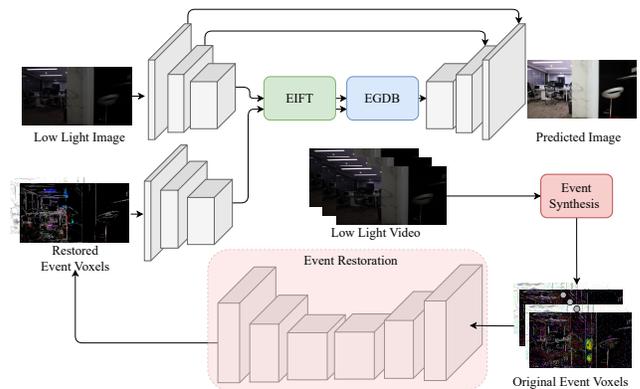
Figure 3: Architecture of our overall network. The original event voxels pass the event restoration network and generate the restored event voxels. And the low-light image and the restored event voxels are combined to obtain the final enhanced normal-light results.

features, fused by the EIFT module (Sec. 3.2) and enhanced by the EGDB module (Sec. 3.3). Finally, the decoder outputs the enhanced images.

3.1 Events Synthesis and Restoration

Event Synthesis. To synthesize original event voxels, we have the following three steps: frame upsampling, event generation, and event voxel generation. In the **first** step, we use an off-the-shelf up-sampling algorithm (Xiang et al. 2020) to get N' up-sampled frames with the same resolution (with N' adaptively chosen) from N frames ($N' > N$). In the **second** step, two adjacent frames (from N' frames) are subtracted to obtain $d_{x,y,t}$, and whether an event e_i is generated is determined by the difference. If d_{x_i,y_i,t_i} exceeds a threshold, we generate an event $e_i = (x_i, y_i, t_i, p_i)$, where (x_i, y_i) and t_i are the location and time of the event and $p_i = \pm 1$ ($p_i d_{x_i,y_i,t_i} > 0$). We use one threshold for low-light events (set to 2) and another for normal-light events (set to 5). This is because setting a higher threshold when generating low-light events will lose some useful events. In the **third** step, in order to make the events processed by neural networks better, it is necessary to convert the discrete event values $\{0, +1, -1\}$ into floating-point event voxels. Following (Zhu et al. 2019; Weng, Zhang, and Xiong 2021), we generate the voxels $E \in \mathbb{R}^{B \times H \times W}$ as:

$$E_k = \sum_{i=0}^{n} p_i \max\left(0, 1 - \left|k - \frac{t_i - t_0}{t_n - t_0}(B-1)\right|\right), \ k \in \{1, ..., B\}, \tag{1}$$

where t_0 and t_n denote the start and end moments of the events, respectively. B equals to $2N \times 3$, where N is the input frame number, 2 corresponds to the positive and negative event voxels, and 3 corresponds to the r-g-b color channels. This equation shows that we use temporal bilinear interpolation to obtain B voxels of the event temporal information.

Event Restoration. In the event restoration, we aim to generate restored events from the original events using an CNN. To the best of our knowledge, there is no available neural network to transfer low-light events to normal-light
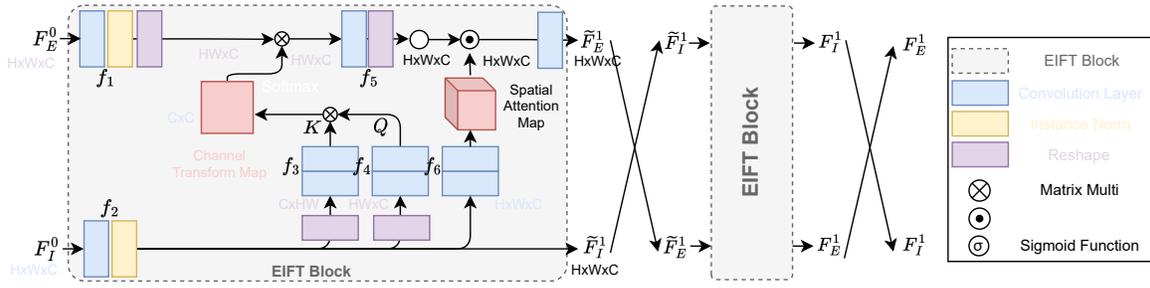
Figure 4: Architecture of the EIFT module. All the ReLU layers are omitted for clarity.

event voxels. We design a network (see the supplementary materials) that is able to predict an event probability map $P \in \mathbb{R}^{B \times H \times W}$ and an event voxel map $V \in \mathbb{R}^{B \times H \times W}$ simultaneously. Finally, the restored event voxels are calculated as:

$$E^r = M(P)V, \quad (2)$$

where the element $M_i(P_i)$ of $M(P)$ is 1 when $P_i \geq 0.5$ and 0 otherwise. The loss functions that constrain M and V will be described in Sec. 3.4.

3.2 Event-Image Fusion

Different from most event-image fusion methods (Pini, Borghi, and Vezzani 2018; Pini et al. 2019; Han et al. 2020; Paikin et al. 2021; He et al. 2022), which combine event and image features by concatenating them simply, we design a novel Event and Image Fusion Transform (EIFT) module for better fusing the sparse event features and dense image features. Our EIFT is based on the mutual attention mechanism, where the event features and the image features modulate each other. Specifically, the event feature modulates the image feature by hinting where fast motion happens and distinguishing between strong noise and image details; the image feature modulates the event feature by introducing the color or semantic information. In our EIFT, such modulation is achieved by generating the channel transform map and the spatial attention map from the modulated features. The maps have values between 0 and 1 indicating the importance of every element.

Fig. 4 is the designed network structure of EIFT. Each EIFT module contains two EIFT blocks. In the first block, F_E^0 and F_I^0 serve as the main feature and the modulation feature, respectively. Inspired by (Zamir et al. 2022), to reduce computational complexity, our modulation is divided into cross-channel transform (CCT) and element-wise product (EWP). In CCT, the modulation feature passes two parallel convolution blocks and generates $Q \in \mathbb{R}^{HW \times C}$ and $K \in \mathbb{R}^{C \times HW}$. The dot product of K and Q is performed, and the Softmax function is applied to generate a $C \times C$ channel transform map. Then it is multiplied by the main feature. The whole CCT can be formulated as:

$$X = f_1(F_E) \cdot \text{Softmax}\{f_3(f_2(F_I)) \cdot f_4(f_2(F_I))\}. \quad (3)$$

In EWP, the modulation feature first passes through some convolution blocks to generate a spatial attention map. And the element-wise multiplication product between the spatial

attention map and the main feature is carried out, which can be formulated as:

$$F = \sigma\left(f_5(X)\right) \odot f_6(f_2(F_I)). \quad (4)$$

f_i in Eqn. 3 and Eqn. 4 are the convolution blocks indicated in Fig. 4. σ and $\odot$ denote the Sigmoid function and the element-wise production, respectively.

In the second EIFT block, the inputs (outputs of the first EIFT block) are swapped, where $\tilde{F}_E^n$ is the modulation feature and $\tilde{F}_I^n$ is the main feature. The final outputs of the EIFT module are F_E^n and F_I^n for the n-th EIFT module. In this paper, n is set to 2.

3.3 Low-Light Enhancement

In low-light image/video enhancement, illumination estimation/enhancement is important. The previous LLVE methods often process the whole image using convolution networks. However, the difficulties for illumination enhancement in areas with little motion and in areas with fast motion are quite different. Especially when there is fast motion, where the illumination will change, the estimated illumination is often inaccurate. We make two improvements: 1) we use two branches to deal with these two kinds of areas (with fast motion, and without fast motion) respectively for illumination enhancement; 2) we use a Transformer to enhance the brightness.

Mask Generation. Given the restored event voxels, $E^r \in \mathbb{R}^{(2N \times 3) \times H \times W}$, generated from the first stage (Sec. 3.1), we use the positive voxels $E^{r,+} \in \mathbb{R}^{(N \times 3) \times H \times W}$ for generating the mask M', which is computed by:

$$M_c = \max\{E_{c,1}^{r,+}, E_{c,2}^{r,+}, ..., E_{c,N}^{r,+}\}, \ c \in \{r, g, b\}, \quad (5)$$

$$M' = \max\{M_r, M_g, M_b\} \in \mathbb{R}^{H \times W}, \quad (6)$$

where M' is then resized to the same resolution as the input of EGDB. Finally, M' is changed to a binary 0-1 mask with a threshold 0.9.

Network Details. In Fig. 5, EGDB consists of a global branch and a local branch. In the global branch, we first concatenate the F_E^n and the masked image feature $(1 - M')F_I^n$ from the event-image fusion stage and perform the adaptive average pooling to get the feature $F \in \mathbb{R}^{32 \times 32 \times 2C}$. Because of the color uniformity and texture sparsity of these
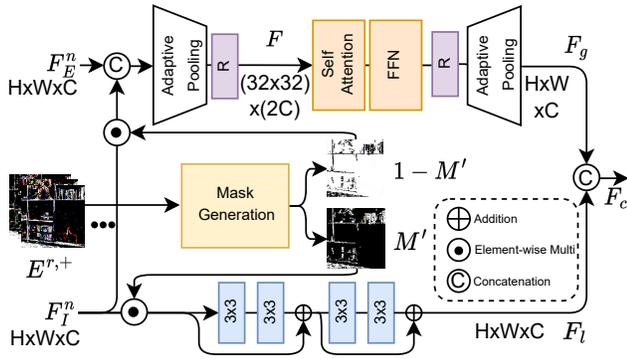
Figure 5: Architecture of the EGDB module with the global branch (upper part) and the local branch (bottom part). All the ReLU layers are omitted for clarity.

kinds of areas, down-sampling can reduce network parameters without loss of performance. The feature F is then reshaped, splited into m feature patches $F_i \in \mathbb{R}^{p^2 \times 2C}, i = \{1, \ldots, m\}$, where p is the patch size and $m = \frac{H}{p} \times \frac{W}{p}$, and passes a self-attention layer and a feed-forward layer of the conventional Transformer [1]. The global branch outputs F_g with the same size of F_E^n. The local branch with two residual blocks receives $M'F_I^n$ and outputs the feature F_l. Finally, the EGDB module outputs F_c, which is the concatenation of F_g and F_l.

3.4 Training and Loss Functions

The training contains two stages. In the first stage, with the low-light event voxels and the normal-light event voxels, we train the event restoration network using

$$L_{s1} = L_m + \lambda_1 L_v, \qquad (7)$$

where the first term is the binary cross-entropy loss between the predicted event probability map P and the binary normal-light voxel mask M^G (generated by the normal-light event voxel ground truth G with a threshold 0.1):

$$L_m = -\frac{1}{HW} \sum_{i=1}^{HW} M_i^G \log(P_i) + (1-M_i^G)\log(1-P_i), \qquad (8)$$

where $P_i \in P$ and $M_i^G \in M^G$. The second term L_v is the L_1 loss between E^r and G.

After the first stage, the parameters of the event restoration network are fixed, and the other networks are trained. The loss function is defined as:

$$L_{s2} = L_1 + \lambda_2 L_{vgg}, \qquad (9)$$

where L_1 and L_{vgg} are respectively the L1 loss and the perceptual loss (Justin, Alexandre, and Li 2016) between the predicted normal-light image and its ground truth.

4 Experiments and Analysis

In this section, we conduct an ablation study and compare with state-of-the-art methods.

[1]See the supplementary materials for more network details.

		PSNR↑	SSIM↑	LPIPS↓	Network Size
Image-based method	DeepUPE	21.82	0.68	–	0.59M
	ZeroDCE	20.06	0.61	–	0.08M
	DeepLPF	22.48	0.66	–	1.77M
	DRBN	22.31	0.65	–	1.12M
	Uformer*	23.46	0.72	0.202	20.4M
	STAR*	23.39	0.70	0.283	0.03M
	SCI*	19.67	0.69	0.298	0.01M
	LLFlow*	24.90	0.78	0.182	5.43M
Video-based method	MBLLVEN	21.79	0.65	0.190	1.02M
	SMID	24.09	0.69	0.213	6.22M
	SMOID	23.45	0.69	0.187	3.64M
	SDSDNet	24.92	0.73	0.138	4.43M
	SGZSL*	23.89	0.70	0.308	28.1M
	Ours	**25.81**	**0.80**	**0.126**	3.51M

Table 1: Quantitative frame-based low-light enhancement comparison on SDSD (Wang et al. 2021). The best results are in bold, and * denotes methods implemented by us using their official models and other results are cited from (Wang et al. 2021).

4.1 Datasets and Implementation Details

Datasets. For a real low-light video dataset, we adopt SDSD (Wang et al. 2021) which contains 37,500 low- and normal-light image pairs with dynamic scenes. SDSD has an indoor subset and an outdoor subset. For a fair comparison, we use the same training/test split as in (Wang et al. 2021). We do not consider the real LLVE datasets which are not released (Jiang and Zheng 2019; Wei et al. 2019) or only contain static scenes (Chen et al. 2019). In order to have experiments on more scenes and data, following the work (Triantafyllidou et al. 2020), we also perform experiments on Vimeo90k (Xue et al. 2019). We select video clips whose average brightness is greater than 0.3 (the highest brightness is 1) as normal-light videos, and then use the method in (Lv, Li, and Lu 2021) to synthesize low-light video sequences [2]. We finally get 9,477 training and 1,063 testing sequences, each with 7 frames, from Vimeo90k.

Implementation Details. We implement our method in the MindSpore (MindSpore 2022) framework, and train and test it on two 3090Ti GPUs. The network parameters are randomly initialized with the Gaussian distribution. In the training stage, the patch size and batch size are 256 and 4, respectively. We adopt the Adam (Kingma and Ba 2015) optimizer with momentum set to 0.9. The input number of frames N is set to 5.

4.2 Comparison with State-of-the-Arts

State-of-the-Art Methods. We compare our work with 7 low-light single image enhancement methods (Deep-UPE (Wang et al. 2019a), ZeroDCE (Guo et al. 2020), DeepLPF (Moran et al. 2020), DRBN (Yang et al. 2020b), STAR (Zhang et al. 2021c), SCI (Ma et al. 2022) and LLFlow (Wang et al. 2022a), one SOTA transformer-based general image restoration method (Uformer (Wang et al. 2022b)), and 5 low-light video enhancement methods (MBLLVEN (Lv et al. 2018), SMID (Chen et al. 2019), SMOID (Jiang and Zheng 2019), SDSDNet (Wang et al.

[2]More details are shown in the supplementary materials.
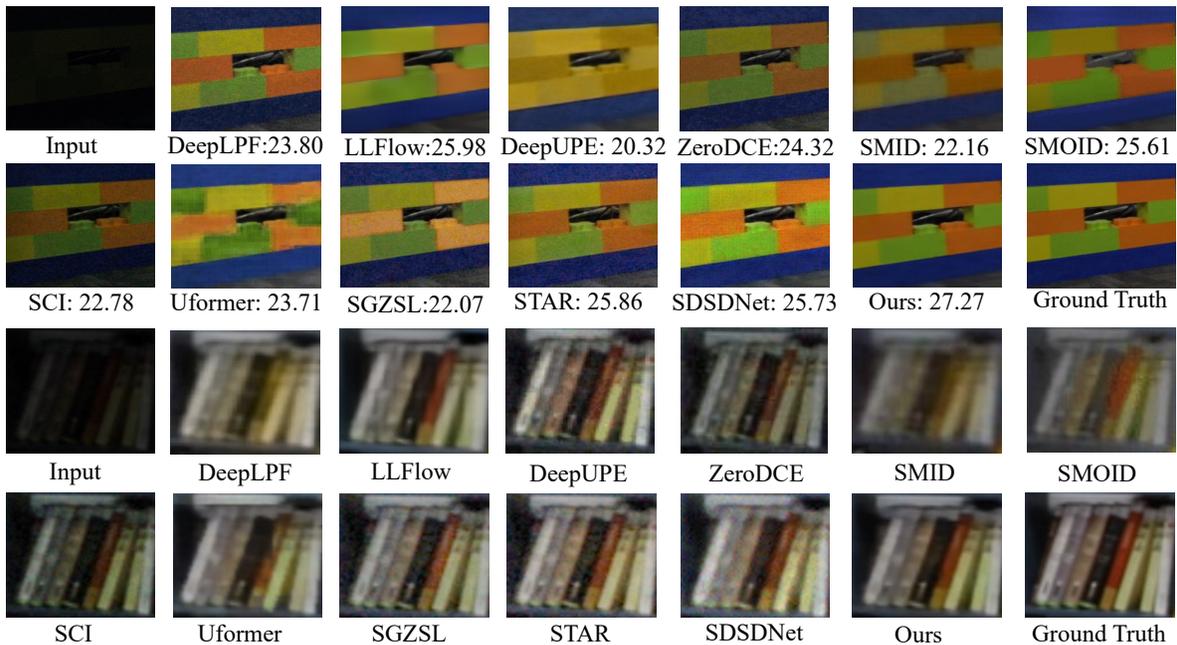
Figure 6: Visual comparison between other SOTA methods and ours on two patches of one example from SDSD. The numbers are PSNR values on the whole image.

Model	PSNR↑	SSIM↑	LPIPS↓	Network size
ZeroDCE	24.56	0.7533	0.202	0.08M
STAR	24.72	0.7595	0.195	0.03M
Uformer	30.29	0.9203	0.058	20.4M
MBLLVEN	27.06	0.8706	0.083	1.02M
SMOID	29.74	0.9239	0.059	3.64M
SDSDNet	29.06	0.9233	0.064	4.43M
Ours	**30.53**	**0.9254**	**0.038**	3.51M

Table 2: Quantitative frame-based low-light enhancement comparison on the Vimeo90K dataset.

Model	MBLLVEN	SMID	SMOID	SDSDNet	SGZSL	Ours
FID↓	0.533	0.552	0.521	**0.368**	0.722	<u>0.370</u>
Warping Error↓	2.98	2.85	<u>2.62</u>	4.36	6.13	**2.61**

Table 3: Quantitative video-based low-light enhancement comparison on the SDSD dataset. The best results are in bold and the second best are underlined.

2021) and SGZSL (Zheng and Gupta 2022)). In these methods, SCI and Zero-DCE are un-supervised, and the models marked by ∗ in Table 1 and all the models in Table 2 are implemented using their official models.

Quantitative Results. For frame-based comparison, we use PSNR and SSIM (Wang et al. 2004), and Learned Perceptual Image Patch Similar (LPIPS) (Zhang et al. 2018) to compare the restored images. For video-based comparison, we adopt FID and Warping Error for video-based methods. These two metrics can well reflect the quality and stability of the predicted videos. As shown in Table 1 and Table

Model	PSNR↑	SSIM↑	LPIPS↓
EIFT → UNet	25.54	0.79	0.135
W/o CCT	25.46	0.78	0.141
W/o EWP	25.52	0.79	0.138
W/o Event Guidance	25.49	0.78	0.143
W/o Global Branch	25.34	0.77	0.167
W/o Local Branch	25.42	0.78	0.162
PCD+EGDB (W/o EG)	24.27	0.73	0.185
Full Model	**25.81**	**0.80**	**0.126**

Table 4: Ablation study on SDSD. EG: event guidance.

2, our method obtains the best performances on both SDSD and Vimeo90K for frame-based comparisons. In Table 3, for video-based comparison, our method gets the 2nd on FID (only 0.002 lower than the best method) and the 1st on Warping Error without using any stability loss like SMID.

Qualitative Results. In Fig. 6, we show visual comparisions on SDSD. Most single-image enhancement methods either produce some noise (DeepLPF, ZeroDCE, Uformer and STAR) or have color shift (DeepUPE and SCI). As for the video-based methods, SMID does not restore the color well, SMOID blurs the details. The deviation of noise map estimation by SDSDNet results in a certain degree of noise in the prediction. With the help of events, our method restores the color and removes the noise simultaneously. We show other visual results on Vimeo90K in the supplementary materials, where ours also outperforms the others.

4.3 Ablation Study

In this section, we do an ablation study based on the SDSD dataset and show the results in Table 4.
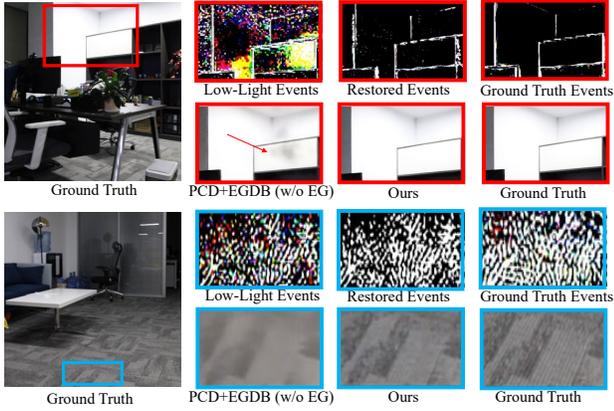
Figure 7: Illustration of how the synthetic events help the restoration of low-light videos. EG: event guidance.

Network Architecture. First, we analyze the effectiveness of the network architecture.

EIFT. We construct three models: 1) EIFT → UNet. We replace the EIFT module with the UNet (Ronneberger, Fischer, and Brox 2015) which has a similar model size with EIFT in the event-image fusion. The two feature maps, F_E^i and F_I^i, $i = 0, 1$, are concatenated and used as the input to the UNet. 2) W/o CCT. We remove the cross-channel transform operation in EIFT. 3) W/o EWP. We remove the element-wise product operation in EIFT. The PSNR value of these models drops 0.27dB, 0.35dB and 0.29dB respectively. This result shows the effectiveness of our EIFT module.

EGDB. In the low-light enhancement stage, we also build three modified models for ablation study. 1) W/o Event Guidance, which means that the map generation block is removed and the input image features are not masked. 2) W/o Global Branch, which removes the global branch in EGDB. Note that we increase the number of the residual blocks in 'W/o Global Branch' to make the network parameter amount equal to the full model. 3) W/o Local Branch, which removes the local branch in this model. The PSNR value of these models drop 0.32dB, 0.47dB, and 0.39dB respectively. The result shows the effectiveness of our EGDB module.

Effectiveness of the Event Guidance. To show the effectiveness of the event guidance, we build a model named 'PCD+EGDB (W/o EG)' in Table 4. The input of this model is five consecutive frames, and the network size is similar to the full model. The first part of 'PCD+EGDB (W/o EG)' is the PCD module in (Wang et al. 2021) and the EGDB module of the second part is without the events (the same as W/o Event Guidance). We can see that the PSNR decreases sharply by 1.54dB, which shows the effectiveness of the event guidance and our pipeline.

4.4 Visualization of the Restored Event Voxels

In Fig. 7, we give some results of restored event voxels to show how the restored events help to enhance the low-light videos. On the low-light voxels in the top row, we can see that the events contain noise and color shift. Our
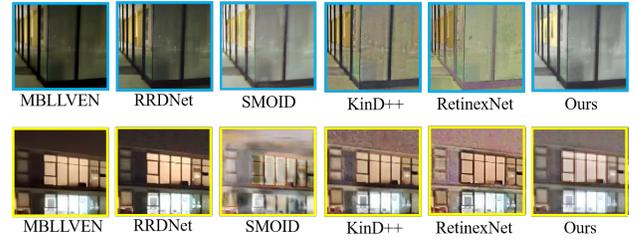
Figure 8: Visual comparison between some SOTA methods and ours on the Loli-Phone dataset.

event restoration network successfully removes these artifacts. Our result has much fewer artifacts compared with the prediction of 'PCD+EGDB (w/o EG)'. The bottom row example demonstrates that with less noise in the restored events, our result is sharper and keeps better details.

4.5 Generalization to Other Real Videos

To verify the generalization ability of our method on other real low-light videos, we also test it and some SOTA low-light enhancement methods (including MBLLVEN, SMOID, KinD++ (Zhang et al. 2021b), RetinexNet (Wei et al. 2018), and RRDNet (Zhu et al. 2020)) on the Loli-Phone dataset (without ground truth) (Li et al. 2021). Two examples are shown in Fig. 8. From the images, we can see that the enhancement of low-light videos by MBLLVEN and RRDNet is limited. KinD++ and RetinexNet produce some color artifacts. The results of SMOID contain multi-frame fusion artifacts (zooming in to see the upper left part of the first result by SMOID). The prediction by our method does not have the issues of those methods. Another advantage of our method is that with the help of events, it better deals with the white balance problem in low light. Under natural light, the floor and wall should appear white or gray. However, the results predicted by other methods are yellowish, while our results are gray.

5 Conclusions

We propose to use synthetic events as the guidance for low-light video enhancement. We design a novel Event and Image Fusion Transform (EIFT) module for event-image fusion and an Event-Guided Dual Branch (EGDB) module for low-light enhancement. Our method outperforms both the state-of-the-art low-light video and single image enhancement approaches. It takes extra time to synthesize the events (about 20% of the total inference time). Future work includes exploring better ways to restore the events in low-light conditions and fusing events and images.

References

Chen, C.; Chen, Q.; Do, M. N.; and Koltun, V. 2019. Seeing Motion in the Dark. *ICCV*.

Chen, C.; Zhou, J.; Wang, F.; Liu, X.; and Dou, D. 2022a. Structure-aware Protein Self-supervised Learning. *ArXiv*, abs/2204.04213.

Chen, H.; Wang, Y.; Guo, T.; Xu, C.; Deng, Y.; Liu, Z.; Ma, S.; Xu, C.; Xu, C.; and Gao, W. 2021. Pre-Trained Image Processing Transformer. *CVPR*.

Chen, R.; Zheng, B.; Zhang, H.; Chen, Q.; Yan, C.; Slabaugh, G.; and Yuan, S. 2023. Improving Dynamic HDR Imaging with Fusion Transformer. *AAAI*.

Chen, X.; Wang, X.; Zhou, J.; and Dong, C. 2022b. Activating More Pixels in Image Super-Resolution Transformer. *ArXiv*, abs/2205.04437.

Cho, H.; and Yoon, K.-J. 2022. Event-image Fusion Stereo using Cross-modality Feature Propagation. *AAAI*.

Dai, P.; Yu, X.; Ma, L.; Zhang, B.; Li, J.; Li, W.; Shen, J.; and Qi, X. 2022. Video Demoireing with Relation-Based Temporal Consistency. *CVPR*.

Dai, T.; Li, W.; Cao, X.; Liu, J.; Jia, X.; Leonardis, A.; Yan, Y.; and Yuan, S. 2021. Wavelet-Based Network For High Dynamic Range Imaging. *ArXiv*, abs/2108.01434.

Ershov, E.; Savchik, A.; Shepelev, D.; Banic, N.; Brown, M. S.; Timofte, R.; and et al. 2022. NTIRE 2022 Challenge on Night Photography Rendering. *CVPRW*.

Gehrig, D.; Gehrig, M.; Hidalgo-Carrió, J.; and Scaramuzza, D. 2020. Video to Events: Recycling Video Datasets for Event Cameras. *CVPR*.

Guo, C.; Li, C.; Guo, J.; Loy, C. C.; Hou, J.; Kwong, S.; and Cong, R. 2020. Zero-Reference Deep Curve Estimation for Low-Light Image Enhancement. *CVPR*.

Han, J.; Yang, Y.; Zhou, C.; Xu, C.; and Shi, B. 2021. Evintsr-net: Event guided multiple latent frames reconstruction and super-resolution. *ICCV*.

Han, J.; Zhou, C.; Duan, P.; Tang, Y.; Xu, C.; Xu, C.; Huang, T.; and Shi, B. 2020. Neuromorphic Camera Guided High Dynamic Range Imaging. *CVPR*.

He, W.; You, K.; Qiao, Z.; Jia, X.; Zhang, Z.; Wang, W.; Lu, H.; Wang, Y.; and Liao, J. 2022. Unlocking the Potential of Event Cameras for Video Interpolation. *CVPR*.

Isobe, T.; Li, S.; Jia, X.; Yuan, S.; Slabaugh, G. G.; Xu, C.; Li, Y.; Wang, S.; and Tian, Q. 2020. Video Super-Resolution With Temporal Group Attention. *CVPR*.

Jiang, H.; and Zheng, Y. 2019. Learning to See Moving Objects in the Dark. *ICCV*.

Jiang, K.; Wang, Z.; Chen, C.; Wang, Z.; Cui, L.; and Lin, C.-W. 2022a. Magic ELF: Image Deraining Meets Association Learning and Transformer. In *ACMMM*.

Jiang, K.; Wang, Z.; Wang, Z.; Yi, P.; Wang, X.; Qiu, Y.; Chen, C.; and Lin, C.-W. 2022b. Degrade is upgrade: learning degradation for low-light image enhancement. In *AAAI*.

Jiang, Z.; Zhang, Y.; Zou, D.; Ren, J.; Lv, J.; and Liu, Y. 2020. Learning event-based motion deblurring. *CVPR*.

Jing, Y.; Yang, Y.; Wang, X.; Song, M.; and Tao, D. 2021. Turning frequency to resolution: Video super-resolution via event cameras. *CVPR*.

Justin, J.; Alexandre, A.; and Li, F.-F. 2016. Perceptual losses for real-time style transfer and super-resolution. *ECCV*.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In Bengio, Y.; and LeCun, Y., eds., *ICLR*.

Li, C.; Guo, C.; Han, L.; Jiang, J.; Cheng, M.-M.; Gu, J.; and Loy, C. C. 2021. Low-Light Image and Video Enhancement Using Deep Learning: A Survey. *TPAMI*.

Li, W.; Xiao, S.; Dai, T.; Yuan, S.; Wang, T.; Li, C.; and Song, F. 2022. SJ-HD^2R: Selective Joint High Dynamic Range and Denoising Imaging for Dynamic Scenes. *ArXiv*, abs/2206.09611.

Liu, L.; Jia, X.; Liu, J.; and Tian, Q. 2020a. Joint Demosaicing and Denoising With Self Guidance. *CVPR*.

Liu, L.; Liu, J.; Yuan, S.; Slabaugh, G.; Leonardis, A.; Zhou, W.; and Tian, Q. 2020b. Wavelet-based dual-branch network for image demoiréing. *ECCV*.

Liu, L.; Xie, L.; Zhang, X.; Yuan, S.; Chen, X.; Zhou, W.; Li, H.; and Tian, Q. 2022a. TAPE: Task-Agnostic Prior Embedding for Image Restoration. *ECCV*.

Liu, L.; Yuan, S.; Liu, J.; Bao, L.; Slabaugh, G.; and Tian, Q. 2020c. Self-adaptively learning to demoiré from focused and defocused image pairs. *NeurIPS*.

Liu, L.; Yuan, S.; Liu, J.; Guo, X.; Yan, Y.; and Tian, Q. 2022b. SiamTrans: Zero-Shot Multi-Frame Image Restoration with Pre-Trained Siamese Transformers. *AAAI*.

Liu, Z.; Ning, J.; Cao, Y.; Wei, Y.; Zhang, Z.; Lin, S.; and Hu, H. 2022c. Video swin transformer. In *CVPR*.

Lv, F.; Li, Y.; and Lu, F. 2021. Attention guided low-light image enhancement with a large scale low-light simulation dataset. *IJCV*.

Lv, F.; Lu, F.; Wu, J.; and Lim, C. 2018. MBLLEN: Low-Light Image/Video Enhancement Using CNNs. *BMVC*.

Ma, L.; Ma, T.; Liu, R.; Fan, X.; and Luo, Z. 2022. Toward fast, flexible, and robust low-light image enhancement. *CVPR*.

MindSpore. 2022. MindSpore. https://www.mindspore.cn/.

Moran, S.; Marza, P.; McDonagh, S.; Parisot, S.; and Slabaugh, G. G. 2020. DeepLPF: Deep Local Parametric Filters for Image Enhancement. *CVPR*.

Mueggler, E.; Rebecq, H.; Gallego, G.; Delbruck, T.; and Scaramuzza, D. 2017. The event-camera dataset and simulator: Event-based data for pose estimation, visual odometry, and SLAM. *IJRR*.

Paikin, G.; Ater, Y.; Shaul, R.; and Soloveichik, E. 2021. EFI-Net: Video Frame Interpolation from Fusion of Events and Frames. *CVPR*.

Pini, S.; Borghi, G.; and Vezzani, R. 2018. Learn to see by events: color frame synthesis from event and rgb cameras. *ArXiv*, abs/1812.02041.

Pini, S.; Borghi, G.; Vezzani, R.; and Cucchiara, R. 2019. Video synthesis from Intensity and Event Frames. *ICIAP*.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. *MICCAI*.

Simon Chane, C.; Ieng, S.-H.; Posch, C.; and Benosman, R. B. 2016. Event-based tone mapping for asynchronous time-based image sensor. *Frontiers in neuroscience*.

Tomy, A.; Paigwar, A.; Mann, K. S.; Renzaglia, A.; and Laugier, C. 2022. Fusing Event-based and RGB camera for Robust Object Detection in Adverse Conditions. *ICRA*.

Triantafyllidou, D.; Moran, S.; McDonagh, S.; Parisot, S.; and Slabaugh, G. 2020. Low Light Video Enhancement using Synthetic Data Produced with an Intermediate Domain Mapping. *ECCV*.

Wang, L.; Gong, Y.; Wang, Q.; Zhou, K.; and Chen, L. 2023. Flora: dual-frequency loss-compensated real-time monocular 3D video reconstruction. In *AAAI*.

Wang, R.; Qing, Z.; Fu, C.-W.; Shen, X.; Zheng, W.-S.; and Jia, J. 2019a. Underexposed Photo Enhancement Using Deep Illumination Estimation. *CVPR*.

Wang, R.; Xu, X.; Fu, C.-W.; Lu, J.; Yu, B.; and Jia, J. 2021. Seeing Dynamic Scene in the Dark: A High-Quality Video Dataset With Mechatronic Alignment. *ICCV*.

Wang, X.; Chan, K. C.; Yu, K.; Dong, C.; and Loy, C. C. 2019b. EDVR: Video Restoration with Enhanced Deformable Convolutional Networks. *CVPR*.

Wang, Y.; Wan, R.; Yang, W.; Li, H.; Chau, L.-P.; and Kot, A. C. 2022a. Low-Light Image Enhancement with Normalizing Flow. *AAAI*.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *TIP*.

Wang, Z.; Cun, X.; Bao, J.; and Liu, J. 2022b. Uformer: A general u-shaped transformer for image restoration. *CVPR*.

Wang, Z. W.; Duan, P.; Cossairt, O.; Katsaggelos, A.; Huang, T.; and Shi, B. 2020. Joint filtering of intensity images and neuromorphic events for high-resolution noise-robust imaging. *CVPR*.

Wei, C.; Wang, W.; Yang, W.; and Liu, J. 2018. Deep retinex decomposition for low-light enhancement. *BMVC*.

Wei, W.; Xin, C.; Cheng, Y.; Xiang, L.; Xuemei, H.; and Tao, Y. 2019. Enhancing Low Light Videos by Exploring High Sensitivity Camera Noise. *ICCV*.

Weng, W.; Zhang, Y.; and Xiong, Z. 2021. Event-based Video Reconstruction Using Transformer. *ICCV*.

Xiang, X.; Tian, Y.; Zhang, Y.; Fu, Y.; Allebach, J. P.; and Xu, C. 2020. Zooming Slow-Mo: Fast and Accurate One-Stage Space-Time Video Super-Resolution. *CVPR*.

Xiao, J.; Fu, X.; Liu, A.; Wu, F.; and Zha, Z.-J. 2022. Image De-raining Transformer. *TPAMI*.

Xue, T.; Chen, B.; Wu, J.; Wei, D.; and Freeman, W. T. 2019. Video Enhancement with Task-Oriented Flow. *IJCV*.

Yang, F.; Guo, B.; Yang, H.; Fu, J.; and Lu, H. 2020a. Learning Texture Transformer Network for Image Super-Resolution. *CVPR*.

Yang, M.-H.; Huang, J.-B.; Shechtman, E.; Yumer, E.; Wang, O.; and Lai, W.-S. 2018. Learning Blind Video Temporal Consistency. *ECCV*.

Yang, W.; Wang, S.; Fang, Y.; Wang, Y.; and Liu, J. 2020b. From Fidelity to Perceptual Quality: A Semi-Supervised Approach for Low-Light Image Enhancement. *CVPR*.

Zamir, S. W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F. S.; and Yang, M.-H. 2022. Restormer: Efficient Transformer for High-Resolution Image Restoration. *CVPR*.

Zeng, Y.; Fu, J.; and Chao, H. 2020. Learning Joint Spatial-Temporal Transformations for Video Inpainting. *ECCV*.

Zhang, F.; Li, Y.; You, S.; and Fu, Y. 2021a. Learning Temporal Consistency for Low Light Video Enhancement from Single Images. *CVPR*.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. *CVPR*.

Zhang, S.; Zhang, Y.; Zhe, J.; Zou, D.; Ren, J.; and Zhou, B. 2020. Learning to See in the Dark with Events. *ECCV*.

Zhang, Y.; Guo, X.; Ma, J.; Liu, W.; and Zhang, J. 2021b. Beyond Brightening Low-light Images. *IJCV*.

Zhang, Z.; Jiang, Y.; Jiang, J.; Wang, X.; Luo, P.; and Gu, J. 2021c. STAR: A Structure-Aware Lightweight Transformer for Real-Time Image Enhancement. *ICCV*.

Zhao, H.; Loy, C. C.; Qiao, Y.; Chan, K. C.; Dong, C.; Wang, X.; and Liu, Y. 2021a. Temporally Consistent Video Colorization with Deep Feature Propagation and Self-regularization Learning. *ArXiv*, abs/2110.04562.

Zhao, H.; Zheng, B.; Yuan, S.; Zhang, H.; Yan, C.; Li, L.; and Slabaugh, G. 2021b. CBREN: Convolutional Neural Networks for Constant Bit Rate Video Quality Enhancement. *TCSVT*.

Zheng, B.; Pan, X.; Zhang, H.; Zhou, X.; Slabaugh, G.; Yan, C.; and Yuan, S. 2022. DomainPlus: cross transform domain learning towards high dynamic range imaging. *ACMMM*.

Zheng, B.; Yuan, S.; Slabaugh, G.; and Leonardis, A. 2020. Image demoireing with learnable bandpass filters. *CVPR*.

Zheng, B.; Yuan, S.; Yan, C.; Tian, X.; Zhang, J.; Sun, Y.; Liu, L.; Leonardis, A.; and Slabaugh, G. 2021. Learning frequency domain priors for image demoireing. *TPAMI*.

Zheng, S.; and Gupta, G. 2022. Semantic-guided zero-shot learning for low-light image/video enhancement. *WACVW*.

Zhu, A.; Zhang, L.; Shen, Y.; Ma, Y.; Zhao, S.; and Zhou, Y. 2020. Zero-Shot Restoration of Underexposed Images via Robust Retinex Decomposition. *ICME*.

Zhu, A. Z.; Yuan, L.; Chaney, K.; and Daniilidis, K. 2019. Unsupervised event-based learning of optical flow, depth, and egomotion. *CVPR*.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. *ICCV*.