Low-Light Video Enhancement with Synthetic Event Guidance

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

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

Abstract

Low-light video enhancement (LLVE) is an important yet challenging task with many applications such as photography 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 \textit{normal-light} images/videos/events.

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

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.

\begin{itemize}
\item Part of the work was done during an internship in Huawei Noah’s Ark Lab.
\item Corresponding author
\end{itemize}

Copyright © 2023, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.
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:

• 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.

• We design an Event and Image Fusion Transform (EIFT) module for event-image fusion and an Event-Guided Dual Branch (EGDB) module for low-light enhancement.

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. (Wang et al. 2021) 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 Chan et al. 2016; Han et al. 2020). Ours is the first to explore the resolution (Han et al. 2021; Jing et al. 2021), joint filter-}
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:

\[E^r = M(P)V,\]

where the element \(M_i(P)\) 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 \(Q \in \mathbb{R}^{H \times W \times C}\) and \(K \in \mathbb{R}^{C \times H \times W}\). 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:

\[X = f_1(F_E) \cdot \text{Softmax}\{f_3(f_2(F_I)) \cdot f_4(f_2(F_I))\}.\]  (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:

\[F = \sigma (f_5(X)) \odot f_6(f_2(F_I)).\]  (4)

\(f_i\) in Eqn. 3 and Eqn. 4 are the convolution blocks indicated in Fig. 4. \(\sigma\) 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 \(F_E^n\) is the modulation feature and \(F_I^n\) 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:

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

\[M' = \max\{M_r, M_g, M_b\} \in \mathbb{R}^{H \times W},\]  (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 \(F_E^2\) and the masked image feature \((1 - M')F_I^2\) from the event-image fusion stage and perform the adaptive average pooling to get the feature \(F' \in \mathbb{R}^{2^{12} \times 32 \times 2C}\). 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, split into $m$ feature patches $F_i \in \mathbb{R}^{p^2 \times 2C}$, $i = 1, \ldots, m$, where $p$ is the patch size and $m = \frac{H \times W}{p}$, 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 $F_E$. The local branch with two residual blocks receives $M^c F_E$ 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

$$L_{s1} = L_m + \lambda_1 L_v,$$  

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):

$$L_m = -\frac{1}{HW} \sum_{i=1}^{HW} M^G_i \log (P_i) + (1 - M^G_i) \log (1 - P_i),$$  

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

$$L_{s2} = L_1 + \lambda_2 L_{egg},$$  

where $L_1$ and $L_{egg}$ are respectively the $L_1$ 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.

---

1See the supplementary materials for more network details.

---

### 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).

<table>
<thead>
<tr>
<th>Method</th>
<th>PSNR</th>
<th>SSIM</th>
<th>LPIPS</th>
<th>Network Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>DeepUPE</td>
<td>21.82</td>
<td>0.68</td>
<td>–</td>
<td>0.59M</td>
</tr>
<tr>
<td>ZeroDCE</td>
<td>20.06</td>
<td>0.61</td>
<td>–</td>
<td>0.08M</td>
</tr>
<tr>
<td>DeepLPF</td>
<td>22.48</td>
<td>0.66</td>
<td>–</td>
<td>1.77M</td>
</tr>
<tr>
<td>DRBN</td>
<td>22.31</td>
<td>0.65</td>
<td>–</td>
<td>1.12M</td>
</tr>
<tr>
<td>Uformer*</td>
<td>23.46</td>
<td>0.72</td>
<td>0.202</td>
<td>20.4M</td>
</tr>
<tr>
<td>STAR*</td>
<td>23.39</td>
<td>0.70</td>
<td>0.283</td>
<td>0.03M</td>
</tr>
<tr>
<td>SCI</td>
<td>19.67</td>
<td>0.69</td>
<td>0.298</td>
<td>0.01M</td>
</tr>
<tr>
<td>LLFlow*</td>
<td>24.90</td>
<td>0.78</td>
<td>0.182</td>
<td>5.43M</td>
</tr>
<tr>
<td>Ours</td>
<td><strong>25.81</strong></td>
<td><strong>0.80</strong></td>
<td><strong>0.126</strong></td>
<td><strong>3.51M</strong></td>
</tr>
</tbody>
</table>

### 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 (DeepUPE (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. 2021)).

---

2More 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.

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

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

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.

<table>
<thead>
<tr>
<th>Model</th>
<th>MBLLVEN</th>
<th>SMOID</th>
<th>SMOID</th>
<th>SDSDNet</th>
<th>SGZSL</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FID↓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EIFT → UNet</td>
<td>0.533</td>
<td>0.552</td>
<td>0.521</td>
<td>0.368</td>
<td>0.722</td>
<td>0.370</td>
</tr>
<tr>
<td>W/o CCT</td>
<td>2.98</td>
<td>2.85</td>
<td>2.65</td>
<td>4.36</td>
<td>6.13</td>
<td>2.81</td>
</tr>
<tr>
<td>W/o EWP</td>
<td>25.46</td>
<td>0.78</td>
<td>0.141</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>25.52</td>
<td>0.79</td>
<td>0.138</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W/o Event Guidance</td>
<td>25.49</td>
<td>0.78</td>
<td>0.143</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W/o Global Branch</td>
<td>25.34</td>
<td>0.77</td>
<td>0.167</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W/o Local Branch</td>
<td>25.42</td>
<td>0.78</td>
<td>0.162</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PCD+EGDB (W/o EG)</td>
<td>24.27</td>
<td>0.73</td>
<td>0.185</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Full Model</td>
<td><strong>25.81</strong></td>
<td><strong>0.80</strong></td>
<td><strong>0.126</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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 SMOID.

**Qualitative Results.** In Fig. 6, we show visual comparisons 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, SMOID 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.
Network Architecture. First, we analyze the effectiveness of the network architecture.

EIFT. We construct three models: 1) EIFT $\rightarrow$ 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, $F_E$ and $F_i$, $i = 0, 1$, 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 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++, 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.

Acknowledgements

Part of this work was supported by the National Natural Science Foundation of China under Contract 61632019. We gratefully acknowledge the support of MindSpore, CANN (Compute Architecture for Neural Networks) and Ascend AI Processor used for this research. We also acknowledge
the support of GPU cluster built by MCC Lab of Information Science and Technology Institution, USTC.

**References**

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


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


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


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.*


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, Z.; Ning, J.; Cao, Y.; Wei, Y.; Zhang, Z.; Lin, S.; and Hu, H. 2022e. 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.*


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.*


Flora: dual-frequency loss-compensated real-time monocular vision.


---

1700