Learning Event-Driven Video Deblurring and Interpolation

Songnan Lin[1]*, Jiawei Zhang[2] **, Jinshan Pan[3], Zhe Jiang[2], Dongqing Zou[2], Yongtian Wang[1], Jing Chen[1] **, and Jimmy Ren[2]

[1] Beijing Institute of Technology, Beijing, China
[2] SenseTime Research, Shenzhen, China
[3] Nanjing University of Science and Technology, Nanjing, China

Abstract. Event-based sensors, which have a response if the change of pixel intensity exceeds a triggering threshold, can capture high-speed motion with microsecond accuracy. Assisted by an event camera, we can generate high frame-rate sharp videos from low frame-rate blurry ones captured by an intensity camera. In this paper, we propose an effective event-driven video deblurring and interpolation algorithm based on deep convolutional neural networks (CNNs). Motivated by the physical model that the residuals between a blurry image and sharp frames are the integrals of events, the proposed network uses events to estimate the residuals for the sharp frame restoration. As the triggering threshold varies spatially, we develop an effective method to estimate dynamic filters to solve this problem. To utilize the temporal information, the sharp frames restored from the previous blurry frame are also considered. The proposed algorithm achieves superior performance against state-of-the-art methods on both synthetic and real datasets.

1 Introduction

Slow-motion analysis of fast-moving objects is crucial for numerous applications but challenging for conventional intensity cameras which only capture low frame-rate blurry videos. To catch the high-speed motion, some recent works, *e.g.* [10,9], attempt to generate a high frame-rate video given a low frame-rate blurry one by deblurring [24,26,28] and interpolation [17,13,1]. Despite their success in certain scenarios, they may fail to deal with severely-blurred videos (see Fig. 1(e)).

Instead of purely relying on an intensity camera, this work utilizes event-based one with a high temporal resolution to compensate for the lost information in intensity frames. Event cameras [5,12] are biologically-inspired sensors capable of asynchronously encoding the changes of pixel intensity, *i.e.*, events, with microsecond accuracy. Significant efforts [3,2,15,21] have been devoted to

* This work was done when Songnan Lin was an intern at SenseTime.
** Corresponding authors: `zhjw1988@gmail.com`; `chen74jing29@bit.edu.cn`

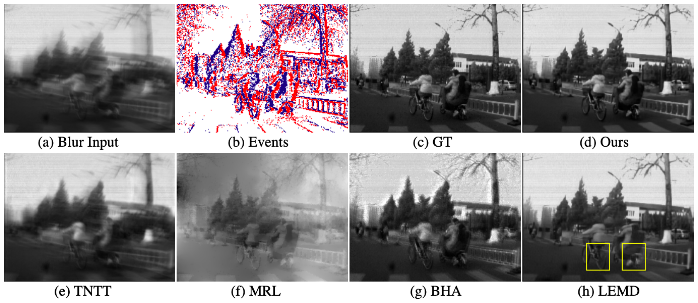


Fig. 1. Challenging case for video reconstruction. (a) The input blurry image. (b) The corresponding event data. The color pair (red, blue) represents its polarity $(1, -1)$ throughout this paper. (c) Ground truth. (d) Our reconstruction result. (e) Result of the image-based video construction [9]. (f) Result of the event-based video generation [15]. (g) Result of conventional BHA [18]. (h) Result of deep learning-based LEMD [8]. The proposed method restores high-quality images via an end-to-end network based on the physical event-based video reconstruction model.

directly converting event streams into intensity videos. However, videos reconstructed from these event-dependent solutions tend to lack textures and seem to be non-photorealistic without intensity information (see Fig. 1(f)).

Therefore, it would be desirable to use both advantages of the intensity and event-based sensors for high-speed video generation. Little attention [6,22,23] has been paid to considering both sources of information. However, as they do not take blur into consideration, the generated videos are blurry sometimes. To solve this problem, Pan *et al.* [18] physically model the relationship among a blurry image, events and latent frames and propose an Event-based Double Integral (EDI) model. Therefore, sharp latent images can be obtained given blurry frames and corresponding event streams. After deblurring, other latent video frames are interpolated from the above initial deblurred one by estimating the residuals between them from the events. This method naturally connects intensity images and event data and shows promising results on high frame-rate video generation. However, as the triggering threshold of an event camera varies spatially and temporally with hardware and scene conditions [6,4,20], it is less effective to consider it as constant as in [18], which introduces strong accumulated noises (see Fig. 1(g)). Jiang *et al.* [8] propose to utilize the large capacity of deep convolutional neural networks (CNNs) to refine the estimated frames from [18] and recover finer details. However, as the deblurring and refinement are separately considered, their method fails to make full use of the model capacity of the CNNs, which makes it less effective for high-speed video generation (see Fig. 1(h)). Moreover, the algorithms [18,8] above bring one blurry frame alive without exploiting the additional information from previous frames.

In this paper, we propose an effective event-driven video deblurring and interpolation algorithm to generate sharp high frame-rate videos based on deep CNNs and the physical model of event-based video reconstruction. Motivated by

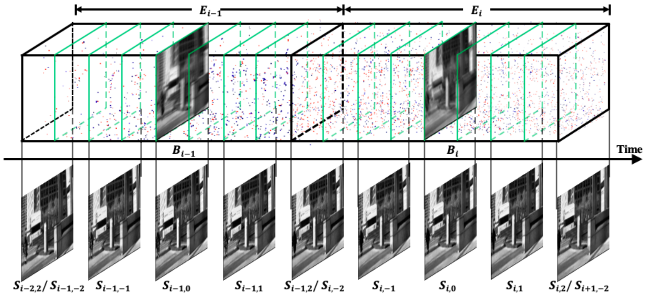


Fig. 2. Sample frames reconstructed from the proposed method. Given a low frame-rate blurry video $\{B_i\}$ and the corresponding event streams $\{E_i\}$ during the exposure time, the proposed method recovers a sharp video $\{S_{i,j}\}_{j\in[-N,N]}$ in a $2N$-time frame rate than the original. $N = 2$ in this example.

[18] which estimates the residual between sharp and blurry images for deblurring as well as that between sharp frames for interpolation, we propose to use a deep CNN to effectively predict them. Moreover, as the triggering threshold is spatially variant, it is inappropriate to use a uniform one as in [18]. Therefore, we propose to use the dynamic filtering layer [7,17,14] to handle this spatially variant threshold. Besides, the proposed network can also help remove the noises from the events when predicting the residuals. To better exploit the additional information across the frames, we further utilize the previously recovered frames together with the previous blurry frame as well as the event stream to estimate current frames, which can enforce the temporal consistency. Our method incorporates the physical properties of event-based video reconstruction compactly and can be trained in an end-to-end manner.

The main contributions of this paper are summarized as follows:

- We propose an end-to-end trainable neural network to generate high-speed videos from the hybrid intensity and event-based sensors. Our algorithm hinges on the physical event-based video reconstruction model with a compact network architecture.
- We propose to use dynamic filtering to handle the events triggered by the spatially variant threshold.
- We quantitatively and qualitatively evaluate our network on both synthetic and real-world videos and show that it performs favorably against state-of-the-art high-speed video generation algorithms.

2 Motivation

Given a low frame-rate blurry video $\{B_i\}_{i\in\mathbb{N}}$ and the corresponding event streams $\{E_i\}_{i\in\mathbb{N}}$ captured during the exposure as shown in Fig. 2, we aim to reconstruct a sharp video with a $2N$-time frame rate than the original. Let $\{S_{i,j}\}_{i,j\in\mathbb{N}}$ denotes the recovered video, where $j \in [-N, N]$ indicates the j^{th} sharp frame within the exposure of the i^{th} blurry frame. The proposed event-based video deblurring and interpolation algorithm is motivated by two observations. First, the intensity residual between the latent sharp images as well as that between sharp and blurry images are both the integral of the events. As a result, we can use the network to estimate accurate integrals from noisy events and then reconstruct high frame-rate sharp videos. Second, even though the intensity residuals can be estimated from the integrals of events, the triggering threshold c_m is spatially and temporally variant. We propose to integrate dynamic filters [7] to handle this spatially variant issue. This section will discuss the above motivations in details.

2.1 Physical Model of Event-based Video Reconstruction

To better motivate our algorithm, we first revisit the physical model of event-based video reconstruction.

Once a log intensity change exceeds a preset threshold c_m, an event e_m[4] is triggered, represented as

$$e_m = (x_m, y_m, t_m, p_m), \tag{1}$$

in which x_m, y_m and t_m denote the spatio-temporal coordinates of the m^{th} event respectively and $p_m \in \{-1, 1\}$ denotes the direction (decrease or increase) of the change. Regardless of quantization, the sum of the events captured in a time interval represents the proportional change in intensity. And thus, given an interval $\Omega_{i,j\to i',j'} = [iT + \frac{j}{2N}T, i'T + \frac{j'}{2N}T]$ of events e_m and a latent sharp frame $S_{i,j}$, we can reconstruct the latent sharp frame $S_{i',j'}$ at pixel (x, y) using:

$$\begin{aligned} S_{i',j'}(x,y) &= S_{i,j}(x,y) \cdot \exp(\sum_{t_m \in \Omega_{i,j\to i',j'}} c_m \cdot p_m \cdot 1(x_m, y_m, x, y)) \\ &= S_{i,j}(x,y) \cdot I_{i,j\to i',j'}(x,y), \end{aligned} \tag{2}$$

in which T denotes the exposure time of blurry frames, $\cdot$ is Hadamard product, the indicator function $1(\cdot)$ equals to 1 if $x_m = x$ & $y_m = y$, and 0 otherwise, and $I_{i,j\to i',j'}$ represents the intensity residual between $S_{i,j}$ and $S_{i',j'}$.

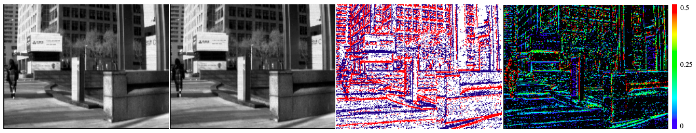
For the blurry image B_i, it can be modeled as the average of discrete latent sharp frames $S_{i,j}$ by:

$$B_i(x,y) = \frac{1}{2N+1} \sum_{j=-N}^{N} S_{i,j}(x,y). \tag{3}$$

Then, we can represent B_i according to Eq. 2 and Eq. 3 as:

$$\begin{aligned} B_i(x,y) &= S_{i,j_0}(x,y) \cdot [\frac{1}{2N+1} \sum_{j=-N}^{N} \exp(\sum_{t_m \in \Omega_{i,j_0\to i,j}} c_m \cdot p_m \cdot 1(x_m, y_m, x, y))] \\ &= S_{i,j_0}(x,y) \cdot D^{-1}_{i\to i,j_0}(x,y), \end{aligned} \tag{4}$$

[4] When $t_m \in \Omega_{i,-N\to i,N}$, e_m is in the event stream E_i.



(a) Frame 1 (b) Frame 2 (c) Events (d) Average Threshold

Fig. 3. Demonstration of the spatially variant event triggering threshold. Given two latent sharp frames (a)(b) and their interval of the events (c) captured with an event camera, we estimate the average threshold of each pixel in this interval using Eq. 2. The valid threshold where events occur is spatially variant across the image plane.

where S_{i,j_0} is the key latent sharp frame related to the blurry frame B_i and $D_{i\to i,j_0}$ is the intensity residual between B_i and S_{i,j_0} and it is actually the discrete version of the Event-based Double Integral (EDI) in [18].

Therefore, it is physically possible to first deblur latent keyframe S_{i,j_0} based on Eq. 4, and then interpolate all other video frames $S_{i,j}$ using Eq. 2. Also, Eq. 2 can be used to generate $S_{i,j}$ from the previously estimated latent frames $S_{i-1,j}$. In [18], they estimate the residual I and D directly according to Eq. 2 and Eq. 4 from the events e. However, the estimation is inaccurate since the events contain severe noises. In this paper, we propose to use a deep neural network to predict the residuals by utilizing its strong capacity and flexibility to compensate for the imperfectness of event data.

2.2 Spatially Variant Triggering Threshold

In previous works, *e.g.* [18], they estimate a fixed triggering threshold and apply it to the whole frame sequence. However, this threshold c_m is both spatially and temporally variant according to [6,4,20]. As can be seen in Fig. 3(d), the estimated thresholds given sharp frames and the respective events by Eq. 2 are not uniform. Therefore, it is inappropriate to use a network composed of convolution layers which are spatially invariant to estimate the residual I and D. We propose to integrate dynamic filters [7], which are estimated at every position, to handle this spatially variant issue.

3 Proposed Methods

3.1 Network Architecture

The overall framework of the proposed video deblurring and interpolation algorithm is illustrated in Fig. 4. It consists of four parts:

- *Residual Estimation:* It aims to estimate the residuals, including $D_{i\to i,0}$ and $I_{i-1,j\to i,0}$ for keyframe deblurring and $I_{i,0\to i,j}$ for video frame interpolation.
- *Keyframe Deblurring:* It utilizes the learned residual $D_{i\to i,0}$ and the blurry frame B_i to estimate a keyframe $C_{i,0}$ via Eq. 4. Then it generates $2N$ keyframes $P_{i,0,j}$ from $I_{i-1,j\to i,0}$ and $2N$ previously recovered sharp frames

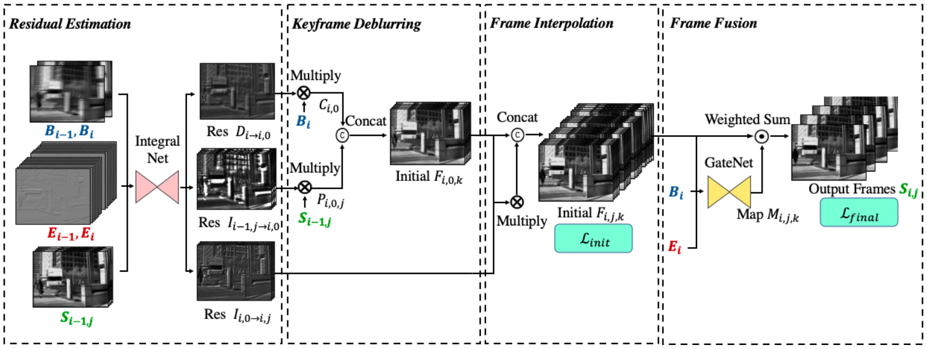


Fig. 4. Overview of our framework. For $2N$-time frame-rate video reconstruction, the previous and current blurry frames B_{i-1}, B_i, their corresponding event streams E_{i-1}, E_i and $2N$ previously recovered sharp frames $S_{i-1,j}$ are fed into an *IntegralNet* to predict the residuals $D_{i\to i,0}$, $I_{i-1,j\to i,0}$ and $I_{i,0\to i,j}$. Given the learned residuals, an initial deblurred keyframe $C_{i,0}$ is estimated from the blurry frame via Eq. 4. Moreover, with $2N$ previously recovered sharp frames, the other initial sharp keyframes $P_{i,0,j}$, where $j \in (-N, N]$, are inferred via Eq. 2. Therefore we obtain $2N+1$ initial keyframes, denoted as $F_{i,0,k}$ by concatenating $C_{i,0}$ and $P_{i,0,j}$. Afterward, the other initial latent sharp frames $F_{i,j,k}$, where $j \in (-N,0) \cup (0,N]$ and $k \in [0,2N]$, are interpolated from $F_{i,0,k}$ via Eq. 2. At last, to adaptively select the initial reconstructed frames, *GateNet* is utilized to predict the weights $M_{i,j,k}$ and the final results are obtained by weight summation of the initial results. Please see the manuscript for more details.

$S_{i-1,j}$ via Eq. 2, where $j \in (-N, N]$. And there are totally $2N + 1$ initial estimated keyframes $F_{i,0,k}$ which are the concatenation of $C_{i,0}$ and $P_{i,0,j}$.

- *Frame Interpolation:* It interpolates the latent sharp frames $F_{i,j,k}$ from every initial deblurred keyframe $F_{i,0,k}$ and $I_{i,0\to i,j}$ according to Eq. 2, where $j \in (-N,0) \cup (0,N]$.
- *Frame Fusion:* It fuses the $2N+1$ initial sharp frames at (i,j) in an adaptive selection manner and restores the final results $S_{i,j}$ with finer details.

Residual Estimation We estimate the residuals $D_{i\to i,0}$, $I_{i-1,j\to i,0}$ and $I_{i,0\to i,j}$ in Eq. 4 and Eq. 2 via an *IntegtalNet*. As discussed above, we need to deal with the spatially and temporally variant triggering contrast threshold. However, the convolution is translation invariant across the feature plane, which is less effective to solve this problem. We apply the dynamic filtering [7] whose pixel-wise filters are estimated by the dynamic filter generation module in the proposed network. Moreover, the proposed network can also help remove the noises from the events when predicting the residuals.

As shown in Fig. 5(a)(b), the *IntegralNet* is composed of three modules: event feature extraction, dynamic filter generation and multi-residual prediction.

As discussed in Sec. 2.1, the residual D and I are the integral of events. Therefore, the events E_{i-1} and E_i are the only input of the event feature ex-

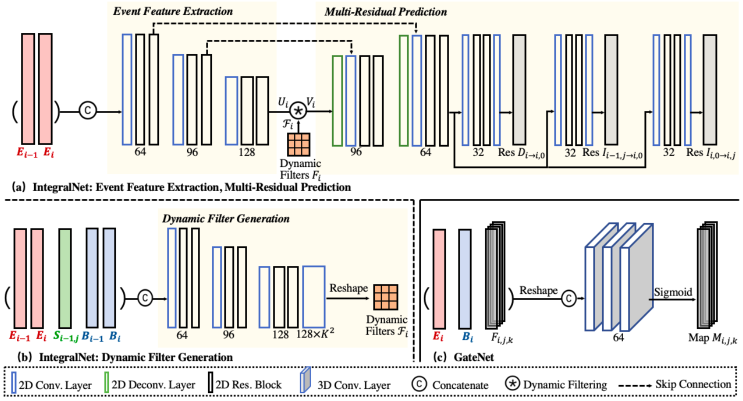


Fig. 5. Structures of the sub-networks. *IntegraltNet* contains the event feature extraction, the dynamic filter generation and the multi-residual prediction. *GateNet* contains three 3D convolution layers. The detailed configurations are provided in the supplemental material.

traction module which extracts features U_i. To feed asynchronous events into the neural network, we divide every event stream E_i into $2N$ equal-time-interval bins. To hold more temporal information, we further divide each bin into M equal-size chunks and stack them as M-channel input images as stated in [25] ($N = 2, M = 2$ in Fig. 2 for example). The stacked event data E_i is passed through three convolution layers followed by two residual blocks. The extracted event features U_i are transformed by dynamic filters in the following process.

We generate different filters for each position in feature maps and perform a spatially variant convolution using the filters. Specifically, for each position (h, w, c) in the extracted feature map $U_i \in \mathbb{R}^{H_U \times W_U \times C_U}$, a specific local filter $\mathcal{F}_i^{(h,w,c)} \in \mathbb{R}^{K\times K\times 1}$ is applied to the region centered around $U_i(h, w, c)$ as

$$V_i(h, w, c) = \mathcal{F}_i^{(h,w,c)} * U_i(h, w, c), \tag{5}$$

where $*$ denotes convolution operation. Filters are dynamically generated given the current and previous blurry frames B_i, B_{i-1}, the corresponding event streams E_i, E_{i-1} and the previously recovered sharp frames $S_{i-1,j}$ by the dynamic filter generation module.

The multi-residual prediction module is used to estimate $D_{i\to i,0}$, $I_{i-1,j\to i,0}$ and $I_{i,0\to i,j}$ taken the transformed event features V_i. As shown in Fig. 5(a), it first upsamples the features back to the full resolution and then generates the residuals respectively. The skip-connections are also adopted in *IntegralNet*.

Keyframe Deblurring With the predicted residuals, we can obtain keyframes from both the current blurry image and the previously recovered images. Specifically, given the predicted $D_{i\to i,0}$, which represents the difference between the blurry image and the keyframe, we can get a keyframe $C_{i,0}$ based on Eq. 4 using:

$$C_{i,0} = B_i \cdot D_{i \to i,0}(B_{i-1}, B_i, E_{i-1}, E_i, S_{i-1,j}; \theta), \tag{6}$$

where θ is the parameters of the *IntegralNet*.

Moreover, using the $2N$ learned residuals $I_{i-1,j \to i,0}$, which indicate the differences between the $2N$ previously recovered sharp frames and the current sharp keyframe, we further estimate $2N$ keyframes according to Eq. 2. Let $P_{i,0,j}$ denote the j^{th} estimated keyframe inferred from the previous j^{th} sharp frame $S_{i-1,j}$, where $j \in (-N, N]$, it is formulated as:

$$P_{i,0,j} = S_{i-1,j} \cdot I_{i-1,j \to i,0}(B_{i-1}, B_i, E_{i-1}, E_i, S_{i-1}; \theta). \tag{7}$$

After all, we obtain a keyframe $C_{i,0}$ from the current blurry frame and $2N$ ones $P_{i,0,j}$ from the previously recovered frames. For simplification, the estimated keyframes are concatenated and represented as $F_{i,0,k}$, in which $k \in [0, 2N]$ indicates the index of the keyframe.

Frame Interpolation Given the $2N+1$ initial deblurred keyframes $F_{i,0,k}$ and the $2N-1$ learned residuals $I_{i,0 \to i,j;j \neq 0}$ between the latent keyframe and the interpolated frames, the interpolated frames can be estimated according to Eq. 2:

$$F_{i,j,k;j \neq 0} = F_{i,0,k} \cdot I_{i,0 \to i,j;j \neq 0}(B_{i-1}, B_i, E_{i-1}, E_i, S_{i-1}; \theta). \tag{8}$$

Frame Fusion After frame interpolation, there are $2N+1$ latent images for each frame. To utilize the merits and remove the flaws of all these latent images $F_{i,j,k}$, we conduct the frame fusion module to integrate them by an adaptive selection scheme. We feed them into *GateNet* to generate a soft gate map $M_{i,j,k}$ together with the blurry frame B_i and the corresponding event stream E_i. We first transform the inputs into four dimensions. The initial results $F_{i,j,k}$ are divided into $2N$ chunks by the timestamps j, generating a feature with the size $(2N+1) \times 2N \times H \times W$, in which H and W represent the resolution of the video frame. As for event data E_i, the events in the intervals between two adjacent sharp frames are stacked together as $M \times 2N \times H \times W$. The blurry input is expanded to $2N$ times along a new dimension as $1 \times 2N \times H \times W$. After that, these transformed features are fed into three 3D convolution layers to generate a gate map $M_{i,j,k}$, as shown in Fig. 5(c). Thus, the final reconstructed frames can be estimated by:

$$S_{i,j} = \sum_{k=0}^{2N} F_{i,j,k} \cdot M_{i,j,k}(B_i, E_i, F_{i,j,k}; \mu), \tag{9}$$

where μ represents the parameters of the *GateNet*.

3.2 Loss Function

We consider two loss functions to measure the differences between the reconstructed frames and the ground-truth ones $G_{i,j}$ for both intermediate and final estimations. Specifically, as for the initial recovered frames $F_{i,j,k}$, we constrain $IntegralNet$ using MSE loss:

$$\mathcal{L}_{init}(B_{i-1}, B_i, E_{i-1}, E_i, S_{i-1}, \theta) = \frac{1}{(2N+1)HW} \sum_{k=0}^{2N} \|F_{i,j,k} - G_{i,j}\|^2. \tag{10}$$

The other one is defined between the final results $S_{i,j}$ and the ground-truth ones $G_{i,j}$ to constrain both $IntegralNet$ and $GateNet$:

$$\mathcal{L}_{final}(B_{i-1}, B_i, E_{i-1}, E_i, S_{i-1}, \theta, \mu) = \frac{1}{HW} \|S_{i,j} - G_{i,j}\|^2. \tag{11}$$

The overall loss function is:

$$\mathcal{L} = w_1 \mathcal{L}_{init} + w_2 \mathcal{L}_{final}, \tag{12}$$

where w_1, w_2 are set to 0.01, 1 in our experiment, respectively.

4 Experiment

4.1 Implementation Details

Training Dataset. We train the proposed method on two synthetic datasets: GoPro [16] and the synthetic subset of Blur-DVS [8]. Low frame-rate blurred inputs, high frame-rate sharp videos, and event streams are required during training. GoPro [16], a widely used video deblurring dataset, provides ground-truth sharp videos and we use them to generate blurred frames. We simulate event data by first increasing the video frame rate from 240 fps to 960 fps via a high-quality frame interpolation algorithm [17] and then applying an event simulator ESIM [20] to the videos. To add noise diversity, we set different contrast thresholds for each pixel from a Gaussian distribution $\mathcal{N}(0.18, 0.03)$ similar to [21]. We also use the synthetic subset of Blur-DVS [8] for training, which is captured with slow camera movement in relatively static scenes and thus provides ground-truth sharp videos and event streams. Blurry images are obtained in the same manner as on GoPro. We split the training and testing datasets as suggested.

Experimental Settings. Our network is implemented using Pytorch [19] and trained in an end-to-end manner supervised by Eq. 12 on a GeForce GTX 1080 GPU. For both datasets, we utilize a batch size of 4 training pairs and Adam [11] optimizer with momentum and momentum2 as 0.9 and 0.999. The network is trained for 60 epochs with the learning rate initialized as 0.0001 for the first 10 epochs and then decayed to zero linearly. We set the parameters M and N as 4, 5 on the GoPro dataset, and 3, 3 on the synthetic Blur-DVS. As for initialization, we recover the first blurry frame B_0 of a video sequence by replacing B_{i-1} and $S_{i-1,j}$ in Fig. 4 with B_0. Moreover, we repeatedly input E_0 to substitute E_{i-1}.

Table 1. Video deblurring and reconstruction performance on GoPro [16] dataset, in terms of average PSNR, SSIM and parameter numbers($\times 10^6$) of different networks.

	Average results of video deblurring						Average results of video Deblurring and interpolation					
Methods	STFAN[28]	STFAN*	E2V*	BHA[18]	LEMD[8]	Ours	TNTT	TNTT*	E2V*	BHA[18]	LEMD[8]	Ours
PSNR	30.28	38.17	35.38	29.06	31.79	**38.74**	32.47	35.90	34.89	28.49	29.67	**37.99**
SSIM	0.901	0.973	0.959	0.943	0.949	**0.982**	0.936	0.965	0.953	0.920	0.927	**0.981**
Params	5.36	5.38	10.71	-	5.37	**4.80**	10.68	10.88	10.71	-	9.13	**5.00**

* denotes the enhanced version of the corresponding single-sensor algorithm. See text for more details.

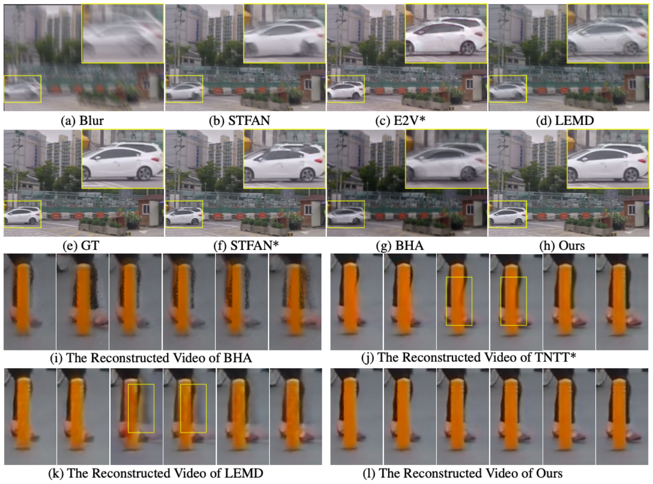

(a) Blur
(b) STFAN
(c) E2V*
(d) LEMD
(e) GT
(f) STFAN*
(g) BHA
(h) Ours
(i) The Reconstructed Video of BHA
(j) The Reconstructed Video of TNTT*
(k) The Reconstructed Video of LEMD
(l) The Reconstructed Video of Ours

Fig. 6. Visual comparisons on video deblurring (above) and high frame-rate video reconstruction (below) with the state-of-the-art on GoPro [16] datasets. The proposed method generates much clearer frames with fewer noises and artifacts. Zoom in for a better view.

4.2 Experimental Results

We quantitatively and qualitatively evaluate our video deblurring network (*i.e.* recovering videos with the original frame rate) and the simultaneous deblurring and interpolation network on both GoPro and Blur-DVS.

We conduct extensive comparisons with state-of-the-art algorithms including image-based methods on video deblurring [28] and high-speed video generation [9], event-based video generation methods [15,21], conventional video reconstruction methods from hybrid intensity and event-based sensors [22,18] and a deep learning-based method with hybrid sensors [8]. To demonstrate the effectiveness of the proposed framework, we also compare the enhanced versions of the single-sensor algorithms. As for image-based STFAN [28], we feed additional event data into the spatio-temporal filter adaptive network to assist frame alignment and deblurring (denoted as 'STFAN*'). TNTT [9] inputs events and blurry images for both keyframe deblurring network and frame interpolation network (denoted as 'TNTT*'). We also feed events together with intensity frames into the event-based E2V [21] for each of its recurrent reconstruction step (denoted as 'E2V*').

Table 2. Video deblurring and reconstruction performance on the synthetic subset of Blur-DVS [8], in terms of average PSNR, SSIM.

Average results of video deblurring									
Methods	E2V[21]	E2V*	STFAN[28]	STFAN*	MRL[15]	CIE[22]	BHA[18]	LEMD[8]	Ours
PSNR	16.89	24.81	19.03	30.18	10.59	19.02	22.43	26.48	**30.57**
SSIM	0.597	0.790	0.518	0.897	0.195	0.478	0.715	0.839	**0.904**
Average results of video deblurring and interpolation									
Methods	E2V[21]	E2V*	TNTT[9]	TNTT*	MRL[15]	CIE[22]	BHA[18]	LEMD[8]	Ours
PSNR	16.60	24.10	19.05	29.02	10.57	18.94	22.06	25.33	**29.65**
SSIM	0.587	0.777	0.521	0.875	0.194	0.473	0.699	0.827	**0.890**

* denotes the enhanced version of the corresponding single-sensor algorithm. See text for more details.

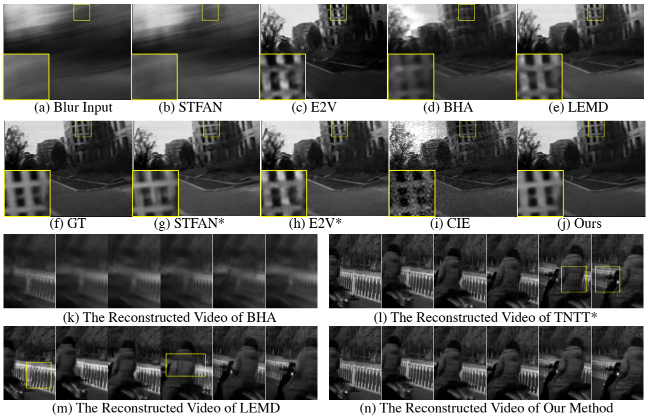

(a) Blur Input (b) STFAN (c) E2V (d) BHA (e) LEMD
(f) GT (g) STFAN* (h) E2V* (i) CIE (j) Ours
(k) The Reconstructed Video of BHA (l) The Reconstructed Video of TNTT*
(m) The Reconstructed Video of LEMD (n) The Reconstructed Video of Our Method

Fig. 7. Visual comparisons on video deblurring (above) and high frame-rate video reconstruction (below) on the synthetic subset of Blur-DVS [8]. The proposed method generates much sharper results with fewer noises and artifacts. More results are provided in our supplementary material. Zoom in for a better view.

We evaluate PSNR and SSIM on the video deblurring task on two synthetic datasets in Table 1 and Table 2. The proposed network performs favorably against state-of-the-art methods. Fig. 6 and Fig. 7 show some examples in the testing sets. The image-based method [28] purely relies on intensity images, thus it is less effective on severely-blurred videos. As event data encodes dense temporal information, it facilitates STFAN* to capture motion information across the frames and makes it more effective on video deblurring. E2V [21], which purely relies on event data, restores images with wrong contrast. However, its enhanced version E2V* keeps the correct contrast with the assistance of intensity frames. These significant improvements demonstrate the inherent advantage of each sensor and the effectiveness of utilizing both advantages for video deblurring. As for existing intensity and event-based algorithms, CIE [22] and BHA [18] adopt simplified physical models without considering the blur or the non-uniform threshold, which leads to blurry results and introduces accumulated noises. The CNN-based method [8] conducts the deblurring and refinement separately, which makes the approach sensitive to deblurring and leads to limited performance. On the contrary, the proposed method hinges on the physical event-based video re-

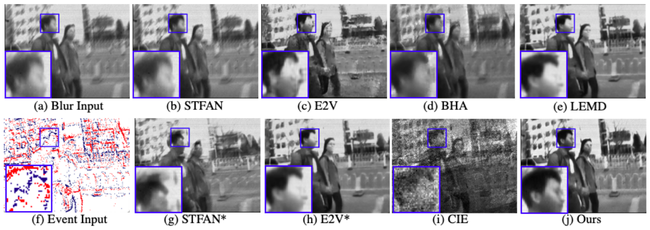


Fig. 8. Visual comparisons with the state-of-the-art on real-word blurry videos. The recovered results of the proposed method have fewer noises and more details. More results are provided in our supplementary material. Zoom in for a better view.

construction model via an end-to-end architecture. The restored video frames present finer details and fewer noises.

We also report the results on simultaneous video deblurring and interpolation in Table 1, Table 2, Fig. 6 and Fig. 7. The conventional method BHA [18] is prone to noises during interpolation, especially at the object edges. Besides, its threshold choosing scheme is not robust, which introduces unaddressed blur when estimating a wrong threshold. The deep learning-based LEMD and TNTT* neglect to utilize the physical constraint between two adjacent frames and thus interpolate blurry frames with undesirable artifacts, especially at occlusion. Fig. 6(l) and Fig.7(n) show that the proposed method can restore sharp and artifact-free frames.

To validate the generalization capacity of the proposed method, we qualitatively compare the proposed network with other algorithms on real-world blurry videos in the real subset of Blur-DVS [8]. As shown in Fig. 8, our method restores more visually pleasing frames than the state-of-the-art.

5 Ablation Study

We have shown that the proposed algorithm performs favorably against state-of-the-art methods. In this section, we further analyze the effectiveness of each component in video deblurring and interpolation.

5.1 Effectiveness of Physical-Based Framework

The proposed algorithm is designed based on the physical model of the event-based video reconstruction. We predict the residuals I and D and apply multiply operation to them according to Eq. 2 and Eq. 4. To demonstrate the effectiveness of the physical-based framework, we compare the method that adds the residuals and the intensity images up (denoted as 'Addition'), as already used in pure image-based algorithms [28,9,27]. The results in Table 3 show that using multiplication achieves higher performance than 'Addition'. As shown in Fig. 9(c), 'Addition' predicts a blurry addition residual and thus generates a smooth result but with more artifacts (Fig. 9(h)). However, as the proposed method is

Table 3. Ablation Study. 'Addition' replaces the multiplication with addition to verify the effectiveness of the physical-based network. 'w/o DF', 'w/o Pre' and 'w/o ASF' represent removing the dynamic filtering, the previous information in keyframe estimation and the adaptively-selected fusion. Our method achieves the highest quantitative results, which demonstrates the effectiveness of each component. See text for details.

Methods	Addition	w/o DF	w/o Pre	w/o ASF	Ours
PSNR	29.24	28.64	29.42	29.34	**29.65**
SSIM	0.882	0.872	0.855	0.885	**0.890**

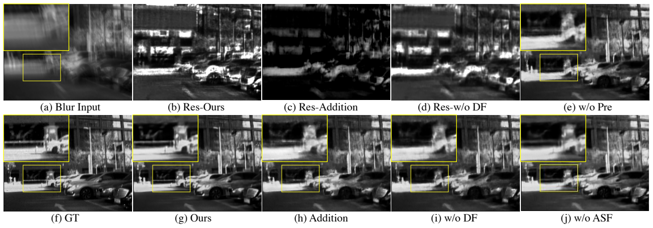

(a) Blur Input (b) Res-Ours (c) Res-Addition (d) Res-w/o DF (e) w/o Pre
(f) GT (g) Ours (h) Addition (i) w/o DF (j) w/o ASF

Fig. 9. Ablation Study. 'Res-' in (b)(c)(d) denotes the learned residual between the keyframe and the interpolated frame. 'Addition' replaces the multiplication with addition to verify the effectiveness of the physical-based framework. 'w/o DF', 'w/o Pre' and 'w/o ASF' represent removing the dynamic filtering, the previous information in keyframe estimation step and the adaptively-selected fusion. The proposed method restores clearer images with more details and fewer artifacts. See text for details.

based on the physical model, which makes it easy to calculate the multiplication residuals (Fig. 9(b)) from event data, it is robust to severely-blurred frames and restores images with more details and fewer artifacts (Fig. 9(g)).

5.2 Effectiveness of Dynamic Filtering

To handle the events triggered by the spatially variant threshold, we propose to integrate the dynamic filters when estimating residuals. To validate the above discussions, we remove the dynamic filter generation module and feed its inputs $(B_{i-1}, B_i, E_{i-1}, E_i, S_{i-1})$ into the event feature extraction directly for a fair comparison (denoted as 'w/o DF'). Table 3 shows that 'w/o DF' is less effective. Due to the lack of compensation for the spatially variant triggering threshold, it provides an overly-smooth residual (Fig. 9(d)) compared to ours (Fig. 9(b)). And thus, it cannot restore the lost details in the final results (Fig. 9(i)), which demonstrates that using dynamic filtering facilitates to minimize the effects of the non-uniform threshold. Besides, generated filters are illustrated in the supplementary materials for visual interpretation.

5.3 Effectiveness of Previous Information

We note that the existing event-based video deblurring and interpolation algorithms [18,8] bring one blurry frame alive without considering additional information that exists across adjacent frames. To verify the effectiveness of utilizing previous information, we compare a method that only estimates the keyframes $C_{i,0}$ from current blurry inputs without the ones $P_{i,0,j}$ from the previously recovered frames (denoted as 'w/o Pre'). The final results shown in Table 3 and Fig. 9(e) indicate that involving previous information is more effective for video deblurring and reconstruction.

5.4 Effectiveness of Frame Fusion

To integrate the $2N + 1$ initial recovered results $F_{i,j,k}$ in an adaptive selection manner, the proposed frame fusion step utilizes the information from the blurry frame, event data and the initial results to generate a gate map and then obtains the final results by weighted summation. To demonstrate the effectiveness of this design, we compare the method that removes the estimation of the gate map but feeds the initial results into three 3D convolution layers to estimate the final results directly (denoted as 'w/o ASF'). The final results in Table 3 and Fig. 9(j) indicate that the proposed frame fusion module can integrate the initial results in an adaptive selection scheme and keep more details, which is more effective for video deblurring and interpolation.

6 Concluding Remarks

In this paper, we propose to learn event-driven video frame deblurring and interpolation to solve high frame-rate video generation. The whole framework hinges on the physical model of the event-based video reconstruction, which estimates the residual between the latent sharp frames as well as that between sharp and blurry frames, and integrates the model into a compact architecture. Benefiting from this design, the proposed method can generate physically-correct results and handle severely-blurred videos. Furthermore, we show that using dynamic filters when predicting residuals can deal with event data triggered by the spatially variant threshold. By training the proposed network in an end-to-end manner, the proposed algorithm is able to reconstruct high-quality and high frame-rate videos. Experiments on the synthetic datasets and real images demonstrate that the proposed method achieves superior performance against the existing image and event-based approaches.

We note that one limitation of the proposed method is that the network need be retrained if we aim to further increase the frame rate. However, we can solve it by applying an additional interpolation network recursively between pairs of restored sharp frames. Further research will be devoted to arbitrary frame-rate video reconstruction.

Acknowledgments

This project was supported by the 863 Program of China (No. 2013AA013802), NSFC (Nos. 61872421, 61922043) and NSF of Jiangsu Province (No. BK20180471).

References

1. Bao, W., Lai, W., Zhang, X., Gao, Z., Yang, M.: Memc-net: Motion estimation and motion compensation driven neural network for video interpolation and enhancement. TPAMI (2019) 1
2. Bardow, P., Davison, A.J., Leutenegger, S.: Simultaneous optical flow and intensity estimation from an event camera. In: CVPR (2016) 1
3. Barua, S., Yoshitaka, M., Ashok, V.: Direct face detection and video reconstruction from event cameras. In: WACV (2016) 1
4. Brandli, C.: Event-Based Machine Vision. Ph.D. thesis, ETH Zurich (2015) 2, 5
5. Brandli, C., Berner, R., Yang, M., Liu, S.C., Delbruck, T.: A 240× 180 130 db 3 μs latency global shutter spatiotemporal vision sensor (2014) 1
6. Brandli, C., Muller, L., Delbruck, T.: Real-time, high-speed video decompression using a frame-and event-based davis sensor. In: ISCAS (2014) 2, 5
7. Jia, X., De Brabandere, B., Tuytelaars, T., Gool, L.V.: Dynamic filter networks. In: NIPS (2016) 3, 4, 5, 6
8. Jiang, Z., Zhang, Y., Zou, D., Ren, J., Lv, J., Liu, Y.: Learning event-based motion deblurring. In: CVPR (2020) 2, 9, 10, 11, 12, 13
9. Jin, M., Hu, Z., Favaro, P.: Learning to extract flawless slow motion from blurry videos. In: CVPR (2019) 1, 2, 10, 11, 12
10. Jin, M., Meishvili, G., Favaro, P.: Learning to extract a video sequence from a single motion-blurred image. In: CVPR (2018) 1
11. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization (2014) 9
12. Lichtsteiner, P., Christoph, P., Tobi, D.: A 128×128 120 db 15 μs latency asynchronous temporal contrast vision sensor. (2008) 1
13. Meyer, S., Djelouah, A., McWilliams, B., Sorkine-Hornung, A., Gross, M., Schroers, C.: Phasenet for video frame interpolation. In: CVPR (2018) 1
14. Mildenhall, B., Barron, J.T., Chen, J., Sharlet, D., Ng, R., Carroll, R.: Burst denoising with kernel prediction networks. In: CVPR (2018) 3
15. Munda, G., Reinbacher, C., Pock, T.: Real-time intensity-image reconstruction for event cameras using manifold regularisation (2018) 1, 2, 10, 11
16. Nah, S., Hyun Kim, T., Mu Lee, K.: Deep multi-scale convolutional neural network for dynamic scene deblurring. In: CVPR (2017) 9, 10
17. Niklaus, S., Mai, L., Liu, F.: Video frame interpolation via adaptive separable convolution. In: ICCV (2017) 1, 3, 9
18. Pan, L., Scheerlinck, C., Yu, X., Hartley, R., Liu, M., Dai, Y.: Bringing a blurry frame alive at high frame-rate with an event camera. In: CVPR (2019) 2, 3, 5, 10, 11, 12, 13
19. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch (2017) 9
20. Rebecq, H., Gehrig, D., Scaramuzza, D.: Esim: an open event camera simulator. In: Conference on Robot Learning (2018) 2, 5, 9
21. Rebecq, H., Ranftl, R., Koltun, V., Scaramuzza, D.: Events-to-video: Bringing modern computer vision to event cameras. In: CVPR. pp. 3857–3866 (2019) 1, 9, 10, 11
22. Scheerlinck, C., Barnes, N., Mahony, R.: Continuous-time intensity estimation using event cameras. In: ACCV (2018) 2, 10, 11
23. Shedligeri, P., Mitra, K.: Photorealistic image reconstruction from hybrid intensity and event-based sensor (2019) 2

24. Su, S., Delbracio, M., Wang, J., Sapiro, G., Heidrich, W., Wang, O.: Deep video deblurring for hand-held cameras. In: CVPR (2017) 1
25. Wang, L., Ho, Y.S., Yoon, K.J., et al.: Event-based high dynamic range image and very high frame rate video generation using conditional generative adversarial networks. In: CVPR (2019) 7
26. Wang, X., Chan, K.C., Yu, K., Dong, C., Change Loy, C.: Edvr: Video restoration with enhanced deformable convolutional networks. In: CVPR (2019) 1
27. Zhang, H., Dai, Y., Li, H., Koniusz, P.: Deep stacked hierarchical multi-patch network for image deblurring. In: CVPR (2019) 12
28. Zhou, S., Zhang, J., Pan, J., Xie, H., Zuo, W., Ren, J.: Spatio-temporal filter adaptive network for video deblurring. In: ICCV (2019) 1, 10, 11, 12