Unifying Motion Deblurring and Frame Interpolation with Events

Xiang Zhang, Lei Yu†
Wuhan University, Wuhan, China.
{xiangz, ly.wd}@whu.edu.cn

Abstract

Slow shutter speed and long exposure time of frame-based cameras often cause visual blur and loss of interframe information, degenerating the overall quality of captured videos. To this end, we present a unified framework of event-based motion deblurring and frame interpolation for blurry video enhancement, where the extremely low latency of events is leveraged to alleviate motion blur and facilitate intermediate frame prediction. Specifically, the mapping relation between blurry frames and sharp latent images is first predicted by a learnable double integral network, and a fusion network is then proposed to refine the coarse results via utilizing the information from consecutive blurry inputs and the concurrent events. By exploring the mutual constraints among blurry frames, latent images, and event streams, we further propose a self-supervised learning framework to enable network training with real-world blurry videos and events. Extensive experiments demonstrate that our method compares favorably against the state-of-the-art approaches and achieves remarkable performance on both synthetic and real-world datasets. Codes are available at https://github.com/XiangZ-0/EVDI.

1. Introduction

Highly dynamic scenes, e.g., fast-moving targets or nonlinear motions, pose challenges for high-quality video generation as the captured frame is often blurred and target information is missing between consecutive frames [29]. Existing frame-based methods attempt to tackle these problems by developing motion deblurring [11], frame interpolation [1] or blurry video enhancement techniques [10, 25]. However, it is difficult for frame-based deblurring methods to predict sharp latent frames from severely blurred videos because of motion ambiguities and the erasure of intensity textures [11]. Besides, current frame-based interpolation approaches generally assume the motion between neighboring frames to be linear [1], which is not always valid in real-world scenarios especially when encountering nonlinear motions, thus often leading to incorrect predictions.

Recent works have revealed the advantages of event cameras [5] in motion deblurring and frame interpolation. On one hand, the output of event camera inherently embeds precise motions and sharp edges [2] since it reports asynchronous event data with extremely low latency (in the order of $\mu$s) [5, 13], which is effective in alleviating motion blur [14, 21, 22, 31, 34]. On the other hand, event camera is able to record almost continuous brightness changes to compensate the missing information between consecutive frames, making it feasible to recover accurate intermediate frames even under non-linear motions [14, 30]. However, existing works generally treat motion deblurring and frame interpolation as separate tasks, while the problems of...
motion blur and missing information between frames have strong co-occurrence in real scenes and thus need to be considered simultaneously. In real-world scenarios, the aforementioned methods face two main challenges as follows.

- **Limitations of Separate Tasks**: The performance of interpolation methods [30] is often highly dependent on the quality of reference frames, and it is difficult to interpolate clear results when the reference frames are degraded by motion blur. For deblurring task, most methods [31, 34] focus on recovering sharp images inside the exposure time of blurry inputs, neglecting these latent images between blurry frames (see Fig. 1).

- **Data Inconsistency**: Most previous works employ well-labeled synthetic datasets for supervised [30, 31] or semi-supervised learning [34], which often causes performance drop in real scenes due to the inconsistency between synthetic and real-world data [34].

In this paper, we present a unified framework of Event-based Video Deblurring and Interpolation (EVDI) for blurry video enhancement. The proposed method consists of two main modules: a learnable double integral (LDI) network and a fusion network. The LDI network is designed to automatically predict the mapping relation between blurry frames and sharp latent images from the corresponding events, where the timestamp of the latent image can be chosen arbitrarily inside the exposure time of blurry frames (deblurring task) or between consecutive blurry frames (interpolation task). The fusion network receives the coarse reconstruction of latent images and generates a fine result by utilizing all the information from consecutive blurry frames and event streams. For training, we take advantage of the mutual constraints among blurry frames, sharp latent images and event streams, and propose a fully self-supervised learning framework to help the network fit the distribution of real-world data without the need of ground truth images.

The main contributions of this paper are three-fold:

- We present a unified framework of event-based video deblurring and interpolation that generates arbitrarily high frame-rate sharp videos from blurry inputs.
- We propose a fully self-supervised framework to enable network training in real-world scenarios without any labeled data.
- Experiments on both synthetic and real-world datasets show that our method achieves state-of-the-art results while maintaining an efficient network design.

2. Related Work

2.1. Frame Interpolation

Existing frame-based interpolation methods can be roughly categorized into two classes: warping-based and kernel-based approaches. Warping-based approaches [1, 9, 18, 35] generally combine optical flow [8, 28] with image warping to predict intermediate frames, and several techniques have been proposed to enhance the interpolation performance, e.g., forward warping [18], spatial transformer networks [35], and depth information [1]. However, these methods often assume linear motion and brightness constancy between two reference frames, thus failing to handle arbitrary motions. Rather than warping reference frames with optical flow, kernel-based methods [19, 20] model the frame interpolation as local convolution on the reference frames, where the kernel is directly estimated from the input frames. Despite kernel-based methods are more robust to complex motions and brightness changes, their scalability is often limited by the fixed sizes of convolution kernels.

The common challenge of frame-based interpolation is the missing information between reference frames, which can be alleviated by leveraging the extremely low latency of events. Recent approach [30] takes the merits of frames and events and achieves excellent interpolation results even under non-linear motions, but its performance is also closely related to the quality of reference frames and thus cannot be directly applied for blurry video enhancement.

2.2. Motion Deblurring

One of the most popular frame-based deblurring methods is to employ neural networks to learn the blur feature and predict sharp images from blurry inputs [7, 11, 17, 37]. Several techniques have been developed to exploit the temporal information inside blurry frames, including dynamic temporal blending mechanism [7], spatio-temporal filter adaptive networks [37], and intra-frame iterations [17]. Recent works have also revealed the potential of events in motion deblurring. Event streams inherently embed motion information and sharp edges, which can be exploited to tackle the temporal ambiguity and texture erasure caused by motion blur. Pioneer event-based methods achieve motion deblurring by relating blurry frames, sharp latent images and the corresponding events according to the physical event generation model [21, 22], but their performance is often degraded due to the imperfection of physical circuits, e.g., intrinsic camera noises. To alleviate this, learning-based approaches [31, 34] have been proposed to fit the distribution of event data, achieving better deblurring performance.

However, most deblurring methods focus solely on restoring sharp latent images inside the exposure time of blurry frames, while the information between blurry frames is also important in practical applications, motivating the combination of deblurring and interpolation.

2.3. Joint Deblurring and Interpolation

Previous frame-based methods have approached the joint deblurring and interpolation task [10, 25]. The work of [10] performs frame interpolation based on the keyframes
pre-processed by a deblurring module, and the work of
[25] treats deblurring and interpolation as a unified task and
achieves better enhancement performance. For event-
based methods, LEDVDI is the closest related work [14],
but LEDVDI is categorized into the cascaded scheme as it
achieves deblurring and interpolation with different stages.
Besides, all the aforementioned methods require supervised
training on the synthetic datasets, limiting their performance
on real-world scenarios due to data inconsistency.

Exploiting the information from both frames and events,
our method achieves blurry video enhancement without distin-
guishing the deblurring and interpolation tasks. Moreover,
a self-supervised learning framework is proposed to
enable network training with real events and blurry videos,
guaranteeing the performance in real-world scenarios.

3. Problem Statement

Videoing highly dynamic scenes often suffers from
blurry artifacts and the Blurry Video Enhancement (BVE)
plays an important role for visual perception. Existing
frame-based methods often struggle to achieve BVE due
to motion ambiguity and loss of inter-frame information,
while this can be effectively mitigated with the aid of events.
Given two consecutive blurry frames $B_i, B_{i+1}$ captured
within the exposure time $T_i, T_{i+1}$ and the corresponding
event streams $E_{i+1}$ triggered inside $T_{i+1}$, where
$T_{i+1} \triangleq T_i \cup T_{i+1} \cup T_{i+1}$ indicating the time interval
between $B_i$ and $B_{i+1}$, the task of EVDI is to achieve BVE
directly from blurry inputs, i.e.,

$$L(t) = EVDI(t; B_i, B_{i+1}, E_{i+1}), \quad \forall t \in T_{i+1},$$

where $L(t)$ indicates the latent image of arbitrary time
t $t \in T_{i+1}$. According to Eq. (1), EVDI degrades to Motion
Deblurring (MD) when $t \in T_i$ or $T_{i+1}$, or Frame Interpo-
lation (FI) when $t \in T_{i+1}$ is. Thus, EVDI is more general
than MD and FI, and provides a unified formulation to the
task of BVE.

EVDI vs. Frame Interpolation. Conventional FI task aims
at recovering the intermediate latent images $\{L(t)\}_{t \in T_{i+1}}$
from sharp reference frames $I_i, I_{i+1}$. Providing the concurrent
event streams $E_{i+1}$ emitted within $T_{i+1}$, we have

$$L(t) = \text{Interp}(t; I_i, I_{i+1}, E_{i+1}), \quad t \in T_{i+1},$$

where $\text{Interp}(\cdot)$ represents an FI operator. Most FI meth-
ods [1, 9, 30] are designed to restore inter-frame latent images
from high-quality (sharp and clear) reference frames
$I_i, I_{i+1}$, while EVDI directly accepts blurry inputs, which
is more challenging than conventional FI.

EVDI vs. Motion Deblurring. The MD aims at recon-
structing the sharp latent images $\{L(t)\}_{t \in T_i}$ from the corre-
ponding blurry frame $B_i$. Providing the concurrent event
streams $E_i$ triggered within $T_i$, we have

$$L(t) = \text{Deblur}(t; B_i, E_i), \quad t \in T_i,$$

where $\text{Deblur}(\cdot)$ indicates an MD operator. Existing de-
blurring methods [4, 27] mainly focus on recovering the
latent frames inside the exposure time $T_i$, while EVDI is
able to predict the latent images of time instance both inside
the exposure time $T_i$ (or $T_{i+1}$) and between blurry frames
$T_{i+1}$ as shown in Fig. 1.

Ideally, EVDI can approach the BVE task by unifying
MD and FI in Eq. (1). However, to efficiently realize EVDI
in real-world scenarios, challenges still exist.

- MD and FI should be simultaneously addressed in a
unified framework to fulfill the EVDI. Previous at-
tempts for BVE [10, 14] employ a cascaded scheme
that performs frame interpolation after deblurring, but
this approach often propagates deblurring error to the
interpolation stage, leading to sub-optimal results.

- Existing related methods are generally developed
within a supervised learning framework [10, 14, 25], of
which the supervision is usually provided by synthetic
blurry images and events. Thus the performance might
degradate in real scenes due to the different distribution
between synthetic and real-world data.

4. Method

In this work, we propose to approximate the optimal
EVDI model with trainable neural networks, and develop
a self-supervised learning framework by exploiting the mu-
tual constraints among blurry frames, sharp latent frames
and event streams.

4.1. Unified Deblurring and Interpolation

We first review the physical generation model of events,
which are triggered whenever the log-scale brightness
change exceeds the event threshold $c > 0$, i.e.,

$$\log(L(t, x)) - \log(L(\tau, x)) = p \cdot c,$$

where $L(t, x)$ and $L(\tau, x)$ denote the instantaneous inten-
sity at time $t$ and $\tau$ at the pixel position $x$, and polarity
$p \in \{-1, +1\}$ indicates the direction of brightness changes.
With the aid of events, we can formulate the following rela-
tion (the pixel position is omitted for readability):

$$L(t) = L(f) \exp(c \int_{t}^{f} e(s)ds),$$

where $L(t)$ and $L(f)$ are latent images at instant time $t$ and
$f$, and $e(t) \triangleq p \cdot \delta(t - \tau)$ denotes the continuous representation
of events with $\delta(\cdot)$ indicating the Dirac function. On
4.1.1 Feasibility Analysis

Previous works of [21, 22] focus on exploiting Eq. (7) to restore the sharp latent image inside the exposure period $T$, while this formulation can be also extended to recover the latent frames at arbitrary time outside $T$ (please see the supplementary material for proof). However, directly applying Eq. (7) for unified deblurring and interpolation often meets the following obstacles: First, the computation of $E(f, T)$ requires the knowledge of event threshold $c$, which is critical to the recovery performance [22] but hard to accurately estimate due to its temporal instability. Second, real-world events are noisy due to the non-ideality of physical sensors [5], e.g., limited read-out bandwidth, and thus often lead to degraded results, especially when encountering long-term integral of events where $E(f, T)$ is severely contaminated by noises. Hence, we propose to employ learning-based approaches to fit the statistics of real-world events.

4.1.2 Network Architecture

Our network receives a latent image timestamp $f \in T_i$, two consecutive blurry frames $B_i, B_{i+1}$ and the corresponding event streams $E_{i+1}$ as input, and outputs a sharp latent image $L(f)$. There are two main modules in our network: a learnable double integral (LDI) network and a fusion network, where the LDI network learns to approximate the double integral behavior of Eq. (8) and the fusion network is designed to refine the results generated by the blurry images and the outputs of LDI network, as shown in Fig. 3.

**LDI Network.** Suppose the LDI network is trained to approximate a specific case $E(0, T_{[0,T]}) \approx \text{LDI}(E_{[0,T]})$ i.e.,

$$\text{LDI}(E_{[0,T]}) \approx \frac{1}{T} \int_0^T \exp(c \int_0^t e(s)ds)dt,$$  \hspace{1cm} (9)

where $T_{[0,T]}$ indicates the time interval from 0 to $T > 0$ and $E_{[0,T]}$ is the corresponding event streams. Now we consider a more general case of $E(f, T)$, which can be written as

$$E(f, T) = \frac{1}{T} \int_{t_o}^f \exp(c \int_0^t e(s)ds)dt + \frac{1}{T} \int_{t_o + T}^f \exp(c \int_0^t e(s)ds)dt,$$ \hspace{1cm} (10)

where $t_o$ indicates the starting time of $T$. Applying $t' = t - f$ and $s' = s - f$ to Eq. (10), we have

$$E(f, T) = -\frac{1}{T} \int_0^{t_o - f} \exp(c \int_0^{t'} e(s' + f)ds')dt' + \frac{1}{T} \int_0^{t_o + T - f} \exp(c \int_0^{t'} e(s' + f)ds')dt'$$

$$= w_1 G(E_{[f,t_o]} + G(E_{[f,t_o + T]}),$$ \hspace{1cm} (11)

where $w_1 = (f - t_o)/T$, $w_2 = (t_o + T - f)/T$ are weights and $G(\cdot)$ is a general formula defined as

$$G(E_{[f,t_o]} = \frac{1}{t_r - f} \int_0^{t_r - f} \exp(c \int_0^t e(s + f)ds)dt \hspace{1cm} (12)$$

and $t_r$ denotes the reference time. Based on the above definition, we can calculate $E(f, T)$ by approximating Eq. (12) with the LDI network, i.e., Eq. (9). For the case of $t_r - f \geq 0$, $G(\cdot)$ can be directly approximated by

$$G(E_{[f,t_o]} \approx \text{LDI}(S(E_{[f,t_o]})) \hspace{1cm} (13)$$

with $S(E_{[f,t_o]} \triangleq \{e(t + f), t \in [0, t_r - f]\}$ representing the event operator of time shift. For the case of $t_r - f < 0$,

$$G(E_{[f,t_o]} = \frac{1}{f - t_r} \int_0^{f - t_r} \exp(c \int_0^t -e(-s + f)ds) \hspace{1cm} (14)$$

where $R(E_{[f,t_o]} \triangleq \{-e(-t + f), t \in [0, f - t_r]\}$ indicates the event operator composed of time shift, flip and polarity.
The double integral of events on the mutual constraints among blurry frames, sharp latent images, and the final result L(f) at arbitrary f ∈ T^t_{i+1} without any network modification or re-training process.

**Fusion Network.** After obtaining E(f, T) from LDI network, the latent image L(f) can be coarsely restored by Eq. (7). We denote the latent images reconstructed from B^t_i, B^t_{i+1} as L_i(f), L_{i+1}(f), respectively, and manually generate an extra result L_{i+1}(f) by

\[
L^\text{t}_{i+1}(f) = \omega(f)L_i(f) + (1 - \omega(f))L_{i+1}(f),
\]

where the weighting function ω(f) with f ∈ [0, T^t_{i+1}] is defined as

\[
\omega(f) = \begin{cases} 
1 & \text{if } f \in T_i, \\
1 - \frac{T_{i+1}}{T_{i+1} - T_i} & \text{if } f \in T_{i+1} - T_i, \\
0 & \text{if } f \in T_{i+1},
\end{cases}
\]

since a weighted reconstruction is helpful for frame interpolation in our observation. Finally, our fusion network receives \(L_i(f), L_{i+1}(f), L^\text{t}_{i+1}(f), E(f, T_i), E(f, T_{i+1})\) and produces the final latent image \(\bar{L}(f)\), as illustrated in Fig. 3.

### 4.2. Self-supervised Learning Framework

The proposed self-supervised learning framework consists of three different losses which are formulated based on the mutual constraints among blurry frames, sharp latent images and event streams.

**Blurry-event Loss.** The double integral of events \(E(f, T)\) corresponds to the mapping relation between blurry frames and sharp latent images. For multiple blurry inputs, we propose to formulate the consistency between the latent images reconstructed from different blurry frames, e.g., \(L_i(f) =\)

\[
\mathcal{P}(E_{[f,t_i]}(E_{[f,t_i]})) = \begin{cases} 
S(E_{[f,t_i]}) & \text{if } t_r \geq f, \\
R(E_{[f,t_i]}) & \text{if } t_r < f.
\end{cases}
\]

Thus, Eq. (11) can be reformulated as

\[
E(f, T) \approx w_1 \text{LDI}(\mathcal{P}(E_{[f,t_i]})) + w_2 \text{LDI}(\mathcal{P}(E_{[f,t_i]+T})),
\]

meaning that arbitrary E(f, T) can be approximated by a weighted combination of LDI outputs, where the LDI network only needs to be trained once to fit the case of Eq. (9).

For the input of LDI network, we introduce a spatio-temporal event representation. With a pre-defined number, e.g., N, we fairly divide N temporal bins from \(t = 0\) to \(t = T_{i+1}\) where \(T_{i+1}\) denotes the total duration of \(T^t_{i+1}\). We then accumulate the events pre-processed by \(\mathcal{P}(\cdot)\) inside each temporal bin, and form a \(2N \times H \times W\) tensor as the LDI input with 2, H, W indicating event polarity, image height and width, respectively. Therefore, our event representation enables flexible choice of the target timestamp \(f\) while maintaining a fixed input format, which allows network to restore the latent images \(L(f)\) at arbitrary \(f \in T^t_{i+1}\) without any network modification or re-training process.

\begin{align*}
\bar{L}(f) & \approx \log\left(\frac{B_i}{E(f, T_i)}\right) \\
& + \log\left(\frac{B_{i+1}}{E(f, T_{i+1})}\right),
\end{align*}

where \(E(f, T_i), E(f, T_{i+1})\) are generated by the LDI network. We convert Eq. (19) to the logarithmic domain and rewrite it as the blurry-event loss \(\mathcal{L}_{B,E}\).

\[
\mathcal{L}_{B,E} = \|\tilde{B}_{i+1} - B_i - (\Delta f, T_{i+1}) - \Delta f, T_i\|_1,
\]

where the top tilde denotes logarithm, e.g., \(\tilde{B}_i = \log(B_i)\). With the blurry-event constraint, the LDI network can learn to perform event double integral through utilizing the brightness difference between blurry frames.

**Blurry-sharp Loss.** Providing the reconstructed latent images \(\tilde{L}(t)\) with \(t \in T_i\), the blurring process Eq. (6) can be reformulated as the discrete version, i.e.,

\[
\tilde{B}_i = \frac{1}{T} \int_{t \in T_i} \tilde{L}(t) dt \approx \frac{1}{M} \sum_{m=0}^{M-1} \tilde{L}_i[m],
\]

where the height and width are denoted as \(H, W\), respectively. Therefore, our event representation enables flexible choice of the target timestamp \(f\) while maintaining a fixed input format, which allows network to restore the latent images \(L(f)\) at arbitrary \(f \in T^t_{i+1}\) without any network modification or re-training process.

![Data flow of the proposed method. With two consecutive blurry frames B_i, B_{i+1} and the concurrent events \(E_{i+1}\), we first split the events into four subsets according to the target timestamp \(f\), and then feed the events pre-processed by \(\mathcal{P}(\cdot)\) to four weight-sharing LDI networks. Following that, three coarse results \(L_i(f), L_{i+1}(f), L_{i+1}(f)\) are generated based on \(B_i, B_{i+1}\) and the corresponding double integral of events \(E(f, T_i), E(f, T_{i+1})\), and the final result \(L(f)\) is produced by refining them with the fusion network.](image-url)
Figure 4. Qualitative comparisons of the deblurring task on the RBE dataset. Details are zoomed in for a better view.

Table 1. Quantitative comparisons of the proposed method to the state-of-the-arts on the deblurring task. Note that LEDVDI only produces 6 frames for each blurry frame while the others output 7 frames. The number of network parameters (#Param.) is also provided.

<table>
<thead>
<tr>
<th>Method</th>
<th>GoPro</th>
<th>HQF</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PSNR ↑</td>
<td>SSIM ↑</td>
</tr>
<tr>
<td>LEVS [11]</td>
<td>20.84</td>
<td>0.5473</td>
</tr>
<tr>
<td>EDI [22]</td>
<td>21.29</td>
<td>0.6402</td>
</tr>
<tr>
<td>LEDVDI [14]</td>
<td>25.38</td>
<td>0.8567</td>
</tr>
<tr>
<td>eSL-Net [31]</td>
<td>17.80</td>
<td>0.5655</td>
</tr>
<tr>
<td>RED [34]</td>
<td>25.14</td>
<td>0.8587</td>
</tr>
<tr>
<td>EVDI (Ours)</td>
<td>30.40</td>
<td>0.9058</td>
</tr>
</tbody>
</table>

where \(\hat{L}_i[m]\) indicates the \(m\)-th latent image inside the exposure time of \(B_i\) and \(M\) is the total number of reconstructions. Previous attempts reduce the discretization error by assuming linear [3] or piece-wise linear motion [34] between latent frames and interpolating more intermediate frames, while this assumption might be violated in real-world scenarios especially in the case of complex non-linear motions.

In contrast, we restore \(\hat{L}_i[m]\) all by our network to exploit the real motion embedded in the event streams, and formulate the blurry-sharp loss \(L_{B-S}\) between the reblurred images \(\hat{B}_i, \hat{B}_{i+1}\) and the original blurry inputs \(B_i, B_{i+1}\) as

\[
L_{B-S} = \|\hat{B}_i - B_i\|_1 + \|\hat{B}_{i+1} - B_{i+1}\|_1,
\]

which guarantees the brightness consistency by learning from the blurry inputs.

**Sharp-event Loss.** Apart from the above constraints, the relation between sharp latent images and events can be also leveraged to supervise the reconstruction of consecutive latent frames. Based on Eq. (5), we have

\[
\mathcal{N}(\Delta \hat{L}) = \mathcal{N}(J),
\]

where \(\Delta \hat{L} = \hat{L}(t) - \hat{L}(f), J = \int_a^b e(s)ds\) and \(\mathcal{N}(\cdot)\) is the min/max normalization operator adopted in [23]. Therefore, we can avoid the estimation of threshold \(c\) and formulate the sharp-event loss \(L_{S-E}\) as

\[
L_{S-E} = \|\mathcal{M}(\mathcal{N}(\Delta \hat{L})) - \mathcal{M}(\mathcal{N}(J))\|_1,
\]

where \(\mathcal{M}(\cdot)\) denotes a pixel-wise masking operator for \(\mathcal{M}(\cdot) = 0\) only when there are no events. Finally, the total self-supervised framework can be summarized as follows.

\[
\mathcal{L} = \alpha L_{B-E} + \beta L_{B-S} + \gamma L_{S-E},
\]

with \(\alpha, \beta, \gamma\) denoting the balancing parameters.

5. Experiments and Analysis

5.1. Experimental Settings

**Datasets.** We evaluate the proposed method with three different datasets, including synthetic and real-world ones.

**GoPro:** We build a purely synthetic dataset based on the REDS dataset [16]. We first downsample and crop the images to size 160 \(\times\) 320 and then increase the frame rate by interpolating 7 images between consecutive frames using RIFE [6]. Finally, we generate both blurry frames and events based on the high frame-rate sequences, where the blurry frames are obtained by averaging a specific number of sharp images, and events are simulated by ESIM [24].

**HQF:** The HQF dataset [26] contains real-world events and high-quality frames captured simultaneously by a
Table 2. Quantitative results on the interpolation task. We compute PSNR and SSIM on the reconstruction results of the skipped frames, and use the official models provided by the authors for comparison. The column #Param. indicates the number of network parameters.

<table>
<thead>
<tr>
<th>Method</th>
<th>GoPro 1 frame skip</th>
<th>GoPro 3 frame skips</th>
<th>HQF 1 frame skip</th>
<th>HQF 3 frame skips</th>
<th>#Param.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PSNR ↑</td>
<td>SSIM ↑</td>
<td>PSNR ↑</td>
<td>SSIM ↑</td>
<td>PSNR ↑</td>
</tr>
<tr>
<td>Jin [10]</td>
<td>20.47</td>
<td>0.5244</td>
<td>20.48</td>
<td>0.5958</td>
<td></td>
</tr>
<tr>
<td>BIN [25]</td>
<td>19.54</td>
<td>0.4645</td>
<td>18.25</td>
<td>0.4576</td>
<td></td>
</tr>
<tr>
<td>DAIN [1]</td>
<td>20.89</td>
<td>0.5297</td>
<td>20.97</td>
<td>0.5980</td>
<td></td>
</tr>
<tr>
<td>EDI [22]</td>
<td>18.72</td>
<td>0.5059</td>
<td>16.62</td>
<td>0.4266</td>
<td></td>
</tr>
<tr>
<td>LEDVDI [14]</td>
<td>24.42</td>
<td>0.8198</td>
<td>19.24</td>
<td>0.6034</td>
<td></td>
</tr>
<tr>
<td>Time Lens</td>
<td>21.56</td>
<td>0.5809</td>
<td>21.21</td>
<td>0.6090</td>
<td></td>
</tr>
<tr>
<td>EVDI (Ours)</td>
<td>29.17</td>
<td>0.8797</td>
<td>23.09</td>
<td>0.6929</td>
<td>28.77</td>
</tr>
</tbody>
</table>

DAVIS240C camera where the images are minimally blurred. We up-convert the frame rate and synthesize blurry frames using the same manner as the GoPro dataset, and form a semi-synthetic dataset of blurry videos.

**RBE:** The RBE dataset [34] employs a DAVIS346 camera to collect real-world blurry videos and the corresponding event streams, which can be used for training with the proposed self-supervised learning framework and verifying the performance of our method in real-world scenarios.

**Implementation details.** We implement the LDI network with 5 convolution layers and the fusion network with 6 convolution layers, 2 residual blocks and 1 CBAM [33] block, forming a lightweight network architecture (detailed in the supplementary material). Our network is implemented using Pytorch and trained on NVIDIA GeForce RTX 2080 Ti GPUs with batch size 4 by default. The Adam optimizer [12] is employed accompanied with the SGDR [15] schedule where the parameter $T_{max}$ is set to 100 (reset the learning rate every 100 epochs). We set the number of temporal bins $N = 16$ for LDI inputs and randomly crop the images to $128 \times 128$ patches for training. The training process is divided into two stages: We first train our model in the deblurring setting with the weighting factors $[\alpha, \beta, \gamma] = [512, 1, 1 \times 10^{-1}]$ and the initial learning rate $1 \times 10^{-3}$ for 100 epochs, and then continue training under the setting of unified deblurring and interpolation with the weighting factors $[\alpha, \beta, \gamma] = [128, 1, 1 \times 10^{-1}]$ and the initial learning rate $1 \times 10^{-4}$ for another 100 epochs. We train a model for each dataset and evaluate it on the corresponding dataset, which is convenient as we do not need ground-truth images for supervision.

5.2. Results of Deblurring

For the setting of deblurring, we synthesize 1 blurry image using 49 frames on the GoPro and HQF datasets and evaluate the performance by restoring 7 original frames (before up-converting the frame rate) per blurry image. We compare to the state-of-the-art frame-based deblurring approach LEVS [11] and event-based methods including EDI [22], LEDVDI [14], eSL-Net [31], RED [34], and evaluate the results by metrics PSNR, SSIM [32], and LPIPS [36].

As demonstrated in Tab. 1, the proposed method achieves remarkable deblurring results compared to the state-of-the-arts. The performance of LEVS is limited under highly dynamic scenes, e.g., Fig. 4, due to the motion ambiguity. For event-based approaches, the model-driven
method EDI provides comparable performance to LEVS by exploiting the precise motion embedded in events. LEDVDI further enhances this advantage on the GoPro dataset through introducing learning-based techniques, but this overwhelming performance is not maintained in the HQF dataset due to the inconsistency between datasets. RED achieves the most competing results with semi-supervised learning, but it still pays performance losses to balance different data distributions. Our EVDI method tackles this problem by fitting the particular data distribution with the self-supervised framework, and thus achieves the best performance on each dataset. Meanwhile, our model only contains 0.393M network parameters, which is an order of magnitude smaller than other methods except eSL-Net. Note that eSL-Net requires 122.8G FLOPs to infer a 160×320 image due to its recursive structure, while our model only needs 13.45G FLOPs, maintaining the overall efficiency.

5.3. Results of Interpolation

For the task of interpolation, we collect consecutive 97 frames (which are 13 original frames before up-converting) as a set of input, and synthesize 1 blurry images using 41 frames at both ends, leaving 1 latent original frame in the middle for evaluation (noted as 1 frame skip). Similarly, we design another case of 3 frame skips by synthesizing 1 blurry image with 33 frames and leaving 3 original middle frames. The frame-based interpolation methods Jin’s work [10], BIN [25], DAIN [1] and event-based approaches EDI [22], LEDVDI [14], Time Lens [30] are compared.

The results in Fig. 5 and Tab. 2 demonstrate the difficulty of blurry video interpolation for frame-based approaches. The optical flow used in DAIN often projects motion blur to the interpolation results as shown in Fig. 5. Jin’s work employs a cascaded scheme for deblurring and interpolation, which tends to propagate the deblurring error to the interpolation stage. Despite BIN achieves joint deblurring and interpolation, it is difficult to synthesize the accurate intermediate frames due to the missing information between frames. For event-based methods, LEDVDI and Time Lens are able to correctly estimate the intermediate frames using the precise motion inside events. However, the performance of Time Lens is highly dependent on the quality of reference frames, and LEDVDI often faces performance drop when inferring on other datasets due to data inconsistency. The proposed EVDI method utilizes both frames and events to guarantee the interpolation quality, and tackles the inconsistency problem by learning on the target scenarios with the self-supervised framework, thus producing better results.

<table>
<thead>
<tr>
<th>$L_{B,S}$</th>
<th>$L_{B,E}$</th>
<th>$L_{S,E}$</th>
<th>Fusion</th>
<th>PSNR / SSIM / LPIPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>22.66 / 0.6769 / 0.0954</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>9.152 / -0.0631 / 0.3213</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>6.847 / 0.0192 / 0.7840</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>29.97 / 0.8998 / 0.0182</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>28.07 / 0.8734 / 0.0274</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>29.36 / 0.8924 / 0.0221</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>30.15 / 0.9026 / 0.0162</td>
</tr>
</tbody>
</table>

5.4. Ablation Study

We study the importance of each loss in our self-supervised framework and investigate the contribution of the fusion network. The following conclusions are drawn:

Necessity of loss combination. As depicted in Fig. 6, blurry-sharp loss $L_{B,S}$ contributes to brightness consistency but cannot produce sharp results. Blurry-event loss $L_{B,E}$ and sharp-event loss $L_{S,E}$ are able to deal with motion ambiguity by gaining supervision from blurry frames and events, respectively, but do not constrain brightness. With the combination of loss functions, the brightness inconsistency and motion ambiguity can be simultaneously addressed by taking the complementary advantage of $L_{B,S}$ and $L_{B,E}, L_{S,E}$.

Importance of information fusion. Although $L_{B,E}$ and $L_{S,E}$ are both capable of tackling motion ambiguity, their supervision comes from different information sources: $L_{B,E}$ exploits blurry frames $B_i, B_{i+1}$ to supervise the estimation of $E(f, T_i), E(f, T_{i+1})$, while $L_{S,E}$ utilizes events to constrain the generation of $\hat{L}(f)$. Hence, combining $L_{B,E}$ and $L_{S,E}$ will achieve the best performance, as demonstrated in Tab. 3. Moreover, the fusion network also improves the results by fusing the information from different blurry frames and events.

6. Conclusion

This paper introduces a unified framework of event-based video deblurring and interpolation that generates high frame-rate sharp videos from low-frame-rate blurry inputs. Through analyzing the mutual constraints among blurry frames, sharp latent images and events, a self-supervised learning framework is also proposed to enable network training in real-world scenarios without any labeled data. Evaluation on both synthetic and real-world datasets demonstrates that our method competes favorably against state-of-the-arts while maintaining an efficient network design, showing potential for practical applications.


