Motion Deblurring with Real Events

Fang Xu[1], Lei Yu[1]*, Bishan Wang[1], Wen Yang[1]*, Gui-Song Xia[2], Xu Jia[3]*, Zhendong Qiao[4], Jianzhuang Liu[4]
[1]School of Electronic Information, Wuhan University
[2]School of Computer Science, Wuhan University
[3]School of Artificial Intelligence, Dalian University of Technology
[4]Noah's Ark Lab, Huawei Technologies
{xufang, ly.wd, bswang, yangwen, guisong.xia}@whu.edu.cn, xjia@dlut.edu.cn, {qiaozhendong, liu.jianzhuang}@huawei.com

Abstract

In this paper, we propose an end-to-end learning framework for event-based motion deblurring in a self-supervised manner, where real-world events are exploited to alleviate the performance degradation caused by data inconsistency. To achieve this end, optical flows are predicted from events, with which the blurry consistency and photometric consistency are exploited to enable self-supervision on the deblurring network with real-world data. Furthermore, a piecewise linear motion model is proposed to take into account motion non-linearities and thus leads to an accurate model for the physical formation of motion blurs in the real-world scenario. Extensive evaluation on both synthetic and real motion blur datasets demonstrates that the proposed algorithm bridges the gap between simulated and real-world motion blurs and shows remarkable performance for event-based motion deblurring in real-world scenarios.

1. Introduction

Due to motion ambiguities as well as the erasure of intensity textures [12], the task of motion deblurring is severely ill-posed [11, 25]. With the help of an event camera which can "continuously" emit events asynchronously with extremely low latency (in the order of μs) [15, 6], inherently embedding motions and textures [2], the task of motion deblurring [24, 22, 28] can be essentially alleviated. Many event-based motion deblurring methods have been proposed by learning from synthesized dataset composed of simulated events and blurry images as well as sequences of sharp clear ground-truth images [10, 31]. However, the inconsistency between synthetic and real data degrades the performance of inference on real-world event cameras [29].

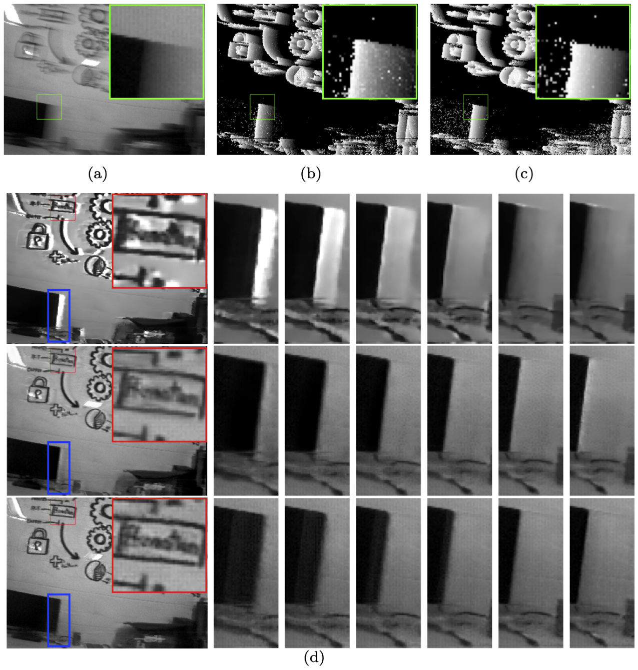

Figure 1. Illustrative example of inconsistency between synthesized and real-world motion blurs with respect to the event time-surface [14]: (a) a real-world motion blurred image; (b) the time-surface of real-world events corresponding to (a); (c) the time-surface of events collected with the same trajectory as (a) but at a slow motion speed; (d) deblurred results respectively by eSL-Net [31] (top row), LEDVDI [16] (middle row) and our proposed RED-Net (bottom row) that trained with real-world events and real-world motion blurs, sequence predictions of the blue area are shown on the right. Our proposed RED-Net generates less *halo artifacts* and achieves the best visualization performance.

The physical intrinsic noise of event cameras raises the difficulty of simulating labeled events that highly match the real event data [26]. Even though the event simulator to some extent reduces the gap by considering the pixel-to-pixel variation in the event threshold [26], additional noise effects such as background activity noise and false negatives [1] still exist, leading to tremendous discrepancy between the virtual events synthesized from event simulators

*Corresponding author

The research was partially supported by the National Natural Science Foundation of China, No. 61871297 and the Fundamental Research Funds for the Central University of China, WHU No. 2042020kf0019 and DUT No. 82232026. And the numerical calculations have been done on the supercomputing system in the Supercomputing Center of Wuhan University.

[26] and the real events emitted by event cameras. An alternative approach is to build a labeled dataset composed of real-world events accompanying with synthesized blurry images, and then train networks on it [10]. Unfortunately, obtaining such pairs is not always easy, which needs to be captured with a slow motion speed as well as under good lighting conditions to avoid motion blur. Subsequent blurry image synthesis and alignment on temporal domain is also tedious but indispensable. Furthermore, inconsistency still exists between the events associated with synthesized and real-world motion blur in that limited read-out bandwidth leads to more event timing variations, as shown in Fig. 1.

In this paper, we propose a novel framework of learning the event-based motion deblurring network in a self-supervised manner, where real-world events and real-world motion blurred images are exploited to alleviate the performance degradation caused by data inconsistency and bridge the gap between simulations and real-world scenario. The proposed framework consists of two neural networks, an event-based motion deblurring network (Deblur-Net) and an event-based optical flow estimation network (OF-Net). The Deblur-Net is fed with both events and a single motion blurred image, and outputs a sequence of sharp clear images, while the OF-Net receives events and provides motions between the reconstructed sharp clear images. We relate motions and sharp clear images according to the photometric constancy [36]. The real-world motion blurs with non-linearities are considered by a piece-wise linear motion (PLM) model which improves the accuracy of optical flow and thus provides a more precise blurring model between the reconstructed intensity images from Deblur-Net and the blurry input. The overall network is jointly trained end-to-end over partially labeled dataset composed of synthetic data using ESIM [26] with ground-truth sharp clear images and real-world data only containing real-world events and real-world motion blurred images, which can be captured simultaneously by the DAVIS [3] or with a dual camera set connecting the event camera and the RGB camera with a beam splitter [32].

The main contributions of our work are in three folds:

- We propose a framework of learning the event-based motion deblurring network with real-world events, which remarkably improves the performance of motion deblurring in the real-world scenario.
- We propose a piece-wise linear motion model to consider the motion nonlinearity, based on which the OF-Net is ameliorated for real events and outputs accurate and dense motion flows.
- Extensive experiments show that the proposed method can yield high quality sharp frames and achieve state-of-the-art results on the real blurry event dataset.

2. Related Work

Motion Deblurring. Even though motion blur degenerates photographs by averaging over the exposure time intervals, it inherently embeds both motions and textures of the moving objects and thus enables the possibility of motion deblurring for visualization of the dynamic scene behind blurred photographs [12, 11, 25]. Early attempts address this problem by assuming spatially uniform motion where blurs can be modelled as the convolution of a blur kernel with a sharp image, leading to the kernel-based motion deblurring approaches by means of regularized deconvolution [13, 30]. For motion blurs caused by complex motion behaviors, pixel/patch-wise motion flows are often required to flexibly depict non-uniform motions. To achieve this end, the optical flow is often estimated from a single blurry image [7] or consecutive frames [34] and provides photometric constancy [8] between the recovered latent sharp images, leading to flow-based motion deblurring approaches. Motion blurs bring the ambiguity of temporal ordering as well as the erasure of spatial textures [12], thus several priors have been successfully employed for blur kernels, *e.g*., sparsity [33], Gaussian scale mixture [5], and dark channel [21], and for optical flows, *e.g*., linear motion [17], and order invariance [12], which however may generate artifacts and degrade deblurring performance if the above assumptions are not fulfilled [35].

An alternative approach is to learn a neural network for motion deblurring directly from the data, which often achieves prominent performance [12, 11, 25, 7, 20]. Jin *et al*. pioneer to recover a sequence of sharp frames from a single motion-blurred image [12], where they sequentially train multiple neural networks with a temporal invariant loss. Purohit *et al*. utilize a single recurrent neural network to generate the entire sequence [25] which, however, may suffer from ambiguities in temporal ordering. Take this into account, Jin *et al*. tackle the problem of the temporal ambiguity by feeding multiple motion blurred images [11], while Rengarajan *et al*. additionally use two consecutive images over a short exposure time [27]. On the other hand, Chen *et al*. propose a Reblur2Deblur network [4] for motion deblurring via self-supervised learning to leverage the physical model and the data prior, where multiple blurry images and a linear motion assumption between them are essentially required.

Event-Based Motion Deblurring. Event camera measures per-pixel brightness change and outputs an event once the change exceeds a threshold [6]. The triggered events can be "continuously" emitted asynchronously with extremely low latency and thus provide missing information during exposure intervals if the motion blur happens [24]. The temporal ambiguity and texture erasure can be easily tackled by introducing events into the deblurring algorithms [24, 16]. Event-based motion deblurring methods can be categorized into two groups, *i.e.*, *model driven* and *data driven* algorithms. Model-based algorithms relate events, blurry images and the corresponding latent sharp clear images according to the physical event generation principle [24, 23, 28]. However, due to the imperfections of physical implementation including intrinsic noise and limited read-out bandwidth [9, 6], real events are essentially with noise both in temporal and spatial domains [15] which inevitably degrades performance [28].

Data driven algorithms relax the above limitations by utilizing neural networks and directly learn the relation from a blurry image to a sequence of sharp clear images with the aid of events [16, 31, 10]. For training purpose, a synthesized dataset composed of labeled events and blurry images is commonly simulated from sharp clear video sequences [10, 31] which, however, may have inconsistency to the real events due to event noise in the spatio-temporal domain. Even though data variations have been considered by manually adding noise [31, 29], the generalizability remains limited for real event cameras.

Jiang *et al.* [10] build the Blur-DVS dataset based on real events triggered at a slow motion speed (to obtain a sequence of sharp clear intensity images as the ground truth) and train their network on it, where the sim-to-real gap is reduced to some extent [16]. However, building such a dataset is strenuous and requires strict conditions to obtain blur-free ground-truths, *e.g.*, relative slow motion between camera and scenario and good lighting environment. Besides, due to the limited read-out bandwidth, the inconsistency between events emitted at a slow motion speed and their counterparts at a fast motion speed also exists.

Learning with real-world data is more adaptive to the real-world scenario than synthetic data, and moreover, real-world events and motion blurs can be easily captured at extremely low cost without sophisticated procedures, which motivates us to propose a new framework to learn event-based motion deblurring by leveraging the *real-world events* and *real-world motion blurred images*.

3. Problem Statement

Event-based motion deblur (ED) aims at reconstructing a sequence of sharp and clear latent images $\{\mathbf{I}_t\}_{t\in T}$ from a single motion blurred image $\mathbf{B}$ captured with exposure time T and corresponding events $\mathbf{E}_T \triangleq \{(\mathbf{x}_i, p_i, t_i)\}_{t_i \in T}$ triggered in T where t_i and $\mathbf{x}_i$ respectively denote the timestamp and the pixel location of the i-th event, and $p_i \in \{+1, -1\}$ is the polarity. Many algorithms are proposed to tackle the problem of ED by learning-based approaches [16, 10, 31], *i.e.*,

$$\mathbf{I}_t = \text{ED-Net}\left(\mathbf{B}, \mathbf{E}_T\right), \quad t \in T \tag{1}$$

which are commonly trained over a synthetic dataset $\mathcal{D}_s \triangleq \{\hat{\mathbf{B}}_k, \hat{\mathbf{E}}_{T_k}, \tilde{\mathcal{G}}_k\}_k$ where $\hat{\mathbf{B}}_k$ and $\hat{\mathbf{E}}_{T_k}$ are synthesized from the ground-truth sequence $\tilde{\mathcal{G}}_k$, *i.e.*, $\hat{\mathbf{E}}_{T_k} = \text{ESIM}(\tilde{\mathcal{G}}_k)$ [26] and $\hat{\mathbf{B}}_k = \text{Avg}(\tilde{\mathcal{G}}_k)$.

Real-World Events. Real-world events $\tilde{\mathbf{E}}_{T_k}$ are different from the simulated events $\hat{\mathbf{E}}_{T_k}$ in two aspects: (1) *event noises in spatial domain* are principally induced by physical intrinsic camera imperfections (*etc.*, variant event threshold), background activities, false negatives, *etc.*, [6]; (2) *event noises in temporal domain* are produced owing to the limited read-out bandwidth and bring timing variances [9], which commonly exist at a fast motion speed when the motion blur happens. Statistics of noises from both aspects are too complicated to be addressed in current event simulators [26], leading to inconsistency between synthetic dataset $\mathcal{D}_s$ and its real-world counterparts.

For spatial event noises, [16] has built a real event dataset $\mathcal{D}_e \triangleq \{\hat{\mathbf{B}}_k, \tilde{\mathbf{E}}_{T_k}, \tilde{\mathcal{G}}_k\}_k$ composed of the real-world events $\tilde{\mathbf{E}}_{T_k}$, the real-world sharp clear ground-truth images $\tilde{\mathcal{G}}_k$ captured at an ultra slow speed to avoid motion blurs, and synthesized blurry images $\hat{\mathbf{B}}_k = \text{Avg}(\tilde{\mathcal{G}}_k)$. Theoretically, training on $\mathcal{D}_e$ is able to take into account the spatial noises, while the temporal noises cannot be considered since they always accompany with real-world motion blurs where no ground-truth images can be obtained. It raises the problem of learning on unlabeled real-world dataset $\mathcal{D}_r \triangleq \{\tilde{\mathbf{B}}_k, \tilde{\mathbf{E}}_{T_k}\}_k$, with real-world blurry images $\tilde{\mathbf{B}}_k$ and real-world events $\tilde{\mathbf{E}}_{T_k}$.

Real-World Motions. Motions and intensity textures are coupled together based on the photometric consistency and thus commonly exploited for the motion deblur [11, 27]. For simplicity, most existing motion deblur algorithms assume linear motion during the exposure time T [23, 4], where the motion flow stays constant. However, it is not always correct for dynamic scenes since real-world motions might be nonlinear especially at a fast motion speed.

Thus the main obstacles of leveraging the power of events to learn motion deblurring in real world are in two folds. First, the network should be trained on real-world data with real events and real motion blurs. Second, the motion non-linearity should be considered to deal with motion blurs in real-world complex dynamic scenes.

4. Method

Existing event-based deblurring methods are usually developed within a supervised learning framework, of which performance is limited to the specific training data (simulated events or synthesized blurred image from the dataset captured with slow motion). Taking into account both deblurring and the physical image blur formation process, we construct a semi-supervised learning framework which can generalize well to real-world motion-blurred images caused by fast motion.

Fig. 2 illustrates the overall architecture of our framework. The Deblur-Net takes a single motion-blurred image

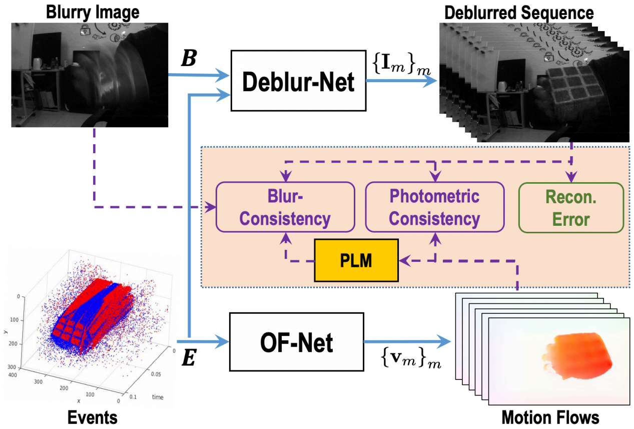

Figure 2. Overview of the proposed learning framework for Real-world Event based motion Deblurring Network (RED-Net), where the blur-consistency and phtometric-consistency provide self-supervised losses for real-world datasets and the reconstruction error provides the supervised loss for synthetic datasets.

$\mathbf{B}$ and its event data $\mathbf{E}$ as input and outputs a sequence of sharp frames $\{\mathbf{I}_m\}_{m=0}^{M-1}$, where M is the number of estimated sharp frames. The supervised loss that compares the estimated sharp frames with ground truth only works for labeled synthesized dataset. To guide the semi-supervised learning of Deblur-Net, our framework consists of an additional OF-Net for motion estimation, which takes event data $\mathbf{E}$ as input and outputs optical flow serving as motion information. On the strength of the physical blur formation process, a motion-blurred image can be re-rendered on the basis of estimated sharp images and motion information, which provides a *blur-consistency* constraint by comparing with the original motion-blurred image. Besides, estimated sharp images are propagated by the estimated optical flow, which provides an additional constraint of *photometric-consistency* by comparing the propagated sharp images with the directly recovered ones.

Considering the motion non-linearity in real-world complex dynamic scenes, we take full advantage of high temporal resolution of event camera. Since event data is able to provide continuous motion information in a period, we do not need strong assumption on linear constant motion in a large time interval as other deblurring methods [7, 23, 4, 17]. Event data can be used in modeling highly nonlinear motion happened in the exposure time.

4.1. Blur-consistency Constraint

Providing the latent sharp image $\mathbf{L}_t(\mathbf{x})$ during the exposure time interval T, the blurring process can be physically formulated as the average of them in the duty cycle,

$$\bar{\mathbf{B}}(\mathbf{x}) = \frac{1}{|T|}\int_{t\in T}\mathbf{L}_t(\mathbf{x})dt \approx \frac{1}{N}\sum_{n=0}^{N-1}\mathbf{L}_n(\mathbf{x}), \tag{2}$$

where $\{\mathbf{L}_n\}_{n=0}^{N-1}$ is the discrete version of $\mathbf{L}_t$ and a large enough $N \gg M$ is required to achieve small discretization error. Providing recovered sequences $\{\mathbf{I}_m\}_{m=0}^{M-1}$ output from the Deblur-Net, one can resolve $\mathbf{L}_n$ by interpolating with $\{\mathbf{I}_m\}_{m=0}^{M-1}$ over the motion field $\mathbf{u}_n(\mathbf{x})$ representing per-pixel flow from $\mathbf{L}_0$ to $\mathbf{L}_n$ *i.e.*,

$$\mathbf{L}_n(\mathbf{x}) = \mathbf{L}_0(\mathbf{x} + \mathbf{u}_n(\mathbf{x})). \tag{3}$$

For simplicity, a linear motion (LM) model is commonly assumed [17, 23, 7], *i.e.*,

$$\mathbf{u}_n(\mathbf{x}) = n\mathbf{v}(\mathbf{x}), \tag{4}$$

with $\mathbf{v}$ denoting the per-pixel unit motion flow which is assumed to be constant during the exposure time interval T.

To consider the motion non-linearity for real-world scenario, we propose a piece-wise linear motion (PLM) model by making full use of rich motion information encoded in events. We divide event stream within the exposure time T into $M-1$ equal time intervals and the LM model is assumed within each interval. Finally, we can get the following PLM model, for $n \in [mK, (m+1)K)$,

$$\mathbf{u}_n(\mathbf{x}) = (n - mK)\mathbf{v}_m(\mathbf{x}) + \mathbf{u}_{mK}(\mathbf{x}), \tag{5}$$

with $\mathbf{v}_m$ the per-pixel unit motion flow that associates with the m-th time interval, $m = 0, 1, ..., M-1$ and $K \triangleq N/M$ (N can be chosen as integer times of M). And moreover, $\mathbf{v}_m$ can be predicted from the OF-Net using events in the m-th time interval and resolve a nonlinear motion field $\mathbf{u}_n(\mathbf{x})$.

Then providing the sequence of sharp frames $\{\mathbf{I}_m\}_{m=0}^{M-1}$ output from the Deblur-Net, the n-th latent image $\mathbf{L}_n$ with $n \in [mK, (m+1)K)$ can be resolved by warping the m-th frame $\mathbf{I}_m$ using the motion field defined in (5),

$$\begin{aligned}\mathbf{L}_n(\mathbf{x}) &= \text{Warp}\left(\mathbf{I}_m(\mathbf{x}), \mathbf{u}_n(\mathbf{x})\right)\\ &\triangleq \mathbf{I}_m(\mathbf{x} + (n - mK)\mathbf{v}_m(\mathbf{x})).\end{aligned} \tag{6}$$

Hence, we can re-render a corresponding reblurred image based on (2) using the generated latent sharp images $\{\mathbf{L}_n\}_{n=0}^{N-1}$. A *blur-consistency loss* is constructed between the reblurred image $\bar{\mathbf{B}}$ and the original blurry input $\mathbf{B}$:

$$\mathcal{L}_{blur} = \|\bar{\mathbf{B}} - \mathbf{B}\|_1. \tag{7}$$

By imposing this loss in the whole framework, we build a bridge between the motion-blurred image and its corresponding sharp clear images, which provides a self-supervision to the network. For real-world motion blurs with motion non-linearities, PLM is important to ensure the model accuracy and thus provide faithful supervision even with noise disturbances.

4.2. Photometric-consistency Constraint

In addition to the *blur-consistency* constraint on the sequence level, further *photometric-consistency* constraint is imposed, which utilizes the inter-frame connection between the temporal consistency of estimated sharp images and motion information encodes in events. We warp $\mathbf{I}_{m+1}$ to $\mathbf{I}_m$ with the predicted optical flow $\mathbf{v}_m$ via the OF-Net by feeding events in the m-th time interval and obtain the warped image:

$$\bar{\mathbf{I}}_m(\mathbf{x}) = \mathbf{I}_{m+1}(\mathbf{x} + K\mathbf{v}_m(\mathbf{x})). \tag{8}$$

Then *photometric-consistency* loss is constructed by comparing the propagated sharp images with the recovered ones directly predicted by Deblur-Net:

$$\mathcal{L}_{photo} = \frac{1}{M-1}\sum_{m=0}^{M-2} \|\bar{\mathbf{I}}_m - \mathbf{I}_m\|_1. \tag{9}$$

4.3. Optimization

We propose to train the Deblur-Net and OF-Net over a partially labeled dataset, composed of the synthetic dataset $\mathcal{D}_s$ with ground-truth sharp clear images and the real-world dataset $\mathcal{D}_r$ without ground truth. Apparently, the blur-consistency and the photometric-consistency are applicable for both $\mathcal{D}_s$ and $\mathcal{D}_r$, and provide self-supervised losses respectively defined as $\mathcal{L}^s_{blur}, \mathcal{L}^s_{photo}$ and $\mathcal{L}^r_{blur}, \mathcal{L}^r_{photo}$ according to (7) and (9). Besides, for the synthetic dataset $\mathcal{D}_s$ with a sequence of ground-truth sharp clear images $\tilde{\mathcal{G}}$, we use it to supervise the network via a reconstruction error loss,

$$\mathcal{L}_{error} = \frac{1}{M}\sum_{m=0}^{M-1} ||\mathbf{I}_m - \mathbf{G}_m||_1, \tag{10}$$

where $\mathbf{G}_m \in \tilde{\mathcal{G}}$. Thus the overall function is as follows:

$$\mathcal{L} = \mathcal{L}^s_{error} + \alpha\mathcal{L}^s_{blur} + \beta\mathcal{L}^s_{photo} + \gamma\mathcal{L}^r_{blur} + \delta\mathcal{L}^r_{photo}, \tag{11}$$

with α, β, γ and δ denoting the balancing parameters.

5. Experiments

5.1. Experimental Settings

Dataset. The proposed Real-world Event-based motion Deblurring network (RED-Net) is trained in a self-supervised manner, where one synthetic dataset (*GoPro*) is provided for training with ground-truth and two datasets respectively captured by a DAVIS240C camera (*HQF* provided in [29] containing real events but blur-free intensity frames) and a DAVIS346 camera (*RBE* built in this paper for the real-world scenario with real events and real motion blurs).

GoPro: Based on the GoPro dataset [18], we build a synthetic dataset as [31] composed of simulated events and synthetic blurry images as well as sharp clear ground-truth images. We first increase the frame rate by interpolating 7 images between consecutive frames [19] and then generate both events and blurry images based on the interpolated high frame-rate sequences. ESIM [26] is exploited to simulate events with consideration of per-pixel threshold variation. And blurry images are simply obtained by averaging over 49 consecutive images.

HQF: We construct a similar dataset as Blur-DVS [16] based on HQF dataset [29], which contains real-world events and sharp clear ground-truth frames captured simultaneously from a DAVIS240C that are well-exposed and minimally motion-blurred. Providing the ground-truth frames, motion blurs can be synthesized using the same manner as the GoPro dataset. Finally, the HQF dataset contains real-world events and synthesized blurry images as well as the ground-truth frames. The ground-truth frames are provided only for quantitative evaluation. In training stage of our proposed RED-Net over the HQF dataset, only real events and synthesized blurry images are used.

RBE: A Real-world Blurry images and Events (RBE) dataset is built with a DAVIS346 camera and only contains real-world events and real-world motion blurs, which can be collected in a facilitated manner. Thus the RBE dataset can provide a large number of paired real-world data, which can be fed into the RED-Net and improve the adaptivity to the real-world scenario.

Implementation Details For the Deblur-Net and OF-Net modules, we take advantage of existing neural network architectures which have performed well in the past for the respective learning tasks. Inspired by that a sharp image can be mapped from a blurry image using a residual term encoded with events [24], we select a residual network to predict a sequence of sharp frames from a single motion-blurred image and its event data. In particular, we adopt the residual network from Jin *et al.* [11]. For the OF-Net, we adopt the EV-FlowNet architecture from Zhu *et al.* [36], which is widely used to predict optical flow from events.

Our network is implemented using Pytorch on a single NVIDIA Geforce RTX 3090 GPU. During training, we randomly crop the samples into 128 × 128 patches. Adam optimizer is used and the maximum epoch of training iterations is set to 30. The learning rate starts at 10^{-4}, then decays by 25% every five epochs from the 15-th epoch. The weighting factors α, β, γ and δ are all set to 1, the number of deblurred images $\{\mathbf{I}_m\}_{m=0}^{M-1}$ and latent sharp images $\mathbf{L}_n$ warped from $\mathbf{I}_m$ are respectively $M = 7$ and $K = 11$. The optical flow module is pretrained on the Multi-Vehicle Stereo Event Camera dataset (MVSEC) [36], where the reference ground-truth optical flow is provided.

RED-Nets. To validate the effectiveness of exploiting real-world data, RED-Net is trained respectively over three different datasets, *i.e.*, GoPro, GoPro+HQF and GoPro+RBE, and final networks are respectively denoted as *RED-GoPro,* *RED-HQF and RED-RBE*. Specifically, RED-GoPro is trained only over synthesized GoPro dataset with the supervision of ground-truth frames. Both RED-HQF and RED-RBE are trained in a self-supervised manner, where RED-HQF uses real-world events but synthetic motion blurs, while RED-RBE uses real-world events and motion blurs.

To validate the superiority of piece-wise linear motion model (PLM), we replace it with the linear motion model (LM) as (4), and train RED-Net respectively on GoPro and GoPro+RBE datasets. Specifically, the resulted network is denoted as RED-(GoPro/RBE)-LM and its counterpart with PLM is denoted as RED-(GoPro/RBE)-PLM. By default, RED-(GoPro/RBE) means RED-(GoPro/RBE)-PLM in the following sections.

5.2. Results of Optical Flow

Fig. 3 presents the optical flow (OF) outputs of OF-Net from different RED-Nets and their corresponding deblur results of a real-world motion blurred image containing a fast moving *magic cube*. For comparison, the output of OF-Net pretrained on MVSEC dataset (*i.e*. EV-FlowNet) is also given. Apparently, the output of OF-Net from RED-RBE trained on real data achieves the best performance with consistent motion and sharp clear edge. Thus the deblur image of RED-RBE has the best qualitative appearance as well.

To reveal the advantage of PLM, we compare the outputs of OF-Net from RED-GoPro-LM as shown in Fig. 3(b) and RED-GoPro-PLM as shown in Fig. 3(c). It is obvious that RED-GoPro-PLM gives better OF result with clear edges and stable background while RED-GoPro-LM suffers from blurry effects and background movement. Furthermore, when violating the linear motion assumption, the result of OF-Net from RED-GoPro-LM is even worse than its origin, i.e. EvFlowNet as shown in Fig. 3(a) and 3(b).

Besides, we compare the OF results of EvFlowNet and two OF-Nets respectively from RED-GoPro-PLM and RED-RBE as shown in Fig. 3(a), 3(c) and 3(d). With the supervision of blurry image and the intensity sequence according to the blur-consistency loss (7), OF-Nets from both RED-GoPro-PLM and RED-RBE can give better results than EvFlowNet. It reveals that the motion information embedded in the blurry images helps the prediction of OF within our framework. Moreover, OF-Net from RED-RBE trained over the real dataset outperforms that of RED-GoPro-PLM trained over the synthetic dataset and gives more consistent motions of the cube surface as shown in Fig. 3(c) and 3(d). And it validates the advantage and necessity of training on real-world dataset.

5.3. Comparisons with State-of-the-art Methods

Since our method is able to remove motion blur and reconstruct a sequence of latent sharp clear images, the deblur performance is evaluated on both single frame and sequence reconstructions. We compare the proposed RED-Nets to state-of-the-art conventional deblurring methods including blur2mflow [7] and LEVS [12], and event-based motion deblurring methods including EDI [24], eSL-Net [31] and LEDVDI [16]. Due to the lack of ground-truth for real-world motion blurs, we exploit the HQF dataset constructed from blur-free frames for quantitative evaluations, where the effectiveness of training with real-world events are validated. The quantitative results are presented in Tab.1 and show that the proposed RED-Nets bring remarkable improvements comparing to the state-of-the-arts, on average 3.5 dB gain on single frame prediction and 2.5 dB gain on sequence prediction over the HQF dataset with real-world events. Detailed analyses are presented in the following.

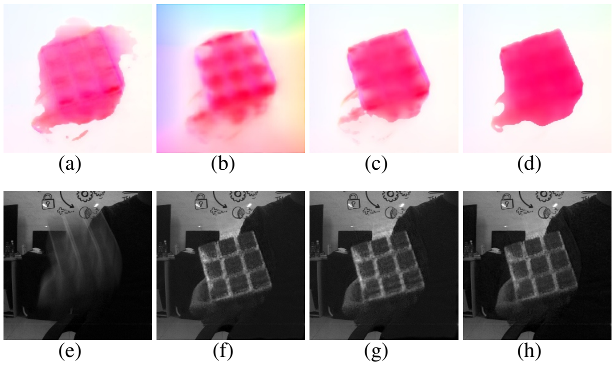

Figure 3. Optical flow output of OF-Net: (a) pre-trained on MVSEC (equivalent to EV-FlowNet), (b) RED-GoPro-LM, (c) RED-GoPro-PLM and (d) RED-RBE. The deblurred results (f), (g) and (h) of the motion blurred image (e) are respectively corresponding to (b), (c) and (d).

Inconsistency of Simulated and Real-world Events. We first compare the deblurring performance of conventional methods, *i.e*., blur2mflow and LEVS to event-based methods, *i.e*., EDI and eSL-Net, over the synthetic GoPro dataset and the HQF dataset with real-world events. It is shown that EDI provides comparable performance to learning based conventional methods with the help of events. For the learning based approaches, eSL-Net enhanced with simulated events outperforms blur2mflow and LEVS by a large margin on the synthetic GoPro dataset. However, the overwhelming performance of eSL-Net is not retained on the HQF dataset with real-world events, which reflects the inconsistency of simulated and real-world events.

Effectiveness of Training with Real-world Events. Comparing to RED-GoPro trained only on synthetic dataset, the other two RED-Nets, *i.e*., RED-HQF and RED-RBE, trained respectively over the HQF and RBE dataset with real events, bring a performance drop on the synthetic GoPro dataset but achieve the PSNR gain on the HQF dataset with real-world events. Specifically, RED-HQF obtains on average 1.1 dB gain on PSNR, which shows the effectiveness of

Table 1. Quantitative comparisons of proposed RED-Nets trained over different datasets to the state-of-the-arts. RED-GoPro, RED-HQF, and RED-RBE are respectively trained on GoPro, GoPro+HQF and GoPro+RBE. Note that LEDVDI only outputs 6 frames for sequence prediction, while the others output 7 frames.

Method	Single frame prediction comparison				Sequence prediction comparison			
	GoPro		HQF		GoPro		HQF	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
blur2mflow [7]	21.818	0.6454	21.539	0.6298	/	/	/	/
LEVS [12]	21.950	0.6406	21.900	0.6367	19.893	0.5546	19.068	0.5403
EDI [24]	21.497	0.6510	20.321	0.6212	20.945	0.6326	19.081	0.5873
eSL-Net [31]	24.791	0.8009	20.438	0.6017	23.955	0.7578	19.866	0.5851
LEDVDI [16]	22.856	0.7334	22.221	0.7567	22.673	0.7329	21.558	**0.7355**
RED-GoPro	**28.984**	**0.8499**	24.149	0.7332	**28.343**	**0.8359**	23.118	0.7116
RED-HQF	27.137	0.8361	25.650	**0.7661**	26.640	0.8204	23.872	0.7292
RED-RBE	26.672	0.8372	**25.717**	0.7629	26.302	0.8247	**24.076**	0.7340

training with real-world events and bridges the sim-to-real gap. Furthermore, we validate this point by training RED-RBE with the real-world data in the RBE dataset which can be constructed in a facilitated manner.

Event-based Motion Deblur with Real-World Events. As shown in Tab. 1, RED-HQF/RBE outperforms EDI and eSL-Net over the HQF dataset with real-world events by a large margin. Furthermore, we exploit the recently proposed network, *i.e.*, LEDVDI trained with real-world events for fair comparisons. Different from our proposed method, LEDVDI is trained with real-world events in a supervised manner. Our approach requires neither blurry synthesis nor strict slow moving speed to collect blur-free ground-truth images, and thus is much easier to be adopted in real-world scenario. From the quantitative results in Tab. 1, it is shown that both RED-HQF and RED-RBF achieve better performance than LEDVDI for single frame prediction and sequence prediction, which validates the superiority of our proposed framework. Note that LEDVDI deals with single frame prediction and sequence prediction separately with different networks, while our proposed RED-HQF/RBE tackle both problems in one network.

Real-World Motion Deblur. The above quantitative comparisons are all based on synthesized motion blurs. Due to the lack of ground-truth frames, the performance with real-world motion blurs are qualitatively investigated, as shown in Fig. 4. We choose two scenes captured manually with a DAVIS346 camera, *i.e.*, *labfloor* and *window*. Without the aid of events, blur2mflow and LEVS fail to recover sharp clear images. For event-based methods, EDI fails to cope with severe motion blurs as illustrated in *window*. Among the learning based approaches, RED-RBE achieves the best visualization performance, while the others are likely to give over smoothed effects or generate *halo artifacts* around black edges, for instance the bottom of garbage can in *labfloor* and the top-left corner of TV in *window*. Compared with RED-HQF and LEDVDI trained with synthesized motion blurs, RED-RBE is trained with real-world motion blurs which can alleviate the problem of event noise in temporal domain, *e.g.*, *timing chaos* caused by limited read-out bandwidth and thus give less halo artifacts.

Table 2. Ablation study of proposed framework w/o *synthesized events* (SynEv), *blurring with LM* (LM) , *blurring with PLM* (PLM) and *real-world events* (ReEv).

Methods	SynEv	LM	PLM	ReEv	GoPro	HQF
Deblur-Net [11]					23.563	21.373
Deblur-Net-GoPro	✓				28.451	23.553
RED-GoPro-LM	✓	✓			28.752	23.988
RED-GoPro	✓		✓		**28.984**	24.149
RED-RBE-LM	✓	✓		✓	26.451	25.266
RED-RBE	✓		✓	✓	26.672	**25.717**

5.4. Ablation Study

The proposed RED-Net improves the performance of motion deblurring by self-supervising with real-word events. To achieve this end, optical flows are predicted from events with which the piece-wise linear motion model (PLM) is proposed to accurately model the blurry procedure even with motion non-linearities. To find out what contributes to the superior performance of the approach, we compare a few variants with and without use of synthesized events (SynEv), blurring with LM (LM), blurring with PLM (PLM) and real-world events (ReEv), as shown in Tab. 2. From the table, we can draw the following conclusions:

Importance of Events. We validate the importance of events by training the Deblur-Net without events over the synthetic GoPro dataset, and the quantitative deblur results over 2 dataset are shown in the first row of Tab. 2. Comparing to RED-Nets with events, there is a large performance gap. It shows that the rich motion information encoded in events can effectively improve the deblurring performance.

Linear *vs*. piece-wise Linear Motion Model. With the optical flow output of the OF-Net, motions provide the inter-frame relationship inside the reconstructed sequence as well as the blurry consistency which relies on the physical blurry procedure. Inaccurate motion model will violate the blurry consistency and thus degrades the deblurring performance. We validate this point by respectively apply LM and PLM

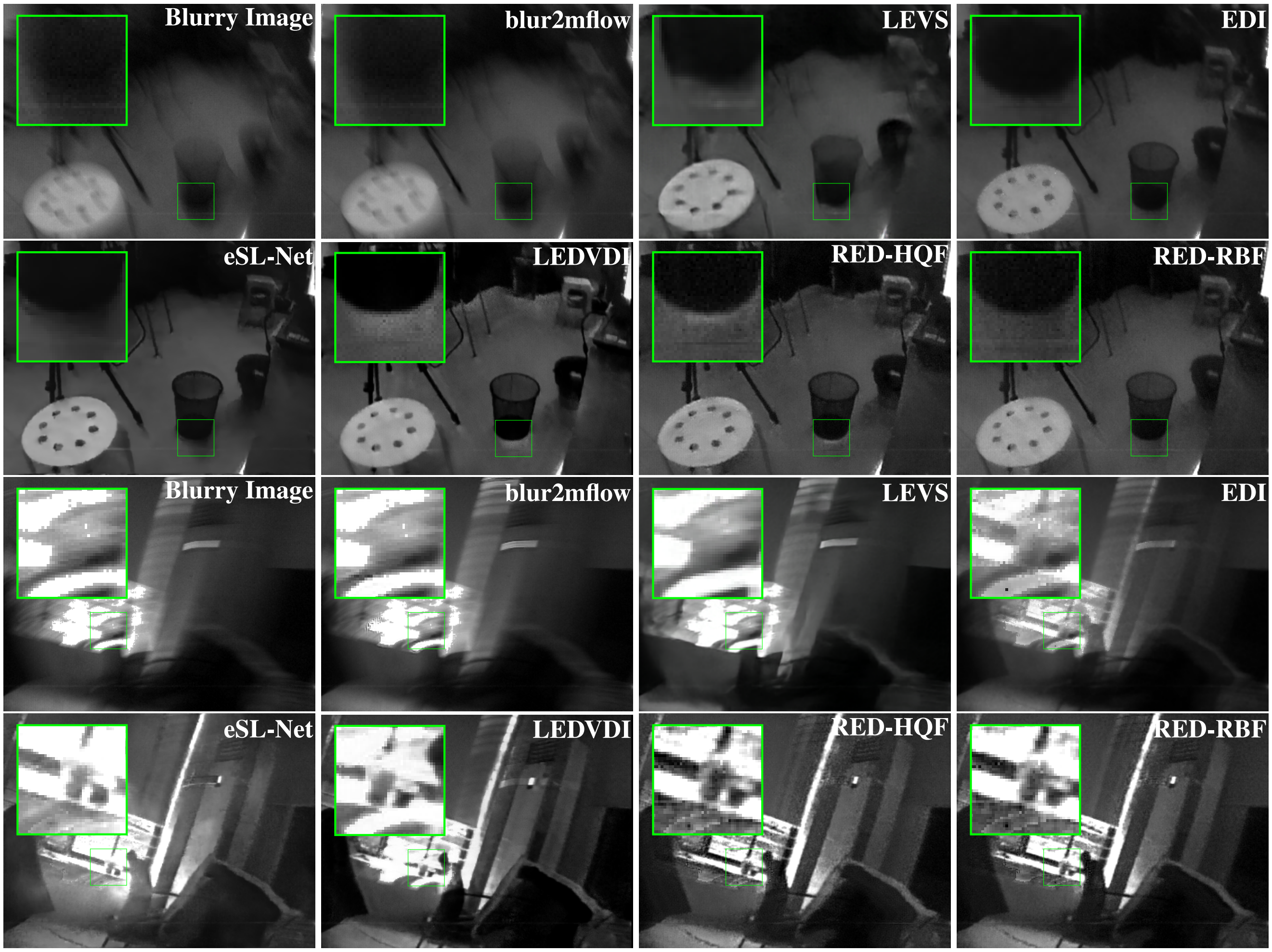

Figure 4. Qualitative results of motion deblur for 2 blurry scenes by 7 different methods where the top-two rows correspond to *labfloor* and the bottom-two correspond to *window*. For each scene, from top-left to bottom-right are respectively the blurry image and its deblurred result by blur2mflow [7], LEVS [12], EDI [24], eSL-Net [31], LEDVDI [16] and our proposed RED-HQF and RED-RBE.

in (2) and their performance can be referred from Tab. 2 in the 3rd, 4th, 5th and 6th rows. Apparently, RED-Nets with PLM outperform that with LM and achieve PSNR gains up to 0.5 dB on the real-world dataset after training with real-world events, while it only gains 0.16 dB without real-world events. It demonstrates that PLM plays an important role especially for real-world events.

Synthetic *vs*. Real-world Events. To validate this point, we train RED-Net respectively on the synthetic GoPro dataset, *i.e*., RED-GoPro and on the real-world RBE dataset, *i.e*., RED-RBE. Their results are shown in the 4th and 6th rows of Tab. 2. And we also trained RED-GoPro and RED-RBE with LM and present their results respectively in the 3rd and 5th rows of Tab. 2. By comparisons, we can easily find that the self-supervision with real-world events can achieve PSNR gains 1.6 dB with PLM (1.3 dB with LM) on real-world dataset. It shows that the supervision of real-world events plays a major role in improving the deblurring performance on real-world dataset.

6. Conclusion

A self-supervised learning framework for event-based motion deblurring is proposed where real-world events and real-world motion blurs are exploited to alleviate the performance degradation caused by data inconsistency. Motion non-linearities are also considered through a PLM model to improve the accuracy of physical blur-consistency. Within the proposed learning framework, we evaluate RED-Net over different datasets and validate the effectiveness of the proposed PLM-based blur-consistency and photometric-consistency over the real-world dataset. With a blurry image and corresponding events, the proposed RED-Net can produce a sequence of sharp clear intensity images as well as motion flows between them. Extensive experiments demonstrate that the proposed method can achieve the state-of-the-art with real-world events.

References

[1] R Baldwin, Mohammed Almatrafi, Vijayan Asari, and Keigo Hirakawa. Event probability mask (epm) and event denoising convolutional neural network (edncnn) for neuromorphic cameras. In *CVPR*, pages 1701–1710, 2020. 1

[2] Ryad Benosman, Charles Clercq, Xavier Lagorce, Sio-Hoi Ieng, and Chiara Bartolozzi. Event-based visual flow. *IEEE Transactions on Neural Networks and Learning Systems*, 25(2):407–417, 2013. 1

[3] Christian Brandli, Raphael Berner, Minhao Yang, Shih-Chii Liu, and Tobi Delbruck. A 240×180 130 db 3 μs latency global shutter spatiotemporal vision sensor. *IEEE Journal of Solid-State Circuits*, 49(10):2333–2341, 2014. 2

[4] Huaijin Chen, Jinwei Gu, Orazio Gallo, Ming-Yu Liu, Ashok Veeraraghavan, and Jan Kautz. Reblur2deblur: Deblurring videos via self-supervised learning. In *IEEE International Conference on Computational Photography (ICCP)*, pages 1–9, 2018. 2, 3, 4

[5] Rob Fergus, Barun Singh, Aaron Hertzmann, Sam T Roweis, and William T Freeman. Removing camera shake from a single photograph. In *ACM SIGGRAPH*, pages 787–794. 2006. 2

[6] Guillermo Gallego, Tobi Delbruck, Garrick Michael Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew Davison, Jorg Conradt, Kostas Daniilidis, and Davide Scaramuzza. Event-based vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 1, 2, 3

[7] Dong Gong, Jie Yang, Lingqiao Liu, Yanning Zhang, Ian Reid, Chunhua Shen, Anton Van Den Hengel, and Qinfeng Shi. From motion blur to motion flow: A deep learning solution for removing heterogeneous motion blur. In *CVPR*, pages 2319–2328, 2017. 2, 4, 6, 7, 8

[8] Tae Hyun Kim and Kyoung Mu Lee. Generalized video deblurring for dynamic scenes. In *CVPR*, pages 5426–5434, 2015. 2

[9] IniVation. *Understanding the Performance of Neuromorphic Event-Based Vision Sensors*. IniVation, https://inivation.com/, 1 edition, 05 2020. 3

[10] Zhe Jiang, Yu Zhang, Dongqing Zou, Jimmy Ren, Jiancheng Lv, and Yebin Liu. Learning event-based motion deblurring. In *CVPR*, pages 3320–3329, 2020. 1, 2, 3

[11] Meiguang Jin, Zhe Hu, and Paolo Favaro. Learning to extract flawless slow motion from blurry videos. In *CVPR*, pages 8112–8121, 2019. 1, 2, 3, 5, 7

[12] Meiguang Jin, Givi Meishvili, and Paolo Favaro. Learning to extract a video sequence from a single motion-blurred image. In *CVPR*, pages 6334–6342, 2018. 1, 2, 6, 7, 8

[13] Dilip Krishnan, Terence Tay, and Rob Fergus. Blind deconvolution using a normalized sparsity measure. In *CVPR*, pages 233–240, 2011. 2

[14] Xavier Lagorce, Garrick Orchard, Francesco Galluppi, Bertram E Shi, and Ryad B Benosman. Hots: a hierarchy of event-based time-surfaces for pattern recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(7):1346–1359, 2016. 1

[15] Patrick Lichtsteiner, Christoph Posch, and Tobi Delbruck. A 128 × 128 120 dB 15 μs Latency Asynchronous Temporal Contrast Vision Sensor. *IEEE Journal of Solid-State Circuits*, 43(2):566–576, 2008. 1, 3

[16] Songnan Lin, Jiawei Zhang, Jinshan Pan, Zhe Jiang, Dongqing Zou, Yongtian Wang, Jing Chen, and Jimmy Ren. Learning event-driven video deblurring and interpolation. In *ECCV*, 2020. 1, 2, 3, 5, 6, 7, 8

[17] Peidong Liu, Joel Janai, Marc Pollefeys, Torsten Sattler, and Andreas Geiger. Self-supervised linear motion deblurring. *IEEE Robotics and Automation Letters*, 5(2):2475–2482, 2020. 2, 4

[18] Seungjun Nah, Sungyong Baik, Seokil Hong, Gyeongsik Moon, Sanghyun Son, Radu Timofte, and Kyoung Mu Lee. Ntire 2019 challenge on video deblurring and super-resolution: Dataset and study. In *CVPRW*, pages 1974–1984, 2019. 5

[19] Simon Niklaus, Long Mai, and Feng Liu. Video frame interpolation via adaptive separable convolution. In *ICCV*, pages 261–270, 2017. 5

[20] Thekke Madam Nimisha, Akash Kumar Singh, and Ambasamudram N Rajagopalan. Blur-invariant deep learning for blind-deblurring. In *ICCV*, pages 4752–4760, 2017. 2

[21] Jinshan Pan, Deqing Sun, Hanspeter Pfister, and Ming-Hsuan Yang. Blind image deblurring using dark channel prior. In *CVPR*, pages 1628–1636, 2016. 2

[22] Liyuan Pan, Richard Hartley, Cedric Scheerlinck, Miaomiao Liu, Xin Yu, and Yuchao Dai. High frame rate video reconstruction based on an event camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 1

[23] Liyuan Pan, Miaomiao Liu, and Richard Hartley. Single image optical flow estimation with an event camera. In *CVPR*, pages 1669–1678, 2020. 3, 4

[24] Liyuan Pan, Cedric Scheerlinck, Xin Yu, Richard Hartley, Miaomiao Liu, and Yuchao Dai. Bringing a blurry frame alive at high frame-rate with an event camera. In *CVPR*, pages 6820–6829, 2019. 1, 2, 3, 5, 6, 7, 8

[25] Kuldeep Purohit, Anshul Shah, and AN Rajagopalan. Bringing alive blurred moments. In *CVPR*, pages 6830–6839, 2019. 1, 2

[26] Henri Rebecq, Daniel Gehrig, and Davide Scaramuzza. Esim: an open event camera simulator. In *Conference on Robot Learning*, pages 969–982, 2018. 1, 2, 3, 5

[27] Vijay Rengarajan, Shuo Zhao, Ruiwen Zhen, John Glotzbach, Hamid Sheikh, and Aswin C Sankaranarayanan. Photosequencing of motion blur using short and long exposures. In *CVPRW*, pages 510–511, 2020. 2, 3

[28] Cedric Scheerlinck, Nick Barnes, and Robert Mahony. Continuous-time intensity estimation using event cameras. In *ACCV*, pages 308–324, 2018. 1, 3

[29] Timo Stoffregen, Cedric Scheerlinck, Davide Scaramuzza, Tom Drummond, Nick Barnes, Lindsay Kleeman, and Robert Mahony. Reducing the sim-to-real gap for event cameras. In *ECCV*, 2020. 1, 3, 5

[30] Libin Sun, Sunghyun Cho, Jue Wang, and James Hays. Edge-based blur kernel estimation using patch priors. In *IEEE International Conference on Computational Photography (ICCP)*, pages 1–8, 2013. 2

[31] Bishan Wang, Jingwei He, Lei Yu, Gui-Song Xia, and Wen Yang. Event enhanced high-quality image recovery. In *ECCV*, 2020. 1, 3, 5, 6, 7, 8

[32] Zihao W Wang, Peiqi Duan, Oliver Cossairt, Aggelos Katsaggelos, Tiejun Huang, and Boxin Shi. Joint filtering of intensity images and neuromorphic events for high-resolution noise-robust imaging. In *CVPR*, pages 1609–1619, 2020. 2

[33] Li Xu, Shicheng Zheng, and Jiaya Jia. Unnatural ℓ_0 sparse representation for natural image deblurring. In *CVPR*, pages 1107–1114, 2013. 2

[34] Haichao Zhang and Jianchao Yang. Intra-frame deblurring by leveraging inter-frame camera motion. In *CVPR*, pages 4036–4044, 2015. 2

[35] Shangchen Zhou, Jiawei Zhang, Jinshan Pan, Haozhe Xie, Wangmeng Zuo, and Jimmy Ren. Spatio-temporal filter adaptive network for video deblurring. In *ICCV*, pages 2482–2491, 2019. 2

[36] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Ev-flownet: Self-supervised optical flow estimation for event-based cameras. In *Robotics: Science and Systems*, 2018. 2, 5