High Frame Rate Video Reconstruction Based on an Event Camera

Liyuan Pan, Richard Hartley, *Fellow, IEEE*, Cedric Scheerlinck, Miaomiao Liu, *Member, IEEE*, Xin Yu, and Yuchao Dai

Abstract—Event-based cameras measure intensity changes (called '*events*') with microsecond accuracy under high-speed motion and challenging lighting conditions. With the 'active pixel sensor' (APS), the 'Dynamic and Active-pixel Vision Sensor' (DAVIS) allows the simultaneous output of intensity frames and events. However, the output images are captured at a relatively low frame rate and often suffer from motion blur. A blurred image can be regarded as the integral of a sequence of latent images, while events indicate changes between the latent images. Thus, we are able to model the blur-generation process by associating event data to a latent sharp image. Based on the abundant event data alongside a low frame rate, easily blurred images, we propose a simple yet effective approach to reconstruct high-quality and high frame rate sharp videos. Starting with a single blurred frame and its event data from DAVIS, we propose the *Event-based Double Integral (EDI)* model and solve it by adding regularization terms. Then, we extend it to *multiple Event-based Double Integral (mEDI)* model to get more smooth results based on multiple images and their events. Furthermore, we provide a new and more efficient solver to minimize the proposed energy model. By optimizing the energy function, we achieve significant improvements in removing blur and the reconstruction of a high temporal resolution video. The video generation is based on solving a simple non-convex optimization problem in a single scalar variable. Experimental results on both synthetic and real datasets demonstrate the superiority of our *mEDI* model and optimization method compared to the state-of-the-art.

Index Terms—Event camera (DAVIS), motion blur, high temporal resolution reconstruction, mEDI model, fibonacci sequence

1 Introduction

Event cameras (such as the Dynamic Vision Sensor (DVS) [6] and the DAVIS [7]) are sensors that asynchronously measure intensity changes at each pixel independently with microsecond temporal resolution (if nothing moves in the scene, no events are triggered). The event stream encodes the motion information by measuring the precise pixel-by-pixel intensity changes. Event cameras are more robust to low lighting and highly dynamic scenes than traditional cameras since they are not affected by under/over exposure associated with a synchronous shutter.

Due to the inherent differences between event cameras and standard cameras, existing computer vision algorithms designed for standard cameras cannot be applied to event cameras directly. Although the DAVIS [7] can provide simultaneous output of intensity frames and events, there still exist major limitations with current DAVIS cameras:

- *Low Frame Rate Intensity Images:* In contrast to the high temporal resolution of event data ($\geq 3\,\mu$s frame rate), the current DAVIS only output low frame rate intensity images (≥ 20 ms temporal resolution).
- *Inherent Blur Effects:* When recording highly dynamic scenes, motion blur is a common issue due to the relative motion between the camera and the scene. The output of the intensity image from the APS tends to be blurry.

To address these challenges, various methods have been proposed by reconstructing high frame rate videos. Existing methods can be in general categorized as:

1) Event-only solutions [4], [8], [9], [10], where the results tend to lack the texture and consistency of natural videos (especially for scenes with a static background or a slowly moving background/foreground), as they fail to use the complementary information contained in low frame rate intensity images;
2) Events and intensity images combined solutions [3], [11], [12], which build upon the interaction between both sources of information. However, these methods fail to address the blur issue associated with the captured image frame. Therefore, the reconstructed high frame rate videos can be degraded by blur.

Contrary to existing 'image + event' based methods that ignore the blur effect in the image, or discard it entirely, we give an alternative insight into the problem. While blurred frames cause undesired image degradation, they inherently encode the relative motion between the camera and the observed scene, and the integral of multiple images during the exposure time. Taking full advantage of the encoded

- *Liyuan Pan, Richard Hartley, Cedric Scheerlinck, and Miaomiao Liu are with the Research School of Engineering, Australian National University, Canberra, Australia and Australian Centre for Robotic Vision, Brisbane, QLD 4000, Australia. E-mail: {liyuan.pan, richard.hartley, cedric.scheerlinck, miaomiao.liu}@anu.edu.au.*
- *Xin Yu is with the Faculty of Engineering and Information Technology, University of Technology Sydney, Sydney 2007, Australia. E-mail: xin.yu@uts.edu.au.*
- *Yuchao Dai is with the School of Electronics and Information, Northwestern Polytechnical University, Xi'an 710129, China. E-mail: daiyuchao@gmail.com.*

Manuscript received 17 Apr. 2019; revised 30 July 2020; accepted 26 Oct. 2020. Date of publication 9 Nov. 2020; date of current version 1 Apr. 2022.
(Corresponding author: Liyuan Pan.)
Recommended for acceptance by E. Shechtman.
Digital Object Identifier no. 10.1109/TPAMI.2020.3036667

information in the blurred image would benefit the reconstruction of high frame rate videos.

To tackle above problems, in our previous work [5], we propose an *Event-based Double Integral (EDI)* model to fuse an image (even with blur) with its event sequence to reconstruct a high frame rate, blur-free video. Our *EDI* model naturally relates the desired high frame rate sharp video, the captured intensity frame and event data. Based on the *EDI* model, high frame rate video generation is as simple as solving a non-convex optimization problem in a single scalar variable.

As the *EDI* model is based on a single image, noise from the event data can easily degrade the quality of reconstructed videos, especially at transitions between images. To mitigate accumulated noise from events, we limit the integration to a small time interval around the centre of the exposure time, allowing us to reconstruct a small video segment associated with one image. The final video is obtained by stitching all the video segments together. However, this still result in flickering, especially when the camera and objects have larger relative motion. In addition, the regularization terms (with extra weight parameters) are included in the energy function when solving the contrast threshold for our EDI model. Thus, we extended our *EDI* model to a *multiple Event-based Double Integral (mEDI)* one to handle discontinuities at the boundaries of reconstructed video segments and develop a simple yet effective optimization solution. Later in our experiments, it shows the significant improvement in the smoothness and quality of the reconstructed videos.

In this paper, we first introduce our previous approach (the EDI model) in Section 3. Then, we build an extension framework based on multiple images and describe the approach in Section 4. Jointly optimizing *multi-frames for generating long video sequences* significantly alleviates the flickering problem for the generated videos, whereas EDI treats each image individually and may suffer flicking artifacts.

The extensions are as follows:

1) We propose a *multiple Event-based Double Integral (mEDI)* model to restore better high frame rate sharp videos. The model is based on multiple images (even blurred) and their corresponding events.
2) Our *mEDI* is able to generate a sharp video under various types of blur by solving a single variable non-convex optimization problem, especially in low lighting condition and complex dynamic scene.
3) We develop a simple yet effective optimization solution. In doing so, we significantly reduce the computational complexity with the Fibonacci sequence.
4) The frame rate of our reconstructed video can theoretically be as high as the event rate (200 times greater than the original frame rate in our experiment). With multiple images, the reconstructed videos preserve more abundant texture and the consistency of natural images.

2 Related Work

Event cameras such as the DAVIS and DVS [6], [7] report log intensity changes, inspired by human vision. The result is a continuous, asynchronous stream of events that encodes non-redundant information about local brightness change. Estimating intensity images from events is important. The reconstructed images grant computer vision researchers a readily available high temporal resolution, high-dynamic-range imaging platform that can be used for tasks such as face-detection [13], moving object segmentation [14], SLAM [15], [16], [17], [18], [19], localization [20], [21] and optical flow estimation [22], [23], [24], [25]. Although several works try to explore the advantages of the high temporal resolution provided by event cameras [26], [27], [28], [29], [30], [31], how to make the best use of the event camera has not yet been fully investigated. In this section, we review image reconstruction from event-based methods, and images and event combined methods. We further discuss works on image deblurring.

Event-Based Image Reconstruction. A typical way is done by processing a spatio-temporal window of events. Taking a spatio-temporal window of events imposes a latency cost at minimum equal to the length of the time window, and choosing a time-interval (or event batch size) that works robustly for all types of scenes is not trivial. Barua *et al.* [13] generate image gradients by dictionary learning and obtain a logarithmic intensity image via Poisson reconstruction. Bardow *et al.* [8] simultaneously optimise optical flow and intensity estimates within a fixed-length, sliding spatio-temporal window using the primal-dual algorithm [32]. Cook *et al.* [15] integrate events into interacting maps to recover intensity, gradient, and optical flow while estimating global rotating camera motion. Kim *et al.* [16] reconstruct high-quality images from an event camera under a strong assumption that the only movement is pure camera rotation, and later extend their work to handle 6-degree-of-freedom motion and depth estimation [17]. Reinbacher *et al.* [33] integrate events over time while periodically regularising the estimate on a manifold defined by the timestamps of the latest events at each pixel. Optimization based event-only methods (*i.e.* without the process of learning from training data) will generate artifacts and lack of texture when event data is sparse, because they cannot integrate sufficient information from the available sparse events. Recently, learning-based approaches have improved the image reconstruction quality significantly with powerful event data representations via deep learning [4], [9], [34], [35]. Rebecq *et al.* propose E2VID [4], a fully convolutional, recurrent UNet architecture to encode events in a spatio-temporal voxel grid. In [34], Rebecq *et al.* propose a recurrent network to reconstruct videos from a stream of events and they incorporate stacked ConvLSTM gates, which prevent vanishing gradients during backpropagation for long sequences. Wang *et al.* [9] form a 3D event volume by stacking event frame in a time interval. A reconstructed intensity frame is generated by summing events at each pixel in a smaller time interval. To achieve more image details in the reconstructed images, several methods trying to combine events with intensities have been proposed. The DAVIS [7] uses a shared photo-sensor array to simultaneously output events (DVS) and intensity images (APS). Brandli *et al.* [12] combine images and event streams from the DAVIS camera to create inter-frame intensity estimates by dynamically estimating the contrast threshold (temporal contrast) of each event. Each new image frame resets the intensity estimate, preventing excessive growth of

integration error. However, it also discards important accumulated event information. Scheerlinck *et al.* [3] propose an asynchronous event-driven complementary filter to combine APS intensity images with events, and obtain continuous-time image intensities. Shedligeri *et al.* [11] first exploit two intensity images to estimate depth. Second, they only use events to reconstruct a pseudo-intensity sequence (using method [33]) between the two intensity images. They, taking the pseudo-intensity sequence, they estimate the ego-motion using visual odometry. With the estimated 6-DOF pose and depth, they directly warp the intensity image to the intermediate location. Liu *et al.* [36] assume a scene should have static background. Thus, their method needs an extra sharp static foreground image as input and the event data are used to align the foreground with the background.

Image Deblurring. Recently, significant progress has been made in blind image deblurring. Traditional deblurring methods usually make assumptions on the scenes (such as a static scene) or exploit multiple images (such as stereo, or video) to solve the deblurring problem. Significant progress has been made in the field of single image deblurring. Methods using gradient based regularizers, such as Gaussian scale mixture [37], $l_1\backslash l_2$ norm [38], edge-based patch priors [39], [40] and l_0-norm regularizer [41], [42], have been proposed. Non-gradient-based priors such as the color line based prior [43], and the extreme channel (dark/bright channel) prior [44], [45] have also been explored. Since blur parameters and the latent image are difficult to be estimated from a single image, the single-image-based approaches are extended to use multiple images [46], [47], [48], [49], [50].

Driven by the success of deep neural networks, Sun *et al.* [51] propose a convolutional neural network (CNN) to estimate locally linear blur kernels. Gong *et al.* [52] learn optical flow from a single blurred image through a fully-convolutional deep neural network. The blur kernel is then obtained from the estimated optical flow to restore the sharp image. Nah *et al.* [53] propose a multi-scale CNN that restores latent images in an end-to-end learning manner without assuming any restricted blur kernel model. Tao *et al.* [1] propose a light and compact network, SRN-DeblurNet, to deblur the image. However, deep deblurring methods generally need a large dataset to train the model and usually require sharp images provided as supervision. In practice, blurred images do not always have corresponding ground-truth sharp images.

Blurred Image to Sharp Video. Recently, two deep learning based methods [2], [54] propose to restore a video from a single blurred image with a fixed sequence length. However, their reconstructed videos do not obey the 3D geometry of the scene and camera motion. Although deep-learning based methods achieve impressive performance in various scenarios, their success heavily depend on the consistency between the training datasets and the testing datasets, thus hinder the generalization ability for real-world applications.

3 Formulation

Our goal is to reconstruct a high frame rate, sharp video from a single or multiple (blurred) images and their corresponding events. In this section, we first introduce our *EDI* model. Then, we extend it to the *mEDI* model that includes multiple blurred images. Our models, both *EDI* and *mEDI*, can tackle various blur types and work stably in highly dynamic scenarios and low lighting conditions.

3.1 Event Camera Model

Event cameras are bio-inspired sensors that asynchronously report logarithmic intensity changes [6], [7]. Unlike conventional cameras that produce full images at a fixed frame rate, event cameras trigger events whenever the change in intensity at a given pixel exceeds a preset threshold. Event cameras do not suffer from limited dynamic ranges typical of sensors with the synchronous exposure time, and capture the high-speed motion with microsecond accuracy.

Inherent in the theory of event cameras is the concept of the latent image $\mathbf{L}_{xy}(t)$, denoting the instantaneous intensity at pixel (x, y) at time t, related to the rate of photon arrival at that pixel. The latent image $\mathbf{L}_{xy}(t)$ is not directly output by the camera. Instead, the camera outputs a sequence of *events*, denoted by (x, y, t, σ). Here, (x, y) denote image coordinates, t denotes the time the event takes place, and polarity $\sigma = \pm 1$ denotes the direction (increase or decrease) of the intensity change at that pixel and time. Polarity is given by,

$$\sigma = \mathcal{T}\left(\log\left(\frac{\mathbf{L}_{xy}(t)}{\mathbf{L}_{xy}(t_{\text{ref}})}\right), c\right), \tag{1}$$

where $\mathcal{T}(\cdot,\cdot)$ is a truncation function,

$$\mathcal{T}(d, c) = \begin{cases} +1, & d \geq c, \\ -1, & d \leq -c. \end{cases}$$

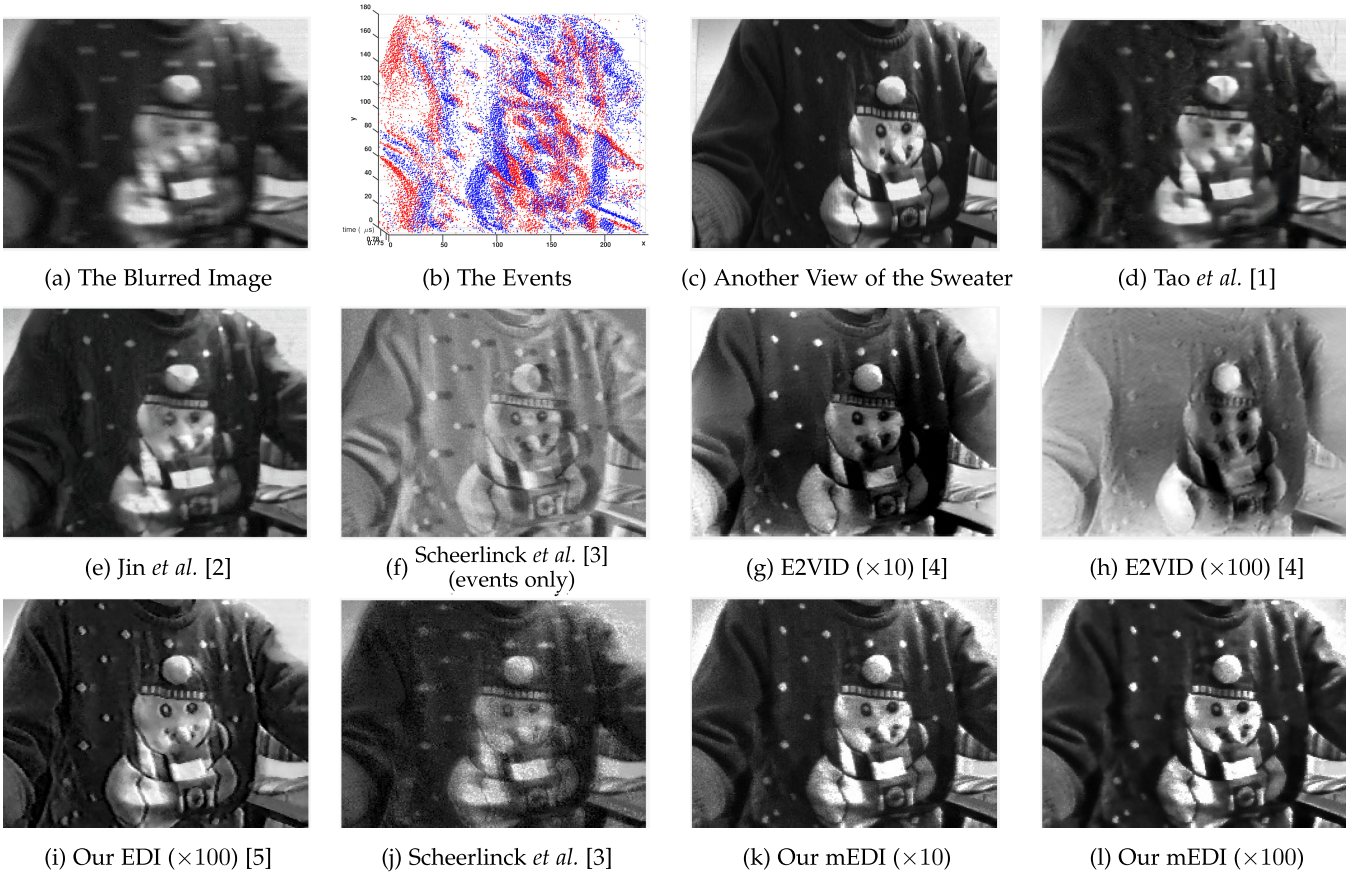
Here, c is a threshold parameter determining whether an event should be recorded or not, $\mathbf{L}_{xy}(t)$ and $\mathbf{L}_{xy}(t_{\text{ref}})$ denote the intensity of the pixel (x, y) at time instances t and t_{ref}, respectively. When an event is triggered, $\mathbf{L}_{xy}(t_{\text{ref}})$ at that pixel is updated to a new intensity level. As described by [6], the DVS only uses a global threshold c. However, the contrast threshold of an event camera is not constant, but normally distributed. Several methods [55], [56] assume that the positive and negative contrast thresholds (*i.e.*, c_+ and c_-) exhibit different distribution noise. We observed using a global threshold c, (*i.e.*, $c_+ = c_-$) also yields satisfying video deblurring and high-frame rate reconstruction results while significantly simplifying the optimization procedure. Thus, we adopt a global c in the following section.

3.2 Intensity Image Formation

In addition to event streams, event cameras can provide full-frame grey-scale intensity images, at a much lower rate than the event sequence. Grey-scale images may suffer from motion blur due to their long exposure time. A general model of the blurred image formation is given by,

$$\mathbf{B} = \frac{1}{T}\int_{f-T/2}^{f+T/2} \mathbf{L}(t)\, dt, \tag{2}$$

where $\mathbf{B}$ is the blurred image, equal to the average of latent images during the exposure time $[f - T/2, f + T/2]$. Let



(a) The Blurred Image (b) The Events (c) Another View of the Sweater (d) Tao *et al.* [1]

(e) Jin *et al.* [2] (f) Scheerlinck *et al.* [3] (events only) (g) E2VID ($\times$10) [4] (h) E2VID ($\times$100) [4]

(i) Our EDI ($\times$100) [5] (j) Scheerlinck *et al.* [3] (k) Our mEDI ($\times$10) (l) Our mEDI ($\times$100)

Fig. 1. Deblurring and reconstruction results of our method compared with the state-of-the-art methods on our real *blur event dataset*. (a) The input blurred image. (b) The corresponding event data. (c) A sharp image for the sweater captured as a reference for colour and shape (a real blurred image can hardly have its ground truth sharp image). (d) Deblurring result of Tao *et al.* [1]. (e) Deblurring result of Jin *et al.* [2]. Jin uses video as training data to train a supervised model to perform deblur, where the video can also be considered as similar information as the event data. (f) Reconstruction results of Scheerlinck *et al.* [3] from only events. (g) Reconstruction results of Rebecq *et al.* [4] from only events. Based on their default settings, the time resolution of the reconstructed video is around $\times 10$ times higher than the time resolution of the original video. (h) Reconstruction results of Rebecq *et al.* [4] from only events. The time resolution here is around $\times 100$. (i) Reconstruction result of Pan *et al.* [5] from combining events and a single blurred frame. (j)Reconstruction results of Scheerlinck *et al.* [3] from events and images. (k)-(l) Our reconstruction result from combining events and multiple blurred frame at different time resolution. Our result preserves more abundant and faithful texture and the consistency of the natural image. (Best viewed on screen).

$\mathbf{L}(f)$ be the snapshot of the image intensity at time f, the latent sharp image at the centre of the exposure period.

3.3 Event-Based Double Integral Model

We aim to recover the latent sharp intensity video by exploiting both the blur model and the event model. We define $e_{xy}(t)$ as a function of continuous time t such that,

$$e_{xy}(t) = \sigma\ \delta_{t_0}(t),$$

whenever there is an event (x, y, t_0, σ). Here, $\delta_{t_0}(t)$ is an impulse function, with unit integral, at time t_0, and the sequence of events is turned into a continuous time signal, consisting of a sequence of impulses. There is such a function $e_{xy}(t)$ for every point (x, y) in the image. Since each pixel can be treated separately, we omit the subscripts x, y.

Given a reference timestamp f, we define $\mathbf{E}(t)$ as the sum of events between time f and t,

$$\mathbf{E}(t) = \int_f^t e(s)ds,$$

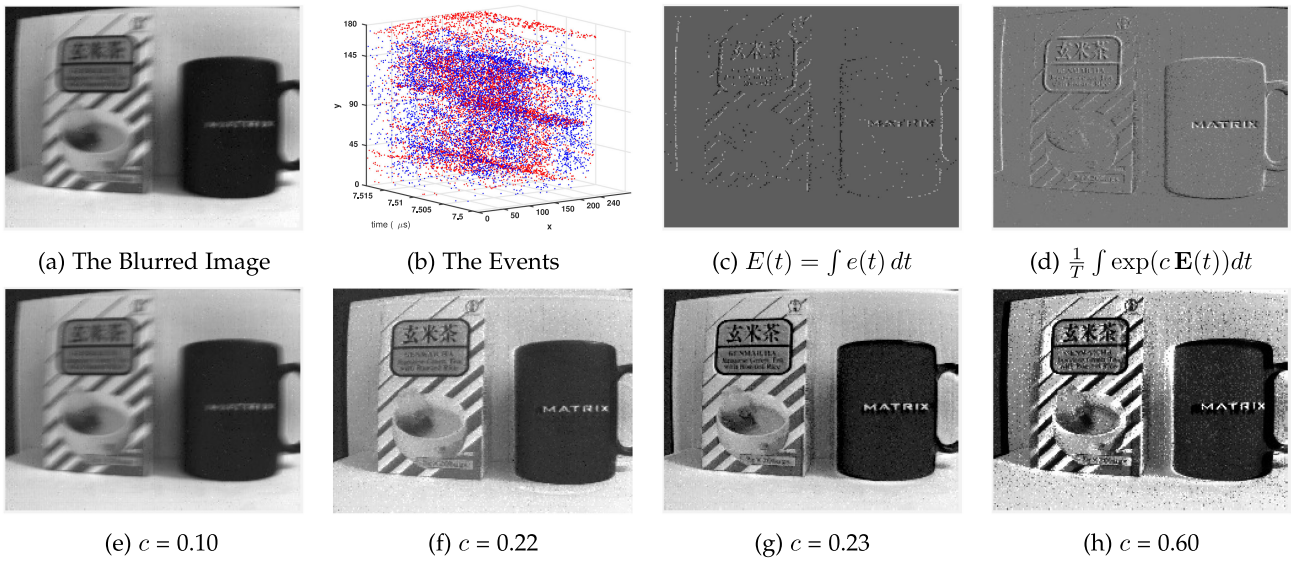
which represents the proportional change in intensity between time f and t. Except under extreme conditions, such as glare and no-light conditions, the latent image sequence $\mathbf{L}(t)$ is expressed as,

$$\mathbf{L}(t)\ = \mathbf{L}(f)\exp(c\,\mathbf{E}(t)).$$

In particular, an event (x, y, t, σ) is triggered when the intensity of a pixel (x, y) increases or decreases by an amount c at time t. With a high enough temporal resolution, the intensity changes of each pixel can be segmented to consecutive event streams with different amounts of events. We put a tilde on top of things to denote logarithm, *e.g.* $\widetilde{\mathbf{L}}(t) = \log(\mathbf{L}(t))$. Thus, we have,

$$\widetilde{\mathbf{L}}(t)\ = \widetilde{\mathbf{L}}(f) + c\,\mathbf{E}(t). \tag{3}$$

Given a sharp frame, we can reconstruct a high frame rate video from the sharp starting point $\mathbf{L}(f)$ by using Eq. (3). When an input image is blurred, a trivial solution would be to first deblur the image with an existing deblurring method and then to reconstruct a video using Eq. (3) (see Fig. 4 for details). However, in this way, the event data between intensity images are not fully exploited, thus resulting in inferior performance. Moreover, none of



(a) The Blurred Image (b) The Events (c) $E(t) = \int e(t)\,dt$ (d) $\frac{1}{T}\int \exp(c\,\mathbf{E}(t))dt$

(e) c = 0.10 (f) c = 0.22 (g) c = 0.23 (h) c = 0.60

Fig. 2. The event data and our reconstructed result, where (a) and (b) are the input of our method. (a) The intensity image from the DAVIS. (b) Events from the event camera plotted in 3D space-time (x, y, t) (blue: positive event; red: negative event). (c) The first integral of several events during a small time interval. (d) The second integral of events during the exposure time. (e)-(h) Samples of reconstructed image with different c. The value is from low (0.10), to proper (around 0.23) and high (0.60). Note, c = 0.23 in (g) is the chosen automatically by our optimization process.

existing deblurring methods can be guaranteed to work stably in a complex dynamic scenery. Instead, we propose to reconstruct the video by exploiting the inherent connection between events and blur, and present the following model.

As for the blurred image,

$$\begin{aligned}\mathbf{B} &= \frac{1}{T}\int_{f-T/2}^{f+T/2} \mathbf{L}(f)\exp\Big(c\,\mathbf{E}(t)\Big)\;dt \\ &= \frac{\mathbf{L}(f)}{T}\int_{f-T/2}^{f+T/2} \exp\Big(c\int_{f}^{t} e(s)ds\Big)\;dt.\end{aligned} \tag{4}$$

In this manner, we build the relation between the captured blurred image $\mathbf{B}$ and the latent image $\mathbf{L}(f)$ through the double integral of the event. We name Eq. (4) the *Event-based Double Integral (EDI)* model.

We denote

$$\mathbf{J}(c) = \frac{1}{T}\int_{f-T/2}^{f+T/2} \exp(c\,\mathbf{E}(t))dt.$$

Taking the logarithm on both sides of Eq. (4) and rearranging it yields

$$\widetilde{\mathbf{L}}(f) = \widetilde{\mathbf{B}} - \widetilde{\mathbf{J}}(c), \tag{5}$$

which shows a linear relationship between the blurred image, the latent image and integrated events in the log space.

3.4 High Frame Rate Video Generation

The right-hand side of Eq. (5) is known, apart from perhaps the value of the contrast threshold c, the first term from the grey-scale image, the second term from the event sequence, so it is possible to compute $\widetilde{\mathbf{L}}$, and hence $\mathbf{L}$ by exponentiation. Subsequently, from Eq. (3) the latent image $\mathbf{L}(t)$ at any time may be computed.

To avoid accumulated errors of constructing a video from many frames of a blurred video, it is more suitable to construct each frame $\mathbf{L}(t)$ using the closest blurred frame.

Theoretically, we could generate a video with a frame rate as high as the DVS's event rate. However, since each event carries little information and is subject to noise, several events must be processed together to yield a reasonable image. We generate a reconstructed frame every $50 - 100$ events, so for our experiment, the frame rate of the reconstructed video is usually 200 times greater than the input low frame rate video. Furthermore, as indicated by Eq. (5), the challenging blind motion deblurring problem has been reduced to a single variable optimization problem of how to find the best value of the contrast threshold c.

3.5 Finding c With Regularization Terms

As indicated by Eq. (5), the blind motion deblurring problem has been reduced to a single variable optimization problem of how to find the best value of the threshold c. To this end, we need to build an evaluation metric (energy function) that can evaluate the quality of the deblurred image $\mathbf{L}(t)$. Specifically, we propose to exploit different prior knowledge for sharp images and the event data.

Edge Constraint for Event Data. As mentioned before, when a proper c is given, our reconstructed image $\mathbf{L}(c, t)$ will contain much sharper edges compared with the original input intensity image. Furthermore, event cameras inherently yield responses at moving intensity boundaries, so edges in the latent image may be located where (and when) events occur. We convolve the event sequence with an exponentially decaying window, to obtain a denoised yet wide edge boundary,

$$\mathbf{M}(t) = \int_{-T/2}^{T/2} \exp(-(|t-s|))\,e(s)\,ds,$$

Then, we use the Sobel filter $\mathcal{S}$ to get a sharper binary edge map, which is also applied to $\mathbf{L}(c, t)$. Here, we use $\mathbf{L}(c, t)$ to present the latent sharp image $\mathbf{L}(t)$ with different c.

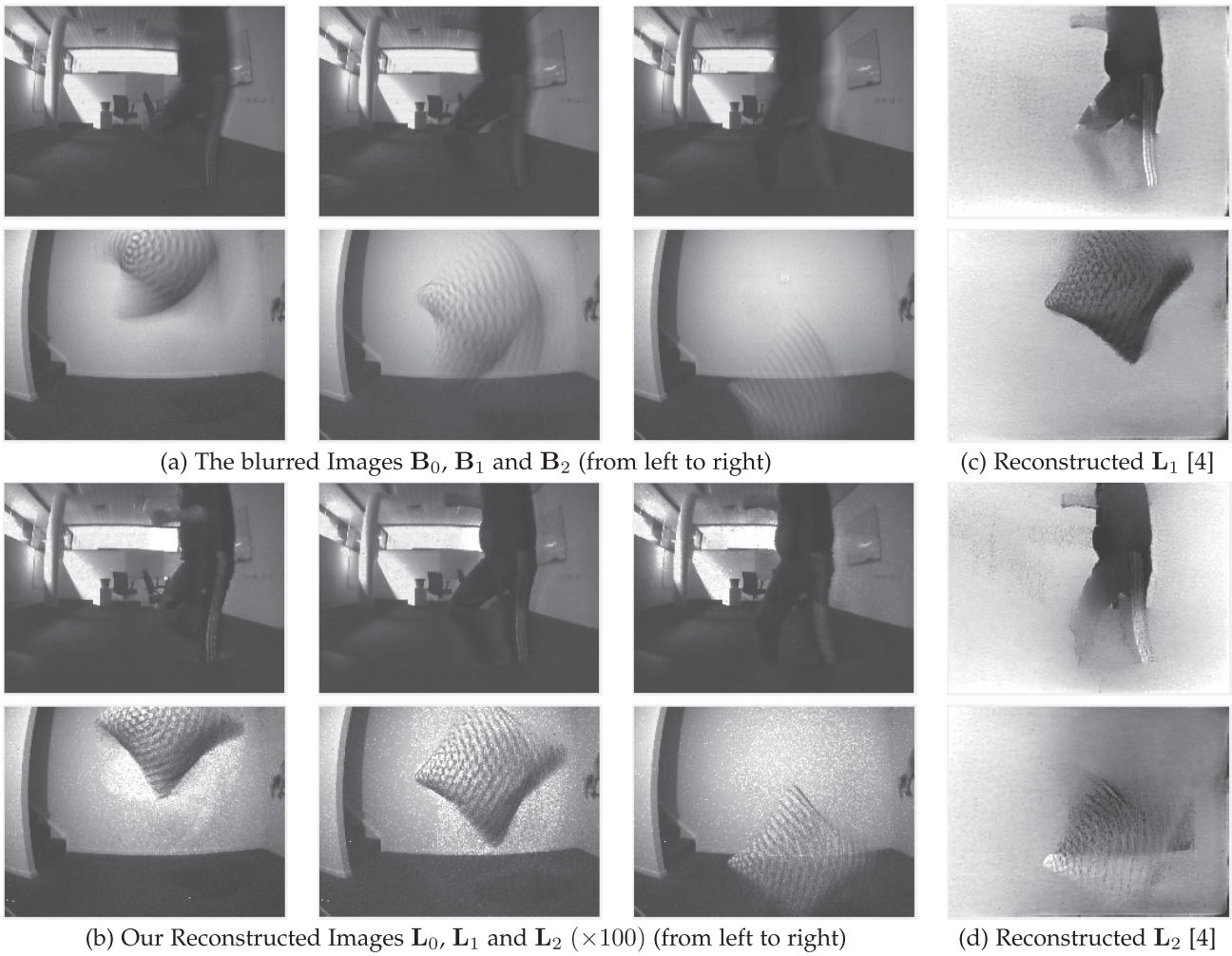


Fig. 3. The examples of our reconstructed results are based on our real event dataset. The threshold c is estimated automatically from three blurred images and their events based on our mEDI model. (a), (b) Blur image and our reconstructed Images $\mathbf{L}_0$, $\mathbf{L}_1$ and $\mathbf{L}_2$ (c), (d) Reconstruction results of $\mathbf{L}_1$ and $\mathbf{L}_2$ by Rebecq *et al.* [4] from only events. The time resolution here is around $\times 6$ based on their default settings. The time resolution of the reconstructed video by E2VID [4] is around $\times 8$ to 15 times higher than the time resolution of the original video. (Best viewed on screen).

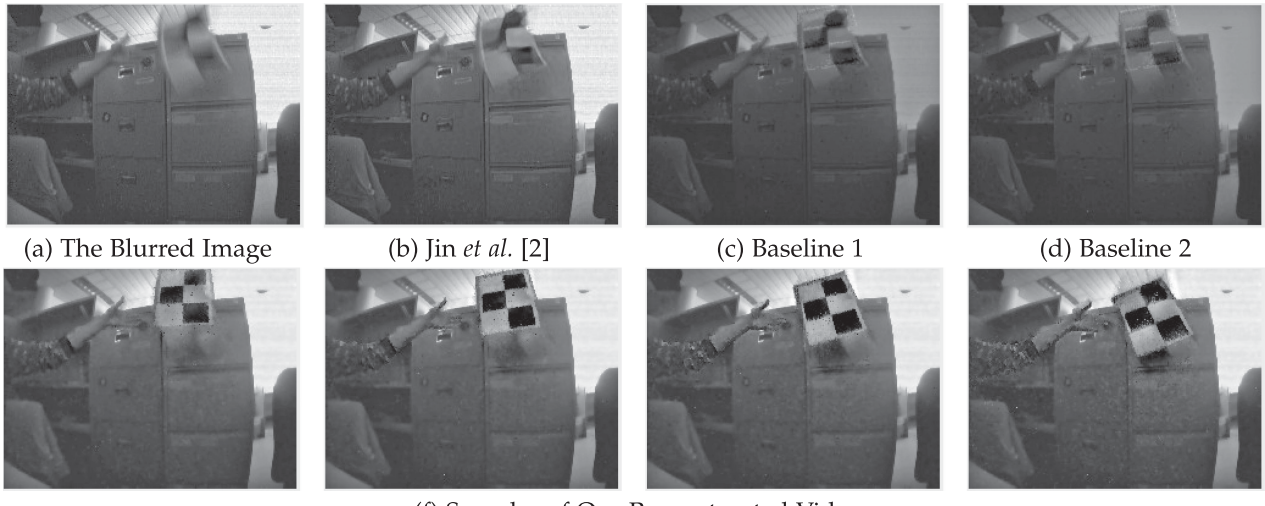


Fig. 4. Deblurring and reconstruction results on our real *blur event dataset*. (a) Input blurred images. (b) Deblurring result of [2]. (c) Baseline 1 for our method. We first use the state-of-the-art video-based deblurring method [2] to recover a sharp image. Then use the sharp image as input to a state-of-the-art reconstruction method [3] to get the intensity image. (d) Baseline 2 for our method. We first use method [3] to reconstruct an intensity image. Then use a deblurring method [2] to recover a sharp image. (e) Samples from our reconstructed video from $\mathbf{L}(0)$ to $\mathbf{L}(150)$.

Here, we use cross-correlation between $\mathcal{S}(\mathbf{L}(c,t))$ and $\mathcal{S}(\mathbf{M}(t))$ to evaluate the sharpness of $\mathbf{L}(c,t)$.

$$\phi_{\text{edge}}(c) = \sum_{x,y} \mathcal{S}(\mathbf{L}(c,t))(x,y) \cdot \mathcal{S}(\mathbf{M}(t))(x,y). \tag{6}$$

Intensity Image Constraint. Total variation is used to suppress noise in the latent image while preserving edges, and to penalize spatial fluctuations [57].

$$\phi_{\text{TV}}(c) = |\nabla \mathbf{L}(c,t)|_1, \tag{7}$$

where ∇ represents the gradient operators.

Energy Minimization. The optimal c can be estimate by solving Eq. (8),

$$\min_c \phi_{\text{TV}}(c) + \lambda \phi_{\text{edge}}(c), \tag{8}$$

where λ is a trade-off parameter. The response of cross-correlation reflects the matching rate of $\mathbf{L}(c,t)$ and $\mathbf{M}(t)$ which makes $\lambda < 0$. This single-variable minimization problem can be solved by Golden Section Search.

4 Using More Than One Frame

Though our EDI can reconstruct high frame rate videos efficiently, noise from events can easily degrade the quality of reconstructed videos with low temporal consistency. In addition, regularization terms in the energy function introduce unexpected weight parameters. Therefore, we propose a multiple images based approach to tackle the above problems with a simple yet effective optimization solution.

4.1 Multiple Event-Based Double Integral Model

Suppose an event camera captures a continuing sequence of events, and also blurred images, $\mathbf{B}_i$ for $i = 0, \ldots, n$. Assume that the exposure time is T and the reference frame $\mathbf{B}_i$ is at time f_i. Each $\mathbf{B}_i$ is associated with a latent image $\mathbf{L}_i(f_i)$ and is generated as an integral of $\mathbf{L}_i(t)$ over the exposure interval $[f_i - T/2, f_i + T/2]$. In addition, we rewrite $\mathbf{E}(t)$, $\mathbf{L}(t)$ and $\mathbf{J}(c)$ for the i^{th} frame as

$$\begin{aligned}
\mathbf{E}_i(t) &= \int_{f_i}^{t} e(s)ds \\
\mathbf{L}_i(t) &= \mathbf{L}_i(f_i)\exp(c\,\mathbf{E}_i(t)) \\
\mathbf{J}_i(c) &= \frac{1}{T}\int_{f_i - T/2}^{f_i + T/2} \exp(c\,\mathbf{E}_i(t))dt.
\end{aligned}$$

The **EDI** model in Eq. (5) in Section 3 gives

$$\widetilde{\mathbf{B}}_i = \widetilde{\mathbf{L}}_i(f_i) + \widetilde{\mathbf{J}}_i(c), \tag{9}$$

for each blurred image in the sequence.

We use $\mathbf{L}_i$ to represent $\mathbf{L}_i(f_i)$ in the following section. Then, Eq. (9) is written as

$$\widetilde{\mathbf{B}}_i = \widetilde{\mathbf{L}}_i + \widetilde{\mathbf{J}}_i(c) = \widetilde{\mathbf{L}}_i + a_i. \tag{10}$$

The latent image $\mathbf{L}_{i+1}$ is formed from latent image $\mathbf{L}_i$ by integrating events over the period $[f_i, f_{i+1}]$, which gives

$$\begin{aligned}
\widetilde{\mathbf{L}}_{i+1} &= \widetilde{\mathbf{L}}_i + c\int_{f_i}^{f_{i+1}} e(s)ds \\
&= \widetilde{\mathbf{L}}_i + b_i.
\end{aligned} \tag{11}$$

This describes the *mEDI* model based on multiple images and their events.

$$\begin{aligned}
\widetilde{\mathbf{L}}_i &= \widetilde{\mathbf{B}}_i - a_i \\
\widetilde{\mathbf{L}}_{i+1} - \widetilde{\mathbf{L}}_i &= b_i.
\end{aligned} \tag{12}$$

The known values are $\widetilde{\mathbf{B}}_i$, whereas the unknowns are $\widetilde{\mathbf{L}}_i$, a_i and b_i. These quantities are associated with a single pixel and we solve for each pixel independently.

We therefore obtain a set of linear equations based on Eq. (12) as

$$\begin{bmatrix}
1 & -1 & & & & \\
 & 1 & -1 & & & \\
 & & \ddots & \ddots & & \\
 & & & 1 & -1 & \\
1 & & & & & \\
 & 1 & & & & \\
 & & 1 & & & \\
 & & & \ddots & & \\
 & & & & & 1
\end{bmatrix}
\begin{bmatrix}
\widetilde{\mathbf{L}}_1 \\ \widetilde{\mathbf{L}}_2 \\ \vdots \\ \widetilde{\mathbf{L}}_n
\end{bmatrix}
=
\begin{bmatrix}
-b_1 \\ \vdots \\ -b_{n-1} \\ \widetilde{\mathbf{B}}_1 - a_1 \\ \vdots \\ \widetilde{\mathbf{B}}_n - a_n
\end{bmatrix}, \tag{13}$$

where a_i and b_i depend on c, but particularly a_i depends on c in a non-linear way. Writing Eq. (13) as $\mathbf{Ax} = \mathbf{w}$, the least-squares solution is given by solving $\mathbf{A}^T\mathbf{Ax} = \mathbf{A}^T\mathbf{w}$.

4.2 LU Decomposition

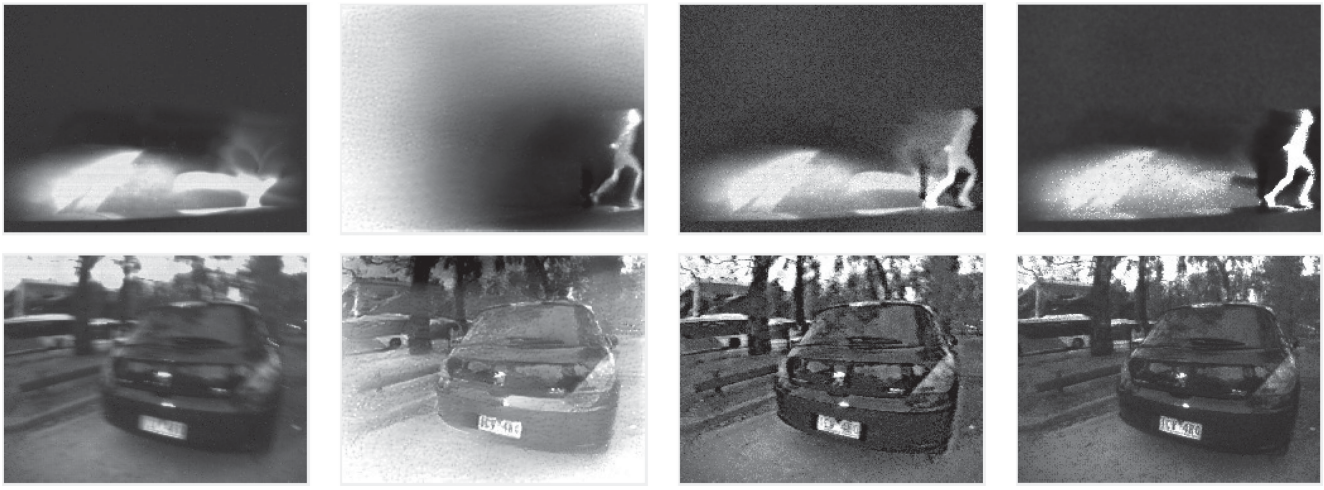
Because of their particular form, these equations can be solved very efficiently as will now be shown. Expanding the equations

$$\mathbf{A}^T\mathbf{Ax} = \mathbf{A}^T\mathbf{w},$$

gives

$$\begin{bmatrix}
2 & -1 & & & & & \\
-1 & 3 & -1 & & & & \\
 & -1 & 3 & -1 & & & \\
 & & & \ddots & & & \\
 & & & & -1 & 3 & -1 \\
 & & & & & -1 & 2
\end{bmatrix}
\begin{bmatrix}
\widetilde{\mathbf{L}}_1 \\ \widetilde{\mathbf{L}}_2 \\ \vdots \\ \widetilde{\mathbf{L}}_n
\end{bmatrix}
=
\begin{bmatrix}
\widetilde{\mathbf{B}}_1 - a_1 - b_1 \\
\widetilde{\mathbf{B}}_2 - a_2 - b_2 + b_1 \\
\vdots \\
\widetilde{\mathbf{B}}_{n-1} - a_{n-1} - b_{n-1} + b_{n-2} \\
\widetilde{\mathbf{B}}_n - a_n + b_{n-1}
\end{bmatrix}. \tag{14}$$

This is a particularly easy set of equations to solve. Since it has to be solved for each pixel, it is important to do it efficiently. The best way to solve Eq. (14) is to take the LU



(a) The Blurred Image (b) E2VID ($\times 8$) [4] (c) Our EDI ($\times 100$) [5] (d) Our mEDI ($\times 100$)

Fig. 5. Examples of reconstruction results on real event dataset. (a) The intensity image from the event camera. (b) Reconstruction result of our E2VID *et al.* [4] from only events. The temporal resolution is around $\times 8$ based on their default settings, while ours are $\times 100$ times higher than the original videos'. (c) Reconstruction result of our EDI model *et al.* [5] from combining events and a single blurred frame. (d) Reconstruction result of our mEDI model from combining events and multiple blurred frames. Our method based on multiple images gets better results than our previous one based only on one single image, especially on large motion scenery and extreme light conditions. (Best viewed on screen).

decomposition of the left-hand-side matrix, which has a particularly simple form.

Let $\mathbf{A}^T\mathbf{w} = \mathbf{r}$, we writing Eq. (14) as $\mathbf{A}^T\mathbf{A}\mathbf{x} = \mathbf{r}$. The LU decomposition of $\mathbf{A}^T\mathbf{A}$ (with appropriate reordering of rows) is given by

$$\mathtt{LU} = \begin{bmatrix} -2 & -5 & -13 & \cdots & 1 \\ 1 & & & & 0 \\ & 1 & & & 0 \\ & & \ddots & & \vdots \\ & & & 1 & 0 \end{bmatrix} \begin{bmatrix} -1 & 3 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 3 & -1 \\ & & & -1 & 2 \\ & & & & \phi_{2n-1} \end{bmatrix}.$$

More precisely, if the Fibonacci sequence is $1, 2, 3, 5, 8, \ldots$ and ϕ_k denotes the $k-$th entry of this sequence (thus $\phi(0) = 1$, $\phi(2) = 2$), then the top line of the left-hand matrix is

$$\begin{bmatrix} \phi_2 & \phi_4 & \ldots & \phi_{2(n-1)} & 1 \end{bmatrix},$$

consisting of the even numbered entries of the Fibonacci sequence. The entry at the bottom right of the right-hand matrix is ϕ_{2n-1}, the next odd-numbered Fibonacci number, which is also the determinant of the original matrix. Solving equations by LU decomposition and back-substitution is particularly simple in this case. The procedure in solving equations $\mathtt{LU}\mathbf{x} = \mathbf{r}$ is done by solving

$$\mathtt{L}\mathbf{y} = \mathbf{r}\mathtt{U}\mathbf{x} = \mathbf{y}.$$

The solution of $\mathtt{L}\mathbf{y} = \mathbf{r} = (r_1, r_2, \ldots, r_n)^T$ is simply

$$\mathbf{y} = \left(r_2, r_3, \ldots, r_n, \sum_{i=1}^{n-1} r_i \phi_{2i} \right)^T.$$

The solution of $\mathtt{U}\mathbf{x} = \mathbf{y}$ is given by back-substitution from the bottom:

$$\begin{aligned} x_n &= y_n/\phi_{2n-1} = \sum_{i=1}^{n-1} r_i \phi_{2i}/\phi_{2n-1} \\ x_{n-1} &= 2x_n - r_n \\ x_{n-2} &= 3x_{n-1} - x_n - r_{n-1} \\ x_{n-3} &= 3x_{n-2} - x_{n-1} - r_{n-2} \\ &\cdots \\ x_1 &= 3x_2 - x_3 - r_2 \end{aligned} \tag{15}$$

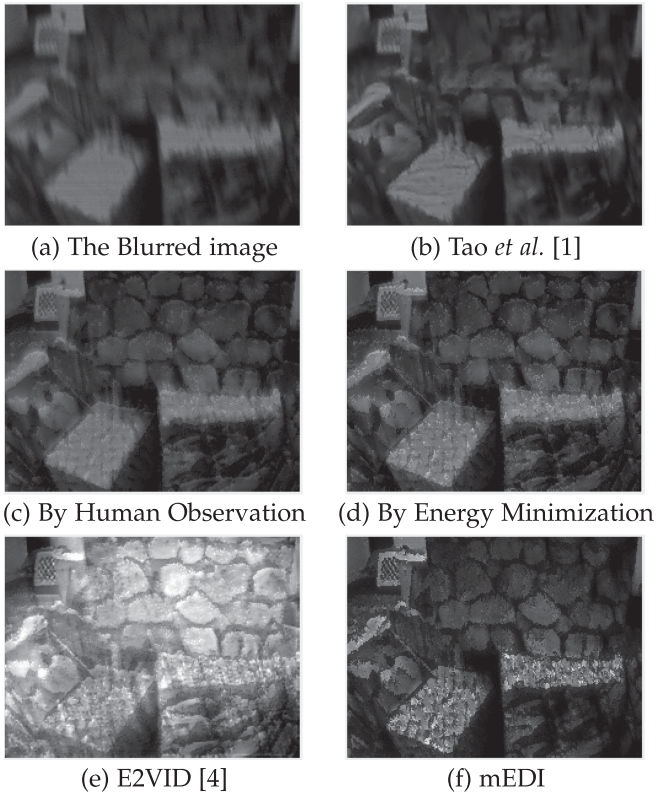
The values x_i is the pixel value for latent image $\mathbf{L}_i$. If c is known, then the values on the right of are dependent on c, and the sequence of $\mathbf{L}_n$ can be computed.

$$\begin{aligned} \mathbf{L}_n &= \sum_{i=1}^{n-1} r_i \phi_{2i}/\phi_{2n-1} \\ \mathbf{L}_{n-1} &= 2\mathbf{L}_n - \widetilde{\mathbf{B}}_n - a_n + b_{n-1} \\ \mathbf{L}_{n-2} &= 3\mathbf{L}_{n-1} - \mathbf{L}_n - \widetilde{\mathbf{B}}_{n-1} - a_{n-1} - b_{n-1} + b_{n-2} \\ \mathbf{L}_{n-3} &= 3\mathbf{L}_{n-2} - \mathbf{L}_{n-1} - \widetilde{\mathbf{B}}_{n-2} - a_{n-2} - b_{n-2} + b_{n-3} \\ &\cdots \\ \mathbf{L}_1 &= 3\mathbf{L}_2 - \mathbf{L}_3 - \widetilde{\mathbf{B}}_2 - a_2 - b_2 + b_1. \end{aligned} \tag{16}$$

Furthermore, the problem has been reduced to a single variable optimization problem of how to find the best value of the contrast threshold c.

5 Optimization

The unknown contrast threshold c represents the minimum change in log intensity required to trigger an event. With an appropriate c in Eq. (12), we can generate a sequence of



(a) The Blurred image (b) Tao *et al.* [1]

(c) By Human Observation (d) By Energy Minimization

(e) E2VID [4] (f) mEDI

Fig. 6. An example of our reconstruction result using different methods to estimate c, on a real sequence from the *Event-Camera Dataset* [58]. (a) The blurred image. (b) Deblurring result of [1]. (c) Our result where c is chosen by manual inspection. (d) Our result where c is computed automatically by our proposed energy minimization Eq. (19). (e) Reconstruction results of Rebecq *et al.* [4] from only events. The temporal resolution of the reconstructed video is around $\times 8$ times higher than the original videos' based on their default settings. (f) Our mEDI result where the temporal resolution is the same as (e).

sharper images. Here, we propose two different methods to estimate the unknown variable c, which are manually chosen and automatically optimized by our approach.

5.1 Manually Chosen c

According to our *mEDI* model in Eq. (12), given a value for c, we obtain sharp images.

Therefore, we develop a method for deblurring by manually inspecting the visual effect of the deblurred image. In this way, we incorporate human perception into the reconstruction loop and the deblurred images should satisfy human observation. In Figs. 2 and 6, we give examples for manually chosen results on our dataset, and the *Event-Camera Dataset* [58].

5.2 Automatically Chosen c

To automatically find the best c, we need to build an evaluation metric (energy function) that can evaluate the quality of the deblurred image $\mathbf{L}_i(t)$. Different from our EDI that including regularization terms (with extra weight parameters) in the energy function, we develop a simple yet effective optimization solution. More specifically, we adopt the Fibonacci sequence search to solve the optimization which significantly reduces the computational complexity.

5.2.1 Energy Function

The values on the right-hand side of Eq. (12) depend on an unknown parameter c. In particular, we write

$$\begin{aligned} b_i &= c \int_{f_i}^{t} e(s)ds \\ a_i &= \log\left(\frac{1}{T}\int_{f_i-T/2}^{f_i+T/2} \exp(c\,\mathbf{E}(t))dt\right). \end{aligned} \tag{17}$$

Given c, x_i can be solved by `LU` decomposition in Section 4.2. Subsequently, from Eq. (12) the blur image $\mathbf{B}_i$ can be computed.

$$\widetilde{\mathbf{B}}_i(c) = x_i + a_i \tag{18}$$

Here, we use $\mathbf{B}_i(c)$ to present the blurred image $\mathbf{B}_i$ with different c. In this case, the optimal c can be estimated by solving Eq. (19),

$$\min_c ||\mathbf{B}_i(c) - \mathbf{B}||_2^2. \tag{19}$$

Examples show that as a function of c, the residual error in solving the equations is not convex. However, in most cases (empirically) it seems to be convex, or at least it has a single minimum (See Fig. 8 for an example).

5.2.2 Fibonacci Search

Finding the minimum of a function along a single line is easy if that function has a single minimum. In the case of least-squares minimization problems, various strategies for determining the line-search direction are currently used, such as conjugate gradient methods, gradient descent, and the Levenberg-Marquardt method.

When the function has only one stationary point, the maximum/minimum, and when it depends on a single variable in a finite interval, the most efficient way to find the maximum is based on the Fibonacci numbers. The procedure, now known widely as 'Fibonacci search', was discovered and shown optimal in a minimax sense by Kiefer [60], [61].

In this work, we use Fibonacci search for the value of c that gives the least error. In Fig. 8, we illustrate the clearness of the reconstructed image (in PSNR value) as a function of the value of c. As demonstrated in the figure, our proposed reconstruction metric could locate/identify the best deblurred image with peak PSNR properly.

In our proposed method, we assume that $c_+ = c_-$ and use a global c based on the following reasons:

1) As illustrated in Fig. 8 our deblurring performance has a relatively flat crest against different values of c. Experimental results demonstrated that the quality of our reconstructed videos is robust to the estimation of c within a certain range.
2) We have conducted the experiments with $c_+ \neq c_-$, namely, optimising two variables in our formulation. We observed that the improvement on PSNR is less than 0.1 dB in comparison to the results of

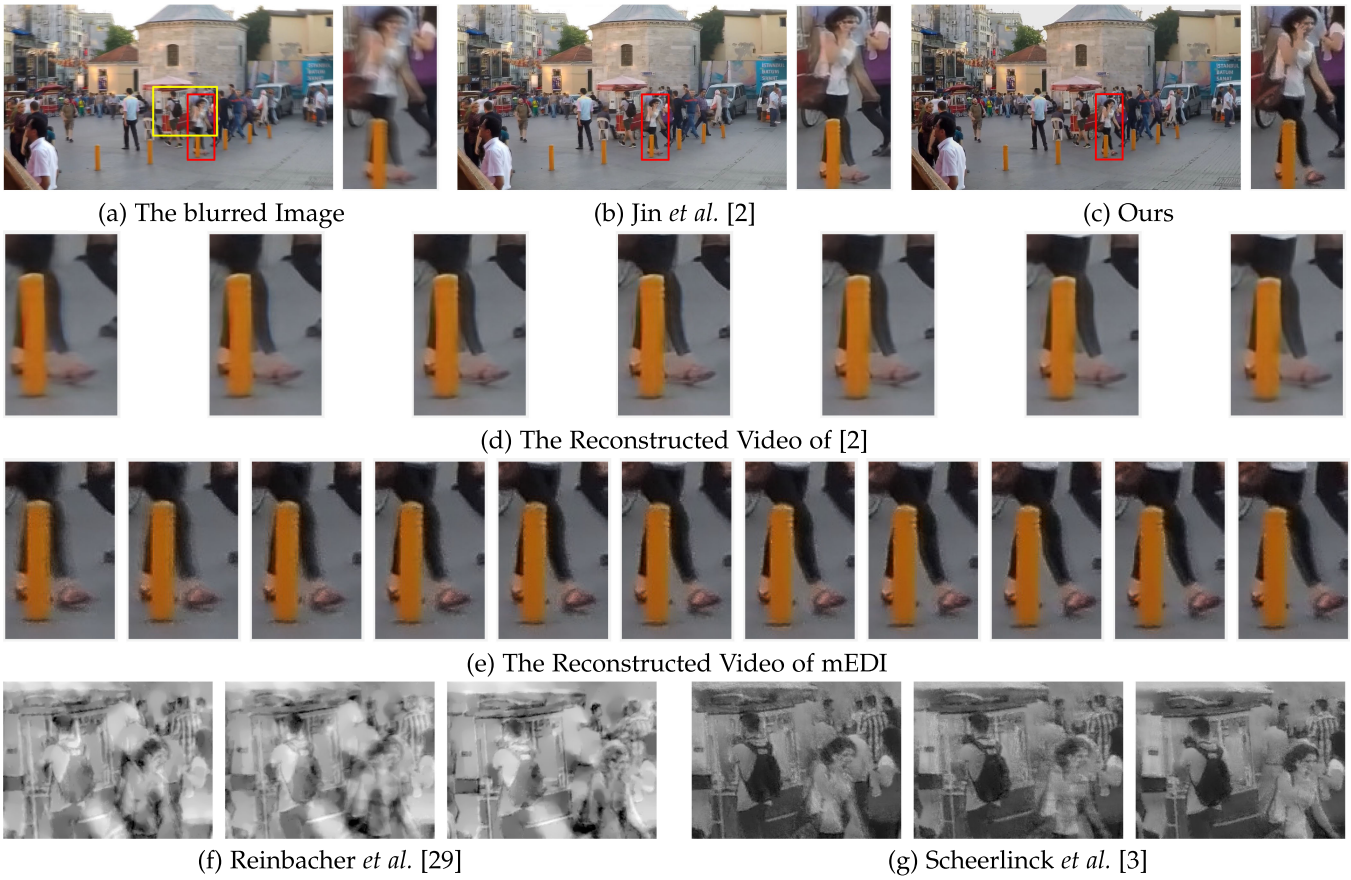


Fig. 7. An example of the reconstructed result on our synthetic event dataset based on the GoPro dataset [53]. [53] provides videos to generate blurred images and event data. (a) The blurred image. The red close-up frame is for (b)-(e), the yellow close-up frame is for (f)-(g). (b) The deblurring result of Jin *et al.* [2]. (c) Our deblurring result. (d) The crop of their reconstructed images and the frame number is fixed at 7. Jin *et al.* [2] uses the GoPro dataset added with 20 scenes as training data and their model is supervised by 7 consecutive sharp frames. (e) The crop of our reconstructed images. (f) The crop of Reinbacher [33] reconstructed images from only events. (g) The crop of Scheerlinck [3] reconstructed image, they use both events and the intensity image. For (e)-(g), the shown frames are the chosen examples, where the length of the reconstructed video is based on the number of events. (Best viewed on screen).

optimising a global c. However, the computational complexity increases from $\mathcal{O}(n)$ to $\mathcal{O}(n^2)$.

Therefore, we believe it is worthy of trading off between computational simplicity and performance accuracy.

6 Experiment

In all of our experiments, unless otherwise specified, the parameter c for reconstructing images is chosen automatically by our optimization process.

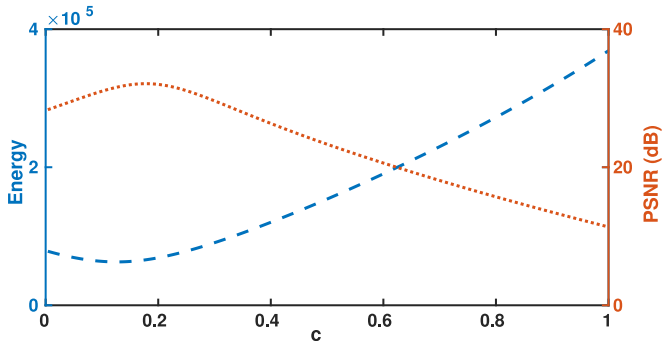
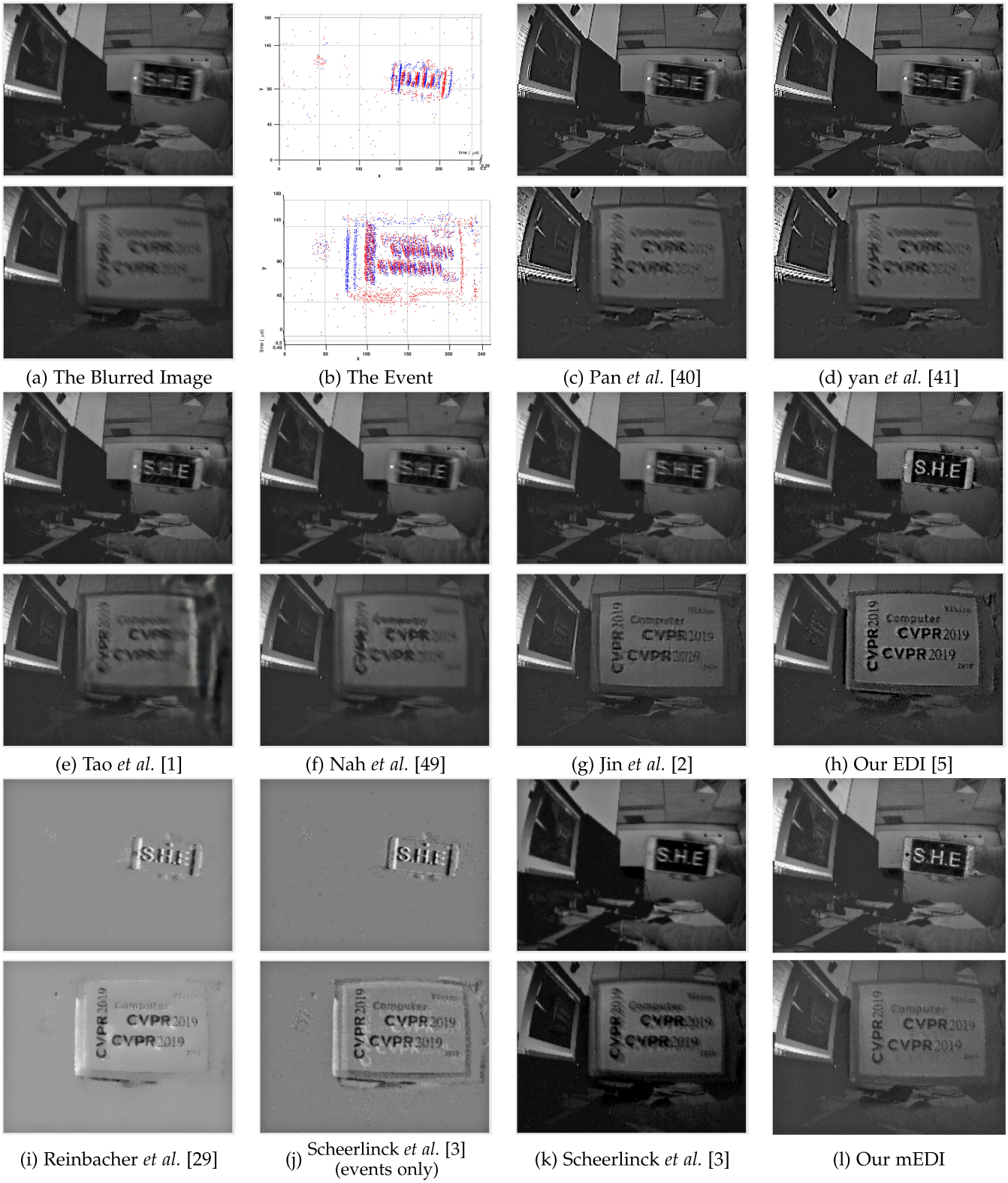


Fig. 8. Deblurring performance plotted against the value of c. The image is clearer with higher PSNR value.

6.1 Experimental Setup

Synthetic Dataset. In order to provide a quantitative comparison, we build a synthetic dataset based on the GoPro blur dataset [53]. It supplies ground truth videos which are used to generate the blurred images. Similarly, we employ the ground-truth images to generate event data based on the methodology of *event camera model*. In this GoPro dataset, we did not notice obvious rolling shutter artifacts because images in this dataset were requested to be captured with low speed of camera motions for providing ground-truth latent sharp images.

Real Dataset. We evaluate our method on a public Event-Camera dataset [58], which provides a collection of sequences captured by the event camera for high-speed robotics. Furthermore, we present our real *blur event dataset*, where each real sequence is captured with the DAVIS240 [7] under different conditions, such as indoor, outdoor scenery, low lighting conditions, and different motion patterns (*e.g.*, camera shake, objects motion) that naturally introduce motion blur into the APS intensity images. We also evaluate our method on a newly published Color Event Camera Dataset (CED) [62] built with DAVIS346 Red Color sensor. They present an extension of the event simulator ESIM [63] that enables simulation of color events. In contrast to GoPro cameras, event cameras, such as DAVIS, employ global shutters, where an entire scene is captured at the same



(a) The Blurred Image (b) The Event (c) Pan *et al.* [40] (d) yan *et al.* [41]

(e) Tao *et al.* [1] (f) Nah *et al.* [49] (g) Jin *et al.* [2] (h) Our EDI [5]

(i) Reinbacher *et al.* [29] (j) Scheerlinck *et al.* [3] (events only) (k) Scheerlinck *et al.* [3] (l) Our mEDI

Fig. 9. Examples of reconstruction result on our real *blur event dataset* in low lighting and complex dynamic conditions (a) Input blurred images. (b) The event information. (c) Deblurring results of [44]. (d) Deblurring results of [45]. (e) Deblurring results of [1]. (f) Deblurring results of [53]. (g) Deblurring results of [2] and they use video as training data. (h) Reconstruction result of [5] from combining events and frames. (i) Reconstruction result of [33] from only events. (j)-(k) Reconstruction results of [3], (j) from only events, (k) from combining events and frames. (l) Our reconstruction result. Results in (c)-(g) show that real high dynamic settings and low light conditions are still challenging in the deblurring area. Results in (i)-(j) show that while intensity information of a scene is still retained with an event camera recording, color, and delicate texture information cannot be recovered. (Best viewed on screen).

TABLE 1
Quantitative Comparisons on the Synthetic Dataset [53]

Average result of the deblurred **images** on dataset [49]									
	Pan [40]	Sun [47]	Gong [48]	Jin [2]	Tao [1]	Zhang [53]	Nah [49]	EDI [5]	mEDI
PSNR(dB)	23.50	25.30	26.05	26.98	30.26	29.18	29.08	29.06	**30.29**
SSIM	0.8336	0.8511	0.8632	0.8922	0.9342	0.9306	0.9135	**0.9430**	0.9194
Average result of the reconstructed **videos** on dataset [49]									
	Baseline 1 [1] + [3]		Baseline 2 [3] + [1]		Scheerlinck *et al.* [3]		Jin *et al.* [2]	EDI [5]	mEDI
PSNR(dB)	25.52		26.34		25.84		25.62	28.49	**28.83**
SSIM	0.7685		0.8090		0.7904		0.8556	**0.9199**	0.9098

The provided videos are able to generate not only blurred images but also event data. All methods are tested under the same blur condition, where methods [1], [2], [53], [59] use GoPro dataset [53] to train their models. Jin [2] achieves their best performance when the image is down-sampled to 45 percent mentioned in their paper. In this dataset, blurry images are generated by averaging every 11 frames, and treat the middle clean one (the 6^{th} frame) as the ground truth. The top part in this figure aims to compare with deblurring methods and only the blurry image (the 6^{th} frame) is evaluated. The bottom part shows the measures of whole reconstructed videos.

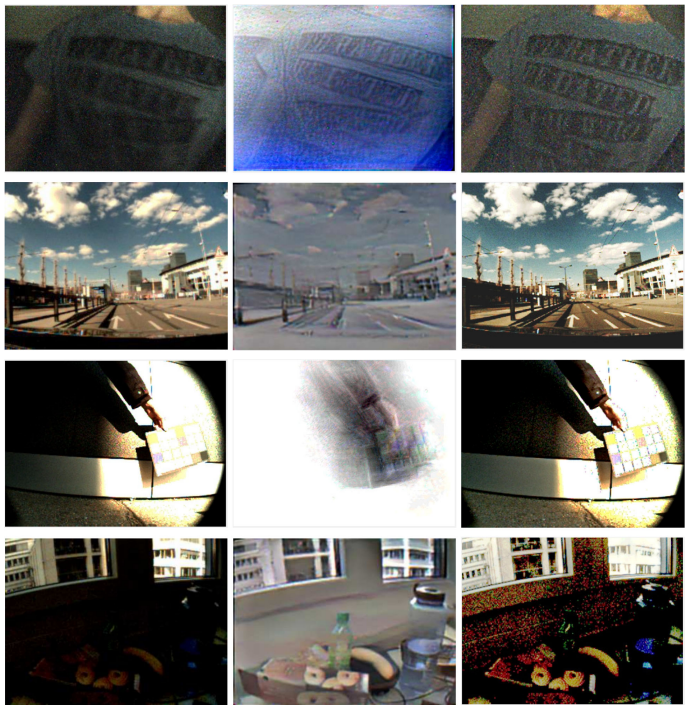
instant. Therefore, global shutter cameras, *e.g.*, our event camera, do not have rolling shutter effects.

Implementation Details. For all our real experiments, we use the DAVIS [7] that shares photosensor array to simultaneously output events (DVS) and intensity images (APS). The framework is implemented using MATLAB. It takes around 1.5 seconds to process one image on a single i7 core running at 3.6 GHz.

6.2 Experimental Results

We compare our proposed approach with state-of-the-art blind deblurring methods, including conventional deblurring methods [44], [45], deep based dynamic scene deblurring methods [1], [2], [51], [53], [59], and event-based image reconstruction methods [3], [4], [33]. Moreover, Jin *et al.* [2] can restore a video from a single blurred image based on a deep network, where the middle frame in the restored odd-numbered sequence is the best.

To prove the effectiveness of our model, we show some baseline comparisons in Fig. 4 and Table 1. For baseline 1, we first apply a state-of-the-art deblurring method [1] to recover a sharp image, and then feed the recovered image as input to a reconstruction method [3]. For baseline 2, we first use the video reconstruction method [3] to reconstruct a sequence of intensity images, then apply the deblurring



(a) DAVIS frame (b) E2VID [4] (c) Ours

Fig. 10. An example of our reconstruction result on the color event camera dataset CED [62]. (a) The input image. (b) Reconstruction results of Rebecq *et al.* [4] from only events. The temporal resolution of the reconstructed video is around $\times 12$ times higher than the original videos' based on their default settings. (c) Our mEDI result where the temporal resolution is the same as (b). From top to bottom, a scene with a low lighting condition, an outdoor scene, a scene with slow-moving objects (static background), and an HDR scene. Our mEDI model performs well in the top two rows, while E2VID is able to provide vivid color textures in the HDR scene. Note that our method focuses on reconstructing high-frame rate videos rather than changing the dynamic range of input videos. In order to illustrate our detailed textures in the HDR scene, we employ an HDR enhancement method [64].

method [1] to each frame. As seen in Table 1, our approach obtains higher PSNR and SSIM in comparison to both baseline 1 and baseline 2. This also implies that our approach better exploits the event data to not only recover sharp images but also reconstruct high frame rate videos.

In Table 1 and Fig. 7, we show quantitative and qualitative comparison on our synthetic dataset, respectively. As indicated in Table 1, our approach achieves the best performance on PSNR and competitive results on SSIM compared to state-of-the-art methods, and attains significant performance improvements on high-frame video reconstruction.

In Figs. 3, 5 and 10, we qualitatively compare our generated videos with state-of-the-art event-based image reconstruction methods [3], [4], [5]. Experimental results indicate that event-only methods work well on scenes of fast camera motions since the distribution of events has a wide coverage of scene content. Also, E2VID [4] is enabled to provide more vivid color textures in the HDR scene. However, for scenes with a static background or a slowly moving background/foreground, the reconstructed images by event-only methods will lose texture details in the areas without events. On the contrary, our 'image and event' combined method achieves superior performance on scenes with high dynamic motions and works robustly even with static backgrounds and sparse events.

We also report our reconstruction (and deblurring) results on real dataset, including text images and low-lighting images, in Figs. 1, 6, and 9.

Compared with state-of-the-art deblurring methods, our method achieves superior results. In comparison to existing event-based image reconstruction methods [3], [4], [5], [33], our reconstructed images are not only more realistic but also contain richer details. For more deblurring results and *high-temporal resolution videos*, please visit our home page.

7 Limitation

Though event cameras record continuous, asynchronous streams of events that encode non-redundant information for our *mEDI* model, there are still some limitations when doing reconstruction.

1) Extreme lighting changes, such as suddenly turning on/off the light, moving from dark indoor scenes to outdoor scenes. The relatively low dynamic range of the intensity image might degrade the performance of our method in high dynamic scenes;
2) Event error accumulation, such as noisy event data, small object motions with fewer events. Though we integrate over small time intervals from the centre of the exposure time to mitigate this error, accumulated noise can reduce the quality of reconstructed images.

8 Conclusion

In this paper, we have proposed a *multiple Event-based Double Integral (mEDI)* model to naturally connect intensity images and events recorded by an event camera (DAVIS), which also takes the blur generation process into account. In this way, our model can be used to not only recover the latent sharp images but also reconstruct intermediate frames at a high frame rate. We also propose a simple yet effective method to solve our *mEDI* model. Due to the simplicity of our optimization process, our method is efficient as well. Extensive experiments show that our method can generate high-quality, high frame-rate videos efficiently under different conditions, such as low lighting and complex dynamic scenes.

Acknowledgments

This work was supported in part by the Australian Research Council through the "Australian Centre of Excellence for Robotic Vision" CE140100016, the Natural Science Foundation of China grants (61871325, 61420106007, 61671387, and 61603303), National Key Research and Development Program of China under Grant 2018AAA0102803 and the Australian Research Council (ARC) grants (DE140100180, DE180100628, and DP200102274).

References

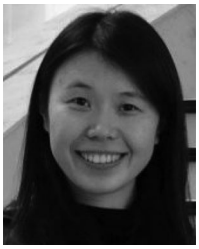
[1] X. Tao, H. Gao, X. Shen, J. Wang, and J. Jia, "Scale-recurrent network for deep image deblurring," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8174–8182.
[2] M. Jin, G. Meishvili, and P. Favaro, "Learning to extract a video sequence from a single motion-blurred image," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6334–6342.
[3] C. Scheerlinck, N. Barnes, and R. Mahony, "Continuous-time intensity estimation using event cameras," in *Proc. Asian Conf. Comput. Vis.*, 2018, pp. 308–324.
[4] H. Rebecq, R. Ranftl, V. Koltun, and D. Scaramuzza, "Events-to-video: Bringing modern computer vision to event cameras," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3852–3861.
[5] L. Pan, C. Scheerlinck, X. Yu, R. Hartley, M. Liu, and Y. Dai, "Bringing a blurry frame alive at high frame-rate with an event camera," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 6813–6822.
[6] P. Lichtsteiner, C. Posch, and T. Delbruck, "A 128 $\times$ 128 120 db 15 μs latency asynchronous temporal contrast vision sensor," *IEEE Journal Solid-State Circuits*, vol. 43, no. 2, pp. 566–576, Feb. 2008.
[7] C. Brandli, R. Berner, M. Yang, S.-C. Liu, and T. Delbruck, "A 240 $\times$ 180 130 db 3 μs latency global shutter spatiotemporal vision sensor," *IEEE J. Solid-State Circuits*, vol. 49, no. 10, pp. 2333–2341, Sep. 2014.
[8] P. Bardow, A. J. Davison, and S. Leutenegger, "Simultaneous optical flow and intensity estimation from an event camera," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 884–892.
[9] L. Wang, S. Mohammad Mostafavi I., Y.-S. Ho, and K.-J. Yoon, "Event-based high dynamic range image and very high frame rate video generation using conditional generative adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 10081–10090.
[10] C. Scheerlinck, N. Barnes, and R. Mahony, "Asynchronous spatial image convolutions for event cameras," *IEEE Robot. Autom. Lett.*, vol. 4, no. 2, pp. 816–822, Apr. 2019.
[11] P. Shedligeri and K. Mitra, "Photorealistic image reconstruction from hybrid intensity and event-based sensor," *J. Electron. Imag.*, vol. 28, no. 6, 2019, Art. no. 063012.
[12] C. Brandli, L. Muller, and T. Delbruck, "Real-time, high-speed video decompression using a frame- and event-based DAVIS sensor," in *IEEE Int. Symp. Circuits Syst. (ISCAS)*, Jun. 2014, pp. 686–689.
[13] S. Barua, Y. Miyatani, and A. Veeraraghavan, "Direct face detection and video reconstruction from event cameras," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2016, pp. 1–9.
[14] T. Stoffregen, G. Gallego, T. Drummond, L. Kleeman, and D. Scaramuzza, "Event-based motion segmentation by motion compensation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 7244–7253.
[15] M. Cook, L. Gugelmann, F. Jug, C. Krautz, and A. Steger, "Interacting maps for fast visual interpretation," in *Proc. Int. Joint Conf. Neural Netw.*, 2011, pp. 770–776.
[16] H. Kim, A. Handa, R. Benosman, S.-H. Ieng, and A. J. Davison, "Simultaneous mosaicing and tracking with an event camera," in *Proc. Brit. Mach. Vis. Conf.*, 2014, Art. no. 1.
[17] H. Kim, S. Leutenegger, and A. J. Davison, "Real-time 3D reconstruction and 6-DOF tracking with an event camera," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 349–364.

[18] H. Rebecq, T. Horstschäfer, G. Gallego, and D. Scaramuzza, "EVO: A geometric approach to event-based 6-DOF parallel tracking and mapping in real-time," *IEEE Robot. Autom. Lett.*, vol. 2, no. 2, pp. 593–600, Apr. 2017.
[19] A. R. Vidal, H. Rebecq, T. Horstschaefer, and D. Scaramuzza, "Ultimate SLAM? Combining events, images, and IMU for robust visual SLAM in HDR and high-speed scenarios," *IEEE Robot. Autom. Lett.*, vol. 3, no. 2, pp. 994–1001, Apr. 2018.
[20] L. Liu, H. Li, and Y. Dai, "Efficient global 2D-3D matching for camera localization in a large-scale 3D map," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2391–2400.
[21] L. Liu, H. Li, and Y. Dai, "Stochastic attraction-repulsion embedding for large scale image localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 2570–2579.
[22] A. Zhu, L. Yuan, K. Chaney, and K. Daniilidis, "Ev-flownet: Self-supervised optical flow estimation for event-based cameras," in *Proc. Robot.: Sci. Syst.*, 2018, doi: 10.15607/RSS.2018.XIV.062.
[23] D. Gehrig, A. Loquercio, K. G. Derpanis, and D. Scaramuzza, "End-to-end learning of representations for asynchronous event-based data," in *Proc. Int. Conf. Comput. Vis.*, 2019, pp. 5633–5643.
[24] T. Stoffregen *et al.*, "Reducing the Sim-to-Real gap for event cameras," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 1–36.
[25] L. Pan, M. Liu, and R. Hartley, "Single image optical flow estimation with an event camera," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 1672–1681.
[26] Y. Zhou, G. Gallego, H. Rebecq, L. Kneip, H. Li, and D. Scaramuzza, "Semi-dense 3D reconstruction with a stereo event camera," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 235–251.
[27] A. Z. Zhu, N. Atanasov, and K. Daniilidis, "Event-based visual inertial odometry," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5816–5824.
[28] D. Gehrig, H. Rebecq, G. Gallego, and D. Scaramuzza, "Asynchronous, photometric feature tracking using events and frames," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 1–22.
[29] B. Kueng, E. Mueggler, G. Gallego, and D. Scaramuzza, "Low-latency visual odometry using event-based feature tracks," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, 2016, pp. 16–23.
[30] G. Gallego *et al.*, "Event-based vision: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Jul. 10, 2020, doi: 10.1109/TPAMI.2020.3008413.
[31] C. Brandli, L. Muller, and T. Delbruck, "Real-time, high-speed video decompression using a frame-and event-based davis sensor," in *Proc. IEEE Int. Symp. Circuits Syst.*, 2014, pp. 686–689.
[32] C. Posch, D. Matolin, and R. Wohlgenannt, "A QVGA 143 dB dynamic range asynchronous address-event PWM dynamic image sensor with lossless pixel-level video compression," in *Proc. IEEE Intl. Solid-State Circuits Conf.*, 2010, pp. 400–401.
[33] C. Reinbacher, G. Graber, and T. Pock, "Real-time intensity-image reconstruction for event cameras using manifold regularisation," in *Proc. Brit. Mach. Vis. Conf.*, 2016, pp. 1–12.
[34] H. Rebecq, R. Ranftl, V. Koltun, and D. Scaramuzza, "High speed and high dynamic range video with an event camera," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Dec. 31, 2019, doi: 10.1109/TPAMI.2019.2963386.
[35] C. Scheerlinck, H. Rebecq, D. Gehrig, N. Barnes, R. Mahony, and D. Scaramuzza, "Fast image reconstruction with an event camera," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2020, pp. 156–163.
[36] H.-C. Liu, F.-L. Zhang, D. Marshall, L. Shi, and S.-M. Hu, "High-speed video generation with an event camera," *Vis. Comput.*, vol. 33, no. 6–8, pp. 749–759, 2017.
[37] R. Fergus, B. Singh, A. Hertzmann, S. T. Roweis, and W. T. Freeman, "Removing camera shake from a single photograph," *ACM Trans. Graph.*, vol. 25, pp. 787–794, 2006.
[38] D. Krishnan, T. Tay, and R. Fergus, "Blind deconvolution using a normalized sparsity measure," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 233–240.
[39] L. Sun, S. Cho, J. Wang, and J. Hays, "Edge-based blur kernel estimation using patch priors," in *Proc. IEEE Int. Conf. Comput. Photography*, 2013, pp. 1–8.
[40] X. Yu, F. Xu, S. Zhang, and L. Zhang, "Efficient patch-wise non-uniform deblurring for a single image," *IEEE Trans. Multimedia*, vol. 16, no. 6, pp. 1510–1524, Oct. 2014.
[41] L. Xu, S. Zheng, and J. Jia, "Unnatural l0 sparse representation for natural image deblurring," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 1107–1114.
[42] L. Pan, R. Hartley, M. Liu, and Y. Dai, "Phase-only image based kernel estimation for single image blind deblurring," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 6034–6043.
[43] W.-S. Lai, J.-J. Ding, Y.-Y. Lin, and Y.-Y. Chuang, "Blur kernel estimation using normalized color-line prior," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 64–72.
[44] J. Pan, D. Sun, H. Pfister, and M.-H. Yang, "Deblurring images via dark channel prior," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 10, pp. 2315–2328, Oct. 2017.
[45] Y. Yan, W. Ren, Y. Guo, R. Wang, and X. Cao, "Image deblurring via extreme channels prior," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6978–6986.
[46] T. H. Kim and K. M. Lee, "Generalized video deblurring for dynamic scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 5426–5434.
[47] A. Sellent, C. Rother, and S. Roth, "Stereo video deblurring," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 558–575.
[48] L. Pan, Y. Dai, M. Liu, and F. Porikli, "Simultaneous stereo video deblurring and scene flow estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6987–6996.
[49] L. Pan, Y. Dai, M. Liu, and F. Porikli, "Depth map completion by jointly exploiting blurry color images and sparse depth maps," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2018, pp. 1377–1386.
[50] L. Pan, Y. Dai, M. Liu, F. Porikli, and Q. Pan, "Joint stereo video deblurring, scene flow estimation and moving object segmentation," *IEEE Trans. Image Process.*, vol. 29, pp. 1748–1761, Oct. 2020.
[51] J. Sun, W. Cao, Z. Xu, and J. Ponce, "Learning a convolutional neural network for non-uniform motion blur removal," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 769–777.
[52] D. Gong *et al.*, "From motion blur to motion flow: A deep learning solution for removing heterogeneous motion blur," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2319–2328.
[53] S. Nah, T. H. Kim, and K. M. Lee, "Deep multi-scale convolutional neural network for dynamic scene deblurring," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 257–265.
[54] K. Purohit, A. Shah, and A. N. Rajagopalan, "Bringing alive blurred moments," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 6823–6832.
[55] T. Delbruck, Y. Hu, and Z. He, "V2e: From video frames to realistic dvs event camera streams," 2020, *arXiv:2006.07722*. [Online]. Available: http://arxiv.org/abs/2006.07722
[56] G. Gallego, J. E. Lund, E. Mueggler, H. Rebecq, T. Delbruck, and D. Scaramuzza, "Event-based, 6-DOF camera tracking from photometric depth maps," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 10, pp. 2402–2412, Oct. 2018.
[57] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D: Nonlinear Phenomena*, vol. 60, no. 1–4, pp. 259–268, 1992.
[58] E. Mueggler, H. Rebecq, G. Gallego, T. Delbruck, and D. Scaramuzza, "The event-camera dataset and simulator: Event-based data for pose estimation, visual odometry, and slam," *Int. J. Robot. Res.*, vol. 36, no. 2, pp. 142–149, 2017.
[59] J. Zhang *et al.*, "Dynamic scene deblurring using spatially variant recurrent neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2521–2529.
[60] J. Kiefer, "Sequential minimax search for a maximum," *Proc. Amer. Math. Soc.*, vol. 4, no. 3, pp. 502–506, 1953.
[61] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in c.* Cambridge, UK: Cambridge University Press, vol. 1, 1988, Art. no. 3.
[62] C. Scheerlinck, H. Rebecq, T. Stoffregen, N. Barnes, R. Mahony, and D. Scaramuzza, "CED: Color event camera dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2019, pp. 1–10.
[63] H. Rebecq, D. Gehrig, and D. Scaramuzza, "ESIM: An open event camera simulator," in *Proc. Conf. Robot Learn.*, 2018, pp. 969–982.
[64] G. Eilertsen, J. Kronander, G. Denes, R. K. Mantiuk, and J. Unger, "HDR image reconstruction from a single exposure using deep cnns," *ACM Trans. Graph.*, vol. 36, no. 6, pp. 1–15, 2017.

Liyuan Pan received the BEng degree from Northwestern Polytechnical University, Xi'an, China, from 2014. She is currently working toward the PhD degree in the College of Engineering and Computer Science, Australian National University, Canberra, Australia. Her interests include deblurring, flow estimation, depth estimation, and event-based vision.

Richard Hartley (Fellow, IEEE) is a member of the computer vision group in the Research School of Engineering, at ANU, where he has been since January, 2001. He is also a member of the computer vision research group in NICTA. He worked at the GE Research and Development Center from 1985 to 2001, working first in VLSI design, and later in computer vision. He became involved with Image Understanding and Scene Reconstruction working with GE's Simulation and Control Systems Division. He is an author (with A. Zisserman) of the book Multiple View Geometry in Computer Vision.

Cedric Scheerlinck received the BSc and MEng degrees from the University of Melbourne, in 2014 and 2016, respectively. He is currently working toward the PhD degree from the College of Engineering and Computer Science, Australian National University, Canberra, Australia. His interests include event-based vision and deep learning.



Miaomiao Liu (Member, IEEE.) received the BEng degree from Yantai Normal University, Yantai, China, in 2004, MEng degree from the Nanjing University of Aeronautics and Astronautics, Nanjing, China in 2007, and the PhD degree from the University of Hong Kong, Hong Kong, China, in 2012. She worked as a researcher in the computer vision group at NICTA (2012-2016) and a research scientist (2017-2018) at CSIRO's Data61 in Canberra, Australia. She joined the Australian National University (ANU) as an ARC DECRA fellow in 2018. She is currently a tenure-track lecturer in the ANU.

Xin Yu received the BS degree in electronic engineering from the University of Electronic Science and Technology of China, Chengdu, China, in 2009, and the PhD degree from the Department of Electronic Engineering, Tsinghua University, Beijing, China, in 2015, and the second PhD degree from the College of Engineering and Computer Science, Australian National University, Canberra, Australia, in 2019. He is currently a lecturer with the University of Technology Sydney. His interests include computer vision and image processing.

Yuchao Dai received the BE, ME, and PhD degrees from Northwestern Polytechnical University, Xi'an, China, in 2005, 2008 and 2012, respectively, all in signal and information processing. He is currently a professor at the School of Electronics and Information at the Northwestern Polytechnical University (NPU). He was an ARC DECRA fellow with the Research School of Engineering at the Australian National University, Canberra, Australia. His research interests include structure from motion, multiview geometry, low-level computer vision, deep learning, compressive sensing, and optimization. He won the Best Paper Award at IEEE CVPR 2012, the Best Paper Award Nominee at IEEE CVPR 2020, the DSTO Best Fundamental Contribution to Image Processing Paper Prize at DICTA 2014, the Best Algorithm Prize in NRSFM Challenge at CVPR 2017, the Best Student Paper Prize at DICTA 2017 and the Best Deep/Machine Learning Paper Prize at APSIPA ASC 2017. He served as area chair for IEEE CVPR, ACCV, ACM MM, etc.