

E-CIR: Event-Enhanced Continuous Intensity Recovery

Chen Song Qixing Huang Chandrajit Bajaj
 The University of Texas at Austin

{song, huangqx, bajaj}@cs.utexas.edu

Abstract

A camera begins to sense light the moment we press the shutter button. During the exposure interval, relative motion between the scene and the camera causes motion blur, a common undesirable visual artifact. This paper presents E-CIR, which converts a blurry image into a sharp video represented as a parametric function from time to intensity. E-CIR leverages events as an auxiliary input. We discuss how to exploit the temporal event structure to construct the parametric bases. We demonstrate how to train a deep learning model to predict the function coefficients. To improve the appearance consistency, we further introduce a refinement module to propagate visual features among consecutive frames. Compared to state-of-the-art event-enhanced deblurring approaches, E-CIR generates smoother and more realistic results. The implementation of E-CIR is available at <https://github.com/chensong1995/E-CIR>.

1. Introduction

The shutter speed, or the length of the exposure interval, controls how much light reaches the image sensor from the environment. If the exposure interval is too short, the camera only has the time to capture very few photons. Consequently, the resulting image is not only unilluminated but also lacks fine details. On the other hand, if the exposure interval is too long, the relative motion between the scene and the camera may potentially be very significant. The resulting image is then the temporal average of a moving trajectory, causing blurry artifacts. Traditionally, it is presumed that any motion during the exposure interval, including both camera shake and subject movement, is unwanted and should therefore be removed. Over the past several decades, researchers have studied extensively how to convert a blurry image into a sharp one [1, 5, 6, 10, 11, 13–16, 27, 32, 34, 41, 42]. It is only until recently when several works that reconstruct the complete motion trajectory have received profound attention [9, 28]. These works introduce algorithms that convert a blurry image into a sharp video describing the exact movement that causes the blurry artifact.

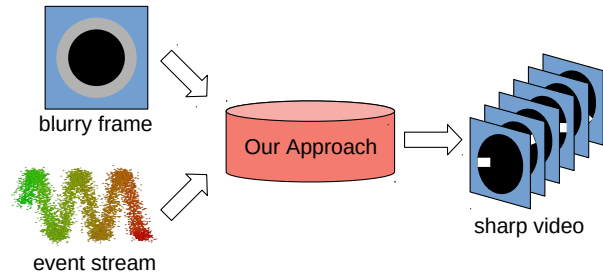


Figure 1. Problem Description. In this imaginary scene, we place a white square along the edge of a black disk. The image taken by the conventional camera is blurry because the disk rotates at a fast speed. It is as if the perimeter of the disk somehow grows into a gray collar. During the exposure interval, the event sensor produces a spiral of events. Our approach takes the blurry frame and the events as input and produces a sharp video sequence as output. The output video explains the motion blur by entailing the complete motion trajectory of the rotating disk.

Sharp video reconstruction is an ill-posed problem because there are infinitely many motion trajectories whose temporal averages correspond to the same blurry frame. To compensate for the ambiguity, previous works [7, 8, 20, 24, 25, 35, 39, 40, 43] exploit event data as an auxiliary input, which provides additional information during the exposure interval at a finer temporal resolution, as shown in Figure 1. Even with the event input, difficult challenges remain. The events fail to capture the complete motion information. The video reconstruction quality is determined not only by the appearance of each individual frame but also the temporal smoothness. The immense density of events creates another obstacle for effective and efficient processing. The success of video deblurring depends on how the blurry image, the events, and priors about video sequences are integrated together. This calls for suitable video representations and prediction algorithms.

This paper makes fundamental contributions in video representations and methodologies for recovering accurate and temporally consistent videos. Specifically, we propose a continuous video representation whose coefficients are highly interpretable and easy to learn, due to their strong correlation to the events. For every pixel (x, y) , we rep-

resent its intensity as a parametric polynomial function $\mathbf{L}_{xy}(t)$, allowing us to render the sharp image at any given timestamp t during the exposure interval. We show how to choose the polynomial bases such that the derivative of $\mathbf{L}_{xy}(t)$ interpolates the significant intensity changes. We also demonstrate how to train a deep neural network that regresses the polynomial coefficients. Instead of processing the video as a volume and implicitly encoding motions in convolutional filters, our approach explicitly asks the model to elaborate the motions that have already been described by the events. To further polish the frame quality, we introduce a refinement module that propagates the visual features among consecutive frames, which can be trained in an end-to-end manner with the rest of the model. The proposed regress-and-refine paradigm nicely combines the strength of recurrent modules for enforcing temporal smoothness and the strength of regression for drifting avoidance.

We quantitatively evaluate our method on the synthetic REDS dataset [21]. In terms of reconstruction quality, ECIR achieves an MSE of 0.114, representing a 37.4% improvement from state-of-the-art algorithms. We also present a qualitative evaluation on the real captures provided by Pan et al. [25]. Compared with baseline approaches, our method is less noisy, more realistic, and temporally smoother.

In summary, our key contributions are:

1. We represent a video by per-pixel parametric polynomials. We discuss why this representation integrates easily with the event mechanism by showing the parallelism between function derivatives and events.
2. From a blurry image and its associated events in the exposure interval, we demonstrate how to use a deep learning model to predict a sharp video represented by the proposed parametric polynomials.
3. To overcome the limitations of the polynomial representation, we discuss how to formulate a refinement objective and encourage the temporal propagation of sharp visual features.
4. We provide source code and documentation for converting the original REDS dataset into the event format. This clears the vagueness of the evaluation dataset in previous works and establishes an open-source benchmark for future comparisons.

2. Related Work

2.1. Event-Enhanced Deblurring

First commercialized in 2006 [18], event cameras are an emerging type of vision sensor that models the environment evolution as intensity changes and represents the scene as *events*. Each event is a 4-tuple (x, y, t, p) that contains the location, time, and polarity of an intensity change. This

simple representation allows event cameras to support a fast data rate (up to 1 MHz), orders of magnitude higher than the frame rate of conventional cameras. The density of events during the exposure interval provides valuable motion information to explain the blurred image.

Pan et al. propose the Event-based Double Integral (EDI) model [24, 25] that analytically reconstructs a high frame-rate sharp video from a blurry frame and its associated events. Jiang et al. [8] formulate a Maximum-a-Posteriori problem and solve for the latent sharp images under the Markov assumption with the help of deep neural networks. Lin et al. [20] believe it is inadequate to calculate the intensity residual between sharp and blurry frames directly from the event threshold and propose to predict the intensity residual using deep learning. Meanwhile, the structural similarity between the EDI model and the blur kernel formulation has inspired Wang et al. [39] to represent sharp images as sparse codes in a learnable dictionary and optimize them using an iterative network. Shang et al. [35] assume that the input sequence contains a mixture of blurry and sharp frames and propose to wrap sharp frames to deblur the blurry frame. Zhang et al. [43] emphasize the temporal correlation among consecutive frames and design a multi-patch convolutional LSTM to exploit such correlation. Han et al. [7] extend this idea by modeling the intensity residual between neighboring sharp frames. Xu et al. [40] also identify the importance of temporal correlation and propose to utilize the optical flow estimation instead.

Closely related to deblurring, event-enhanced frame interpolation has also attracted increasing attention [23, 38]. While both tasks aim at constructing a high frame-rate video, frame interpolation methods typically assume the input frames are free of motion blur. Several works have attempted to reconstruct a high frame-rate video directly from the events without the conventional frame input [4, 30, 31] as well. However, these event-only methods are less robust than their dual-input counterparts [20, 43].

2.2. Video Representation

To the best of our knowledge, most existing works in computer vision process videos as discrete collections of frames. The only exception is Vid-ODE [26], which represent videos by continuous latent states. The latent state can be evaluated at any given timestamp, allowing the video to be rendered with an infinitely high frame rate.

With the help of per-pixel parametric polynomials, our proposed representation also supports infinitely high rate rendering and enjoys two additional advantages. First, the polynomial bases are chosen to closely mimic the event mechanism, which makes the algorithm robust to domain differences between synthetic training data and real testing data. Second, the polynomial coefficients are more interpretable than the latent code hidden inside a deep network.

This allows humans to easily explain and debug the model.

3. Preliminaries

3.1. Event Camera Model

Let $\mathbf{L}_{xy}(t)$ be the latent intensity of pixel (x, y) at time t . In the natural logarithmic space, the temporal contrast between t_{ref} and t is given by [18]:

$$\Delta \ln[\mathbf{L}_{xy}(t)] = \ln[\mathbf{L}_{xy}(t)] - \ln[\mathbf{L}_{xy}(t_{\text{ref}})] \quad (1)$$

where t_{ref} denotes the timestamp of the last event associated with pixel (x, y) . The magnitude of $\Delta \ln[\mathbf{L}_{xy}(t)]$ determines whether the hardware produces an event. Let (x, y, t, p) denote an event, where $p \in \{-1, +1\}$ is the polarity of the intensity change [18]:

$$p = \begin{cases} +1 & \Delta \ln[\mathbf{L}_{xy}(t)] \geq c^+ \\ 0 \text{ (no event)} & c^- < \Delta \ln[\mathbf{L}_{xy}(t)] < c^+ \\ -1 & \Delta \ln[\mathbf{L}_{xy}(t)] \leq c^- \end{cases} \quad (2)$$

Here, c^+ and c^- are thresholds controlling the sensitivity of positive and negative events, respectively. It is commonly assumed that c^+ and c^- are stochastic variables [19, 29].

During an exposure interval $[-\frac{T}{2}, \frac{T}{2}]$ with length T , let $\mathbf{B}_{h \times w} = \{\mathbf{B}_{xy}\}$ be the blurry output from the conventional camera. The relation between the blurry frame and the latent frames is given by temporal averaging [25, 29]:

$$\mathbf{B}_{xy} = \frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} \mathbf{L}_{xy}(t) dt \quad (3)$$

3.2. Task Description

The input to the task has two components:

1. The blurry intensity \mathbf{B}_{xy} for all pixels (x, y) ;
2. A collection of events during the exposure interval $\{e_i = (x_i, y_i, t_i, p_i) \mid -\frac{T}{2} \leq t_i \leq \frac{T}{2}\}$.

Given an arbitrary timestamp $t \in [-\frac{T}{2}, \frac{T}{2}]$, the goal of the task is to construct the corresponding latent frame $\mathbf{L}_{h \times w}(t) = \{\mathbf{L}_{xy}(t)\}$.

3.3. Challenges

Intensity reconstruction is a highly ill-posed problem because there are infinitely many motion trajectories whose temporal averages correspond to the same blurry frame. Specifically, we face three challenges:

First, the events fails to capture complete motion information during the exposure interval. Equation (2) states that when an event happens, the magnitude of $\Delta \ln[\mathbf{L}_{xy}(t)]$ is greater than the event threshold. It remains unclear exactly how much $\Delta \ln[\mathbf{L}_{xy}(t)]$ exceeds the threshold. Imagine a

scene with two edges moving in the same pattern. The first edge has a strong contrast to the background and generates significant intensity changes for its movement. The second edge has a weak contrast to the background and generates small intensity changes. Suppose these two sets of intensity changes both exceed the event threshold. This means the number of events generated by the camera is determined only by the edge length. These two edges will yield the same number of events, as long as their lengths are equal, even though the absolute change in intensity of the first edge is several times higher than that of the second edge.

Second, the reconstruction quality is determined not only by the appearance of each individual frame but also the temporal smoothness. A naive model that independently optimizes each latent frame’s quality may lead to unrealistic motion trajectories and cause frequent jitter. We refer readers to the supplementary animations for a demonstration of how some of our baseline approaches fail to solve this issue.

Third, the event format is incompatible with popular deep learning models. One possible remedy is to aggregate the events into a histogram. This method ignores the event camera model and trains a network as if the inputs are merely some unexplained features. It remains unclear how to properly instill human knowledge about the bio-inspired event mechanism, such as the correlation between an event and an intensity change, into the model design.

While previous works fail to address some or all of these challenges, the next section discusses how the proposed ECIR handles them effectively.

4. Method

4.1. Parametric Intensity Function

For each pixel (x, y) , we propose to approximate the function $\mathbf{L}_{xy}(t)$ as a continuously differentiable mapping from the time domain to the intensity domain: $[-\frac{T}{2}, \frac{T}{2}] \rightarrow [0, 1]$. Inspired by the Taylor’s theorem, we parameterize the mapping as a degree- n polynomial. Let $\alpha_0, \alpha_1, \alpha_2, \dots, \alpha_n$ be the $n + 1$ polynomial coefficients. The simplest parameterization uses the standard polynomial bases:

$$\mathbf{L}_{xy}(t) = \alpha_0 + \alpha_1 t + \alpha_2 t^2 + \dots + \alpha_n t^n \quad (4)$$

Most real-life videos do not have frequent oscillations. This means a realistic intensity curve does not have very high-order derivative information. For a small change in t , the variation in $\mathbf{L}_{xy}(t)$ should not be too big. The fact that $|\alpha_0| \approx |\alpha_1| > |\alpha_2| > \dots \gg |\alpha_n|$ suggests that the standard representation is prone to numerical issues since high-order coefficients are expected to be very close to zero.

The temporal derivative of $\mathbf{L}_{xy}(t)$ reveals how the intensity changes across time. The events associated with pixel (x, y) provide a set of timestamps where the intensity change considerably. The derivatives at these timestamps

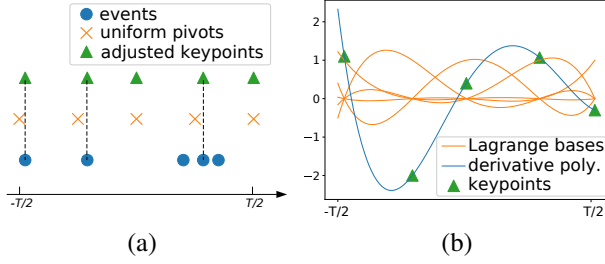


Figure 2. Keypoint Selection Algorithm. (a) Imagine there are $n = 5$ different events scattered across the exposure interval represented by blue circles. We first sample n evenly spaced pivots from the exposure interval $[-\frac{T}{2}, \frac{T}{2}]$ (the orange crosses). We then shift the pivots to their closest events and obtain n keypoints (green triangles). This selection scheme not only caters to the temporal event structure but also provides support to regions not covered by events. (b) At each keypoint, exactly one of the Lagrange bases has a value of 1, while all other Lagrange bases have a value of 0. We use a neural network to predict the value of $\frac{d\mathbf{L}_{xy}}{dt}(t)$ at these n keypoints. Their interpolation gives a degree- $(n - 1)$ polynomial, which is used to recover the primitive intensity signal $\mathbf{L}_{xy}(t)$ through indefinite integral. Under the Lagrangian representation, the polynomial coefficients coincide with predicted derivatives.

are expected to have significant magnitudes. The key idea of our proposed parameterization is to interpolate the temporal derivative of the intensity signal at event timestamps. The number of events associated with each pixel is different, presenting a challenge to efficient computation. To address this issue, we extract a fixed number of n keypoints for each pixel, regardless of how many events the pixel initially possesses. The details of our keypoint extraction algorithm are presented in Figure 2(a). This algorithm ensures the selected keypoints are in correspondence to the event timestamps and as distant to each other as possible. The use of the uniform pivots further establishes spatial consistency in the keypoint choices among different pixels. Let the set of n keypoints for pixel (x, y) be:

$$\mathcal{K}_{xy} = \left\{ \left(t_i, \frac{d\mathbf{L}_{xy}}{dt}(t_i) \right) \mid 1 \leq i \leq n \right\} \quad (5)$$

where $-\frac{T}{2} \leq t_1 < t_2 < \dots < t_n \leq \frac{T}{2}$. As shown in Figure 2(b), we parameterize the intensity derivative $\frac{d\mathbf{L}_{xy}}{dt}(t)$ as the polynomial interpolation of these n keypoints:

$$\frac{d\mathbf{L}_{xy}}{dt}(t) = \sum_{i=1}^n \frac{d\mathbf{L}_{xy}}{dt}(t_i) \cdot \beta_{xyi}(t) \quad (6)$$

Here, $\beta_{xyi}(t)$, $1 \leq i \leq n$ are the Lagrange bases of degree $n - 1$:

$$\beta_{xyi}(t) = \frac{(t - t_1) \cdots (t - t_{i-1})(t - t_{i+1}) \cdots (t - t_n)}{(t_i - t_1) \cdots (t_i - t_{i-1})(t_i - t_{i+1}) \cdots (t_i - t_n)}$$

The Lagrange bases have the characteristic that $\beta_{xyi}(t_i) = 1$ and $\forall j \neq i, \beta_{xyj}(t_i) = 0$. This ensures the

continuous function $\frac{d\mathbf{L}_{xy}}{dt}(t)$ passes through the n discrete keypoints $(t_i, \frac{d\mathbf{L}_{xy}}{dt}(t_i))$, where $1 \leq i \leq n$.

We can then recover the primitive intensity signal $\mathbf{L}_{xy}(t)$ from its derivative $\frac{d\mathbf{L}_{xy}}{dt}(t)$ by taking the indefinite integral:

$$\mathbf{L}_{xy}(t) = \int \frac{d\mathbf{L}_{xy}}{dt}(t) dt + a_{xy} \quad (7)$$

where a_{xy} is a constant that can be solved from Equation (3).

Compared to the conventional frame-based representation, the main advantage of the proposed polynomial representation is that it closely mimics the event mechanism. We leverage event timestamps to construct the polynomial bases and allow the polynomial coefficients to be interpreted as the intensity changes that trigger the input events. The regression target of our model, the polynomial coefficients, is therefore highly correlated to the input events. While the input events characterize the locations of the edge features, the output polynomial coefficients reveal exactly how significant the edges are. Section 5.4 presents an empirical verification of the advantage of our representation.

4.2. Prediction Pipeline

We illustrate the overall prediction pipeline in Figure 3, which consists of the initialization stage and the refinement stage. The initialization stage regresses the polynomial coefficients, evaluates predicted parametric functions, and obtains coarse video reconstruction. The refinement stage further polishes the details in the initial reconstruction by learning and enforcing motion priors. This methodology nicely combines the strength of motion priors for recovering temporally smooth videos and the strength of regressing the volumetric output to avoid drifting.

Initialization: Polynomial Coefficient Regression. We assemble the n keypoint timestamps associated with each pixel into an $n \times h \times w$ tensor \mathbf{K} , where $h \times w$ is the spatial frame resolution. Following eSL-Net [39], we voxelize the event stream by creating an $m \times h \times w$ histogram tensor \mathbf{E} , where $m = 40$ is the number of temporal bins. We adopt the U-Net [33] architecture as the backbone prediction network. As shown in Figure 3, the U-Net takes the blurry frame, the keypoints, and the events as input and regresses the polynomial coefficients for $\mathbf{L}_{xy}(t)$ represented under the Lagrange bases of its derivative. Let $\mathbf{t} = \{t_i\}$ be the d -dimensional vector collecting all timestamps of interest. With the help of Equation (7), the predicted coefficients allow us to reconstruct d initial frames $\{\hat{\mathbf{L}}(t_i)\}$. At training time, $d = 14$ is set to the number of available ground-truth latent frames. At inference time, d can be an arbitrary positive integer.

Refinement: Temporal Feature Propagation. When played as a video, the initialization results show an accurate motion reconstruction. Different parts of the scene move according to their respectively trajectories. Nonetheless, we

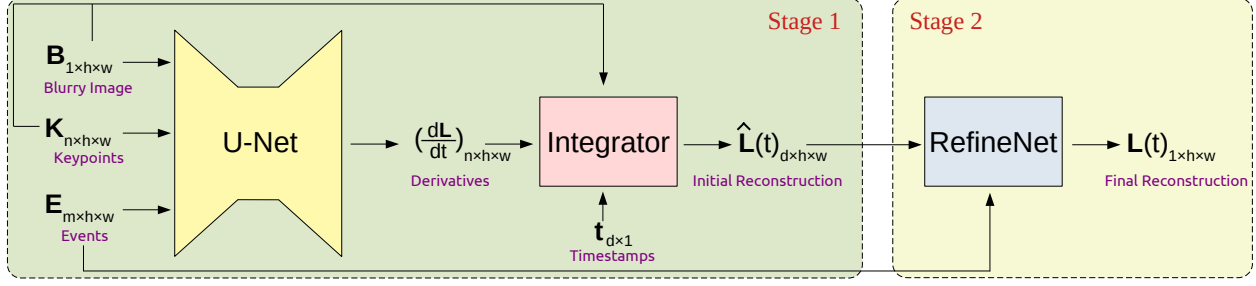


Figure 3. The overall pipeline. We use the U-Net [33] model to regress the polynomial coefficients. The network takes three inputs: the blurry frame \mathbf{B} , the keypoint timestamps \mathbf{K} , and the event histogram \mathbf{E} . The network then outputs the intensity derivatives $\frac{d\mathbf{L}}{dt}$. Given an arbitrary timestamp t , the integrator follows Equation (7) and calculates the initial frame reconstruction $\hat{\mathbf{L}}(t)$. The refinement module takes $\hat{\mathbf{L}}(t)$ and \mathbf{E} as input and outputs the final frame reconstruction $\mathbf{L}(t)$. In this figure, m, n, d, h, w represent the number of histogram bins, the number of keypoints, the number of frames in the output video, the frame height, and the frame width, respectively.

Algorithm 1 Refinement

Input: Initial reconstruction: $\{\hat{\mathbf{L}}(t_i) | 1 \leq i \leq d\}$
Input: All events: $\{(x, y, t, p) | -\frac{T}{2} \leq t \leq \frac{T}{2}\}$
Output: Final reconstruction: $\{\mathbf{L}(t_i) | 1 \leq i \leq d\}$

```

1: loop  $I_{max}$  iterations
2:   for  $i = 1$  to  $d - 1$  do           ▷ Residual Prediction
3:      $\mathbf{E} \leftarrow \text{VOXELIZE}(\text{events from } t_i \text{ to } t_{i+1})$ 
4:     if  $i == 1$  then
5:        $\mathbf{R}_i \leftarrow g_{\theta_1}^{\mathbf{R}}(\mathbf{E}, \hat{\mathbf{L}}(t_i), \hat{\mathbf{L}}(t_{i+1}),$ 
6:          $\nabla \hat{\mathbf{L}}(t_i), \nabla \hat{\mathbf{L}}(t_{i+1}))$ 
7:     else
8:        $\mathbf{R}_i \leftarrow g_{\theta_2}^{\mathbf{R}}(\mathbf{R}_{i-1}, \mathbf{E}, \hat{\mathbf{L}}(t_i), \hat{\mathbf{L}}(t_{i+1}),$ 
9:          $\nabla \hat{\mathbf{L}}(t_i), \nabla \hat{\mathbf{L}}(t_{i+1}))$ 
10:    end if
11:  end for
12:  for  $i = 1$  to  $d$  do           ▷ Apply Updates
13:     $\mathbf{A}_i \leftarrow g_{\theta_1}^{\mathbf{A}}(\hat{\mathbf{L}}(t_i))$ 
14:     $\mathbf{D}_i \leftarrow \frac{\partial f(\mathbf{R}_1, \dots, \mathbf{R}_{d-1}, \hat{\mathbf{L}}(t_1), \dots, \hat{\mathbf{L}}(t_d), \mathbf{L}(t_1), \dots, \mathbf{L}(t_d))}{\partial \mathbf{L}(t_i)}$ 
15:     $\hat{\mathbf{L}}(t_i) \leftarrow \hat{\mathbf{L}}(t_i) - \mathbf{A}_i \odot \mathbf{D}_i$ 
16:  end for
17: end loop
18: for  $i = 1$  to  $d$  do           ▷ Final Polishing
19:    $\mathbf{L}(t_i) \leftarrow g_{\gamma}^{\mathbf{L}}(\hat{\mathbf{L}}(t_i))$ 
20: end for

```

observe that the initial reconstruction occasionally fails to recover temporally consistent features. This initial temporal inconsistency is expected and addressed by the refinement.

There are two factors that contribute to the temporal inconsistency. First, the polynomial function is a continuous signal that smooths sharp features and makes them visually blurry. Second, visual features may only move during a small part of the exposure interval. When the feature is actively moving, the reconstruction is usually sharp because there are input events in the spatial neighborhood depicting the edge locations. When the feature is not moving, how-

ever, the reconstruction becomes blurry due to the lack of associated events and difficulty for volumetric filters used in regression to capture motion priors that are critical for deblurring.

In the refinement stage, we solve the first issue by optimizing the frames independent of polynomial formulation. We solve the second issue by encouraging visual features to propagate between consecutive frames (i.e., enforcing motion priors), a popular technique used extensively in video synthesis via optical flows [3, 40], residuals [7, 17], or deformable convolutional kernels [36, 37]. Details of our refinement process are presented as Algorithm 1. Specifically, we use a recurrent network to predict the residual \mathbf{R}_i between adjacent frames $\mathbf{L}(t_i)$ and $\mathbf{L}(t_{i+1})$. The inputs to the network include the previous residual \mathbf{R}_{i-1} between frames $\mathbf{L}(t_{i-1})$ and $\mathbf{L}(t_i)$, the events from t_i to t_{i+1} , the initial reconstructions $\hat{\mathbf{L}}(t_i)$ and $\hat{\mathbf{L}}(t_{i+1})$, as well as their spatial gradients $\nabla \hat{\mathbf{L}}(t_i)$ and $\nabla \hat{\mathbf{L}}(t_{i+1})$. Algorithm 1 refers this recurrent network as $g_{\theta_1}^{\mathbf{R}}$ (for the residual prediction between the first two frames) and $g_{\theta_2}^{\mathbf{R}}$ (for the rest of residuals).

The recurrent architecture allows the residual to be gradually updated according to the relevant events and intensity reconstruction. We augment the inputs to include $\nabla \hat{\mathbf{L}}(t_i) = (\frac{d\hat{\mathbf{L}}(t_i)}{dx}, \frac{d\hat{\mathbf{L}}(t_i)}{dy})$ and $\nabla \hat{\mathbf{L}}(t_{i+1}) = (\frac{d\hat{\mathbf{L}}(t_{i+1})}{dx}, \frac{d\hat{\mathbf{L}}(t_{i+1})}{dy})$. This is because both the spatial gradients and the temporal residuals are highly related to the edge features.

Consider the objective function in Equation (8), where $\mathcal{L}(\cdot, \cdot)$ represents the distance between two matrices.

$$f = \sum_{i=1}^{d-1} \mathcal{L}(\mathbf{L}(t_i) + \mathbf{R}_i, \mathbf{L}(t_{i+1})) + \lambda \sum_{i=1}^d \mathcal{L}(\mathbf{L}(t_i), \hat{\mathbf{L}}(t_i)) \quad (8)$$

The free variables are the refined frames $\mathbf{L}(t_i)$'s. The first objective term ensures the refinement output follows the residual flow. The second term discourages the refinement from deviating too far from the initialization. The trade-off parameter λ balances these two terms. We expect

both terms to have small residuals throughout the optimization process and the final result to be in local proximity to the initial frames. This leads us to choose the L2-distance as $\mathcal{L}(\cdot, \cdot)$ for its numerical stability. As shown in Algorithm 1, we apply gradient descent to update the refinement result for I_{max} iterations. Different pixels may have different step sizes for the update, which are predicted by a convolutional network referred to as g_ϕ^A . After the gradient descent terminates, we use another convolutional network g_γ^L to perform final polishing on each individual frame.

4.3. Training Objective

Derivative Loss. We use \mathcal{L}_d to supervise the output polynomial coefficients directly in the derivative domain:

$$\mathcal{L}_d = \left| \left(\frac{d\mathbf{L}}{dt} \right)_{\text{gt}} - \left(\frac{d\mathbf{L}}{dt} \right)_{\text{pred}} \right|_1 \quad (9)$$

Primitive Loss. We first recover the primitive intensity signal from Equation (7) and then use \mathcal{L}_p to supervise the polynomial coefficients indirectly in the primitive domain:

$$\mathcal{L}_p = |\mathbf{L}_{\text{gt}} - \mathbf{L}_{\text{pred}}|_1 \quad (10)$$

Refinement Loss. We use \mathcal{L}_{ref} to supervise the refinement output:

$$\mathcal{L}_{\text{ref}} = \sum_t |\mathbf{L}_{\text{gt}}(t) - \mathbf{L}_{\text{pred}}(t)|_1 \quad (11)$$

Residual Loss. We use \mathcal{L}_{res} to supervise the residual prediction in the refinement stage. Note that \mathcal{L}_{res} uses weighted L1-norm with $\rho = 5$ because the intensity residual between consecutive frames is sparse.

$$\mathcal{L}_{\text{res}} = \sum_i (\exp(\rho \cdot |\mathbf{R}_{i_{\text{gt}}}|_1) \odot |\mathbf{R}_{i_{\text{gt}}} - \mathbf{R}_{i_{\text{pred}}}|_1) \quad (12)$$

Total Objective. The final training objective is the weighted sum of the losses introduced above:

$$\mathcal{L}_{\text{total}} = \lambda_d \mathcal{L}_d + \lambda_p \mathcal{L}_p + \lambda_{\text{ref}} \mathcal{L}_{\text{ref}} + \lambda_{\text{res}} \mathcal{L}_{\text{res}} \quad (13)$$

where $\lambda_d, \lambda_p, \lambda_{\text{ref}}, \lambda_{\text{res}}$ are constant trade-off factors. The U-Net and the prediction networks in the refinement stage ($g_{\theta_1}^R, g_{\theta_2}^R, g_\phi^A, g_\gamma^L$) are trained in an end-to-end manner.

5. Evaluation

5.1. Datasets

REDS [21] is a standard deblurring benchmark dataset designed for conventional cameras. The dataset contains 240 training videos and 30 validation videos and is publicly available under the CC BY 4.0 license. These videos are captured at 120 fps and are sharp and clear. We use the frame interpolation algorithm [22] to further increase the frame rate to 960 fps. After that, we convert the videos

| Methods | MSE ↓ | PSNR ↑ | SSIM ↑ |
|---------------------------|--------------|---------------|--------------|
| EDI [25] | 0.182 | 21.663 | 0.664 |
| eSL-Net [39] (official) | 0.203 | 20.640 | 0.601 |
| eSL-Net [39] (re-trained) | 0.201 | 20.748 | 0.646 |
| Ours | 0.114 | 25.531 | 0.819 |

Table 1. Quantitative evaluation on the REDS [21] dataset.

to grayscale and resize the frames to 240×180 , consistent with the DAVIS240¹ [2] sensor resolution. We apply the ESIM [29] simulator to generate events and synthesize blurry frames according to Equation (3) with an exposure interval of 120 milliseconds. We use least-squares to fit polynomial coefficients to the resized 960 fps sharp frames during each exposure interval. These coefficients are used to supervise network training. This process is an effort to reproduce the synthetic dataset described by Wang et al. [39], who have released the model weights after training but have not disclosed data processing or training scripts.

5.2. Baseline Approaches

For quantitative evaluation, we compare our model with eSL-Net [39] using both the official weights released by Wang et al. and the weights re-trained on our synthetic data. We take EDI [25] as an additional baseline model. At the submission time, they are the only two approaches with open-source implementation available online.

For qualitative evaluation, we also visualize the results on real event captures provided by Pan et al. [25]. The blurriness of real captures comes from the physical sensor instead of simulated temporal averaging. Therefore, there is a lack of “ground-truth” sharp images, and we are unable to numerically evaluate the performance of each algorithm.

5.3. Training Details

The trade-off parameters for the derivative (\mathcal{L}_d), primitive (\mathcal{L}_p), refinement (\mathcal{L}_{ref}), and residual (\mathcal{L}_{res}) losses are 1, 10, 10, and 0.5, respectively. We use the Adam [12] optimizer with a batch size of 96 and train the network for 50 epochs. The initial learning rate is 0.0001 and is halved at the end of the 20th and the 40th epochs. The degree of the polynomial functions we use is $n = 10$.

5.4. Analysis of Results

Baseline Comparison. We present a quantitative evaluation of our approach in Table 1. Compared to baseline models, our method obtains lower MSE, higher PSNR, and higher SSIM. Specifically, the proposed E-CIR improves the current state-of-the-art algorithm by 37.4% in MSE, 17.8% in PSNR, and 23.3% in SSIM.

Qualitatively, we compare E-CIR with baseline models in Figure 4 and Figure 5. Visually, our results are not only

¹DAVIS240 is a popular camera that records grayscale conventional frames and events simultaneously.



Figure 4. Qualitative visualization on the REDS [21] dataset.

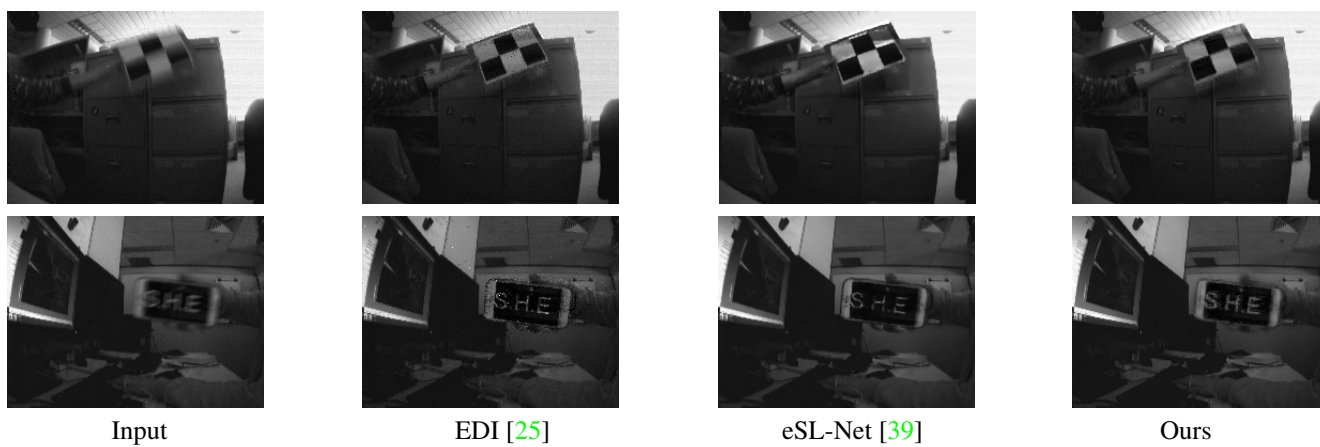


Figure 5. Qualitative visualization on real captures by Pan et al. [25].

| Row | Input Sources | | Video Format | | Stages | | Performance Metrics | | |
|-----|---------------|--------|--------------|-------|----------------|------------|---------------------|---------------|--------------|
| | Frame | Events | Poly. | Frame | Initialization | Refinement | MSE ↓ | PSNR ↑ | SSIM ↑ |
| 1 | ✗ | ✓ | ✓ | ✗ | ✓ | ✗ | 0.134 | 24.210 | 0.767 |
| 2 | ✓ | ✗ | ✓ | ✗ | ✓ | ✗ | 0.180 | 21.721 | 0.654 |
| 3 | ✓ | ✓ | ✓ | ✗ | ✓ | ✗ | 0.125 | 24.807 | 0.787 |
| 4 | ✓ | ✓ | ✗ | ✓ | ✓ | ✗ | 0.136 | 23.504 | 0.723 |
| 5 | ✓ | ✓ | ✓ | ✗ | ✓ | ✓ | 0.114 | 25.531 | 0.819 |

Table 2. On the REDS [21] dataset, we use ablation studies to demonstrate the importance of using a dual-stream input, the power of the polynomial representation, and the strength of the refinement module.

sharper but also less noisy than the EDI [24] reconstruction. The blurriness of EDI can be explained by their use of total variation as an image quality prior, which penalizes spatial edge features. By contrast, our method implicitly learns a deep prior from the sharp images in the training dataset. In terms of the noises, EDI assumes the event threshold can precisely characterize all intensity changes. As discussed in Section 3.3, this assumption is inaccurate and susceptible to artifacts. Compared to eSL-Net [39], our model does not over exaggerate visual features. As shown in the second row of Figure 4, the eSL-Net output contains a distorted patch in the bottom-right corner, an overemphasis of the floor tile. The excessive amount of noise makes eSL-Net unfavorable in the quantitative evaluation, even though subjectively, it reconstructs sharper frames than EDI. We also invite readers to watch our supplementary animations. These animations demonstrate that our method generates temporally smooth reconstruction, while the baseline approaches suffer from trajectory discontinuity to varying degrees.

Ablation Study. The input to our model has two components: the blurry frame and the events. Prior works have approached the video reconstruction problem using both the blurry frame alone [9, 28] and the events alone [4, 30, 31]. We begin our ablation study by examining the benefit of using a dual-stream input. As shown the first three rows in Table 2, removing either input from the pipeline leads to noticeable performance degradation. For example, the MSE of the dual-input model is 6.7% lower than the event-only model and 30.6% lower than the frame-only model. This suggests that conventional frames and events are complementary to each other, and our proposed E-CIR is able to take collaborative advantage of the combined information.

To examine if the polynomial video representation is indeed superior to the traditional frame-based representation, we use the same U-Net architecture to regress the $d = 14$ ground-truth sharp frames directly. The comparison between the third and fourth rows in Table 2 shows that our proposed polynomial representation outperforms the frame-based baseline by 8.1% in MSE. This result demonstrates power of explicit derivative modeling in event data.

Finally, we examine the effectiveness of refinement. Between the third and the fifth row in Table 2, we observe that the refinement module improves the initialization by

8.8% in MSE. The supplementary material includes an animated comparison between the initialization and refinement frames. While the initialization results successfully recovers the motion trajectory, the visual features are occasionally not sharp enough. The refinement algorithm is able to sharpen the initial results by encouraging high-quality visual features to propagate among consecutive frames.

6. Limitations

We point out that standard image quality metrics have a negative bias towards approaches that generate sharp but noisy results, such as eSL-Net [39]. Ideally, we would like to separate the visual signal and the noise and measure them independently. Practically, we have to resort to distance-based metrics without the disentanglement. Readers are encouraged to compare our approach with the baselines visually while using the quantitative evaluation as a reference.

7. Conclusion

This paper introduces E-CIR, a novel event-enhanced deblurring approach that represents the intensity signal as a continuous parametric function. Experiments show that E-CIR outperforms current state-of-the-art models in reconstruction quality. In the future, we plan to extend the approach and integrate spatial continuity into the formulation. Another possible direction is to explore probabilistic inference since each blurry image corresponds to more than one possible realistic motion trajectory.

8. Acknowledgement

We would like to sincerely appreciate Jijia Yu (RPI), Jiacheng Chen (SFU), Zhenpei Yang (UT Austin), Bo Sun (UT Austin), Siming Yan (UT Austin), Haitao Yang (UT Austin), Jiaru Song (UT Austin), Yan He (Rice), and Ao Li (Yext) for their constructive criticism and valuable suggestions. Additionally, Qixing Huang acknowledges the support from NSF Career IIS-2047677 and NSF HDR TRIPODS1934932. Chandrajit Bajaj acknowledges the support from the NIH (DK129979), in part from the Peter O’Donnell Foundation, and in part from a grant from the Army Research Office accomplished under Cooperative Agreement Number W911NF-19-2-0333.

References

- [1] S Derin Babacan, Rafael Molina, Minh N Do, and Aggelos K Katsaggelos. Bayesian blind deconvolution with general sparse image priors. In *European conference on computer vision*, pages 341–355. Springer, 2012. 1
- [2] Christian Brandli, Raphael Berner, Minhao Yang, Shih-Chii Liu, and Tobi Delbruck. A 240×180 130 db 3 μ s latency global shutter spatiotemporal vision sensor. *IEEE Journal of Solid-State Circuits*, 49(10):2333–2341, 2014. 6
- [3] Jose Caballero, Christian Ledig, Andrew Aitken, Alejandro Acosta, Johannes Totz, Zehan Wang, and Wenzhe Shi. Real-time video super-resolution with spatio-temporal networks and motion compensation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4778–4787, 2017. 5
- [4] Pablo Rodrigo Gantier Cadena, Yeqiang Qian, Chunxiang Wang, and Ming Yang. Spade-e2vid: Spatially-adaptive denormalization for event-based video reconstruction. *IEEE Transactions on Image Processing*, 30:2488–2500, 2021. 2, 8
- [5] Rob Fergus, Barun Singh, Aaron Hertzmann, Sam T Roweis, and William T Freeman. Removing camera shake from a single photograph. In *ACM SIGGRAPH 2006 Papers*, pages 787–794, 2006. 1
- [6] DA Fish, AM Brinicombe, ER Pike, and JG Walker. Blind deconvolution by means of the richardson–lucy algorithm. *JOSA A*, 12(1):58–65, 1995. 1
- [7] Jin Han, Yixin Yang, Chu Zhou, Chao Xu, and Boxin Shi. Evintsr-net: Event guided multiple latent frames reconstruction and super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4882–4891, 2021. 1, 2, 5
- [8] Zhe Jiang, Yu Zhang, Dongqing Zou, Jimmy Ren, Jiancheng Lv, and Yebin Liu. Learning event-based motion deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3320–3329, 2020. 1, 2
- [9] Meiguang Jin, Givi Meishvili, and Paolo Favaro. Learning to extract a video sequence from a single motion-blurred image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6334–6342, 2018. 1, 8
- [10] Neel Joshi, C Lawrence Zitnick, Richard Szeliski, and David J Kriegman. Image deblurring and denoising using color priors. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1550–1557. IEEE, 2009. 1
- [11] Sang Ku Kim, Sang Rae Park, and Joon Ki Paik. Simultaneous out-of-focus blur estimation and restoration for digital auto-focusing system. *IEEE Transactions on Consumer Electronics*, 44(3):1071–1075, 1998. 1
- [12] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6
- [13] Dilip Krishnan and Rob Fergus. Fast image deconvolution using hyper-laplacian priors. *Advances in neural information processing systems*, 22:1033–1041, 2009. 1
- [14] Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin, and Jiří Matas. Deblurgan: Blind motion deblurring using conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8183–8192, 2018. 1
- [15] Orest Kupyn, Tetiana Martyniuk, Junru Wu, and Zhangyang Wang. Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8878–8887, 2019. 1
- [16] Anat Levin, Rob Fergus, Frédo Durand, and William T Freeman. Image and depth from a conventional camera with a coded aperture. *ACM transactions on graphics (TOG)*, 26(3):70–es, 2007. 1
- [17] Sheng Li, Fengxiang He, Bo Du, Lefei Zhang, Yonghao Xu, and Dacheng Tao. Fast spatio-temporal residual network for video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10522–10531, 2019. 5
- [18] P. Lichtsteiner, C. Posch, and T. Delbruck. A 128×128 120db 30mw asynchronous vision sensor that responds to relative intensity change. In *2006 IEEE International Solid State Circuits Conference - Digest of Technical Papers*, pages 2060–2069, 2006. 2, 3
- [19] Patrick Lichtsteiner, Christoph Posch, and Tobi Delbruck. A 128×128 120 db 15 μ s latency asynchronous temporal contrast vision sensor. *IEEE Journal of Solid-State Circuits*, 43(2):566–576, 2008. 3
- [20] 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 *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16*, pages 695–710. Springer, 2020. 1, 2
- [21] 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 *CVPR Workshops*, June 2019. 2, 6, 7, 8
- [22] Simon Niklaus, Long Mai, and Feng Liu. Video frame interpolation via adaptive separable convolution. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 261–270, 2017. 6
- [23] Genady Paikin, Yotam Ater, Roy Shaul, and Evgeny Soloveichik. Efi-net: Video frame interpolation from fusion of events and frames. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1291–1301, June 2021. 2
- [24] 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, 2, 8
- [25] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6820–6829, 2019. 1, 2, 3, 6, 7
- [26] Sunghyun Park, Kangyeol Kim, Junsoo Lee, Jaegul Choo, Joonseok Lee, Sookyoung Kim, and Edward Choi. Vid-ode:

- Continuous-time video generation with neural ordinary differential equation. *arXiv preprint arXiv:2010.08188*, 2020. [2](#)
- [27] Daniele Perrone and Paolo Favaro. Total variation blind deconvolution: The devil is in the details. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2909–2916, 2014. [1](#)
- [28] Kuldeep Purohit, Anshul Shah, and AN Rajagopalan. Bringing alive blurred moments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6830–6839, 2019. [1](#), [8](#)
- [29] Henri Rebecq, Daniel Gehrig, and Davide Scaramuzza. ESIM: an open event camera simulator. *Conf. on Robotics Learning (CoRL)*, Oct. 2018. [3](#), [6](#)
- [30] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. Events-to-video: Bringing modern computer vision to event cameras. *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2019. [2](#), [8](#)
- [31] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. High speed and high dynamic range video with an event camera. *IEEE Trans. Pattern Anal. Mach. Intell. (T-PAMI)*, 2019. [2](#), [8](#)
- [32] William Hadley Richardson. Bayesian-based iterative method of image restoration. *JoSA*, 62(1):55–59, 1972. [1](#)
- [33] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. [4](#), [5](#)
- [34] Qi Shan, Jiaya Jia, and Aseem Agarwala. High-quality motion deblurring from a single image. *Acm transactions on graphics (tog)*, 27(3):1–10, 2008. [1](#)
- [35] Wei Shang, Dongwei Ren, Dongqing Zou, Jimmy S Ren, Ping Luo, and Wangmeng Zuo. Bringing events into video deblurring with non-consecutively blurry frames. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4531–4540, 2021. [1](#), [2](#)
- [36] Yapeng Tian, Yulun Zhang, Yun Fu, and Chenliang Xu. Tdan: Temporally deformable alignment network for video super-resolution. *arXiv preprint arXiv:1812.02898*, 2018. [5](#)
- [37] Yapeng Tian, Yulun Zhang, Yun Fu, and Chenliang Xu. Tdan: Temporally-deformable alignment network for video super-resolution. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. [5](#)
- [38] Stepan Tulyakov, Daniel Gehrig, Stamatios Georgoulis, Julius Erbach, Mathias Gehrig, Yuanyou Li, and Davide Scaramuzza. Time lens: Event-based video frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16155–16164, 2021. [2](#)
- [39] Bishan Wang, Jingwei He, Lei Yu, Gui-Song Xia, and Wen Yang. Event enhanced high-quality image recovery. In *European Conference on Computer Vision*. Springer, 2020. [1](#), [2](#), [4](#), [6](#), [7](#), [8](#)
- [40] Fang Xu, Lei Yu, Bishan Wang, Wen Yang, Gui-Song Xia, Xu Jia, Zhendong Qiao, and Jianzhuang Liu. Motion deblurring with real events. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2583–2592, 2021. [1](#), [2](#), [5](#)
- [41] Li Xu and Jiaya Jia. Two-phase kernel estimation for robust motion deblurring. In *European conference on computer vision*, pages 157–170. Springer, 2010. [1](#)
- [42] Li Xu, Shicheng Zheng, and Jiaya Jia. Unnatural l0 sparse representation for natural image deblurring. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1107–1114, 2013. [1](#)
- [43] Limeng Zhang, Hongguang Zhang, Chenyang Zhu, Shasha Guo, Jihua Chen, and Lei Wang. Fine-grained video deblurring with event camera. In *International Conference on Multimedia Modeling*, pages 352–364. Springer, 2021. [1](#), [2](#)