Motion Aware Event Representation-driven Image Deblurring

Zhijing Sun, Xueyang Fu, Longzhuo Huang, Aiping Liu, and Zheng-Jun Zha*

School of Information Science and Technology and MoE Key Laboratory of
Brain-inspired Intelligent Perception and Cognition,
University of Science and Technology of China, Hefei, 230026, China
{sunzhijing, hlz_0}@mail.ustc.edu.cn
{xyfu, aipingl, zhazj}@ustc.edu.cn

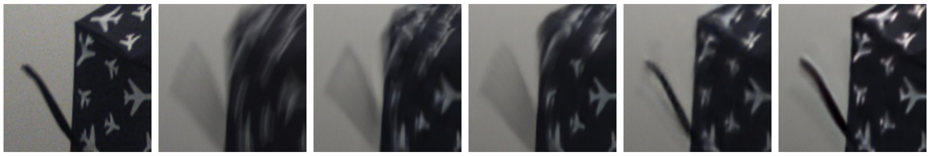
Abstract. Traditional image deblurring struggles with high-quality reconstruction due to limited motion data from single blurred images. Excitingly, the high-temporal resolution of event cameras records motion more precisely in a different modality, transforming image deblurring. However, many event camera-based methods, which only care about the final value of the polarity accumulation, ignore the influence of the absolute intensity change where events generate so fall short in perceiving motion patterns and effectively aiding image reconstruction. To overcome this, in this work, we propose a new event preprocessing technique that accumulates the deviation from the initial moment each time the event is updated. This process can distinguish the order of events to improve the perception of object motion patterns. To complement our proposed event representation, we create a recurrent module designed to meticulously extract motion features across local and global time scales. To further facilitate the event feature and image feature integration, which assists in image reconstruction, we develop a bi-directional feature alignment and fusion module. This module works to lessen inter-modal inconsistencies. Our approach has been thoroughly tested through rigorous experiments carried out on several datasets with different distributions. These trials have delivered promising results, with our method achieving top-tier performance in both quantitative and qualitative assessments. Code is available at https://github.com/ZhijingS/DA_event_deblur.

Keywords: Image deblurring · Event camera · Event representation

1 Introduction

Frame-based cameras often produce blurred images due to factors like camera shake or object motion, given their configured exposure time. Such blur, a common form of image degradation, occurs across various scenes. While photographers typically aim for clear and focused photos, clear inputs are equally critical for numerous computer vision tasks. These tasks include super-resolution, target

*Corresponding author



(a) GT (PSNR |SSIM) **(b)** Blurry (28.62dB |0.819) **(c)** HINet [4] (29.14dB |0.837) **(d)** NAFNet [3] (30.41dB |0.855) **(e)** EFNet [33] (30.76dB |0.856) **(f)** Ours (31.01dB |0.858)

Fig. 1: A deblurring example on HS-ERGB [37] dataset via recent state-of-the-art methods. Compared to other methods, our approach yields a more distinct result owing to its enhanced motion perception capability.

detection, and semantic segmentation, to name a few. Hence, image deblurring serves as a fundamental task in the realm of low-level computer vision, reinforcing the image quality and supporting the proper function of various algorithms.

Traditional image-based techniques often attempt to derive a clear image from a single blurred one. They do this by leveraging natural image priors or making assumptions about blur operations [6, 8, 13, 15, 34, 44]. However, in capturing a single blurred image, much motion information is lost. This lack of motion information during exposure creates undesirable artifacts in the deblurred results, especially in complex motion scenarios, as illustrated in Figures 1c and 1d. To counter this challenge, recent methodologies use more temporal information, such as differently exposed frames [2, 46, 48] and neighboring frames [16, 28, 32]. However, these approaches have their limitations. For instance, the use of differently exposed frames is constrained by the low temporal resolution of frame-based cameras, improving the reconstruction effect primarily in slow-motion scenes. The method that uses adjacent frames to estimate the current frame's motion information can lead to accumulative errors. In summary, traditional deblurring techniques encounter hurdles due to the inability of frame-based cameras to capture motion information during exposure.

As opposed to frame-based cameras, event cameras, inspired by biological systems, offer a promising solution to this issue. These cameras' microsecond-level temporal resolution (μ s) and asynchronous output architecture allow the recording of fast motions. This capability enables event streams to store pixel intensity changes and provide accurate motion information to deblur the corresponding frame. By using event frames as additional input, some methods attempt to guide the deblurring process through these event cameras [9, 25, 39, 43, 49]. One method [45] subdivides the event stream into ternary 2D data and generates a mask to localize high-blur regions, thereby enhancing the network's ability to distinguish different blur levels. Another method [33] symmetrically accumulates event polarities and fuses event and image features using an attention mechanism, enabling the integration of motion information into image reconstruction. However, these methods primarily employ the event stream in a manner akin to polarity accumulation. In essence, they consider only the final accumulation result of relative intensity change between subsequent moments, neglecting the absolute intensity change caused by historical events. Consequently, this approach fails to differentiate between various motion patterns, injecting unnecessary ambiguity into the deblurring process, and leading to edge artifacts as displayed in Figure 1e.

Given the limited perception of motion during exposure offered by polarity accumulation, we posit that the deviation from the initial moment provides more precise motion information for image deblurring. Each time the event generates, the absolute light intensity changes. The deviation from the initial moment enables us to consider the impact of changes in absolute light intensity, even when the absolute light intensity at the initial moment is unknown. Thus, the deviation, to a degree, implies the order of arrival of the events, offering different information for varying motion patterns. As a result, we propose the Deviation Accumulation (DA) method to enhance motion perception. Specifically, every time a new event arrives, our method accumulates the deviation relative to the initial moment, thus encompassing more comprehensive motion information. Furthermore, we find that longer time scales for event accumulation can capture global motion patterns, while shorter scales precisely depict local fast-motion patterns. To better perceive motion at both global and local time scales, we have designed a Recurrent Motion Extraction (RME) module that works with corresponding event representations. Moreover, in line with multi-modal learning [18–20, 29, 36, 42], we present an effective Feature Alignment and Fusion (FAF) module. This module serves to address the issue that simple fusion methods struggle with the disparate nature of event and frame modalities. As a result, these two modalities align, ensuring better information interaction. We integrate these modules into a single image deblurring network. This network extracts the motion features within the exposure time from the event stream and combines them with the texture information supplied by the image to complete the motion blur removal, as illustrated in Figure 1f.

The key contributions of this paper are as follows:

- We introduce an event preprocessing method designed specifically for the deblurring problem. The deviation accumulation is capable of responding distinctively to various motion patterns, thereby adding accurate motion information for image deblurring tasks.
- We develop a network equipped with recurrent motion extraction modules, designed to better extract motion features on local and global time scales. This design enables the network to handle multiple composite motion scenarios effectively.
- We propose a feature alignment and fusion module to mitigate the impacts of inter-modal inconsistencies and ensure that motion information more effectively guides the image deblurring process.

Our experimental results indicate that our method achieves state-of-the-art performance on synthetic, semi-synthetic, and real-world datasets, including a 0.61dB improvement in PSNR on the GoPro dataset compared to previous studies [33]. Moreover, our method produces more distinct subjective results, as demonstrated in Figure 1.

2 Related Work

2.1 Event-based Motion Deblurring

The utilization of event information to assist in deblurring tasks has become increasingly common in recent research [9, 12, 25, 31, 39, 43, 49]. The high temporal resolution and low latency of event cameras not only capture edge positions but also embed moving temporal information within the event stream. This information is crucial to reconstruct the latent sharp image. Many methods have been attempted to explore the relationship between event information, blurred images, and sharp images. One such method [25] includes the Event-based Double Integral (EDI) model, which leverages the principles of event generation to formulate an optimization problem that establishes a correlation between a blurry image and a latent sharp image. In recent studies, the event stream is typically converted into event frames [30, 38], which simplifies the processing of event data within traditional CNN frameworks. [9] combines the motion feature extracted from event data with the image feature, using them as the input for the decoder. [50] employs a Multi-Layer Perceptron (MLP) to predict the fused feature value from event and image features. [33] integrates them through a cross-modal attention method, which generates a query matrix from the image features, but derives key and value matrices from the event features.

Confronting the prevalent issue of integrating image and event features is the problem of inter-modal inconsistencies, which traditional fusion techniques cannot easily mitigate. To address this, our approach seeks to align these two modalities, optimizing the use of event data. Through this alignment, our goal is to minimize the inconsistencies spawned when merging differing modalities, thus amplifying the efficacy of the deblurring process.

2.2 Event Representation

Event data, characterized by its spatial sparsity and strong temporal continuity due to its asynchronous generation, necessitates a custom-designed event representation to deliver the required information. Methods [11, 26, 41] that process events individually maximize the use of high temporal resolution, as suggested by [26], which proposed an adaptive spiking neuron model to manage the fluctuating event input. The effectiveness of these algorithms, however, can be significantly hampered by the noise within the event due to independent processing. A Time Surface [1, 14, 22, 51] is created using the timestamp of the last event at each pixel, which prominently reveals the rich temporal information contained within the event. Furthermore, the event frame [5, 7, 21, 23, 45] transforms the original 4D event data into a 2D frame by accumulating events on the number or the polarity pixel-wise, facilitating extraction with conventional image processing techniques. In [45], the authors utilize ternary data for their event frames to allocate high blur region of the image.

However, existing event frame approaches result in abundant light intensity change information being lost in the squeezing process, making it impossible to distinguish the order of event arrivals and lacking the ability to detect multiple motion patterns. Some techniques attempt to recover some of this lost data by dividing the exposure time into smaller segments, though this essentially equates to sampling the original data. Consequently, we design a motion-aware event representation specifically tailored for motion deblurring.

3 Method

3.1 Overall Framework

The overall framework of our proposed method is depicted in Figure 2. We adopt a classical encoder-decoder architecture for our approach. Initially, the blurred image is inputted into the image branch, where we use the baseline conv block mentioned in [3] as our fundamental block, to extract relevant features. Simultaneously, the corresponding event frames generated by our deviation accumulation are fed into the Recurrent Motion Extraction (RME) module, which facilitates the extraction of features covering motion in local and global time scales. Following each block of the two branches, the Feature Alignment and Fusion (FAF) module is employed to eliminate the inconsistencies between the two modalities. By performing down-sampling operations, we are able to obtain both shallow and semantic information. In the decoder, skip connections are utilized to incorporate the extracted features for image reconstruction. The architecture of the conv block and the decoder is shown in the supplementary materials. The entire training process is conducted in an end-to-end manner, with the PSNR Loss function [4] employed to optimize the network's parameters.

3.2 Deviation Accumulation Event Representation

Event cameras are bio-inspired sensors which asynchronously output signals. They trigger an event while the logarithm of the intensity change exceeds the preset threshold c at any pixel, which is formulated as

$$p = \begin{cases} +1 & \text{if } log(I_{xy}(t)) - log(I_{xy}(t - \mathit{\Delta} t)) \geq c, \\ -1 & \text{if } log(I_{xy}(t)) - log(I_{xy}(t - \mathit{\Delta} t)) \leq -c, \end{cases} \tag{1}$$

where p is the event's polarity, $I_{xy}(t)$ and $I_{xy}(t - \mathit{\Delta} t)$ represent the intensity at the pixel coordinate (x, y) at time t and $t - \mathit{\Delta} t$, respectively. An event is a quaternion value (x, y, t, p) where x, y denotes the pixel coordinate, t represents the time of change, the polarity $p \in \{+1, -1\}$ indicates the increase and decrease. For a given blurry image B, it can be formulated as the average of latent intensity images $I(t)$ during the exposure time [25]:

$$B = \frac{1}{T} \int_0^T I(t) dt, \tag{2}$$

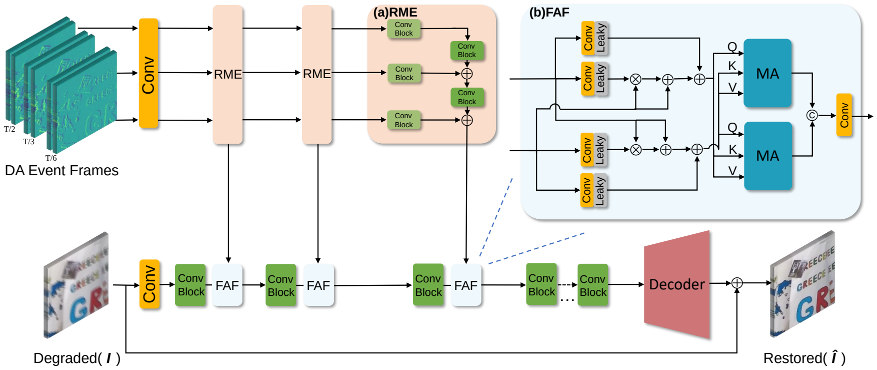


Fig. 2: Overview of the architecture of our proposed method for image deblurring. The blurred image and related event frames within the exposure time are fed into the corresponding network branches respectively. The core module of the event branch: (a) Recurrent Motion Extraction (RME) module processes event frames from longer time scales to shorter, merging old features into new features through weighted sum, which results in perceiving various motion patterns under different time ranges. The image branch extracts the features with conv block and fuses them with event features using (b) Feature Alignment and Fusion (FAF) module. FAF module eliminates the inter-modal inconsistencies in a bi-directional way ensuring better information interaction.

where T is the exposure time. For simplicity of understanding, we now only consider the pixel at (x, y) then combining Equations (1) and (2), we have

$$
\begin{aligned}
B_{xy} &= \frac{1}{T}\int_0^T exp(log(I_{xy}(t)))dt && (3)\\
&= \frac{1}{T}\int_0^T exp(log(I_{xy}(0)) + c\int_0^t p(s)ds)dt. && (4)
\end{aligned}
$$

The internal part of the exponential operation is the logarithmic value of the light intensity of the pixel at time t, while the external part of the exponential operation is the integration and averaging of each moment within the exposure time. Based on the event data, we can discretize Equation (4):

$$
\begin{aligned}
B_{xy} &= \frac{1}{N}\sum_{t'\in\mathcal{T}} exp(log(I_{xy}(0)) + c\sum_{t=0}^{t'} p_t) && (5)\\
&= \frac{I_{xy}(0)}{N}\sum_{t'\in\mathcal{T}} exp(c\sum_{t=0}^{t'} p_t), && (6)
\end{aligned}
$$

where $\mathcal{T} = \{t_1, t_2, ..., t_N\}$ indicates the set of time event occurs, N is the number of events, and p_t means the polarity of the event at time t.

The aforementioned derivation prompts us to consider not only the polarity of the event but also the impact of the alteration in the absolute value after

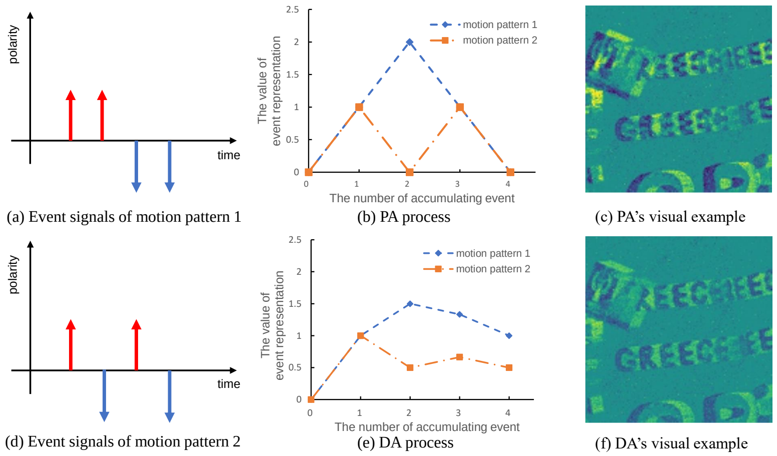
Algorithm 1 Deviation Accumulation (DA)

```
Input: taking pixel at $(x, y)$ for an example, the events stream $(x, y, t_i, p_i)$ in which
    $i \in [1, N]$, deviation from the initial moment $D$, accumulation of deviation $R$,
Output: the value of DA event representation $V$.
$D = 0, R = 0$;
Counting the number of events $N$;
for ($i = 1$ to $N$) do
    $D = D + p_i$
    $R = R + D$
end for
$V = R/N$
```

each accumulation when considering the problem of deblurring. However, since event data cannot provide $I_{xy}(0)$, the absolute light intensity is not accessible. But we can indirectly capture this information through the deviation from the initial moment. Therefore, we design a new event representation method for the deblurring problem presented in Algorithm 1. First, we accumulate all the events in the exposure time, so we get a 2D matrix named frequency matrix representing the number of the event occurs at each pixel position. We deal with the events considering their polarities subsequently, recording the deviation at each pixel by an additional tensor. Upon the arrival of the subsequent event, we update the deviation according to polarity and then add the deviation to the corresponding position. In the end, by dividing the cumulative matrix by the frequency matrix, we obtain the average deviation change over the exposure period as the new event representation named Deviation Accumulation (DA).

Traditional polarity accumulation methods directly accumulating event polarities lead to identical outputs whenever the number of positive and negative events within a given time frame ($[0, T]$) remains unchanged. Nevertheless, by considering the deviation brought by historical events, our method can generate distinct values for events arriving in different orders, thereby improving the ability to perceive motion patterns and providing more precise motion information for the deblurring procedure. For example, as illustrated in Figure 3, (a) shows event signals generated by a motion pattern which represents that the light intensity at that pixel has risen twice and then fallen twice while (d) shows another pattern. Facing these two patterns, (b) and (e) depict the changing process of the outputs of traditional method and our method after each event is accumulated. After processing all four events, traditional polarity accumulation gets the same output of 0 for both patterns, but our DA representation yields 1 and 0.5 respectively which contains richer motion information. More examples of motion patterns our method can distinguish are shown in supplementary materials.

Overall, the phenomenon that polarity accumulation methods represent some different motion patterns as the same value introduces ambiguity into the deblurring process. Our method has a better perception of the motion patterns and offers a more distinct motion edge than conventional approaches from the visual effect as shown in Figure 3 (c) and (f). Inspired by [33], directly segmenting



(a) Event signals of motion pattern 1 (b) PA process (c) PA's visual example

(d) Event signals of motion pattern 2 (e) DA process (f) DA's visual example

Fig. 3: (a) and (d) are event signals of two different motion patterns. (b) shows the Polarity Accumulation (PA) can't distinguish (a) and (d) as it gets the same output of 0 after processing all events while (e) shows our Deviation Accumulation (DA) can. DA perceives more motion information than PA as shown in (c) and (f).

the entire exposure time averagely is not suitable for motion deblurring, which focuses on capturing rapid motion but falls short in adequately modeling the continuous movement throughout the exposure. In this paper, we symmetrically accumulate events along the $T/2$ axis using the DA method. Taking $[0, T/2]$ as an example, we accumulate events on different time scales. The longest accumulating range is $T/2$, followed by $T/3$, and the shortest is $T/6$. A detailed schematic is provided in the supplementary materials.

3.3 Recurrent Motion Extraction Module

The goal of the Recurrent Motion Extraction (RME) module is to cater to both global slow motion and local fast motion in different time scales. Based on our observations of the motion process and event representations, the focus of the event representations varies in different time scales. In shorter time interval, the global slow motion is not significant in the event representation due to the smaller number of events triggered. On the contrary slow motion can be clearly recorded in longer interval, but at the same time, the violent moving part will produce edge spreading artifact.

To enable the network to accurately model the various in-homogeneous motions, we devised an RME module that incorporates a recurrent mechanism. As shown in Figure 2(a), the event features are fed into the processing module from the longest time scales to the shortest and we use $e1$, $e2$, and $e3$ to denote them. Formally, the RME module can be presented as

$$e_i^{'} = \begin{cases} f_{ex}(e_i) & \text{if } i = 1, \\ (1-\alpha) \times f_{ex}(e_i) + \alpha \times f_{re}(e_{i-1}^{'}) & \text{if } i = 2,3, \end{cases} \tag{7}$$

where $f_{ex}(\cdot)$ and $f_{re}(\cdot)$ are two baseline conv blocks, $f_{ex}(\cdot)$ is used to extract information from the current representation, $f_{re}(\cdot)$ is used to retain useful information from previous features, α is the learnable weight as a trade-off between two features. By sharing the parameters of $f_{ex}(\cdot)$ and $f_{re}(\cdot)$ for different time scales, we force the network to be able to perceive motions at varying time scales. Each input feature is symmetric on the time axis so that the priori information about the motion direction is simultaneously fed into the network.

3.4 Feature Alignment and Fusion Module

Although the information captured by the event camera exhibits a high correlation with the frame camera during the equivalent exposure period, there is still an inevitable gap originating from the modal dissimilarities between them. This phenomenon is frequently mentioned in multi-modal learning [18,42], but in event-guided image restoration tasks people usually [9,33,45] fuse features from two modalities directly using convolutions or attention mechanisms. Therefore, we propose an effective Feature Alignment and Fusion (FAF) module to facilitate enhanced feature fusion in a more coherent environment. Concretely, as shown in Figure 2(b), this module consists of a feature alignment module and a multi-head attention module. We use the operations of multiplication and addition successively to merge the raw information of one modality into the changing process of the other modality, which fills the gap of inter-modal inconsistency. The aligned features are fed into the attention module. In each fusion process, attention plays a bi-directional role, we not only generate the Q matrix through the frame features and the K and V matrix through the event features but also use the event features to perform look-ups on the frame features. The weights of the attention module are shared so that we can fuse the motion information under a more consistent space. Therefore, FAF module can be expressed as

$$x^{'} = Conv_2(x) + (Conv_1(x) \times e + e), \tag{8}$$

$$e^{'} = Conv_2(e) + (Conv_1(e) \times x + x), \tag{9}$$

$$F = Conv_3(Concat([Attention(Q_{x'}, K_{e'}, V_{e'}), Attention(Q_{e'}, K_{x'}, V_{x'})])), \tag{10}$$

where $x^{'}$ and $e^{'}$ are aligned image feature x and event feature e respectively, $Attention(\cdot)$ denotes the multi-head attention operation, $Concat(\cdot)$ denotes the concatenation operation along the channel dimension.

4 Experiments and Analysis

4.1 Experimental Settings

Datasets. We use three different datasets containing synthetic, semi-synthetic, and real-world frames and events to evaluate the proposed method.

GoPro: We train the network on the GoPro dataset which is the benchmark dataset for the image motion deblurring [24]. The blurry image is averaged from

Table 1: The quantitative results on GoPro, HS-ERGB, and REBlur test datasets. The best values are highlighted in bold. HINet+ and UFPNet+ are event-enhanced versions of HINet and UFPNet by concatenating our DA to their input.

Method	**Input**	**GoPro**		**HS-ERGB**		**REBlur**		**Params(M)**
		PSNR	**SSIM**	**PSNR**	**SSIM**	**PSNR**	**SSIM**	
LEVS [10]	F	20.84	0.547	22.13	0.555	-	-	18.2
EDI [25]	F+E	29.06	0.943	23.93	0.704	36.52	0.964	0.5
SRN [35]	F	30.26	0.934	-	-	35.10	0.961	10.2
EVDI [49]	F+E	30.40	0.906	25.13	0.707	-	-	**0.4**
HINet [4]	F	32.71	0.959	27.32	0.807	35.58	0.965	88.7
Restormer [47]	F	32.92	0.961	27.55	0.808	35.50	0.959	26.1
MSDI-Net [17]	F	33.28	0.964	27.46	0.809	36.14	0.968	241.3
NAFNet [3]	F	33.71	0.967	27.64	0.811	36.15	0.969	67.8
UFPNet [6]	F	34.06	0.968	27.64	0.809	36.11	0.968	80.3
HINet+ [4]	F+E	34.63	0.968	27.66	0.808	37.92	0.976	88.7
UFPNet+ [6]	F+E	35.22	0.972	27.68	0.809	37.97	0.976	80.3
EFNet [33]	F+E	35.46	0.972	26.68	0.800	38.12	0.975	8.5
Ours	F+E	**36.07**	**0.976**	**27.93**	**0.812**	**38.47**	**0.978**	13.9

adjacent sharp frames, by doing so the dataset contains 3214 pairs of blurry and sharp images. According to the standard division, we adopt 2103 pairs for training and 1111 pairs for testing. For the event data, we use the GoPro raw event dataset provided by [33] which synthesizes events with a random threshold c following the Gaussian distribution by ESIM simulator [27]. Following the DA method, the raw event data is preprocessed into 6 event frames for each image.

HS-ERGB: The HS-ERGB dataset [37] consists of sharp videos and real-world events, we use [50] released normal blur version which synthesizes blurry frames by averaging 49 interpolating images. Each blurry frame in the dataset has a known exposure time, allowing us to generate the DA representation for each frame within its respective exposure duration.

REBlur: The REBlur dataset collects a number of sequences of real-world event data, corresponding with blurry images and reliable ground-truth sharp images [33]. It contains 12 kinds of linear and nonlinear motions. Following the original setting, we use 486 pairs of blurry-sharp images with associated events for training and 983 pairs for testing.

Implementation details. We train the end-to-end network without pre-training on a single Tesla V100 GPU. We extract patches of size 256 × 256 for training images and corresponding event frames, and the batch size is set to 2 by default. The AdamW optimizer is employed ($\beta_1 = 0.9$ and $\beta_2 = 0.99$) with the initial learning rate 0.0001 with the cosine annealing schedule where the T_{max} is 400K iteration. For data augmentation, we perform horizontal and vertical flips, the random noise and the hot pixels are added into event frames in order to simulate the real-world situation. For HS-ERGB and REBlur datasets, We fine-tune the model trained on GoPro with the training set of them. The fine-tuning process

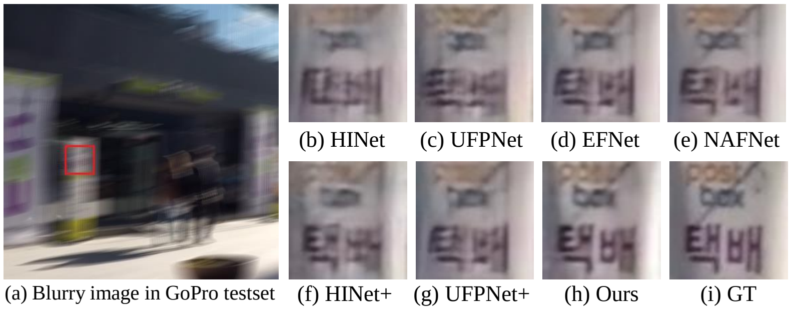

(a) Blurry image in GoPro testset (b) HINet (c) UFPNet (d) EFNet (e) NAFNet (f) HINet+ (g) UFPNet+ (h) Ours (i) GT

Fig. 4: Qualitative comparisons on the GoPro dataset. The notation is the same as in Table 1. Zoom in for better view.

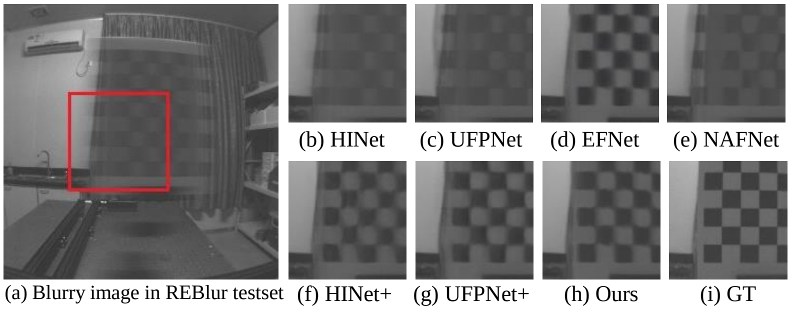

(a) Blurry image in REBlur testset (b) HINet (c) UFPNet (d) EFNet (e) NAFNet (f) HINet+ (g) UFPNet+ (h) Ours (i) GT

Fig. 5: Qualitative comparisons on the REBlur dataset. The notation is the same as in Table 1. Zoom in for better view.

consisted of 4000 iterations while keeping the other configurations consistent with the previous experiments.

4.2 Comparison with State-of-the-Art Methods

Table 1 shows the comparison results of our method and other SOTA methods on GoPro [24], HS-ERGB [50] and REBlur [33]. We compare our network with not only event-based methods, but also with some SOTA image-only deblurring networks enhancing with event input. The results illustrate that: (1) Our method outperforms all other methods in synthetic, semi-synthetic and real-world datasets, which indicates that with the proper utilization of events the designed approach is able to handle blur with different distributions. Specifically, the average gains of our method over the second-best one are 0.61dB, 0.25dB and 0.35dB in terms of PSNR on GoPro, HS-ERGB and REBlur datasets, respectively. (2) Through the enhancement of our proposed event representation, the HINet+ and the UFPNet+ increase 1.92dB and 1.16dB in PSNR on the benchmark GoPro dataset compared with their image-only version, respectively. Taking HINet as an example, with the same operation of [33], HINet with DA

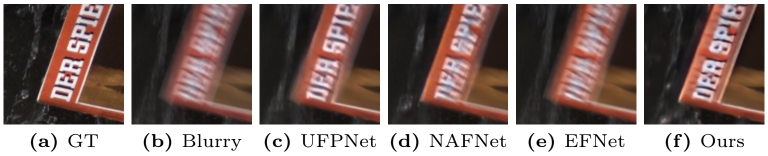

(a) GT **(b)** Blurry **(c)** UFPNet **(d)** NAFNet **(e)** EFNet **(f)** Ours

Fig. 6: Qualitative comparisons on the HS-ERGB dataset.

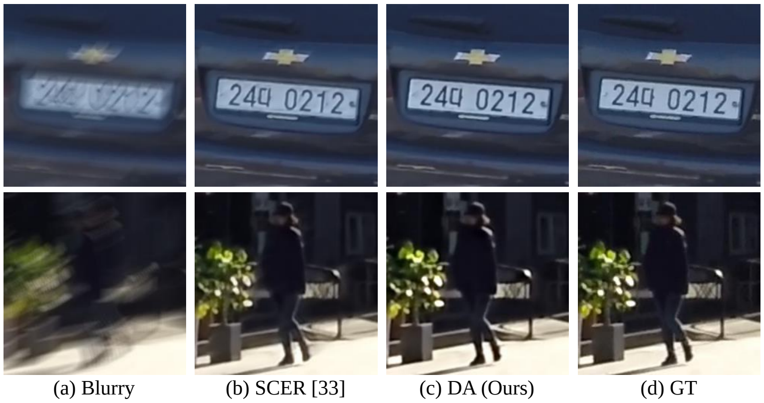

(a) Blurry (b) SCER [33] (c) DA (Ours) (d) GT

Fig. 7: Visual comparisons on GoPro dataset with different event representation. Compared to SCER representation, our DA representation reduces edge artifacts.

event representation exceeds 0.94dB over the result in [33], which presents the fact that DA extracts more conducive information for motion deblurring.

We also show the visual comparison results to verify the effectiveness of our method in Figure 4, Figure 5 and Figure 6 with representative samples from each dataset. As illustrated in the figures, our model reconstructs better details and has minor spatial distortions compared with other methods. This indicates that our event representation perceives more motion information and the modules effectively embed it to reconstruct a high-quality image with enhanced details.

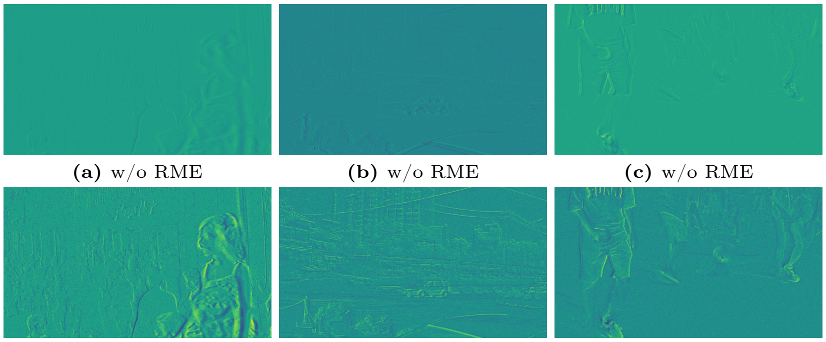
4.3 Ablation Studies

We conduct plenty ablation experiments on GoPro dataset to verify the effectiveness of our event representation and proposed method.

Effectiveness of Event Representation. To demonstrate the adaptability of our event representation for motion deblurring tasks, as shown in Table 2, we replace our DA representation with several other types of representations and trained the network accordingly. The results reveal certain shortcomings of the alternative representations. Firstly, the stack representation simply accumulates all events on their polarities which leaves out the temporal information. Secondly, the SBT approach [40], which subdivides the entire exposure time equally, proves to be inadequate for modeling the varying degrees of blur. Thirdly, the SCER

Table 2: Ablation study of each module effects.

Event Representation	RME	Fusion	PSNR	SSIM
Stack	✓	FAF(Ours)	33.24	0.953
SBT [40]			35.63	0.973
SCER [33]			35.81	0.975
ACA(Ours)	✓	Concat.	34.02	0.967
		Add	34.74	0.970
		Attention	35.75	0.974
		Only Align Event+Attention	35.78	0.975
		Only Align Image+Attention	35.74	0.974
ACA(Ours)	✗	FAF(Ours)	35.42	0.973
	✓		**36.07**	**0.976**



(a) w/o RME **(b)** w/o RME **(c)** w/o RME

(d) w/ RME **(e)** w/ RME **(f)** w/ RME

Fig. 8: Visual comparison results of w/ and w/o RME module. All of these scenarios involve movement in different directions and speeds, RME module can successfully capture motion features in these complex motion scenarios. Zoom in for better view.

method [33], which directly accumulates event polarities, is found to be incapable of capturing the influence of absolute intensity changes. This limitation is illustrated in Figure 3. In contrast, our DA representation not only possesses a solid physical foundation for event generation but also perceives more motion patterns. As a result, it achieves a 0.26dB improvement in performance compared to the best alternative representation. We also provide the visual comparison on the image reconstruction results in Figure 7. Our DA representation reduces motion-generated edge artifacts, resulting in a better deblurring effect.

Effectiveness of Recurrent Motion Extraction Module. To assess the RME module's contribution, we replace it with a baseline conv block and adjust the parameters for fair comparison. In this approach, we concatenate event frames of different time scales and feed them into the conv block. The findings, as presented in the last two rows of Table 2, clearly indicate that the contribution of 0.65dB increment in PSNR stems from the network structure design, rather than an increase in parameter quantity. This shows the importance of incorporating the module for effectively leveraging motion information in multi-time scales and improving performance in motion deblurring tasks. We also present the visual comparisons when removing the RME module in Figure 8, the RME module adeptly captures both fast local and slow global motion characteristics.

Effectiveness of Feature Alignment and Fusion Module. To illustrate the effectiveness of our FAF module, we compare it with several common fusion methods, including concatenation, addition, and cross-modality attention. As shown in Table 2, it is arduous to utilize the motion information contained in event feature through the direct concatenation or addition. While attention mechanisms can capture relative parts in the two features, the inherent inconsistencies within the two modalities still lead to a 0.32dB performance gap when compared to our FAF module. We also use a t-SNE visualization in Figure 9 to demonstrate the effectiveness of our feature alignment, where two originally in-congruent modal features are mixed up together after the FAF module. Furthermore, we also verify the importance of bi-directional alignment by conducting experiments solely on aligning event or image features. Since both events and frames provide crucial information for image restoration, such as edge motion information and textual information, respectively, it is unreasonable to let one modality dominate the alignment process. As shown in Table 2, the bi-directional alignment outperforms the highest score achieved by one-way correction by 0.29dB. This demonstrates the necessity of fully utilizing the mutual information between the two modalities in order to achieve better image reconstruction results.

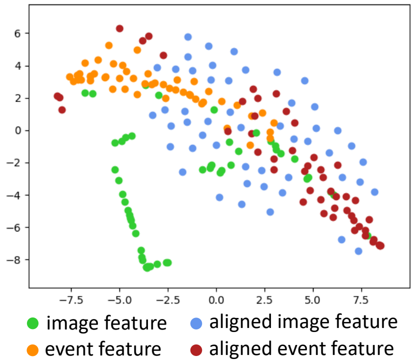


Fig. 9: t-SNE visualization of original and aligned features. The distributions of the aligned two features overlap more compared to the original features. This indicates the effectiveness of FAF module to reduce the inter-modal inconsistency.

5 Conclusion

In this paper, we propose an event-based image deblurring method. Firstly, we rethink the limit of traditional event preprocessing method and propose a new deviation accumulation (DA) technique. With deviation, the new approach can distinguish the order of events and perceive more motion patterns. Moreover, we design an image deblurring network equipped with recurrent motion extraction (RME) module and feature alignment and fusion (FAF) module to fit our proposed event representation. The RME module can better perceive motions in different time scales, and the FAF module promotes better integration of the two modalities. Finally, extensive experimental results over several datasets demonstrate the effectiveness of the proposed algorithm.

Acknowledgements. This work was supported by National Natural Science Foundation of China (NSFC) under Grants 62225207 and 62276243.

References

1. Benosman, R., Clercq, C., Lagorce, X., Ieng, S.H., Bartolozzi, C.: Event-based visual flow. IEEE transactions on neural networks and learning systems **25**(2), 407–417 (2013)
2. Chang, M., Feng, H., Xu, Z., Li, Q.: Low-light image restoration with short-and long-exposure raw pairs. IEEE Transactions on Multimedia **24**, 702–714 (2021)
3. Chen, L., Chu, X., Zhang, X., Sun, J.: Simple baselines for image restoration. In: Proceedings of the European conference on computer vision (ECCV). pp. 17–33. Springer (2022)
4. Chen, L., Lu, X., Zhang, J., Chu, X., Chen, C.: Hinet: Half instance normalization network for image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 182–192 (2021)
5. Cook, M., Gugelmann, L., Jug, F., Krautz, C., Steger, A.: Interacting maps for fast visual interpretation. In: The 2011 International Joint Conference on Neural Networks. pp. 770–776. IEEE (2011)
6. Fang, Z., Wu, F., Dong, W., Li, X., Wu, J., Shi, G.: Self-supervised non-uniform kernel estimation with flow-based motion prior for blind image deblurring. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18105–18114 (2023)
7. Gehrig, D., Rebecq, H., Gallego, G., Scaramuzza, D.: Eklt: Asynchronous photometric feature tracking using events and frames. International Journal of Computer Vision **128**(3), 601–618 (2020)
8. Ji, S.W., Lee, J., Kim, S.W., Hong, J.P., Baek, S.J., Jung, S.W., Ko, S.J.: Xydeblur: divide and conquer for single image deblurring. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17421–17430 (2022)
9. Jiang, Z., Zhang, Y., Zou, D., Ren, J., Lv, J., Liu, Y.: Learning event-based motion deblurring. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3320–3329 (2020)
10. Jin, M., Meishvili, G., Favaro, P.: Learning to extract a video sequence from a single motion-blurred image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6334–6342 (2018)
11. Kim, H., Handa, A., Benosman, R., Ieng, S.H., Davison, A.J.: Simultaneous mosaicing and tracking with an event camera. J. Solid State Circ **43**, 566–576 (2008)
12. Kim, T., Lee, J., Wang, L., Yoon, K.J.: Event-guided deblurring of unknown exposure time videos. In: Proceedings of the European conference on computer vision (ECCV). pp. 519–538. Springer (2022)
13. Krishnan, D., Tay, T., Fergus, R.: Blind deconvolution using a normalized sparsity measure. In: CVPR 2011. pp. 233–240. IEEE (2011)
14. Lagorce, X., Orchard, G., Galluppi, F., Shi, B.E., Benosman, R.B.: Hots: a hierarchy of event-based time-surfaces for pattern recognition. IEEE transactions on pattern analysis and machine intelligence **39**(7), 1346–1359 (2016)
15. Levin, A., Weiss, Y., Durand, F., Freeman, W.T.: Understanding and evaluating blind deconvolution algorithms. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 1964–1971. IEEE (2009)

16. Li, D., Shi, X., Zhang, Y., Cheung, K.C., See, S., Wang, X., Qin, H., Li, H.: A simple baseline for video restoration with grouped spatial-temporal shift. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9822–9832 (2023)
17. Li, D., Zhang, Y., Cheung, K.C., Wang, X., Qin, H., Li, H.: Learning degradation representations for image deblurring. In: Proceedings of the European conference on computer vision (ECCV). pp. 736–753. Springer (2022)
18. Li, Y., Yu, A.W., Meng, T., Caine, B., Ngiam, J., Peng, D., Shen, J., Lu, Y., Zhou, D., Le, Q.V., et al.: Deepfusion: Lidar-camera deep fusion for multi-modal 3d object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17182–17191 (2022)
19. Liang, M., Yang, B., Wang, S., Urtasun, R.: Deep continuous fusion for multi-sensor 3d object detection. In: Proceedings of the European conference on computer vision (ECCV). pp. 641–656 (2018)
20. Liang, X., Qian, Y., Guo, Q., Cheng, H., Liang, J.: Af: An association-based fusion method for multi-modal classification. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(12), 9236–9254 (2021)
21. Liu, M., Delbruck, T.: Adaptive time-slice block-matching optical flow algorithm for dynamic vision sensors. BMVC (2018)
22. Manderscheid, J., Sironi, A., Bourdis, N., Migliore, D., Lepetit, V.: Speed invariant time surface for learning to detect corner points with event-based cameras. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10245–10254 (2019)
23. Maqueda, A.I., Loquercio, A., Gallego, G., García, N., Scaramuzza, D.: Event-based vision meets deep learning on steering prediction for self-driving cars. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5419–5427 (2018)
24. Nah, S., Hyun Kim, T., Mu Lee, K.: Deep multi-scale convolutional neural network for dynamic scene deblurring. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3883–3891 (2017)
25. Pan, L., Scheerlinck, C., Yu, X., Hartley, R., Liu, M., Dai, Y.: Bringing a blurry frame alive at high frame-rate with an event camera. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6820–6829 (2019)
26. Paredes-Vallés, F., Scheper, K.Y., De Croon, G.C.: Unsupervised learning of a hierarchical spiking neural network for optical flow estimation: From events to global motion perception. IEEE transactions on pattern analysis and machine intelligence **42**(8), 2051–2064 (2019)
27. Rebecq, H., Gehrig, D., Scaramuzza, D.: Esim: an open event camera simulator. In: Conference on robot learning. pp. 969–982. PMLR (2018)
28. Ren, D., Shang, W., Yang, Y., Zuo, W.: Aggregating long-term sharp features via hybrid transformers for video deblurring. arXiv preprint arXiv:2309.07054 (2023)
29. Ren, S., Du, Y., Lv, J., Han, G., He, S.: Learning from the master: Distilling cross-modal advanced knowledge for lip reading. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13325–13333 (2021)
30. Shang, W., Ren, D., Zou, D., Ren, J.S., Luo, P., Zuo, W.: Bringing events into video deblurring with non-consecutively blurry frames. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4531–4540 (2021)
31. Song, C., Bajaj, C., Huang, Q.: Deblursr: Event-based motion deblurring under the spiking representation. arXiv preprint arXiv:2303.08977 (2023)

32. Su, S., Delbracio, M., Wang, J., Sapiro, G., Heidrich, W., Wang, O.: Deep video deblurring for hand-held cameras. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1279–1288 (2017)
33. Sun, L., Sakaridis, C., Liang, J., Jiang, Q., Yang, K., Sun, P., Ye, Y., Wang, K., Gool, L.V.: Event-based fusion for motion deblurring with cross-modal attention. In: Proceedings of the European conference on computer vision (ECCV). pp. 412–428. Springer (2022)
34. Sun, L., Cho, S., Wang, J., Hays, J.: Edge-based blur kernel estimation using patch priors. In: IEEE international conference on computational photography (ICCP). pp. 1–8. IEEE (2013)
35. Tao, X., Gao, H., Shen, X., Wang, J., Jia, J.: Scale-recurrent network for deep image deblurring. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8174–8182 (2018)
36. Tian, Y., Krishnan, D., Isola, P.: Contrastive multiview coding. In: Proceedings of the European conference on computer vision (ECCV). pp. 776–794. Springer (2020)
37. Tulyakov, S., Gehrig, D., Georgoulis, S., Erbach, J., Gehrig, M., Li, Y., Scaramuzza, D.: Time lens: Event-based video frame interpolation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16155–16164 (2021)
38. Vitoria, P., Georgoulis, S., Tulyakov, S., Bochicchio, A., Erbach, J., Li, Y.: Event-based image deblurring with dynamic motion awareness. In: Proceedings of the European conference on computer vision (ECCV). pp. 95–112. Springer (2022)
39. Wang, B., He, J., Yu, L., Xia, G.S., Yang, W.: Event enhanced high-quality image recovery. In: Proceedings of the European conference on computer vision (ECCV). pp. 155–171. Springer (2020)
40. Wang, L., Ho, Y.S., Yoon, K.J., et al.: Event-based high dynamic range image and very high frame rate video generation using conditional generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10081–10090 (2019)
41. Weikersdorfer, D., Conradt, J.: Event-based particle filtering for robot self-localization. In: 2012 IEEE International Conference on Robotics and Biomimetics (ROBIO). pp. 866–870. IEEE (2012)
42. Xia, W., Li, X., Deng, A., Xiong, H., Dou, D., Hu, D.: Robust cross-modal knowledge distillation for unconstrained videos. arXiv preprint arXiv:2304.07775 (2023)
43. Xu, F., Yu, L., Wang, B., Yang, W., Xia, G.S., Jia, X., Qiao, Z., Liu, J.: Motion deblurring with real events. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2583–2592 (2021)
44. Xu, L., Jia, J.: Two-phase kernel estimation for robust motion deblurring. In: Proceedings of the European conference on computer vision (ECCV). pp. 157–170. Springer (2010)
45. Yang, D., Yamac, M.: Motion aware double attention network for dynamic scene deblurring. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1113–1123 (2022)
46. Yuan, L., Sun, J., Quan, L., Shum, H.Y.: Image deblurring with blurred/noisy image pairs. In: ACM SIGGRAPH 2007 papers, pp. 1–es (2007)
47. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H.: Restormer: Efficient transformer for high-resolution image restoration. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5728–5739 (2022)

48. Zhang, S., Zhen, A., Stevenson, R.L.: Deep motion blur removal using noisy/blurry image pairs. Journal of Electronic Imaging **30**(3), 033022–033022 (2021)
49. Zhang, X., Yu, L.: Unifying motion deblurring and frame interpolation with events. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17765–17774 (2022)
50. Zhang, X., Yu, L., Yang, W., Liu, J., Xia, G.S.: Generalizing event-based motion deblurring in real-world scenarios. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10734–10744 (2023)
51. Zhou, Y., Gallego, G., Rebecq, H., Kneip, L., Li, H., Scaramuzza, D.: Semi-dense 3d reconstruction with a stereo event camera. In: Proceedings of the European conference on computer vision (ECCV). pp. 235–251 (2018)