Event-Based Fusion for Motion Deblurring with Cross-modal Attention

Lei Sun[1,2] Christos Sakaridis[2] Jingyun Liang[2] Qi Jiang[1] Kailun Yang[3]
Peng Sun[1] Yaozu Ye[1] Kaiwei Wang[1] Luc Van Gool[2,4]

[1]Zhejiang University [2]ETH Zürich [3]KIT [4]KU Leuven

Abstract. Traditional frame-based cameras inevitably suffer from motion blur due to long exposure times. As a kind of bio-inspired camera, the event camera records the intensity changes in an asynchronous way with high temporal resolution, providing valid image degradation information within the exposure time. In this paper, we rethink the event-based image deblurring problem and unfold it into an end-to-end two-stage image restoration network. To effectively fuse event and image features, we design an event-image cross-modal attention module applied at multiple levels of our network, which allows to focus on relevant features from the event branch and filter out noise. We also introduce a novel symmetric cumulative event representation specifically for image deblurring as well as an event mask gated connection between the two stages of our network which helps avoid information loss. At the dataset level, to foster event-based motion deblurring and to facilitate evaluation on challenging real-world images, we introduce the Real Event Blur (REBlur) dataset, captured with an event camera in an illumination-controlled optical laboratory. Our Event Fusion Network (EFNet) sets the new state of the art in motion deblurring, surpassing both the prior best-performing image-based method and all event-based methods with public implementations on the GoPro dataset (by up to 2.47dB) and on our REBlur dataset, even in extreme blurry conditions. The code and our REBlur dataset are available at https://ahupujr.github.io/EFNet/

1 Introduction

Motion blur often occurs in images due to camera shake or object motion during the exposure time. The goal of deblurring is to recover a sharp image with clear edge structures and texture details from the blurry image. This is a highly ill-posed problem because of the infinitely many feasible solutions [2,10,53]. Traditional methods explicitly utilize natural image priors and various constraints [2,11,17,18,22,23,48]. To better generalize when addressing the deblurring problem, modern learning-based methods choose to train Convolutional Neural Networks (CNNs) on large-scale data to learn the implicit relationships between blurry and sharp images [13,27,40,41,52]. Despite their high performance on existing public datasets, these learning-based methods often fail when

facing extreme or real-world blur. Their performance heavily relies on the quality and scale of the training data, which creates the need for a more general and reliable deblurring method.

Event cameras [5,12,30,37] are bio-inspired asynchronous sensors with high temporal resolution (in the order of μs) and they operate well in environments with high dynamic range. Different from traditional frame-based cameras, event cameras capture the intensity change of each pixel (i.e. *event* information) independently, if the change surpasses a threshold. Event cameras encode the intensity change information within the exposure time of the image frame into an event stream, making it possible to deblur an image frame with events [28]. However, because of sensor noise and uncertainty in the aforementioned threshold, it is difficult to use a physical model to deblur images based solely on events. Thus, some methods [15,24,35] utilize CNNs to deal with noise corruption and threshold uncertainty. Nevertheless, these methods only achieve slight performance gains compared to image-only methods, due to rather ineffective event representations and fusion mechanisms between events and images.

In this paper, we first revisit the mechanism of motion blur and how event information is utilized in image reconstruction. To deal with the inherent defect of the event-based motion deblurring equation, we propose EFNet, an Event Fusion Network for image deblurring which effectively combines information from event and frame-based cameras for image deblurring. Motivated by the physical model of event-based image deblurring [28], we design a symmetric cumulative event representation (SCER) specifically for deblurring and formulate our network based on a two-stage image restoration model. Each stage of the model has a U-Net-like architecture [33]. The first stage consists of two branches, an image branch and an event branch, the features of which are fused at multiple levels. In order to perform the fusion of the two modalities, we propose an Event-Image Cross-modal Attention (EICA) fusion module, which allows to attend to the event features that are relevant for deblurring via a channel-level attention mechanism. To the best of our knowledge, this is the first time that a multi-head attention mechanism is applied to event-based image deblurring. We also enable information exchange between the two stages of our network by applying Event Mask Gated Connections (EMGC), which selectively transfer feature maps from the encoder and decoder of the first stage to the second stage. A detailed ablation study shows the effectiveness of our novel fusion module using cross-modal attention, our gated connection module and our multi-level middle fusion design. Additionally, we record a real-world event blur dataset named Real Event Blur (REBlur) in an optical laboratory with stable illumination and a high-precision electronically controlled slide-rail which allows various types of motion. We conduct extensive comparisons against state-of-the-art deblurring methods on the GoPro dataset [27] with synthetic events and on REBlur with real events and demonstrate the superiority of our event-based image deblurring method.

In summary, we make the following main contributions:

- We design a novel event-image fusion module which applies cross-modal channel-wise attention to adaptively fuse event features with image features, and incorporate it at multiple levels of a novel end-to-end deblurring network.
- We introduce a novel symmetric cumulative event voxel representation for deblurring, which is inspired by the physical model that connects blurry image formation and event generation.
- We present REBlur, a real-world dataset consisting of tuples of blurry images, sharp images and event streams from an event camera, which provides a challenging evaluation setting for deblurring methods.
- Our deblurring network, equipped with our proposed modules and event representation, sets the new state of the art for image deblurring on the GoPro dataset and our REBlur dataset.

2 Related Work

Image deblurring. Traditional approaches often formulate deblurring as an optimization problem [11,17,18,22,23,48]. Recently, with the success of deep learning, image deblurring has achieved impressive performance thanks to the usage of CNNs. CNN-based methods directly map the blurry image to the latent sharp image. Several novel components and techniques have been proposed, such as attention modules [39,42], multi-scale fusion [27,41], multi-stage networks [8,50], and coarse-to-fine strategies [9], improving the accuracy and robustness of deblurring. Despite the benefits they have shown for deblurring, all aforementioned deep networks operate solely on images, a modality which does not explicitly capture *motion* and thus inherently limits performance when facing real-world blurry images, especially in extreme conditions.

Event-based deblurring. Recently, events have been used for motion deblurring, due to the strong connection they possess with motion information. Pan *et al*. [28] proposed an Event Double Integral (EDI) deblurring model using the double integral of event data. They established a mathematical event-based model mapping blurry frames to sharp frames, which is a seminal approach to deblurring with events. However, limited by the sampling mechanism of event cameras, this method often introduces strong, accumulated noise. Jiang *et al*. [15] extracted motion information and sharp edges from events to assist deblurring. However, their early fusion approach, which merely concatenates events into the main branch of the network, does not account for higher-level interactions between frames and events. Lin *et al*. [24] fused events with the image via dynamic filters from STFAN [54]. In addition, Shang *et al*. [35] fused event information into a weight matrix that can be applied to any state-of-the-art network. To sum up, most of the above event-based learning methods did not use event information effectively, achieving only minor improvements compared to image-only methods on standard benchmarks.

Event representation. Different from synchronous signals such as images from frame-based cameras, events are asynchronous and sparse. A key point in how to extract information from events effectively is the representation of the events.

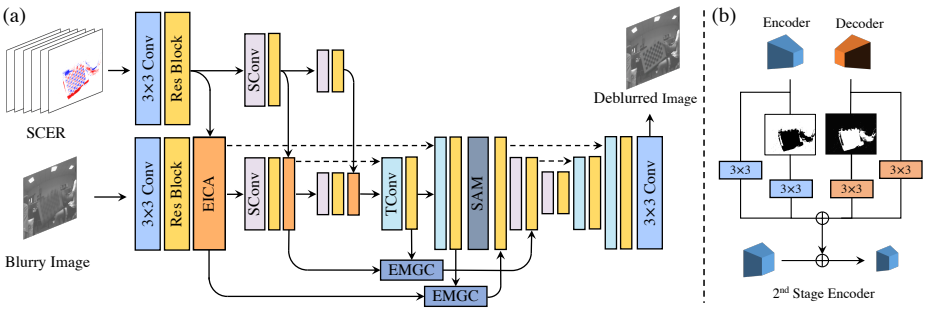

Fig. 1: (a): **The architecture of our Event Fusion Network (EFNet).** EFNet consists of two UNet-like backbones [33] and an event extraction branch. After each residual convolution block ("Res Block"), feature maps from the event branch and the image branch are fused. The second UNet backbone refines the deblurred image further. "SCER": symmetric cumulative event representation, "EICA": event-image cross-modal attention, "SConv": 4×4 strided convolution with stride 2, "TConv": 2×2 transposed convolution with stride 2, "SAM": supervision attention module [50]. (b): **The Event Mask Gated Connection module (EMGC)** transfers features across stages guided by an event mask.

Event representation is an application-dependent problem and different tasks admit different solutions. The event-by-event method is suitable for spiking neural networks owing to its asynchronous architecture [29,34,46]. A Time Surface, which is a 2D map that stores the time value deriving from the timestamp of the last event, has proved suitable for event-based classification [1,21,36]. Some modern learning-based methods convert events to a 2D frame by counting events or accumulating polarities [25,26,35]. This approach is compatible with conventional computer vision tasks but loses temporal information. 3D space-time histograms of events, also called voxel grids, preserve the temporal information of events better by accumulating event polarity on a voxel [4,55]. For image deblurring, most works utilized 2D event-image pairs [35] or borrowed 3D voxel grids like Stacking Based on Time (SBT) from image reconstruction [43]. However, there still is no event representation specifically designed for motion deblurring.

3 Method

We first introduce the mathematical model for the formation of blurry images from sharp images that involves events in Sec. 3.1. Based on this model, we pose the event-based deblurring problem as a deblurring-denoising problem and base the high-level design of our network architecture on this formulation, as explained in Sec. 3.2. We present our symmetric cumulative representation for events, which constitutes a 3D voxel grid in which the temporal dimension is discretized, in Sec. 3.3. This event representation is provided as input together with the blurry image to our two-stage network. We then detail the two main

novel components of our network: our novel event-image cross-modal attention fusion mechanism (Sec. 3.4), which adaptively fuses feature channels associated with events and images, and our event mask gated connection module between the two stages of our network (Sec. 3.5), which helps selectively forward to the second stage the features at sharp regions of the input from the encoder and the features at blurry regions from the decoder of the first stage.

3.1 Problem Formulation

For an event camera, the i-th event e_i is represented as a tuple $e_i = (x_i, y_i, t_i, p_i)$, where x_i, y_i and t_i represent the pixel coordinates and the timestamp of the event respectively, and $p_i \in \{-1, +1\}$ is the polarity of the event [5,30]. An event is triggered at time t only when the change in pixel intensity $\mathcal{I}$ surpasses the threshold compared to the pixel intensity at the time of the last trigger. This is formulated as

$$p_i = \begin{cases} +1, \text{if} \log\left(\frac{\mathcal{I}_t(x_i,y_i)}{\mathcal{I}_{t-\Delta t}(x_i,y_i)}\right) > c, \\ -1, \text{if} \log\left(\frac{\mathcal{I}_t(x_i,y_i)}{\mathcal{I}_{t-\Delta t}(x_i,y_i)}\right) < -c, \end{cases} \tag{1}$$

where c is the contrast threshold of intensity change, which may differ across the sensor plane.

Given the intensity of a latent sharp image $\mathbf{L}$, according to [28], the corresponding blurred image $\mathbf{B}$ can be derived by the Event-based Double Integral (EDI) model:

$$\begin{aligned} \mathbf{B} &= \frac{1}{T}\int_{f-T/2}^{f+T/2} \mathbf{L}(t)dt \\ &= \frac{\mathbf{L}(f)}{T}\int_{f-T/2}^{f+T/2} \exp\left(c\int_f^t p(s)ds\right) dt, \end{aligned} \tag{2}$$

where f is the middle point of the exposure time T, $p(s)$ is the polarity component of the event stream and $\mathbf{L}(f)$ is the latent sharp image corresponding to the blurred image $\mathbf{B}$. The discretized version of (2) can be expressed as

$$\mathbf{B} = \frac{\mathbf{L}(N)}{2N+1}\sum_{i=0}^{2N} \exp\left(c\,\text{sgn}(i-N)\sum_{j:\ m\le t_j\le M} p_j\delta_{x_j y_j}\right), \tag{3}$$

where sgn is the signum function, $m = \min\{f + T/2(i/N - 1), f\}$, $M = \max\{f + T/2(i/N - 1), f\}$ and δ is the Kronecker delta, defined as

$$\delta_{kl}(m,n) = \begin{cases} 1, & \text{if } k = m \text{ and } l = n, \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

In (3), we partition the exposure time T into $2N$ equal intervals. Rearranging (3) yields:

$$\mathbf{L}(N) = \frac{(2N+1)\mathbf{B}}{\sum_{i=0}^{2N}\exp\left(c\,\text{sgn}(i-N)\sum_{j:\ m\le t_j\le M} p_j\delta_{x_j y_j}\right)}. \tag{5}$$

3.2 General Architecture of EFNet

The formulation in (5) indicates that the latent sharp image can be derived from the blurred image combined with the set of events $\mathcal{E} = \{e_i = (x_i, y_i, t_i, p_i) : f - T/2 \leq t_i \leq f + T/2\}$ (*i.e.*, all the events which are triggered within the exposure time), when events in this set are accumulated over time. We propose to learn this relation with a deep neural network, named Event Fusion Network (EFNet), which admits as inputs the blurred image and the events and maps them to the sharp image. The generic form of the learned mapping is

$$\mathbf{L}_{\text{initial}} = f_3\left(f_1(\mathbf{B}; \Theta_1), f_2(\mathcal{E}; \Theta_2); \Theta_3\right), \tag{6}$$

where the blurred image and the events are mapped individually to intermediate representations via f_1 and f_2 respectively and these intermediate representations are afterwards passed to a joint mapping f_3. Θ_1, Θ_2 and Θ_3 denote the respective parameters of the three mappings. The main challenges we need to address given this generic formulation of our model are (i) how to represent the set of events $\mathcal{E}$ in a suitable way for inputting it to the network, and (ii) how and when to fuse the intermediate representations that are generated for the blurred image by f_1 and for the events by f_2, *i.e.*, how to design f_3. We address the issue of how to represent the events in Sec. 3.3 and how to perform fusion in Sec. 3.4.

(3) is the ideal formulation for event-based motion image deblurring. However, in real-world settings, three factors make it impossible to restore the image simply based on this equation:

- Instead of being strictly equal to a fixed value, the values of threshold c for a given event camera are neither constant in time nor across the image [38,45].
- Intensity changes that are lower than the threshold c do not trigger an event.
- Spurious events occur over the entire image.

Most of the restoration errors come from the first two factors, which cause degradation of the restored image in regions with events. We denote these regions as R_e. Taking the above factors into account, we design our network so that it includes a final mapping of the initial deblurred image $\mathbf{L}_{\text{initial}}$ to a denoised version of it, which can correct potential errors in the values of pixels inside R_e:

$$\mathbf{L}_{\text{final}} = f_4(\mathbf{L}_{\text{initial}}; \Theta_4). \tag{7}$$

Two-stage backbone. We design EFNet as a two-stage network to progressively restore sharp images from blurred images and event streams, where the first and second stage implement the generic mappings in (6) and (7) respectively. The detailed architecture of EFNet is illustrated in Fig. 1. Both stages of EFNet have an encoder-decoder structure, based on the UNet [33] architecture. Each stage consists of two down-sampling and two up-sampling layers. Between the encoder and decoder, we add a skip connection with 3×3 convolution. The residual convolution block in UNet consists of two 3×3 convolution layers and leaky ReLUs with a 1×1 convolution shortcut. Recently, the Supervised Attention Module (SAM) in multi-stage methods proved successful in transferring features between different sub-networks [8,50]. Thus, we use SAM to connect the two

stages of EFNet. In the first stage, we fuse features from the event branch and the image branch at multiple levels using a novel cross-modal attention-based block. Between the two stages, we design an Event Mask Gated Connection module to boost feature aggregation with blurring priors from events. The details of the two aforementioned components of EFNet are given in Sec. 3.4 and 3.5.

3.3 Symmetric Cumulative Event Representation

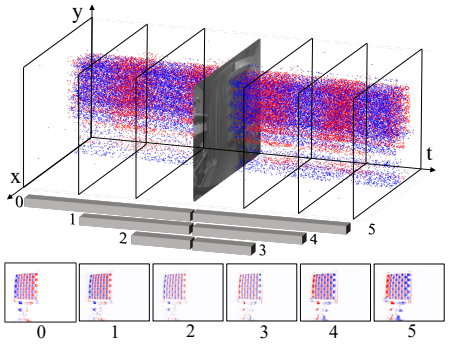

Fig. 2: **Symmetric Cumulative Event Representation (SCER).** Red/blue dots denote events with positive/negative polarity respectively.

To feed the asynchronous events to our network, we design a representation specifically suited for deblurring. In (3), the accumulation of polarities via the inner sum on the right-hand side indicates the relative intensity changes between the target latent sharp image $\mathbf{L}(N)$ and each of the rest of latent sharp images in the exposure time. The accumulation via the outer sum on the right-hand side represents the sum of all latent sharp images. Based on this relationship, we propose the Symmetric Cumulative Event Representation (SCER). As Fig. 2 shows, the exposure time T of the blurry image is divided equally into $2N$ intervals. Assuming $2N+1$ latent sharp images in T, the polarity accumulation from the central target latent image $\mathbf{L}(N)$ to a single latent image turns into a 2D tensor with dimensions (H, W):

$$\mathbf{SCER}_i = \operatorname{sgn}(i-N) \sum_{j:\, m \le t_j \le M} p_j \delta_{x_j y_j}. \tag{8}$$

For $i = N$, $\mathbf{SCER}_N = 0$, so we discard this tensor. The remaining $2N$ tensors are concatenated together, forming a tensor which indicates intensity changes between the central latent sharp image $\mathbf{L}(N)$ and each of the $2N$ other latent images. In this way, $\mathbf{SCER} \in \mathbb{R}^{H \times W \times 2N}$ includes all the relative intensity values corresponding to the center latent sharp frame and it becomes suitable for feature extraction with our image deblurring model. As the accumulation limits change, our SCER also contains both information about the area in which blur occurs (channel 0 and channel $2N-1$) and information about sharp edges (channel $N-1$ and channel N).

Our method discretizes T into $2N$ parts, quantizing temporal information of events within the time interval $\frac{T}{2N}$. However, SCER still holds temporal information, as the endpoints of the time interval in which events are accumulated is different across channels. The larger N is, the less temporal information is lost. In our implementation, we fix $N = 3$.

3.4 Event-Image Cross-modal Attention Fusion

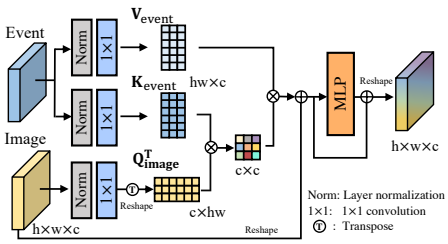

Fig. 3: **The Event-Image Cross-modal Attention fusion module.** The size of the attention map is $c \times c$.

Jointly extracting and fusing information from event streams and images is the key to event-based deblurring. Previous work [15,24] simply multiplies or concatenates low-resolution feature maps from the two modalities, but this fusion approach cannot model the long-range relations between events and images. Other methods estimate optical flow with events and use it for deblurring [35], but this estimation introduces errors.

We instead include a novel cross-modal attention block at multiple levels of EFNet. Contrary to self-attention blocks, in which the queries ($\mathbf{Q}$), keys ($\mathbf{K}$) and values ($\mathbf{V}$) all come from the same branch of the network, our Event-Image Cross-modal Attention (EICA) block admits as inputs the queries $\mathbf{Q}_{\text{image}}$ from the image branch and the keys $\mathbf{K}_{\text{event}}$ and values $\mathbf{V}_{\text{event}}$ from the event branch, as shown in Fig. 3. The input features from the two branches are fed to normalization and 1×1 convolution layers, where the latter have c output channels. We then apply cross-modal attention between vectorized features from the two modalities via

$$\text{Attention}(\mathbf{Q}_{\text{image}}, \mathbf{K}_{\text{event}}, \mathbf{V}_{\text{event}}) = \mathbf{V}_{\text{event}}\, \text{softmax}\left(\frac{\mathbf{Q}^T_{\text{image}}\mathbf{K}_{\text{event}}}{\sqrt{d_k}}\right). \quad (9)$$

We introduce the 1×1 convolution layer to reduce the spatial complexity of the above attention operation. In particular, c is chosen to be much smaller than hw, where h and w are the height and width of the input feature maps, and the soft indexing of $\mathbf{K}_{\text{event}}$ by $\mathbf{Q}_{\text{image}}$ is performed at the *channel* dimension instead of the spatial dimensions. Thus, the resulting soft attention map from (9) is $c \times c$ instead of $hw \times hw$, reducing the spatial complexity from $\mathcal{O}(h^2w^2)$ to $\mathcal{O}(c^2)$ and making the operation feasible even for features with high spatial resolution, as in our case. Finally, the output of the attention operation is added to the input image features and this sum is passed to a multi-layer perceptron consisting of two fully connected layers with a Gaussian Error Linear Unit (GELU) [14] in between. We use the EICA module at multiple levels of EFNEt to fuse event information aggregated across receptive fields of varying size.

3.5 Event Mask Gated Connection Module

Previous work [31] predicts a mask indicating which areas of an image are severely distorted, but this mask is not completely accurate. Apart from information about intensity changes, event data also contain spatial information

about the blurred regions of the input image. Typically, regions in which events occur are more severely degraded in the blurry image. Motivated by this observation, we introduce an Event Mask Gated Connection (EMGC) between the two stages of our network to exploit the spatial information about blurred regions.

In particular, we binarize the sum of the first and last channel of SCER to obtain a binary event mask, in which pixels where an event has occurred are set to 0 and the rest are set to 1. As illustrated in Fig. 1(b), EMGC masks out the feature maps of the encoder at regions where the event mask is 0, which are expected to be more blurry, and masks out the feature maps of the decoder at regions where the event mask is 1 (using the complement of the event mask), which are expected to be less blurry. A skip connection is added beside the mask operation. Feature maps with less artifacts in the encoder and better restored feature maps are combined through the event mask gate. Besides, EMGC eases the flow of information through the network, as it creates a shortcut through which features can be transferred directly from the first to the second stage.

4 REBlur Dataset

Most event-based motion deblurring methods [6,15,24,35,47] train models on blurred image datasets, such as GoPro [27], with synthetic events from ESIM [32]. Although the contrast threshold c in the event simulator varies across pixels as in reality, a domain gap between synthetic and real events still exists because of the background activity noise, dark current noise, and false negatives in refractory period [3,38,45]. Recently, Jiang *et al.* [15] proposed BlurDVS by capturing an image plus events with slow motion, and then synthesizing motion blur by averaging multiple nearby frames. However, motion blur in the ground-truth images is inevitable in this setting and fast motion causes different events from slow motion because of the false negatives in the refractory period of event cameras [3,47]. Thus, a large-scale real-world dataset with blurry images, reliable corresponding events, and ground-truth sharp images is missing.

We present a new event-based dataset for deblurring, Real Event Blur (REBlur), to provide ground truth for blurry images in a two-shot way. To collect REBlur, we built an image collection system in a high-precision optical laboratory with very stable illumination. We fixed an Insightness Seem 1 event camera and a Dynamic and Active Pixel Vision Sensor (DAVIS) to the optical table, outputting time-aligned event streams and 260×360 gray images. To obtain blurry-sharp image pairs under high-speed motion, we also fixed a high-precision electronic-controlled slide-rail system to the optical table. In the first shot, we captured images with motion blur for the pattern on the slide-rail and corresponding event streams. In the second shot, according to the timestamp t_s of the blurry images, we selected events within the time range $[t_s - 125\mu s, t_s + 125\mu s]$ and visualized these events in the preview of the sharp image capture program. Referring to the edge information from high-temporal-resolution events, we could relocate the slide-rail to the coordinate corresponding to the timestamp t_s by an electronic-controlled stepping motor and then capture the latent sharp image.

Table 1: **Comparison of motion deblurring methods on GoPro [27].** †: event-based methods, SRN+ and HINet+: event-enhanced versions of SRN [41] and HINet [8] using our SCER, percentages in brackets: relative reduction in error with EFNet.

Method	PSNR ↑	SSIM ↑
DeblurGAN [19]	28.70 (54.1%)	0.858 (80.3%)
BHA† [28]	29.06 (52.1%)	0.940 (53.3%)
Nah *et al.* [27]	29.08 (52.0%)	0.914 (67.4%)
DeblurGAN-v2 [20]	29.55 (49.4%)	0.934 (57.6%)
SRN [41]	30.26 (45.1%)	0.934 (57.6%)
SRN+† [41]	31.02 (40.0%)	0.936 (56.3%)
DMPHN [51]	31.20 (38.8%)	0.940 (53.3%)
D^2Nets† [35]	31.60 (35.9%)	0.940 (53.3%)
LEMD† [15]	31.79 (34.5%)	0.949 (45.1%)
Suin *et al.* [39]	31.85 (34.0%)	0.948 (46.2%)
SPAIR[31]	32.06 (32.4%)	0.953 (40.4%)
MPRNet [50]	32.66 (27.6%)	0.959 (31.7%)
HINet [8]	32.71 (27.1%)	0.959 (31.7%)
Restormer [49]	32.92 (25.4%)	0.961 (28.2%)
ERDNet† [6]	32.99 (24.8%)	0.935 (56.9%)
HINet+† [8]	33.69 (18.4%)	0.961 (28.2%)
NAFNet [7]	33.69 (18.4%)	0.967 (15.2%)
EFNet (Ours)†	**35.46**	**0.972**

REBlur includes 12 kinds of linear and nonlinear motions for 3 different moving patterns and for the camera itself. It consists of 36 sequences and 1469 blurry-sharp image pairs with associated events, where 486 pairs are used for training and 983 for testing. The supplement includes more details on REBlur.

5 Experiments

5.1 Datasets and Settings

GoPro dataset. We use the GoPro dataset [27], which is widely used in motion deblurring, for training and evaluation. It consists of 3214 pairs of blurry and sharp images with a resolution of 1280×720 and the blurred images are produced by averaging several high-speed sharp images. We use 2103 pairs for training and 1111 pairs for testing, following standard practice [27]. We use ESIM [32], an open-source event camera simulator, to generate simulated event data for GoPro. To make the results more realistic, we set the contrast threshold c randomly for each pixel, following a Gaussian distribution $N(\mu = 0.2, \sigma = 0.03)$.

REBlur dataset. In order to close the gap between simulated events and real events, before evaluating models that are trained on GoPro on REBlur, we fine-tune them on the training set of REBlur. We then evaluate the fine-tuned models on the test set of REBlur. More details on this fine-tuning follow.

Implementation details. Our network requires no pre-training. We train it on 256×256 crops of full images from GoPro. Full details about our network configuration (numbers of channels, kernel sizes *etc.*) are given in the supplement.

Fig. 4: **Visual comparison on GoPro.** SRN+ and HINet+: event-enhanced versions of SRN and HINet using SCER. Compared to image- and event-based state-of-the-art methods, our method restores fine texture and structures better.

For data augmentation, horizontal and vertical flipping, random noise and hot pixels in event voxels [38] are applied. We use Adam [16] with an initial learning rate of 2×10^{-4}, and the cosine learning rate strategy with a minimum learning rate of 10^{-7}. The model is trained with a batch size of 8 for 300k iterations on 4 NVIDIA Titan RTX GPUs, which take 41 hours. Fine-tuning on REBlur involves 600 iterations and a single Titan RTX, the initial learning rate is 2×10^{-5} and other configurations are kept the same as for GoPro. We use the same training and fine-tuning settings for our method and other methods for a fair comparison.
Evaluation protocol. All quantitative comparisons are performed using PSNR and SSIM [44]. Apart from these, we also report the relative reduction in error with the best-performing model compared to each method. This is done by first converting PSNR to RMSE (RMSE $\propto \sqrt{10^{-\mathrm{PSNR}/10}}$) and SSIM to DSSIM (DSSIM $= (1 - \mathrm{SSIM})/2$) and then computing the relative reduction in error.

5.2 Comparisons with State-of-the-Art Methods

We compare our method with state-of-the-art image-only and event-based deblurring methods on GoPro and REBlur. Since most learning-based methods using events do not have publicly available implementations, in the qualitative comparison part, apart from BHA [28], we compare our method with SRN [41] and HINet [8], the latter being the current best model on the GoPro benchmark. To have a fair comparison, we also include event-enhanced versions of these two models by concatenating event voxel grids and images in the input.
GoPro. We report deblurring results in Table 1. Compared to the best existing image-based [8] and event-based [6] methods, EFNet achieves 2.75 dB and

Table 2: **Comparison of motion deblurring methods on REBlur.** Read as Table 1.

Method	PSNR ↑	SSIM ↑	Params (M) ↓
SRN [41]	35.10 (29.4%)	0.961 (35.9%)	10.25
NAFNet [7]	35.48 (26.2%)	0.962 (34.2%)	67.89
Restormer [49]	35.50 (26.0%)	0.959 (39.0%)	26.13
HINet [8]	35.58 (25.4%)	0.965 (28.6%)	88.67
BHA† [28]	36.52 (16.8%)	0.964 (30.6%)	0.51
SRN+† [41]	36.87 (13.4%)	0.970 (16.7%)	10.43
HINet+† [8]	37.68 (4.9%)	0.973 (7.4%)	88.85
EFNet (Ours)†	**38.12**	**0.975**	8.47

2.47dB improvement in PSNR and 0.013 and 0.037 improvement in SSIM resp., with a low parameter count of 8.47M. Despite utilizing an extra modality, other learning-based methods using events such as D^{2}Nets, LEMD, and ERDNet do not improve significantly upon image-only methods. EFNet sets the new state of the art in image deblurring, showing that our principled architecture with attentive fusion leverages event information more effectively for this task. By simply including our SCER to HINet [8], the resulting enhanced version of it also surpasses the best previous event-based method [6]. We show qualitative results on GoPro in Fig. 4. Results of image-based methods are more blurry, losing sharp edge information. BHA [28] restores edges better but suffers from noise around them because of the factors described in Sec. 3.1. Learning-based methods using events cannot fully exploit the motion information from events. By inputting the concatenation of SCER with the image to SRN+ and HINet+, they both achieve large improvements. However, results from SRN+ include artifacts and noise due to the absence of a second stage in the network that would refine the result. HINet+ introduces more artifacts, indicating that concatenating events and images in the input is not sufficient. Based on the physical model for event deblurring, EFNet achieves sharp and faithful results. Both dominant structures and details are restored well thanks to our attentive fusion at multiple levels.

REBlur. We report quantitative results on REBlur in Table 2. Our model outperforms all other methods in this challenging real-world setting. Fig. 5 depicts qualitative results from the test set and the additional set. Even the best image-based method, HINet, does not perform well on these cases of severe real-world motion blur. Event-based methods are more robust to such adverse conditions and less prone to overfitting on synthetic training data. Results from BHA are sharper, but accumulation noise still exists. Simply adding events with our SCER representation to the state-of-the-art image-based method [8] improves performance significantly because of the physical basis of SCER, but still leads to artifacts and ghost structures. EFNet restores both smooth texture and sharp edges, demonstrating the utility of our two-stage architecture and our cross-modal attention for fusion. Thanks to the selective feature connection via EMGC, EFNet restores blurry regions well while also maintaining the content of sharp regions. More results, including failure cases, are provided in the supplement.

Table 3: **Ablation study of various components of our method** on GoPro [27]. "Early": fusion by concatenation of event voxel grid and image, "Multi-level": fusion with our proposed architecture. SCER is used to represent events.

	Architecture	Events	Fusion type	EMGC	Fusion module	PSNR ↑	SSIM ↑
1	1-Stage	✗	n/a	n/a	n/a	29.06	0.936
2	1-Stage	✓	Multi-level	n/a	EICA	34.90	0.968
3	2-Stage	✗	n/a	n/a	n/a	32.15	0.954
4	2-Stage	✓	Early	✗	n/a	33.68	0.960
5	2-Stage	✓	Early	✓	n/a	33.79	0.961
6	2-Stage	✓	Multi-level	✓	Concat.	34.80	0.968
7	2-Stage	✓	Multi-level	✓	Multiply	34.86	0.968
8	2-Stage	✓	Multi-level	✓	Add	34.78	0.968
9	2-Stage	✓	Multi-level	✗	EICA	35.31	0.971
10	2-Stage	✓	Multi-level	✓	EICA	**35.46**	**0.972**

5.3 Ablation Study

Table 4: **Comparison between different event representations** on GoPro. "Stack": temporal event accumulation in a single channel.

Event representation	PSNR ↑	SSIM ↑
None (image-only)	32.15	0.954
Stack	31.90	0.950
SBT [43]	35.12	0.970
SCER (Ours)	**35.46**	**0.972**

We conduct two ablation studies on GoPro to analyze the contribution of different components of our network (Table 3) and our event representation (Table 4). First, our EICA fusion block fuses event and image features effectively, improving PSNR by 0.6 dB or more and SSIM by 0.4% compared to simple strategies for fusion such as multiplication or addition (rows 6–8 and 10 of Table 3). Second, introducing middle fusion at multiple levels and using simple strategies for fusion yields an improvement of ∼1 dB in PSNR and 0.7% in SSIM over early fusion (rows 5–8), evidencing the benefit of using multi-level fusion in our EFNet. Third, adding events as input to the network via early fusion of our SCER voxel grids with the images improves PSNR by 1.53 dB and SSIM by 0.6% (rows 3–4), showcasing the informativeness of the event modality regarding motion, which leads to better deblurring. Fourth, adding a second stage in our network for progressive restoration benefits deblurring significantly, both in the image-only case (rows 1 and 3) and in the case where our fully-fledged EFNet is used (rows 2 and 10). Fifth, connecting the two stages of EFNet with our EMGC improves the selective flow of information between the two stages, yielding an improvement of 0.15 dB (rows 9–10). Finally, all our contributions together yield a substantial improvement of 6.4 dB in PSNR and 3.6% in SSIM over the image-only one-stage baseline, setting the new state of the art in motion deblurring.

Event representation. Introducing events can improve performance due to the high temporal resolution of the event stream, which provides a vital signal for deblurring. Table 4 shows a comparison between SCER and other event

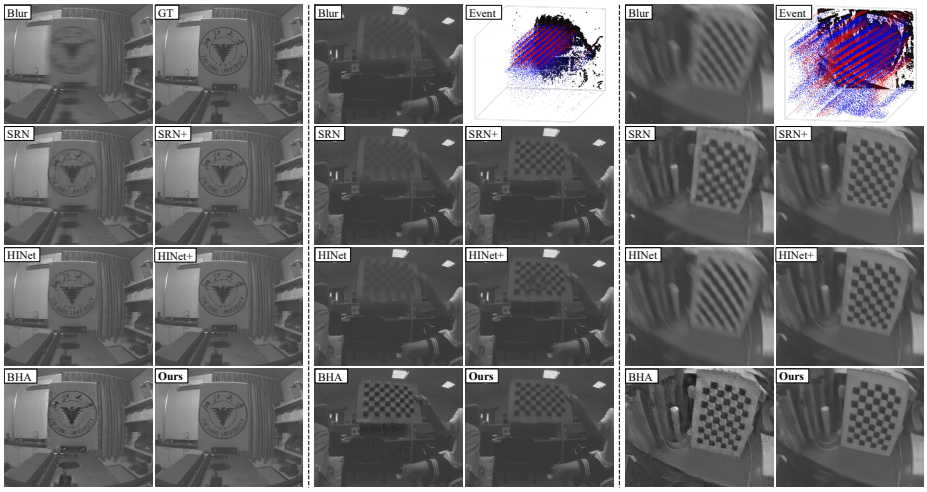

Fig. 5: **Visual comparison on the REBlur dataset.** The first two columns are from the test set of the REBlur dataset, and the rest are from the additional set, for which ground truth is not available. Our method shows superior performance in cases with severe blur both due to object motion and due to camera motion. Best viewed on a screen and zoomed in.

representations, including SBT [43], which accumulates polarities in fixed time intervals. We use the same number of intervals (6) for SBT and SCER for a fair comparison. Based explicitly on physics, SCER utilizes event information for image deblurring more effectively than SBT. Note that simply accumulating all events across the exposure time ("Stack") deteriorates the performance compared to not using events at all, which demonstrates that finding a suitable event representation for deblurring, such as SCER, is non-trivial.

6 Conclusion

In this work, we have looked into single-image motion deblurring from the perspective of event-based fusion. Based on the physical model which describes blurry image formation and event generation, we have introduced EFNet, an end-to-end motion deblurring network with attention-based event-image fusion applied at multiple levels of the network. In addition, we have proposed a novel event voxel representation for deblurring. We have captured a new real-world dataset, REBlur, including several cases of severe motion blur, which provides a challenging evaluation setting. EFNet significantly surpasses the prior state of the art in image deblurring, both on the GoPro dataset and on our new dataset.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (NSFC) under Grant No. 12174341, Sunny Optical Technology (group) co., Ltd, and the China Scholarship Council.

References

1. Ahad, M.A.R., Tan, J.K., Kim, H., Ishikawa, S.: Motion history image: its variants and applications. Machine Vision and Applications (2012) 4
2. Bahat, Y., Efrat, N., Irani, M.: Non-uniform blind deblurring by reblurring. In: ICCV (2017) 1
3. Baldwin, R., Almatrafi, M., Asari, V., Hirakawa, K.: Event probability mask (EPM) and event denoising convolutional neural network (EDnCNN) for neuromorphic cameras. In: CVPR (2020) 9
4. Bardow, P., Davison, A.J., Leutenegger, S.: Simultaneous optical flow and intensity estimation from an event camera. In: CVPR (2016) 4
5. Brandli, C., Berner, R., Yang, M., Liu, S.C., Delbruck, T.: A 240 × 180 130 dB 3 μs latency global shutter spatiotemporal vision sensor. IEEE Journal of Solid-State Circuits (2014) 2, 5
6. Chen, H., Teng, M., Shi, B., Wang, Y., Huang, T.: Learning to deblur and generate high frame rate video with an event camera. arXiv preprint arXiv:2003.00847 (2020) 9, 10, 11, 12
7. Chen, L., Chu, X., Zhang, X., Sun, J.: Simple baselines for image restoration. arXiv preprint arXiv:2204.04676 (2022) 10, 12
8. Chen, L., Lu, X., Zhang, J., Chu, X., Chen, C.: HINet: Half instance normalization network for image restoration. In: CVPRW (2021) 3, 6, 10, 11, 12
9. Cho, S.J., Ji, S.W., Hong, J.P., Jung, S.W., Ko, S.J.: Rethinking coarse-to-fine approach in single image deblurring. In: ICCV (2021) 3
10. Cho, S., Lee, S.: Fast motion deblurring. ACM Transactions on Graphics (2009) 1
11. Fergus, R., Singh, B., Hertzmann, A., Roweis, S.T., Freeman, W.T.: Removing camera shake from a single photograph. ACM Transactions on Graphics (2006) 1, 3
12. Gallego, G., Delbrück, T., Orchard, G., Bartolozzi, C., Taba, B., Censi, A., Leutenegger, S., Davison, A.J., Conradt, J., Daniilidis, K., Scaramuzza, D.: Event-based vision: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence (2022) 2
13. Gong, D., Yang, J., Liu, L., Zhang, Y., Reid, I., Shen, C., Van Den Hengel, A., Shi, Q.: From motion blur to motion flow: A deep learning solution for removing heterogeneous motion blur. In: CVPR (2017) 1
14. Hendrycks, D., Gimpel, K.: Gaussian error linear units (GELUs). arXiv preprint arXiv:1606.08415 (2016) 8
15. Jiang, Z., Zhang, Y., Zou, D., Ren, J., Lv, J., Liu, Y.: Learning event-based motion deblurring. In: CVPR (2020) 2, 3, 8, 9, 10
16. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (2015) 11
17. Kotera, J., Šroubek, F., Milanfar, P.: Blind deconvolution using alternating maximum a posteriori estimation with heavy-tailed priors. In: CAIP (2013) 1, 3
18. Krishnan, D., Tay, T., Fergus, R.: Blind deconvolution using a normalized sparsity measure. In: CVPR (2011) 1, 3
19. Kupyn, O., Budzan, V., Mykhailych, M., Mishkin, D., Matas, J.: DeblurGAN: Blind motion deblurring using conditional adversarial networks. In: CVPR (2018) 10
20. Kupyn, O., Martyniuk, T., Wu, J., Wang, Z.: DeblurGAN-v2: Deblurring (orders-of-magnitude) faster and better. In: ICCV (2019) 10

21. 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 (2017) 4
22. Levin, A., Weiss, Y., Durand, F., Freeman, W.T.: Understanding and evaluating blind deconvolution algorithms. In: CVPR (2009) 1, 3
23. Levin, A., Weiss, Y., Durand, F., Freeman, W.T.: Efficient marginal likelihood optimization in blind deconvolution. In: CVPR (2011) 1, 3
24. Lin, S., Zhang, J., Pan, J., Jiang, Z., Zou, D., Wang, Y., Chen, J., Ren, J.: Learning event-driven video deblurring and interpolation. In: ECCV (2020) 2, 3, 8, 9
25. Liu, M., Delbruck, T.: Adaptive time-slice block-matching optical flow algorithm for dynamic vision sensors. In: BMVC (2018) 4
26. 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: CVPR (2018) 4
27. Nah, S., Hyun Kim, T., Mu Lee, K.: Deep multi-scale convolutional neural network for dynamic scene deblurring. In: CVPR (2017) 1, 2, 3, 9, 10, 13
28. 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: CVPR (2019) 2, 3, 5, 10, 11, 12
29. Paredes-Vallés, F., Scheper, K.Y.W., de Croon, G.C.H.E.: 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 (2020) 4
30. Patrick, L., Posch, C., Delbruck, T.: A 128×128 120 dB 15μ s latency asynchronous temporal contrast vision sensor. IEEE Journal of Solid-State Circuits (2008) 2, 5
31. Purohit, K., Suin, M., Rajagopalan, A.N., Boddeti, V.N.: Spatially-adaptive image restoration using distortion-guided networks. In: ICCV (2021) 8, 10
32. Rebecq, H., Gehrig, D., Scaramuzza, D.: ESIM: an open event camera simulator. In: CoLR (2018) 9, 10
33. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: MICCAI (2015) 2, 4, 6
34. Scheerlinck, C., Barnes, N., Mahony, R.: Continuous-time intensity estimation using event cameras. In: ACCV (2018) 4
35. Shang, W., Ren, D., Zou, D., Ren, J.S., Luo, P., Zuo, W.: Bringing events into video deblurring with non-consecutively blurry frames. In: ICCV (2021) 2, 3, 4, 8, 9, 10
36. Sironi, A., Brambilla, M., Bourdis, N., Lagorce, X., Benosman, R.: HATS: Histograms of averaged time surfaces for robust event-based object classification. In: CVPR (2018) 4
37. Stoffregen, T., Kleeman, L.: Event cameras, contrast maximization and reward functions: An analysis. In: CVPR (2019) 2
38. Stoffregen, T., Scheerlinck, C., Scaramuzza, D., Drummond, T., Barnes, N., Kleeman, L., Mahony, R.: Reducing the sim-to-real gap for event cameras. In: ECCV (2020) 6, 9, 11
39. Suin, M., Purohit, K., Rajagopalan, A.N.: Spatially-attentive patch-hierarchical network for adaptive motion deblurring. In: CVPR (2020) 3, 10
40. Sun, J., Cao, W., Xu, Z., Ponce, J.: Learning a convolutional neural network for non-uniform motion blur removal. In: CVPR (2015) 1
41. Tao, X., Gao, H., Shen, X., Wang, J., Jia, J.: Scale-recurrent network for deep image deblurring. In: CVPR (2018) 1, 3, 10, 11, 12

42. Tsai, F.J., Peng, Y.T., Lin, Y.Y., Tsai, C.C., Lin, C.W.: BANet: Blur-aware attention networks for dynamic scene deblurring. arXiv preprint arXiv:2101.07518 (2021) 3
43. Wang, L., I., S.M.M., Ho, Y., Yoon, K.: Event-based high dynamic range image and very high frame rate video generation using conditional generative adversarial networks. In: CVPR (2019) 4, 13, 14
44. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing (2004) 11
45. Wang, Z., Ng, Y., van Goor, P., Mahony, R.: Event camera calibration of per-pixel biased contrast threshold. In: ACRA (2019) 6, 9
46. Weikersdorfer, D., Conradt, J.: Event-based particle filtering for robot self-localization. In: ROBIO (2012) 4
47. Xu, F., Yu, L., Wang, B., Yang, W., Xia, G.S., Jia, X., Qiao, Z., Liu, J.: Motion deblurring with real events. In: ICCV (2021) 9
48. Xu, L., Zheng, S., Jia, J.: Unnatural L0 sparse representation for natural image deblurring. In: CVPR (2013) 1, 3
49. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H.: Restormer: Efficient transformer for high-resolution image restoration. In: CVPR (2022) 10, 12
50. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H., Shao, L.: Multi-stage progressive image restoration. In: CVPR (2021) 3, 4, 6, 10, 11
51. Zhang, H., Dai, Y., Li, H., Koniusz, P.: Deep stacked hierarchical multi-patch network for image deblurring. In: CVPR (2019) 10
52. Zhang, J., Pan, J., Ren, J., Song, Y., Bao, L., Lau, R.W.H., Yang, M.H.: Dynamic scene deblurring using spatially variant recurrent neural networks. In: CVPR (2018) 1
53. Zhang, K., Zuo, W., Chen, Y., Meng, D., Zhang, L.: Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising. IEEE Transactions on Image Processing (2017) 1
54. Zhou, S., Zhang, J., Pan, J., Xie, H., Zuo, W., Ren, J.: Spatio-temporal filter adaptive network for video deblurring. In: ICCV (2019) 3
55. Zhu, A.Z., Yuan, L., Chaney, K., Daniilidis, K.: Unsupervised event-based learning of optical flow, depth, and egomotion. In: CVPR (2019) 4