CMTA: Cross-Modal Temporal Alignment for Event-guided Video Deblurring

Taewoo Kim*, Hoonhee Cho*, and Kuk-Jin Yoon

Korea Advanced Institute of Science and Technology
{intelpro, gnsgnsgml, kjyoon}@kaist.ac.kr

Abstract. Video deblurring aims to enhance the quality of restored results in motion-blurred videos by effectively gathering information from adjacent video frames to compensate for the insufficient data in a single blurred frame. However, when faced with consecutively severe motion blur situations, frame-based video deblurring methods often fail to find accurate temporal correspondence among neighboring video frames, leading to diminished performance. To address this limitation, we aim to solve the video deblurring task by leveraging an event camera with microsecond temporal resolution. To fully exploit the dense temporal resolution of the event camera, we propose two modules: 1) Intra-frame feature enhancement operates within the exposure time of a single blurred frame, iteratively enhancing cross-modality features in a recurrent manner to better utilize the rich temporal information of events, 2) Inter-frame temporal feature alignment gathers valuable long-range temporal information to target frames, aggregating sharp features leveraging the advantages of the events. In addition, we present a novel dataset composed of real-world blurred RGB videos, corresponding sharp videos, and event data. This dataset serves as a valuable resource for evaluating event-guided deblurring methods. We demonstrate that our proposed methods outperform state-of-the-art frame-based and event-based motion deblurring methods through extensive experiments conducted on both synthetic and real-world deblurring datasets. The code and dataset are available at https://github.com/intelpro/CMTA.

Keywords: Temporal alignment · Video Deblurring · Event cameras

1 Introduction

Motion blur is a common artifact caused by dynamic movements within a scene or camera motion during exposure. Motion deblurring, which aims to reverse the blurring process, presents significant challenges due to variations in blur intensity influenced by scene structure and depth. To achieve high-quality deblurring, video deblurring has emerged, leveraging information from neighboring frames instead of relying solely on a single blurred image. However, identifying temporal correspondence between blurred video frames becomes challenging with

* Equal contribution.

extreme motion blur, hindering the extraction of valuable information from adjacent frames and impeding performance improvement.

Event cameras [9], with their extremely low latency (on the order of microseconds), can offer high-quality guidance for motion deblurring due to their ability to capture high-temporal resolution of brightness change. To effectively utilize the advantages of the events, several event-guided motion deblurring works [15, 36, 50] have been introduced. While these works have typically explored cross-modal feature fusion methods across different modalities, there has been limited work on leveraging the abundant temporal information in videos.

To obtain high-quality results, we emphasize the event camera's temporal continuity, focusing on its interaction with video frames that exhibit long-term temporal dependencies. Unlike previous event-guided motion deblurring works [36,43,50], which relied on a single blurry image and corresponding events on its exposure time, we further design precise temporal feature alignment methods between neighboring video frames by leveraging the advantages of event data. Specifically, we propose novel modules from two perspectives: intra-frame (interaction between events and frames within the exposure time) and inter-frame (interaction between different frames) perspectives.

From an intra-frame perspective within exposure time, we propose a Cross-modal Recurrent Intra-frame Feature Enhancement (CRIFE) module to better leverage the rich temporal information of the events by mutually interacting the blur frame and event features within the duration of exposure time. In this module, we perform recurrent attention-based feature enhancement using a transformer [38] that better captures long-range pixel dependencies.

In temporal feature alignment with the second perspective, we propose a novel event-guided temporal feature alignment module, effectively leveraging rich temporal characteristics of the events. Conventional frame-based temporal feature alignment methods rely on optical flow [28] or deformable convolution [5,58]. While optical flow and deformable convolutions aid in achieving temporal feature alignment, the high computational complexity of these operations makes it challenging to execute them at a higher spatial scale. Therefore, video frame alignment, generally conducted at lower spatial scales, leads to sub-optimal deblurring results due to the lack of spatial contexts of features. To overcome these limitations, our temporal feature alignment module avoids relying on optical flow or deformable convolution by effectively leveraging the temporally dense advantages of the events. Therefore, we can effectively perform temporal feature alignment across multiple visual scales as we do not rely on these complex operations but rather efficiently leverage the temporal information from the events. As a result, our temporal feature alignment module demonstrates a significant performance improvement in event-guided video deblurring tasks. Since we have introduced a pioneering method for aligning temporal features using the advantages of the events, it is expected to be effectively applicable to various event-guided video restoration tasks (*e.g.*, event-guided video super-resolution).

Finally, we propose a novel video deblurring dataset, the EVRB dataset, composed of high-quality RGB and event data. It consists of real-world blurry

videos generated by extreme motion and corresponding sharp videos for generalized applications in real-world scenarios. The network trained on the EVRB dataset can be directly applicable to real-world scenarios, making it a valuable resource for event-guided deblurring research.

2 Related Works

2.1 Video Deblurring

Early works [1, 35] employed CNNs that take the concatenation of adjacent frames as input to address video deblurring. Subsequently, to better leverage temporal information, approaches have emerged that employ 3D convolutions [26, 47], temporal alignment modules with deformable convolutions [39] and optical flow [14, 25, 40, 46]. As an alternative video alignment method, some works utilize RNN [11, 42, 53] and transformer [18, 19] structures to propagate information from long-range video frames. However, as the intensity of motion blur in the video increases, frame-based video alignment methods struggle to achieve accurate video alignment, leading to sub-optimal deblurring results.

2.2 Event-guided Motion Deblurring

An event camera can effectively be used for motion deblurring as it records motion information corresponding to the brightness differences with high temporal resolution. Efforts to utilize event cameras for motion deblurring [8, 13, 15, 20, 27, 36, 37, 43, 48–50] have been actively ongoing. Recent studies have focused on effectively fusing event and image features of different modalities. To this end, Sun *et al*. [36] employed a transformer architecture for feature fusion. Zhang *et al*. [50] effectively combined modalities using a multi-scale architecture. In addition, there are also attempts to address motion deblurring by assuming challenging and general scenarios [8, 15, 33]. However, these studies have primarily focused on single image deblurring and do not exploit the long-range temporal information demonstrated in previous video deblurring tasks. To effectively acquire information that may be missing from sparse events, we introduce a novel method for accurate temporal alignment with events, utilizing information from surrounding adjacent frames in video deblurring.

3 Event-based Video Deblurring Dataset for Real-world Blur

3.1 Limitation of Synthetic Blur Dataset

Typically, the blurred images synthesis procedure adheres to the methodology outlined in [22–24, 34, 35], involving the averaging of consecutive video frames within a fixed-size window. However, as recent studies [31, 52, 54] have discussed, blur synthesis based on discrete signals may result in shutter artifacts even when

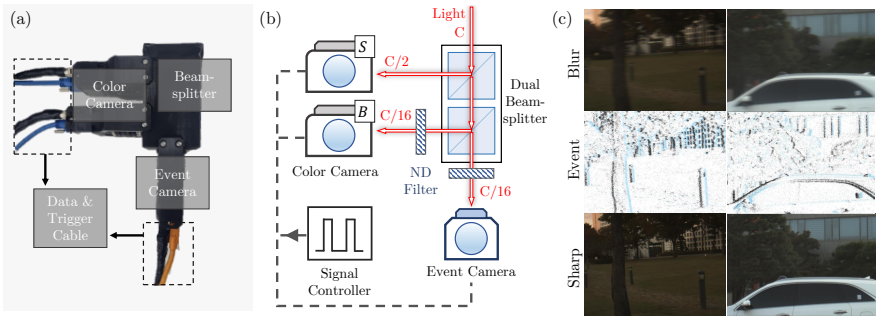

Fig. 1: Illustration of a hybrid camera system for real-world event-based video deblurring dataset. S and B denote the cameras for acquiring sharp and blur videos, respectively. (a): The triple-axis camera system to capture real-world blur. (b): A diagram of our hybrid camera system. (c): Samples from our EVRB dataset with natural blur.

averaging high-frame-rate videos. Furthermore, simply averaging sharp images disregards essential elements in the blurred images such as pixel saturation [30], limitations due to dynamic range, and physical noise [41, 51] in data acquired during exposure time. Furthermore, in the case of the event-guided motion deblurring research, existing works have enhanced the synthetic aspects by using event simulator [29]. To address these limitations, recent works have attempted to acquire real-world blur datasets using hybrid camera systems [3, 31, 52–54]. Meanwhile, to generalize event-based motion deblurring, Sun *et al.* [36] introduced the ReBlur dataset acquired in a high-precision optical laboratory using an electronic-controlled slide-rail system. However, despite these successes, the ReBlur dataset has limitations due to its indoor setting, a lack of dynamic objects, and the absence of high-quality RGB data as it is acquired using the DAVIS sensor [2]. Additionally, the sequences are relatively short(minimum six blurred frames in the sequence), making it challenging for use in event-guided video deblurring. We have acquired a new EVRB dataset for evaluating event-guided image and video deblurring methods in real-world blurry videos. The EVRB dataset was captured using a customized hybrid system consisting of two RGB cameras and one event camera. This system encompasses a range of blur magnitude from slight to extreme in various urban environments.

3.2 Triple-axis Hybrid Camera System

As shown in Fig. 1, we design a hybrid camera system to enable the acquisition of different data sources to be geometrically aligned. Two RGB cameras and one event camera are geometrically aligned using two 50/50 cube beam splitters, resulting in a minimal baseline. We perform pixel-wise alignment between multiple cameras based on the homography calculated by extrinsic calibration for precise alignment. For the synchronization of the multiple cameras, we use an external trigger system. Each camera receives the falling and rising edges of the trigger

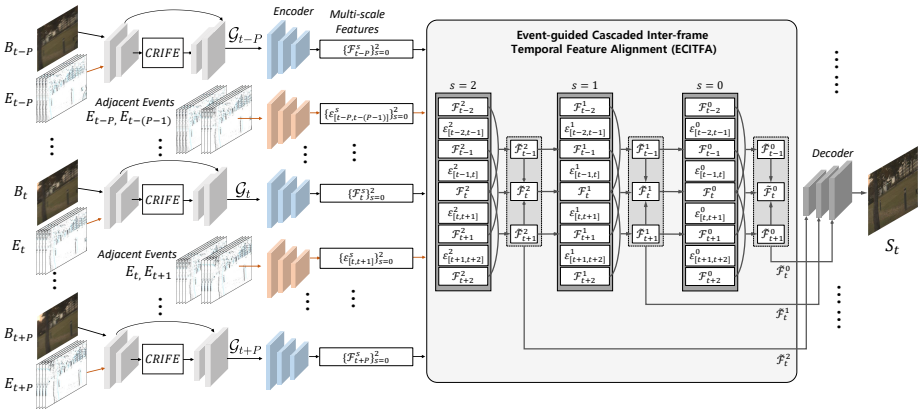

Fig. 2: Overall framework of CMTA is divided into two main components: Cross-modal Recurrent Intra-frame Feature Enhancement (CRIFE) and Event-guided Cascaded Inter-frame Temporal Feature Alignment (ECITFA). s is the scale factor for multi-scale features. In the figure of the ECITFA module, the description was performed for the case of P=2 for simplicity.

signal and acquires data synchronized with the period of the signal. This external trigger can precisely control the exposure times of the two RGB cameras, enabling us to capture paired sharp and blurred video frames. Furthermore, for photometric alignment, we physically adjust the amount of incoming light for both cameras to equalize the total irradiance of the multiple cameras using a neutral density filter.

4 Method

4.1 Overview

The overview of the proposed framework is illustrated in Fig. 2. Given consecutive blurred video frames $\{B_k\}$ and sets of event streams corresponding to the exposure time of each video frame $\{\mathbb{E}_k\}$, where $k \in \{t-P, \ldots, t, \ldots, t+P\}$, our goal is to estimate the latent sharp video frame S_t. To utilize the event stream $\{\mathbb{E}_k\}$ corresponding to the exposure time of $\{B_k\}$ as the input for the networks, we first perform embedding using the event voxel grid representation [56] for the event stream $\{\mathbb{E}_k\}$, resulting in $\{E_k\}$. Our framework consists of two main submodules: (1) Cross-modal Recurrent Intra-frame Feature Enhancement (CRIFE) module and (2) Event-guided Cascaded Inter-frame Temporal Feature Alignment (ECITFA) module. In the first module, we perform cross-modal feature enhancement through recurrent interactions between blurred frame features and event features to leverage the continuous temporal information of events within the exposure time. After obtaining the fused feature $\{\mathcal{G}_k\}$ from the first module, we generate multi-scale features, $\{\mathcal{F}_k^s\}_{s=0}^2$, through a weight-shared pyra-

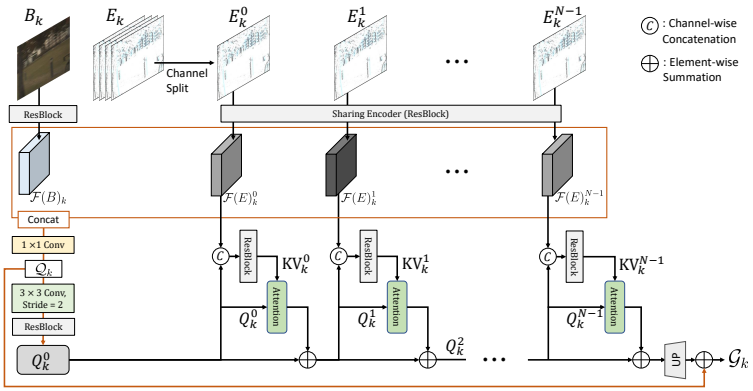

Fig. 3: Illustration of Cross-modal Recurrent Intra-frame Feature Enhancement (CRIFE).

mid encoder. Subsequently, for the temporal feature alignment stage, we encode multi-scale event features, $\{\mathcal{E}^s_{[k,k+1]}\}^2_{s=0}$, by grouping two consecutive events corresponding to the exposure times of each frame to connect adjacent frames. By utilizing the multi-scale pyramid features $\{\mathcal{F}^s_k\}^2_{s=0}$ and the encoded event feature $\{\mathcal{E}^s_{[k,k+1]}\}^2_{s=0}$, we perform temporal feature alignment. Afterward, temporally aligned feature pyramids are fed into the U-Net [32]-based decoder, generating the final sharp video frame S_t. Our framework can be extended to cases where P is an arbitrary positive number. However, for the sake of brevity, we will explain it in the main text with $P = 2$.

4.2 Cross-modal Recurrent Intra-frame Feature Enhancement

Since event cameras provide rich temporal information on brightness changes, it is crucial to effectively utilize this dense temporal information of the events within the duration of the exposure time. Typically, event data captured during the exposure time is transformed into event embeddings [10, 56] as input to the network, followed by feature extraction using 2D CNNs. However, these approaches cannot effectively utilize the continuous temporal information of events, and these limitations can impact the performance of event-guided video deblurring. To address these limitations, we propose a method that leverages the rich temporal nature of events and fuses event and blurred frame features using recurrent-based attention methods.

Recently, researchers have demonstrated the effectiveness of transformer-based architectures [38] by capturing long-range pixel dependencies, which have been proven highly effective in various vision tasks. We leverage transformers' advantages to more effectively utilize the temporal benefits of the events through a recurrent-based approach. For each blur video frame index $k \in \{t-2, \ldots, t+2\}$, we initially partition the exposure time $T_{exp,k}$ into N unit temporal intervals,

denoted as Δt, with $\Delta t = T_{exp,k}/N$. Based on the time interval Δt, we divide the event voxel grid within the exposure time, denoted as $E_k \in \mathbb{R}^{C\times H\times W}$ into N temporally divided event voxel grids $\{E_k^n\}$, where $E_k^n \in \mathbb{R}^{C/N\times H\times W}$ and $n \in \{0, ..., N-1\}$. After, we extract event features $\{\mathcal{F}(E)_k^n\}$ of the temporally divided event voxel grids E_k^n using weight-sharing event feature extractor. To apply cross-attention, we construct a query encompassing global information of blur and temporally divided event features. That is, we concatenate the blur frame feature $\mathcal{F}(B)_k$ with the temporally segmented event sets $\{\mathcal{F}(E)_k^n\}$ and then extract the feature for the encoding query information.

$$\mathcal{Q}_k = \mathrm{Conv_p}(\mathcal{F}(B)_k \| \{\mathcal{F}(E)_k^n\}) \tag{1}$$

$$Q_k^{n=0} = F_R(\mathcal{Q}_k) \tag{2}$$

where $\mathrm{Conv_p}$ represents point-wise convolution, F_R consists of a sequence of blocks with 3×3 convolution and a stride of 2, along with ResBlocks. Additionally, $Q_k^{n=0}$ denotes the initial query feature for the recurrent-based attention method.

As illustrated in Fig. 3, we perform cross-attention to update query information iteratively. To find the key and value of cross-attention for query updating, we recursively input temporally separated event features, helping to better utilize the events' rich temporal information. That is, for iteration n, we extract features to be used as *keys* and *values* by concatenating the previously updated Q_k^n with temporally divided n-th event features $\mathcal{F}(E)_k^n$ as follows:

$$\mathrm{KV}_k^n = F_{KV}(Q_k^n \| \mathcal{F}(E)_k^n) \tag{3}$$

where $\|$ channel-wise concatenation and KV_k^n denotes the output features for key and value projection, and F_{KV} refers to the ResBlocks layer. We then construct K_k^n, V_k^n as $K_k^n = W^K(\mathrm{KV}_k^n)$ and $V_k^n = W^V(\mathrm{KV}_k^n)$ where $W^{(\cdot)}$ denote 1×1 convolution layer. Then, attention can be calculated as:

$$\mathrm{Attn_k^n} = \mathrm{SoftMax}(\frac{Q_k^n (K_k^n)^T}{\alpha}) V_k^n \tag{4}$$

where α is learnable scaling parameter to balance attention weights and $\mathrm{Attn_n}$ is outputs of cross-attention at iteration n. To calculate cross-covariance matrix efficiently, we adopted transposed attention [44] for efficient computations. This method enables the efficient computation of attention values at high resolutions by calculating the cross-covariance matrix along the channel axis, leading to a complexity of $O(C^2)$. Moreover, we reduced the number of channels when encoding step, enabling even more efficient operations. We use the attention value $\mathrm{Attn_k^n}$ to update the query iteratively as follows:

$$Q_k^{n+1} = Q_k^n + \mathrm{Attn_k^n} + \mathrm{MLP}(\mathrm{Attn_k^n}) \tag{5}$$

where MLP denote multi-layer perceptron. After N iterations, we obtain the updated query feature, Q_k^N. The final query feature passes through up-sampling

and skip connections. Then, we generate the final fused features $\{\mathcal{G}_k\}$ as $\mathcal{G}_k = \mathcal{Q}_k + \text{Dconv}_{4\times4}(Q_k^N)$ where $\text{Dconv}_{4\times4}$ denote deconvolution layer with a kernel size of 4. As illustrated in Fig. 2, the fused features $\{\mathcal{G}_k\}$ pass through a pyramid encoder, leading to the creation of multi-scale pyramid features, $\{\mathcal{F}_k^s\}_{s=0}^2$. Through the CRIFE module, we can leverage the advantages of abundant temporal information on the events within the exposure time.

4.3 Event-guided Cascaded Inter-frame Temporal Feature Alignment

Temporal feature alignment aims to extract valuable information from adjacent video frames. Conventional frame-based video deblurring methods face challenges from inaccurate motion estimation due to motion blur. Conversely, event cameras, thanks to their resistance to motion blur, can offer valuable guidance for aligning video frames. The most straightforward way to align adjacent video frames using the events is to utilize event features to estimate optical flows or deformable offsets between neighboring frames [4,5,58]. While this approach can leverage the advantages of event data for offset and optical flow estimation, it is typically employed after spatial down-sampling due to the high computational costs associated with deformable convolutions [58] and optical flows [28]. Therefore, this approach could restrict performance as it inherently hinders the processing of features at high spatial resolutions in network architectures, limiting access to information across multiple visual scales. To address the aforementioned limitations, we propose a new temporal alignment module that combines the advantages of multiple-visual scale pyramids, leveraging the events' rich temporal contexts.

As illustrated in Fig. 2, our proposed temporal feature alignment modules are structured as multi-level networks that gradually perform coarse-to-fine feature alignment. First, we extract event features, incorporating exposure time information from neighboring video frames to reference frames to facilitate video feature alignment. Specifically, we encode event features encompassing the exposure time interval between t and $t+1$, allowing us to connect the frame at time t with the frame at time $t+1$. Through this event encoding step, we obtain event feature pyramid set $\{\varepsilon_{[m,m+1]}^s\}_{s=0}^2$ where $m \in \{t-2, ..., t+1\}$. We use the event feature pyramid containing motion information for adjacent times for the alignment of each blur frame feature pyramid $\{\mathcal{F}_k^s\}_{s=0}^2$ where $k \in \{t-2, ..., t+2\}$.

After the event encoding step for the feature alignment, we gradually perform temporal feature alignment from the bottom pyramid level (scale factor s of 2) to the top pyramid level (s of 0). Specifically, as shown in left side of Fig. 4, in each pyramid level s ($s \in \{0, 1, 2\}$), We first upsample the hidden state of temporally aligned features at the previous scale, $\hat{\mathcal{F}}_i^{s+1}$ through a deconvolution layer and receive them as inputs of alignment module, resulting in h_i^s.

$$h_i^s = \text{Dconv}_{4\times4}(\hat{\mathcal{F}}_i^{s+1}),\ i \in \{t-1, t, t+1\} \tag{6}$$

where $\hat{\mathcal{F}}_i^{s+1}$ denote aligned feature at previous scale $s+1$, $\text{Dconv}_{4\times4}$ denote 4×4 deconvolution layer. Note that there is no hidden state at the bottom pyramid

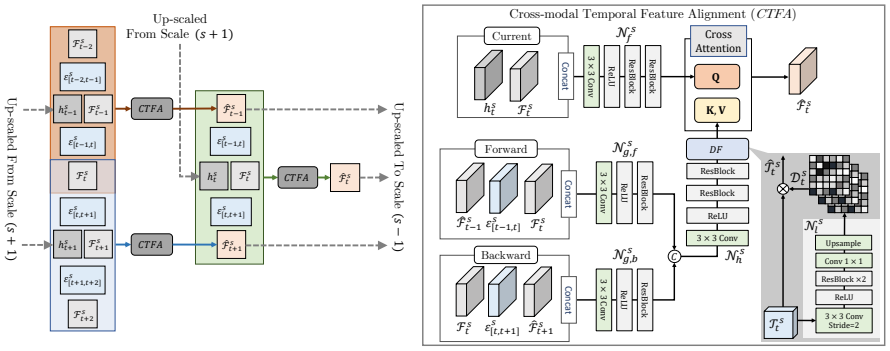

Fig. 4: Overview of the Event-guided Cascaded Inter-frame Temporal Feature Alignment (ECITFA). The left figure illustrates temporal alignment for scale s. The key module for each alignment procedure, Cross-modal Temporal Feature Alignment (CTFA) at time t, is illustrated on the right of the figure. The CTFA module operates similarly for reference times $t-1$ and $t+1$ as well.

level (s of 2) since there are no features aligned at the previous scale, and for brevity, we focus on the case when $s < 2$.

As depicted in the left side of Fig. 4, we progressively group three blurred video frames when performing temporal feature alignment. In other words, we first align with respect to $t-1$ using $\{t-2, t-1, t\}$ and simultaneously align with respect to $t+1$ using $\{t, t+1, t+2\}$ features. Subsequently, we perform the final temporal feature alignment for the previously aligned $t-1$ and $t+1$ with the last target time t. More specifically, when performing frame alignment for the time step $t-1$, we first group the three blur video frame features $\mathcal{F}^s_{t-2}, \mathcal{F}^s_{t-1}, \mathcal{F}^s_t$ and perform alignment. In this case, we make use of two event features $\varepsilon^s_{[t-1,t]}$ and $\varepsilon^s_{[t,t+1]}$, which respectively contain motion information between $t-1$ and t, and between t and $t+1$, respectively.

$$\hat{\mathcal{F}}^s_{t-1} = \text{CTFA}^s(\mathcal{F}^s_t, \mathcal{F}^s_{t-1}, \mathcal{F}^s_{t-2}, h^s_{t-1}, \varepsilon^s_{[t-2,t-1]}, \varepsilon^s_{[t-1,t]}), \tag{7}$$

where CTFA^s denote unit Cross-modal Temporal Feature Alignment (CTFA) module at scale factor s. Across the same scale factor s, CTFA modules are weight-shared. Similarly, we perform alignment for the time step $t+1$ as:

$$\hat{\mathcal{F}}^s_{t+1} = \text{CTFA}^s(\mathcal{F}^s_t, \mathcal{F}^s_{t+1}, \mathcal{F}^s_{i+1}, h^s_{t+1}, \varepsilon^s_{[t+1,t+2]}, \varepsilon^s_{[t,t+1]}). \tag{8}$$

Finally, we utilize the temporally aligned results $\hat{\mathcal{F}}^s_{t+1}$ and $\hat{\mathcal{F}}^s_{t-1}$ to once again group with $\mathcal{F}^s_t$ and perform alignment:

$$\hat{\mathcal{F}}^s_t = \text{CTFA}^s(\hat{\mathcal{F}}^s_{t-1}, \mathcal{F}^s_t, \hat{\mathcal{F}}^s_{t+1}, h^s_t, \varepsilon^s_{[t-1,t]}, \varepsilon^s_{[t,t+1]}). \tag{9}$$

Through this cascaded feature alignment stage, we effectively propagate non-local video frame information, facilitating temporal feature alignment for both

non-local and adjacent video frames. All the results of alignment $\hat{\mathcal{F}}^s_{t-1}, \hat{\mathcal{F}}^s_t, \hat{\mathcal{F}}^s_{t+1}$ are passed to the next scale, $s-1$. The right side of Fig. 4 illustrates the overall alignment process for the time step t in the CTFA module. Since it is applied similarly for all time steps $i \in \{t-1, t, t+1\}$, we will describe it here specifically when $i = t$, where previously aligned $\hat{\mathcal{F}}^s_{t-1}$ and $\hat{\mathcal{F}}^s_{t+1}$ are provided. For leveraging the benefits of coarse-to-fine architecture, we first fuse hidden state aligned feature h^s_t from the previous scale for the time step t at scale factor s, as follows:

$$\mathcal{S}^s_t = \mathcal{N}^s_f(\mathcal{F}^s_t, h^s_t), \tag{10}$$

We simplify the sequential operations of Conv, ReLU, and ResBlock as '$\mathcal{N}_f$'. Additionally, to leverage temporal information, we align the features at times $t-1$ and $t+1$ to the current feature at time t using event features that encapsulate motion between frames in the following manner:

$$\begin{aligned}
\mathcal{T}^s_{t-1\to t} &= \mathcal{N}^s_{g,f}(\hat{\mathcal{F}}^s_{t-1}, \mathcal{F}^s_t, \varepsilon^s_{[t-1,t]}),\\
\mathcal{T}^s_{t+1\to t} &= \mathcal{N}^s_{g,b}(\mathcal{F}^s_t, \hat{\mathcal{F}}^s_{t+1}, \varepsilon^s_{[t,t+1]}),\\
\mathcal{T}^s_t &= \mathcal{N}^s_h(\mathcal{T}^s_{t-1\to t} \| \mathcal{T}^s_{t+1\to t}),
\end{aligned} \tag{11}$$

where $\|$ denotes the channel-wise concatenation, and $\mathcal{N}^s_{g,f}$, $\mathcal{N}^s_{g,b}$, and $\mathcal{N}^s_h$ represent convolution blocks, as illustrated on the right side of Fig. 4. While we obtain temporally aligned features $\mathcal{T}^s_t$ using event features and leveraging the advantages of events, for additional feature refinement, we utilize the spatially variant pixel-wise dynamic filter [21, 55] mechanism. Dynamic filter $\mathcal{D}^s_t$ at scale s can be calculated through filter generation blocks, $\mathcal{N}^s_l$, which consists of convolution and resblock, such as $\mathcal{D}^s_t = \mathcal{N}^s_l(\mathcal{T}^s_t)$ where $\mathcal{D}^s_t \in \mathbb{R}^{(s_k \times s_k) \times H^s \times W^s}$, s_k is kernel size of dynamic convolution filter, H^s, W^s denote height and width of the feature at scale factor s, respectively. Then, we apply the dynamic convolution operation as follows:

$$\hat{\mathcal{T}}^s_t(h, w) = \mathcal{D}^s_t(h, w) \otimes \mathcal{T}^s_t(h, w) \tag{12}$$

where $h \in \{1, \ldots, H^s\}, w \in \{1, \ldots, W^s\}$, and $\otimes$ denotes the convolution operation. Then, we employ a cross-attention mechanism [38,44] to effectively combine the information of two features, $\mathcal{S}^s_t$ and $\hat{\mathcal{T}}^s_t$. The cross-attention generally examines the correlation between input features (query) and the key-value features. Therefore, we conduct correlation analysis by applying cross-attention between the current blurred frame features to project query $\mathcal{S}^s_t$ and the aligned features $\mathcal{T}^s_t$ to project key and values. That is, we generate query, key, and value features, $\mathbf{Q} = W_Q(\mathcal{S}^s_t)$, $\mathbf{K} = W_K(\hat{\mathcal{T}}^s_t)$, $\mathbf{V} = W_V(\hat{\mathcal{T}}^s_t)$, where $W_{(\cdot)}$ is 1×1 convolution and 3×3 depth-wise convolution. Utilizing these $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$, we compute the attention matrix similarly to Eq. (4).

$$\mathcal{A}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathrm{Softmax}(\frac{\mathbf{Q}\mathbf{K}^T}{\alpha}) \cdot \mathbf{V} \tag{13}$$

Table 1: Quantitative results on the GoPro dataset. The asterisk(∗) indicates that the results are not officially on the GoPro so we retrained the official model by us. CMTA-5 and CMTA-7 refer to the results obtained using 5 and 7 input frames, respectively.

Methods	MPRNet [45]	Restormer [44]	NAFNet [6]	EDVR [39]	ESTRNN [53]	RNN-MBP [57]	DSTNet [26]	VRT [17]
PSNRs	32.66	32.92	33.69	26.83	31.02	33.32	34.16	34.81
SSIMs	0.959	0.961	0.967	0.843	0.911	0.963	0.968	0.972
Params(MB)	20.1	26.1	67.8	23.6	2.4	16.4	7.5	18.3
Methods	RVRT [18]	Shift-Net [16]	UEVD* [15]	EFNet [36]	REFID [37]	SpkNet [7]	**CMTA-5**	**CMTA-7**
PSNRs	34.92	35.49	35.48	35.46	35.91	36.12	36.55	**36.78**
SSIMs	0.974	0.976	0.971	0.972	0.973	0.971	**0.977**	**0.977**
Params(MB)	10.8	10.5	27.9	8.5	15.9	13.5	9.7	9.7

Finally, temporally aligned feature, $\hat{\mathcal{F}}_t^s$, can be obtained by:

$$\hat{\mathcal{F}}_t^s = \mathrm{MLP}(\mathcal{A}) + \mathcal{A}. \tag{14}$$

where MLP denote multi-layer perceptrons, $\mathcal{A}$ denote the result of attention operation. Finally, we obtained the aligned feature pyramid $\{\hat{\mathcal{F}}_t^s\}$ through cascaded temporal feature alignment. As illustrated in Fig. 2, these multi-scale aligned features are fed into the decoder.

4.4 Decoder

The decoder is designed based on the standard U-Net [32]. It takes multi-scale temporally aligned features $\{\hat{\mathcal{F}}_t^s\}$ as inputs and produces output feature pyramid $\{\mathcal{F}(D)_k^s\}$. The final deblurred outputs S_t using last scale of output feature $\mathcal{F}(D)_t^{s=0}$ is calculated as follows:

$$S_t = B_t + \mathrm{Conv}_{5\mathrm{x}5}(\mathcal{F}(D)_t^{s=0}) \tag{15}$$

where $\mathrm{Conv}_{5\mathrm{x}5}$ represents a conv. layer with a filter size of 5×5, and S_t denotes the final estimated sharp frame.

5 Experiments

5.1 Datasets

GoPro Dataset [23]. For a fair comparison with other previous event-guided deblurring methods, we utilize the same raw events provided by the authors of recent work [36]. These events were generated using ESIM [29] with a randomly generated contrast threshold set to a Gaussian normal distribution of parameters as $N(\mu = 0.2, \sigma = 0.03)$. We used the official train and test splits.

HighREV Dataset [37]. HighREV consists of high-resolution events and RGB data at 1632×1224 resolution, designed for both deblurring and interpolation tasks. To evaluate motion deblurring exclusively, we use the *11+1* split, excluding the interpolation ground truth.

Real-world Video Deblurring Dataset. EVRB dataset consists of 11 training sequences and 6 test sequences. With each sequence containing 149 frames, it is well-suited for video deblurring tasks.

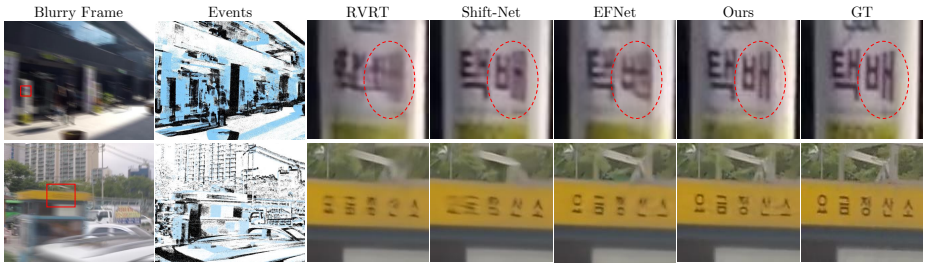

Fig. 5: Visual comparison of deblurring results on the GoPro dataset. The qualitative results of other methods were taken from the results provided by the authors.

Table 2: Quantitative results on the HighREV dataset.

Methods	ESTRNN [53]	DSTNet [26]	UEVD [15]	EFNet [36]	REFID [37]	Ours
PSNRs	30.38	31.77	37.40	37.99	38.37	**39.12**
SSIMs	0.940	0.948	0.974	0.976	0.977	**0.980**

5.2 Comparison on Synthetic Blur Datasets

We present the quantitative results of our frameworks with other frame-based image and video deblurring methods and event-guided motion deblurring methods using the GoPro dataset as depicted in Table 1. When compared to the existing best performance of video deblurring methods, Shift-Net [16], our approach (CMTA-5 model) shows a significant improvement of 1.06 dB in terms of PSNR, demonstrating that our method effectively leverages the temporal dense characteristics of event modality for accurate temporal feature alignment and cross-modality feature enhancement. Moreover, compared to the best-performing SpkNet [7] among existing event-guided motion deblurring methods, CMTA-5 model exhibits a performance improvement of 0.43 dB with a lower model params of 3.8 MB. Furthermore, by using 7 input video frames (CMTA-7 model) instead of 5 (CMTA-5 model), we achieved an impressive state-of-the-art performance with a PSNR of 36.78dB in the GoPro, an increase of 0.23. We further demonstrated the superiority of our approach through the qualitative results in Fig. 5. Also, we conduct experiments on HighREV [37], which consists of real events, and report the results in Table 2. Our method still achieves the best performance.

5.3 Comparison on Real-world Blur Datasets

For comparisons in the EVRB dataset, we trained representative video deblurring methods [12, 16, 18, 26, 39, 53] and event-guided motion deblurring methods [15, 36, 37] on the same training set. Tab 3 presents the quantitative results on the EVRB dataset. The EVRB dataset includes extremely blurred videos captured during exposure times. As a result, frame-based video deblurring methods exhibit subpar performance. For instance, the best-performing network, Shift-Net [16], achieves only 30.56 dB, which is 0.42 dB lower than the best-performing

Table 3: Quantitative results on the EVRB dataset.

Methods	EDVR [39]	ESTRNN [53]	ERDN [12]	DSTNet [26]	RVRT [18]
PSNRs	29.02	29.79	28.32	29.15	30.24
SSIMs	0.886	0.911	0.893	0.898	0.906
Methods	Shift-Net [16]	UEVD [15]	EFNet [36]	REFID [37]	Ours
PSNRs	30.56	30.55	30.98	30.33	**31.38**
SSIMs	0.922	0.915	**0.927**	0.918	**0.927**

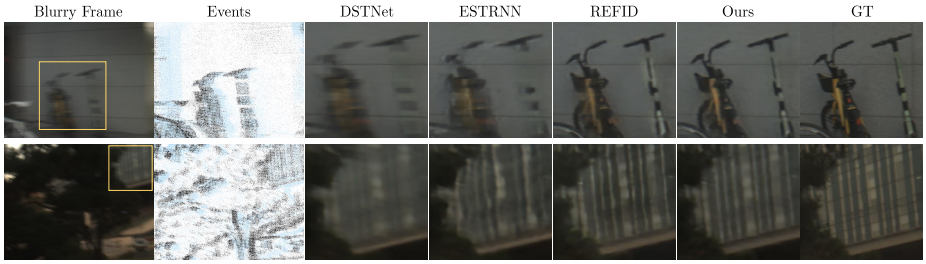

Fig. 6: Visual comparison of deblurring results on the EVRB dataset.

event-guided deblurring method, EFNet [36]. In contrast, our approach leverages video and event characteristics, effectively restoring even severe motion blur, which may be difficult to recover. Our method outperforms all approaches, achieving the best performance. We show our qualitative comparison with other methods in Fig. 6.

5.4 Ablation Study

We analyzed the performance contribution of the various modules in our frameworks. For a fair ablation study, we trained all the models for 600 epochs with 5 video frame inputs, conducting all experiments on the GoPro dataset.

CRIFE module. To demonstrate the effectiveness of the CRIFE module, we replaced it with concatenation and convolution for comparison. When comparing the first column(Ver.1) and the second column(Ver.2) of Tab 4, we observed a performance gain of +0.35 dB with a small additional parameter (+0.08 MB). Similarly, when comparing the performance of the third and fourth rows, we observed performance improvement.

ECITFA module is the most crucial component of our model, performing feature alignment by leveraging information of the events. As in the Tab. 4, when comparing the first (Ver.1) with the third column (Ver.3) of the table, we observed a significant performance gain (+**1.69** dB) upon the insertion of the ECITFA module. This trend is similarly maintained when the CRIFE module is incorporated. When comparing Ver.2 with Ver.4 of the Tab. 4, we observed a notable performance improvement (+**1.55** dB).

Effectiveness of components in ECITFA module. In the Tab. 5, we demonstrated an effectiveness analysis for each component of ECITFA. Comparing the 2nd column with the last column labeled 'Ours', we observed a performance

Table 4: Ablation study on the GoPro dataset.

Methods	Ver.1	Ver.2	Ver.3	Ver.4
CRIFE		✓		✓
ECITFA			✓	✓
PSNRs / Params	34.65 / 4.57M	35.00 / 4.65M	<u>36.34</u> / 9.59M	**36.55** / 9.68M

Table 5: Effect of the component of ECITFA module on the GoPro dataset. DF denotes dynamic filter operation in the Eq.(12).

Methods	w/o Cascaded	w/o DF	w/o Attention	Ours
PSNRs	36.08	36.18	<u>36.29</u>	**36.55**

Table 6: Comparison of CRIFE with various module variants. "Baseline" refers to Ver.3 of Tab.4. E and F represent event and frame features, respectively.

Methods	Baseline	w/o concat (E+F)	w/concat (E+F)	
			w/o recurrent	w/ recurrent(Ours)
PSNRs	<u>36.34</u>	36.04	35.99	**36.55**

gain of our method (+0.37 dB) when employing spatial pixel-wise dynamic filters, in contrast to not using DF. When utilizing cross-attention (Eq.13) to better leverage long-range pixel-dependencies in the alignment blocks, we observed a performance improvement (+0.29 dB) compared to aggregate feature using ResBlocks. Finally, we confirmed the effectiveness of leveraging non-local video frame information through a cascaded-based temporal feature alignment method for video deblurring. After using the proposed cascaded structure for temporal feature alignment, we can observe a performance gain (+0.47 dB).

Effectiveness of components in CRIFE module. To evaluate each component's effectiveness of the CRIFE module, we removed the concatenation between RGB and event features, directly matching their features. However, as shown in the 3rd column of Tab. 6, this approach yielded sub-optimal performance. Additionally, combining event features without a recurrent structure degraded performance (see 4th column). These ablations confirm that the recurrent structure effectively utilizes temporal information from events within exposure time.

6 Conclusions

This paper proposes a video deblurring framework, CMTA, that elaborately considers the characteristics of an event and video. Specifically, we achieve significant performance improvement through intra-frame feature enhancement and inter-frame temporal feature alignment. Furthermore, we construct a real-world deblurring dataset, the EVRB dataset, which will be valuable for evaluating event-guided deblurring methods. Finally, CMTA demonstrates state-of-the-art performance across various deblurring datasets.

Acknowledgements. This work was supported by the Technology Innovation Program (1415187329,20024355, Development of autonomous driving connec-

tivity technology based on sensor-infrastructure cooperation) funded By the Ministry of Trade, Industry & Energy(MOTIE, Korea) and the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (NRF2022R1A2B5B03002636).

References

1. Aittala, M., Durand, F.: Burst image deblurring using permutation invariant convolutional neural networks. In: Proceedings of the European conference on computer vision (ECCV). pp. 731–747 (2018)
2. 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 **49**(10), 2333–2341 (2014)
3. Cao, M., Zhong, Z., Wang, J., Zheng, Y., Yang, Y.: Learning adaptive warping for real-world rolling shutter correction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17785–17793 (2022)
4. Chan, K.C., Wang, X., Yu, K., Dong, C., Loy, C.C.: Basicvsr: The search for essential components in video super-resolution and beyond. In: Proceedings of the IEEE conference on computer vision and pattern recognition (2021)
5. Chan, K.C., Zhou, S., Xu, X., Loy, C.C.: Basicvsr++: Improving video super-resolution with enhanced propagation and alignment (2021)
6. Chen, L., Chu, X., Zhang, X., Sun, J.: Simple baselines for image restoration. In: European Conference on Computer Vision. pp. 17–33. Springer (2022)
7. Chen, S., Zhang, J., Zheng, Y., Huang, T., Yu, Z.: Enhancing motion deblurring in high-speed scenes with spike streams. Advances in Neural Information Processing Systems **36** (2024)
8. Cho, H., Jeong, Y., Kim, T., Yoon, K.J.: Non-coaxial event-guided motion deblurring with spatial alignment. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12492–12503 (2023)
9. Gallego, G., Delbrück, T., Orchard, G., Bartolozzi, C., Taba, B., Censi, A., Leutenegger, S., Davison, A.J., Conradt, J., Daniilidis, K., et al.: Event-based vision: A survey. IEEE transactions on pattern analysis and machine intelligence **44**(1), 154–180 (2020)
10. Gehrig, D., Loquercio, A., Derpanis, K.G., Scaramuzza, D.: End-to-end learning of representations for asynchronous event-based data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5633–5643 (2019)
11. Hyun Kim, T., Mu Lee, K., Scholkopf, B., Hirsch, M.: Online video deblurring via dynamic temporal blending network. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4038–4047 (2017)
12. Jiang, B., Xie, Z., Xia, Z., Li, S., Liu, S.: Erdn: Equivalent receptive field deformable network for video deblurring. In: European Conference on Computer Vision. pp. 663–678. Springer (2022)
13. 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)
14. Kim, T.H., Sajjadi, M.S., Hirsch, M., Scholkopf, B.: Spatio-temporal transformer network for video restoration. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 106–122 (2018)

15. Kim, T., Lee, J., Wang, L., Yoon, K.J.: Event-guided deblurring of unknown exposure time videos. In: European Conference on Computer Vision. pp. 519–538. Springer (2022)
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 (CVPR). pp. 9822–9832 (June 2023)
17. Liang, J., Cao, J., Fan, Y., Zhang, K., Ranjan, R., Li, Y., Timofte, R., Van Gool, L.: Vrt: A video restoration transformer. arXiv preprint arXiv:2201.12288 (2022)
18. Liang, J., Fan, Y., Xiang, X., Ranjan, R., Ilg, E., Green, S., Cao, J., Zhang, K., Timofte, R., Gool, L.V.: Recurrent video restoration transformer with guided deformable attention. Advances in Neural Information Processing Systems **35**, 378–393 (2022)
19. Lin, J., Cai, Y., Hu, X., Wang, H., Yan, Y., Zou, X., Ding, H., Zhang, Y., Timofte, R., Van Gool, L.: Flow-guided sparse transformer for video deblurring. arXiv preprint arXiv:2201.01893 (2022)
20. Lin, S., Zhang, J., Pan, J., Jiang, Z., Zou, D., Wang, Y., Chen, J., Ren, J.S.: Learning event-driven video deblurring and interpolation. In: ECCV (8). pp. 695–710 (2020)
21. Mildenhall, B., Barron, J.T., Chen, J., Sharlet, D., Ng, R., Carroll, R.: Burst denoising with kernel prediction networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2502–2510 (2018)
22. Nah, S., Baik, S., Hong, S., Moon, G., Son, S., Timofte, R., Mu Lee, K.: Ntire 2019 challenge on video deblurring and super-resolution: Dataset and study. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 0–0 (2019)
23. 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)
24. Oh, J., Kim, M.: Demfi: deep joint deblurring and multi-frame interpolation with flow-guided attentive correlation and recursive boosting. In: European Conference on Computer Vision. pp. 198–215. Springer (2022)
25. Pan, J., Bai, H., Tang, J.: Cascaded deep video deblurring using temporal sharpness prior. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3043–3051 (2020)
26. Pan, J., Xu, B., Dong, J., Ge, J., Tang, J.: Deep discriminative spatial and temporal network for efficient video deblurring. In: The IEEE Conference on Computer Vision and Pattern Recognition(CVPR) (Feb 2023)
27. 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)
28. Ranjan, A., Black, M.J.: Optical flow estimation using a spatial pyramid network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4161–4170 (2017)
29. Rebecq, H., Gehrig, D., Scaramuzza, D.: Esim: an open event camera simulator. In: Conference on Robot Learning. pp. 969–982. PMLR (2018)
30. Rim, J., Kim, G., Kim, J., Lee, J., Lee, S., Cho, S.: Realistic blur synthesis for learning image deblurring. In: European conference on computer vision. pp. 487–503. Springer (2022)

31. Rim, J., Lee, H., Won, J., Cho, S.: Real-world blur dataset for learning and benchmarking deblurring algorithms. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16. pp. 184–201. Springer (2020)
32. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. pp. 234–241. Springer (2015)
33. 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 (ICCV). pp. 4531–4540 (October 2021)
34. Shen, W., Bao, W., Zhai, G., Chen, L., Min, X., Gao, Z.: Video frame interpolation and enhancement via pyramid recurrent framework. IEEE Transactions on Image Processing **30**, 277–292 (2020)
35. 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)
36. Sun, L., Sakaridis, C., Liang, J., Jiang, Q., Yang, K., Sun, P., Ye, Y., Wang, K., Van Gool, L.: Event-based fusion for motion deblurring with cross-modal attention. In: European Conference on Computer Vision (ECCV) (2022)
37. Sun, L., Sakaridis, C., Liang, J., Sun, P., Cao, J., Zhang, K., Jiang, Q., Wang, K., Van Gool, L.: Event-based frame interpolation with ad-hoc deblurring. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18043–18052 (2023)
38. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: NIPS. pp. 5998–6008 (2017)
39. Wang, X., Chan, K.C., Yu, K., Dong, C., Change Loy, C.: Edvr: Video restoration with enhanced deformable convolutional networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (June 2019)
40. Wang, Y., Lu, Y., Gao, Y., Wang, L., Zhong, Z., Zheng, Y., Yamashita, A.: Efficient video deblurring guided by motion magnitude. In: Proceedings of the European Conference on Computer Vision (ECCV) (2022)
41. Wei, K., Fu, Y., Yang, J., Huang, H.: A physics-based noise formation model for extreme low-light raw denoising. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2758–2767 (2020)
42. Wieschollek, P., Hirsch, M., Scholkopf, B., Lensch, H.: Learning blind motion deblurring. In: ICCV. pp. 231–240 (2017)
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 (ICCV). pp. 2583–2592 (October 2021)
44. 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)
45. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H., Shao, L.: Multi-stage progressive image restoration. arXiv preprint arXiv:2102.02808 (2021)
46. Zhang, H., Xie, H., Yao, H.: Spatio-temporal deformable attention network for video deblurring. In: ECCV (2022)

47. Zhang, K., Luo, W., Zhong, Y., Ma, L., Liu, W., Li, H.: Adversarial spatio-temporal learning for video deblurring. IEEE Transactions on Image Processing (2018)
48. Zhang, L., Zhang, H., Zhu, C., Guo, S., Chen, J., Wang, L.: Fine-grained video deblurring with event camera. In: ICMM. pp. 352–364. Springer (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. Zhang, Y., Qin, H., Wang, X., Li, H.: Rethinking noise synthesis and modeling in raw denoising. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4593–4601 (2021)
52. Zhong, Z., Cao, M., Ji, X., Zheng, Y., Sato, I.: Blur interpolation transformer for real-world motion from blur. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5713–5723 (2023)
53. Zhong, Z., Gao, Y., Zheng, Y., Zheng, B.: Efficient spatio-temporal recurrent neural network for video deblurring. In: European Conference on Computer Vision. pp. 191–207. Springer (2020)
54. Zhong, Z., Zheng, Y., Sato, I.: Towards rolling shutter correction and deblurring in dynamic scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9219–9228 (2021)
55. Zhou, J., Jampani, V., Pi, Z., Liu, Q., Yang, M.H.: Decoupled dynamic filter networks. In: IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR) (Jun 2021)
56. Zhu, A.Z., Yuan, L., Chaney, K., Daniilidis, K.: Unsupervised event-based learning of optical flow, depth, and egomotion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 989–997 (2019)
57. Zhu, C., Dong, H., Pan, J., Liang, B., Huang, Y., Fu, L., Wang, F.: Deep recurrent neural network with multi-scale bi-directional propagation for video deblurring. In: Proceedings of the AAAI conference on artificial intelligence. vol. 36, pp. 3598–3607 (2022)
58. Zhu, X., Hu, H., Lin, S., Dai, J.: Deformable convnets v2: More deformable, better results. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9308–9316 (2019)