Frequency-aware Event-based Video Deblurring for Real-World Motion Blur

Taewoo Kim*
KAIST
intelpro@kaist.ac.kr

Hoonhee Cho*
KAIST
gnsgnsgml@kaist.ac.kr

Kuk-Jin Yoon
KAIST
kjyoon@kaist.ac.kr

Abstract

Video deblurring aims to restore sharp frames from blurred video clips. Despite notable progress in video deblurring works, it is still a challenging problem because of the loss of motion information during the duration of the exposure time. Since event cameras can capture clear motion information asynchronously with high temporal resolution, several works exploit the event camera for deblurring as they can provide abundant motion information. However, despite these approaches, there were few cases of actively exploiting the long-range temporal dependency of videos. To tackle these deficiencies, we present an event-based video deblurring framework by actively utilizing temporal information from videos. To be specific, we first introduce a frequency-based cross-modal feature enhancement module. Second, we propose event-guided video alignment modules by considering the valuable characteristics of the event and videos. In addition, we designed a hybrid camera system to collect the first real-world event-based video deblurring dataset. For the first time, we build a dataset containing synchronized high-resolution real-world blurred videos and corresponding sharp videos and event streams. Experimental results validate that our frameworks significantly outperform the state-of-the-art frame-based and event-based deblurring works in the various datasets. The project pages are available at https://sites.google.com/view/fevd-cvpr2024.

1. Introduction

Motion blur often occurs due to the abrupt motion of the object and/or the camera during the exposure time of a frame-based camera. Motion deblurring aims to restore the latent sharp frame from blurred frame, which is a highly ill-posed problem [1, 16, 20]. With the recent development of deep learning methods [27, 40, 59, 63], considerable progress has been made in the field of motion deblurring. However, it still fails in cases of substantial motion magnitude.

Event cameras record per-pixel brightness changes with micro-second-level temporal resolutions. Thanks to their high dynamic range and motion blur resilience, event cameras can be effectively utilized in extreme scenarios across diverse fields [4, 9, 10, 53]. In particular, event cameras provide degradation information of blurred frames for motion deblurring. Thus, recent event-based motion deblurring works [21, 42, 43, 62] utilized the advantages of event cameras to restore the sharp frames from motion blurred ones.

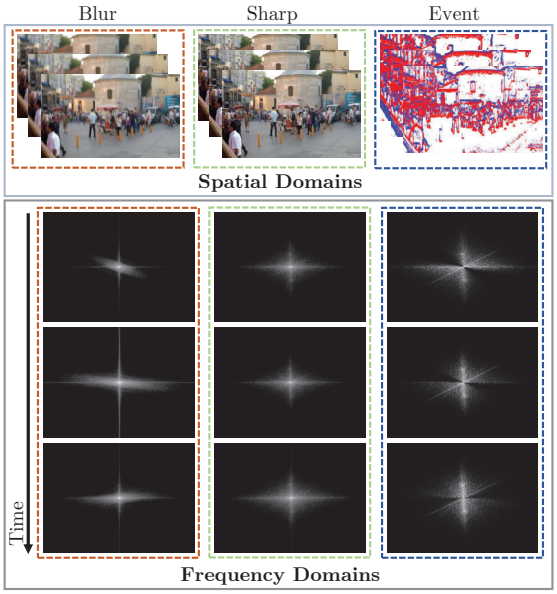

Figure 1. The visualization of blur and sharp videos with event sequences in spatial and frequency domains. We utilize the discrete fourier transform (DFT) to visualize the frequency spectrum.

However, there are still unexplored aspects in event-based deblurring research. In conventional frame-based video deblurring, temporal information from video has been actively utilized to improve restoration quality. In contrast, event-based motion deblurring still predominantly focuses on single-image deblurring despite the rich temporal information from event cameras. Although previous works [21, 43] use a pair of consecutive images for deblurring, there is currently no research actively leveraging the temporal information in larger continuous video sequences.

In this paper, we first explore the potential of rich temporal information of the events for video deblurring. To effectively leverage the characteristics of both the video

*The first two authors contributed equally.

and the events, we propose two temporal feature alignment modules leveraging the events. Specifically, we first propose an Event-guided Local-windowed Temporal Propagation (ELTP) module. The ELTP module effectively leverages multi-frame information by utilizing temporal information of the events from adjacent neighbor frames. In addition, we propose a Bidirectional Temporal Feature Fusion (BTFF) module to fully exploit information across the entire video inputs by fusing long-term temporal features. These modules effectively extract temporal information from the videos, resulting in the generation of high-quality sharp features. Finally, we propose a Frequency-aware Cross-Modal Feature Enhancement (FCFE) module to achieve more reliable feature enhancement between images and events. In contrast to feature fusion methods that solely occur in the spatial domain, the FCFE module fuses cross-modality features based on the spectral domain analysis [8, 15, 23, 26, 54, 67]. The events record bright changes of the scene, providing information about the high-frequency components of the latent sharp frame we aim to restore. Therefore, the high-frequency components of the latent sharp frame are similar to those of the events, as shown in Fig. 1. Therefore, to actively leverage these characteristics, we propose a method that efficiently fuses the frame and event modality by observing them in the spectral domain and analyzing their correlations between modalities. Moreover, the FCFE module performs adaptively enhance the features along with the spatial and channel dimension between different modalities in the frequency domain. As a result, we obtain a more robust cross-modal feature representation with the assistance of the frequency domain.

Furthermore, to contribute a real-world dataset to the community, we built a new dataset consisting of real-world motion blur with events, called the Real-world Event Video Deblurring (REVD) dataset. Our dataset contains high-quality and high-resolution motion-blurred images, corresponding sharp images, and events.

Our contributions can be summarized as: 1) We propose a novel frequency-aware event-based video deblurring framework. 2) We propose a Frequency-aware Cross-modal Feature Enhancement (FCFE) module to effectively enhance cross-modality features in the frequency domain. 3) We propose two temporal alignment modules, named Event-guided Local-windowed Temporal Propagation (ELTP) and Bidirectional Temporal Feature Fusion (BTFF), effectively leveraging the rich temporal information of events. 4) To the best of our knowledge, we provide the first real blur dataset for event-based video deblurring.

2. Related Works

2.1. Frame-based Image and Video Deblurring

The deep learning-based approaches [27, 29, 52, 57, 59] have demonstrated the potential to find non-uniform and variant blur kernels using a single image. They intricately designed networks using CNNs with various techniques such as attention [23, 41, 45, 58], multi-stage design [7, 57, 59], and multi-scale fusion [3, 12, 27]. Video deblurring [6, 28, 34, 66] has improved the quality of deblurring by leveraging temporal information of videos. It aggregates sharp patches from adjacent frames to fill in missing information from a single image. Most methods utilize optical flow [14, 32] and deformable convolution [19, 60] to perform temporal feature alignment between adjacent frames. However, when severe and continuous blur is present, it becomes challenging to consider temporal correlation among video frames, leading to limitations in temporal alignment performance improvement.

2.2. Event-based Motion Deblurring

Event cameras, a bio-inspired sensor, can record temporally dense information with pixel-wise intensity changes. Early works [17, 35, 36] succeeded in modeling relationships between a sharp image and a blurry image using the physical model-based formulation. Following event-based deblurring works [20, 24, 49], deep learning-based deblurring methods have been introduced to leverage the advantages of event data. In recent works, efforts have been directed towards designing more advanced architecture [42, 61] and addressing real-world scenarios more effectively [11, 43, 55, 56, 62], including challenges such as non-consecutive blurry videos [38] and videos with unknown exposure time [21]. Existing methods use dynamic filters [21, 24] or attention mechanisms [42] to fuse images and events from different modalities. Unlike these works, we propose a method that combines information between the two modalities in the frequency spectrum. Furthermore, while previous approaches mainly used a single image with corresponding events occurring during the exposure to restore a sharp image, we propose a method that effectively extracts the long-term temporal dependencies in videos by leveraging the rich temporal information from events.

3. Method

3.1. Overall Framework

Figure 2 illustrates the overall framework of the proposed method. We convert the event stream into voxel grid [65] representation, where voxel grid $\mathrm{E} \in \mathbb{R}^{B\times H\times W}$ with bin size B. Given T sequential blur frames $\{\mathrm{B}_i\}_{i=1}^{T}$ and corresponding voxel grid $\{\mathrm{E}_i\}_{i=1}^{T}$, we first extract features through 3D convolution with kernel size of $1 \times 3 \times 3$. Then, we reduce the spatial dimension of features to one-fourth using two 3D convolution layers with a kernel size of $3 \times 3 \times 3$ and a stride size of $1 \times 2 \times 2$. $\mathcal{F}(\mathrm{E})_i^s$ and $\mathcal{F}(\mathrm{B})_i^s$ represent the event and blur features, respectively, where s is the scale factor. The spatial dimensions of the

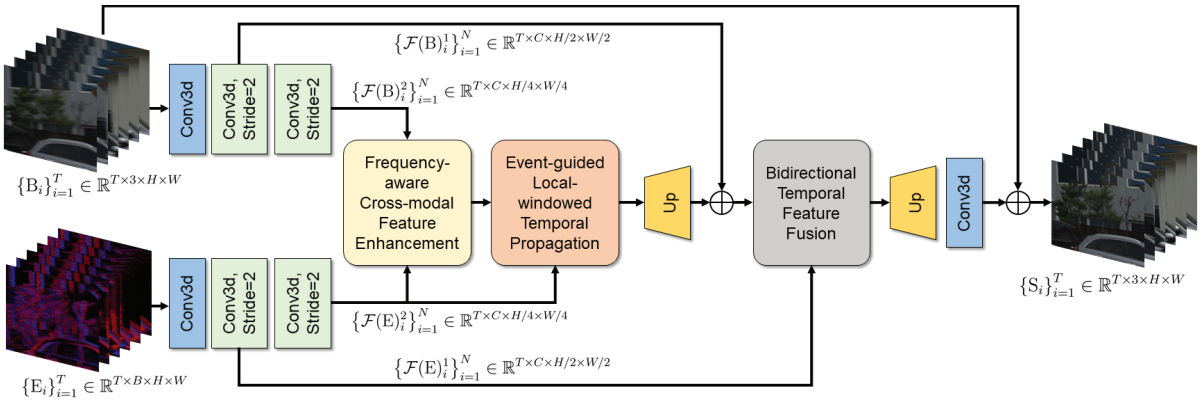

Figure 2. Overall framework of the proposed methods. Sequential blur frames and voxel grids are first processed through 3D convolutions to extract multiple-scale features. The Frequency-aware Cross-modal Feature Enhancement (FCFE) module performs cross-modal fusion in the frequency spectrum at the lowest resolution. Then, the Event-guided Local-windowed Temporal Propagation (ELTP) module aligns adjacent neighbor frames through temporal propagation. Finally, after upsampling to increase spatial resolution, the Bidirectional Temporal Feature Fusion (BTFF) facilitates interaction among all frames within a rich spatial context.

features at scale factor s are denoted as H_s and W_s, and $H_s = H/2^s, W_s = W/2^s$. After the encoding step, we extract useful information from each modality's features in the frequency domain through the Frequency-aware Cross-modal Feature Enhancement (FCFE) module. After that, at a scale factor s of 2, the Event-guided Local-windowed Temporal Propagation (ELTP) module aligns temporal features between adjacent video frames. Finally, at a scale factor s of 1, we perform temporal feature alignment with the help of the events through the Bidirectional Temporal Feature Fusion (BTFF) modules.

3.2. Frequency-aware Cross-Modal Feature Enhancement Module

Since the events provide clear motion information within the exposure time, effectively integrating this information with blur frames can significantly aid in restoring sharp frames. However, due to the inherent modality differences between events and blur frames, cross-modality feature fusion is challenging yet crucial in event-based deblurring tasks. The most straightforward way of cross-modality fusion is to use standard convolution with a fixed-sized kernel by utilizing two modality features. However, standard convolution can lead to sub-optimal feature fusion results as it does not consider the context differences between the two modalities as it applies a fixed-sized filter across the entire pixels. Therefore, for more reliable cross-modality fusion, we require a content-adaptive dynamic filter-based fusion method that can aggregate features based on spatial or channel content instead of using standard convolution. To achieve this goal, we first revisit the convolution theorem [31]. According to this theorem, the element-wise multiplication of two signals in the frequency domain is equivalent to the convolution between two signals. That is, performing frequency-domain filtering can be interpreted

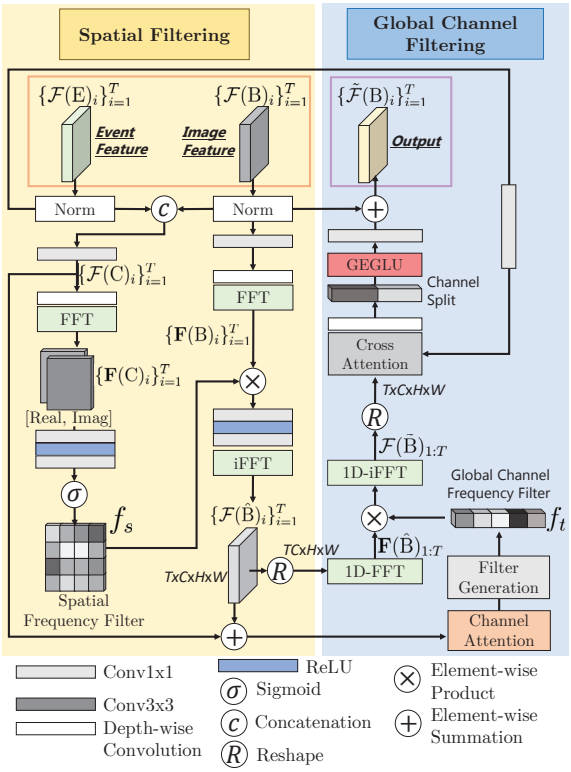

Figure 3. Frequency-aware Cross-modal Feature Enhancement (FCFE) modules.

as conducting dynamic filtering with a large receptive field size. Based on this fact, we propose a Frequency-aware Cross-Modal Feature Enhancement (FCFE) module to overcome the inherent differences between the two modalities.

Within the FCFE module, we aim to filter the blur features by leveraging event features in the frequency domain. Our module can be classified into two branches: spatial frequency filtering and global channel filtering. As shown in

Fig. 3, given T sequential blur features $\{\mathcal{F}(\mathrm{B})_i^s\}_{i=1}^T$ and event features $\{\mathcal{F}(\mathrm{E})_i^s\}_{i=1}^T$ with a scale factor s of 2, we first apply layer normalization [2] layers to each modality feature. We then concatenate these two features in the channel dimension to consider the correlation between blur and event features and apply a 1×1 convolution layer to get the correlated features $\{\mathcal{F}(\mathrm{C})_i\}_{i=1}^T$ (omit the scale factor for the simplicity), where $\mathcal{F}(\mathrm{C})_i \in \mathbb{R}^{C \times H_s \times W_s}$.

For spatial filtering, we perform a 2D Fast Fourier Transform (FFT) [30] for correlated features and concatenate the real and imaginary parts of the FFT results to obtain the frequency representation such as $\mathbf{F}(\mathrm{C})_i = \mathrm{FFT}(\mathcal{F}(\mathrm{C})_i)$, where $\mathbf{F}(\mathrm{C})_i \in \mathbb{R}^{2C \times H_s \times (\lfloor \frac{W_s}{2} \rfloor + 1)}$. Then, through convolution, ReLU, and Sigmoid layers, we generate a spatial frequency filter $f_s \in \mathbb{R}^{C \times H_s \times (\lfloor \frac{W_s}{2} \rfloor + 1)}$ to extract valuable information observed in the frequency spectrum. Similarly, by applying FFT with convolution layers to the blur feature, we obtain $\mathbf{F}(\mathrm{B})_i \in \mathbb{R}^{C \times H_s \times (\lfloor \frac{W_s}{2} \rfloor + 1)}$, and the proposed spatial filtering is calculated as follows:

$$\mathbf{F}(\hat{\mathrm{B}})_i = \mathbf{F}(\mathrm{B})_i \otimes f_s, \tag{1}$$

where $\otimes$ denotes element-wise multiplication. After applying convolution and ReLU layers, we transfer the spatially filtered feature $\mathcal{F}(\hat{\mathrm{B}})_i$ in the frequency space to the spatial domain using an inverse FFT as:

$$\mathcal{F}(\hat{\mathrm{B}})_i = \mathrm{iFFT}(\mathrm{ReLU}(\mathrm{Conv}_{1\times1}(\mathbf{F}(\hat{\mathrm{B}})_i))). \tag{2}$$

In spatial frequency filtering, f_s can be interpreted as a dynamic filtering operation with a large receptive field that extracts valuable information from cross-modalities. Hence, Eq. (1) more reliably enhances blur features by leveraging event features in the frequency domain.

We have extracted crucial cross-modality information from the spatial domain through spatial frequency filtering. However, the feature's channels also contain a substantial amount of information. To address this, we design a global channel frequency filtering mechanism that learns a specialized frequency filter. This module modulates the frequency feature spectrum based on the global channel context. For this, we first reshape the spatially filtered feature along with the channel dimension as $\{\mathcal{F}(\hat{\mathrm{B}})_i\}_i^T \in \mathbb{R}^{T \times C \times H_s \times W_s} \rightarrow \mathcal{F}(\hat{\mathrm{B}})_{1:T} \in \mathbb{R}^{(T \times C) \times H_s \times W_s}$, denoted $(T \times C)$ as M for simplicity. We apply 1D FFT along the flattened channel dimension to obtain the channel-wise frequency domain features as:

$$\mathbf{F}(\hat{\mathrm{B}})_{1:T} = \text{1D-FFT}(\mathcal{F}(\hat{\mathrm{B}})_{1:T}), \tag{3}$$

where $\mathbf{F}(\hat{\mathrm{B}})_{1:T} \in \mathbb{R}^{(\lfloor \frac{M}{2} \rfloor + 1) \times H_s \times W_s}$. To generate frequency filters for the global channel dimension, we add $\mathcal{F}(\mathrm{C})_i$ and $\mathcal{F}(\hat{\mathrm{B}})_i$ through skip connections to generate $\mathcal{F}(\mathrm{C})'_i = \mathcal{F}(\mathrm{C})_i + \mathcal{F}(\hat{\mathrm{B}})_i$ followed by channel attention [66] block, $\mathcal{A}_i \in \mathbb{R}^{C \times 1 \times 1}$.

The channel attended feature $\mathcal{F}(\hat{\mathrm{C}})_i$ can be obtained by multiplication $\mathcal{F}(\hat{\mathrm{C}})_i = \mathcal{F}(\mathrm{C})'_i \otimes \mathcal{A}_i$. Then, we also flatten the attended feature such as $\{\mathcal{F}(\hat{\mathrm{C}})_i\}_i^T \in \mathbb{R}^{T \times C \times H_s \times W_s} \rightarrow \mathcal{F}(\hat{\mathrm{C}})_{1:T} \in \mathbb{R}^{M \times H_s \times W_s}$. We apply a 1×1 convolution layer to the attended feature to calculate the channel filter, $f_t \in \mathbb{R}^{(\lfloor \frac{M}{2} \rfloor + 1) \times H_s \times W_s}$. The global channel-filtered feature can be calculated as:

$$\mathbf{F}(\tilde{\mathrm{B}})_{1:T} = \mathbf{F}(\hat{\mathrm{B}})_{1:T} \otimes f_t. \tag{4}$$

After that, we apply inverse 1D FFT as follows:

$$\mathcal{F}(\tilde{\mathrm{B}})_{1:T} = \text{1D-iFFT}(\mathbf{F}(\tilde{\mathrm{B}})_{1:T}). \tag{5}$$

f_t is also an adaptive filter that performs correlation in the global channel dimension through a 1×1 convolution, enabling discriminative channel filtering. Similar to f_s, according to the convolution theorem, Eq. (4) is equivalent to performing dynamic kernel operations for the global channel dimension. We reshape the dimensions of the feature as

$$\mathcal{F}(\tilde{\mathrm{B}})_{1:T} \in \mathbb{R}^{(T \times C) \times H_s \times W_s} \rightarrow \{\mathcal{F}(\tilde{\mathrm{B}})_i\}_i^T \in \mathbb{R}^{T \times C \times H_s \times W_s},$$

As spatial feature enhancement branch, we adopt a widely used transformer architecture [42, 58] as cross-attention to fuse $\mathcal{F}(\mathrm{E})_i$ and $\mathcal{F}(\tilde{\mathrm{B}})_i$ Finally, we generate the output feature $\tilde{\mathcal{F}}(\mathrm{B})_i$ by adding the results of the GEGLU function [39] to the original blur feature $\mathcal{F}(\mathrm{B})_i$ using a residual connection.

3.3. Event-guided Temporal Feature Alignment

Temporal alignments aim to mine valuable information from neighboring video frames. Conventional feature alignment methods explicitly use optical flows [51] and deformable convolutions [13, 50]. Despite the notable progress in frame-based feature alignment methods, there needs to be more research for aligning video frames by leveraging information from event streams. To better leverage the advantages of events, we first revisit the concept of event-based video interpolation works and utilize this concept to perform temporal alignment. Recently, event-based video interpolation works [22, 46, 47] have introduced the concept of "synthesis-based alignment" for interpolating supporting frames to the reference frames using both event and image modalities without optical flows. This concept allows us to leverage the advantages of the events without extensive computation of optical flows, enabling direct alignment of supporting features using the events. Inspired by these concepts, we devise two temporal feature alignment modules, Event-guided Local windowed Temporal Propagation (ELTP) and Bidirectional Temporal Feature Fusion (BTFF). The ELTP module performs temporal alignment using the information of neighboring video frames at relatively lower spatial resolution (scale s of 2). We then enhance the adjacent and current features through frequency

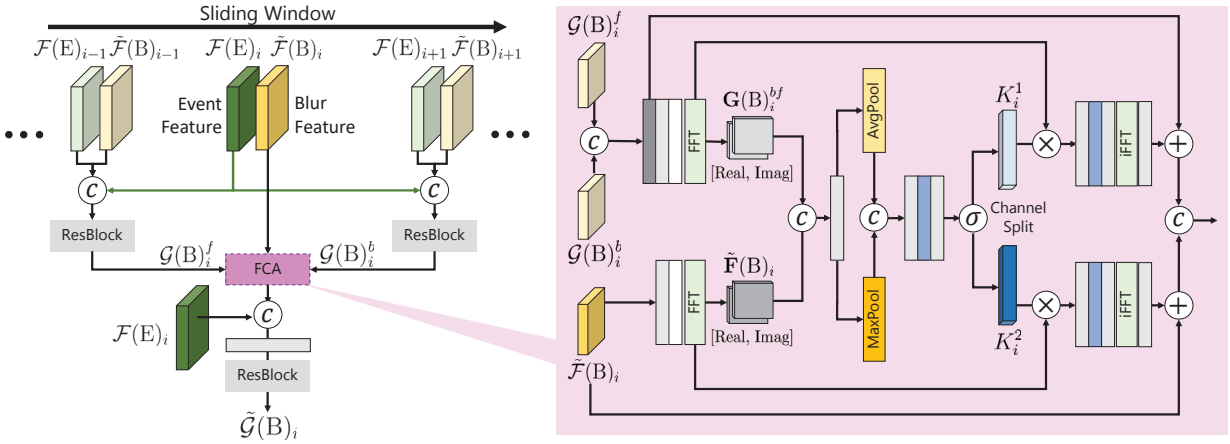

Figure 4. Event-guided Local-windowed Temporal Propagation (ELTP) module.

domain analysis to improve alignment. After the ELTP module, we introduce the BTFF module to improve deblurring effectiveness. It achieves this by conducting temporal alignment to leverage the abundant temporal information from events at elevated spatial resolutions (scale s of 1), utilizing bi-directional structures.

Event-guided Local-windowed Temporal Propagation. As illustrated in Fig. 4, we aim to simultaneously align the left and right video frame features $\tilde{\mathcal{F}}(\mathrm{B})_{i+1}, \tilde{\mathcal{F}}(\mathrm{B})_{i-1}$ with the reference blur video frame features $\tilde{\mathcal{F}}(\mathrm{B})_i$. In this module, we consider both left and right features to align a single current frame feature. In this way, we can simultaneously integrate information from both forward and backward directions, thereby reducing the negative influence of occlusion, compared to using information from only one direction. To this end, we first align the supporting blur frame features $\tilde{\mathcal{F}}(\mathrm{B})_{i+1}, \tilde{\mathcal{F}}(\mathrm{B})_{i-1}$. To align a single supporting blur feature, we need event features that correspond to the time duration from the starting point of the exposure time of the reference frame feature to the end of the exposure time of the supporting frame feature. Therefore, we perform synthesis-based feature alignment utilizing the supporting blur frame feature along with two event features ($\mathcal{F}(\mathrm{E})_i$ and $\mathcal{F}(\mathrm{E})_{i+k}$). That is, we concatenate the supporting blur feature $\tilde{\mathcal{F}}(\mathrm{B})_{i+k}$ with the current and supporting event features ($\mathcal{F}(\mathrm{E})_i$, $\mathcal{F}(\mathrm{E})_{i+k}$) with containing motion information of both the current and supporting blur frames during exposure time where $k \in \{-1, +1\}$. We then apply ResBlocks [18] to produce aligned features as:

$$\begin{aligned} \mathcal{G}(\mathrm{B})_i^f &= Res[\tilde{\mathcal{F}}(\mathrm{B})_{i-1}, \mathcal{F}(\mathrm{E})_i, \mathcal{F}(\mathrm{E})_{i-1}] \\ \mathcal{G}(\mathrm{B})_i^b &= Res[\tilde{\mathcal{F}}(\mathrm{B})_{i+1}, \mathcal{F}(\mathrm{E})_i, \mathcal{F}(\mathrm{E})_{i+1}] \end{aligned} \tag{6}$$

where $[\cdot]$ denotes channel-wise concatenation and Res denotes Resblocks and $\mathcal{G}(\mathrm{B})_i^f$, $\mathcal{G}(\mathrm{B})_i^b$ denote aligned feature from the forward and backward direction within local windows, respectively. In this way, we align the features of the left and right frame frames by leveraging the rich temporal contexts of the events. To achieve better alignment, we perform feature fusion between the current frame feature and the aligned frame features in the frequency domain. We concatenate the features of the two aligned features and apply several convolutions: $\mathcal{G}(\mathrm{B})_i^{bf} = \mathrm{Conv}_{3\times3}([\mathcal{G}(\mathrm{B})_i^b, \mathcal{G}(\mathrm{B})_i^f])$. We then apply FFT transforms from the aligned feature $\mathcal{G}(\mathrm{B})_i^{bf}$ to $\mathbf{G}(\mathrm{B})_i^{bf}$ and from the current feature $\tilde{\mathcal{F}}(\mathrm{B})_i$ to $\tilde{\mathbf{F}}(\mathrm{B})_i$. Following this step, we concatenate the real and imaginary parts of each transformed feature $\tilde{\mathbf{F}}(\mathrm{B})_i$, $\mathbf{G}(\mathrm{B})_i^{bf}$ to make the feature Y_F as $Y_F = \mathrm{Conv}_{1\times1}[\mathcal{R}(\mathbf{G}(\mathrm{B})_i^{bf}), \mathcal{I}(\mathbf{G}(\mathrm{B})_i^{bf}), \mathcal{R}(\tilde{\mathbf{F}}(\mathrm{B})_i), \mathcal{I}(\tilde{\mathbf{F}}(\mathrm{B})_i)]$ where $\mathcal{R}$ and $\mathcal{I}$ represent the real and imaginary parts of the transformed feature. To apply the channel attention for each feature in the frequency domain, we aim to estimate the frequency-domain channel attention map.

$$K_i = \sigma(F_{\mathrm{conv}}([\mathrm{AvgPool}(Y_F), \mathrm{MaxPool}(Y_F)])), \tag{7}$$

where F_{conv} represents "$\mathrm{Conv}_{1\times1} - \mathrm{ReLU} - \mathrm{Conv}_{1\times1}$" function. We then perform a split feature K_i along the channel dimension resulting in two independent channel-wise features, K_i^1 and K_i^2. We apply channel attention to each blur frame feature $\mathbf{G}(\mathrm{B})_i^{bf}$ and $\tilde{\mathbf{F}}(\mathrm{B})_i$ in the frequency domain. Additionally, we apply F_{conv} operation for each feature and then conduct an inverse FFT. In this way, we can effectively fuse the advantages of the frequency domain for each feature to combine temporal features. Afterward, we apply resblocks to the enhanced two features in the frequency domain and perform concatenation with the current event feature to carry out the final alignment results, $\tilde{\mathcal{G}}(\mathrm{B})_i$.

Bidirectional Temporal Feature Fusion. The BTFF module serves two distinct purposes compared to the ELTP module: (1) The ELTP module focuses solely on adjacent local information. In contrast, BTFF aims to convey long-range and non-local information to all blurred images. (2) While ELTP extracts high channel-dimensional features at the lowest resolution, leveraging the advantages of fre-

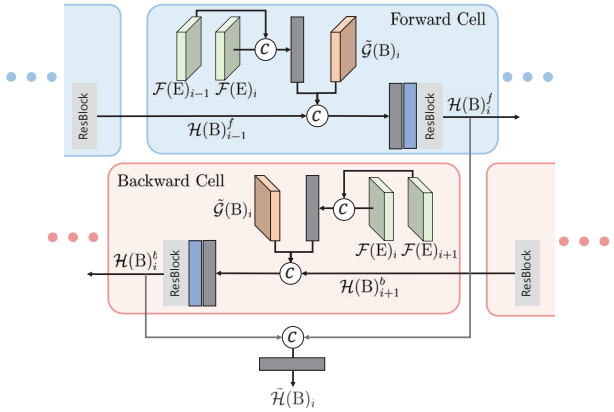

Figure 5. Bidirectional Temporal Feature Fusion (BTFF) module.

quency domain for video alignment to restore highly fine and detailed portions, BTFF operates at more higher resolution, targeting overall content-based temporal consistency maintenance, utilizing rich spatial context. To effectively leverage long-term temporal information, BTFF module is designed with a bidirectional structure that allows for the better utilization of non-local frame information. We illustrate the process of BTFF module at time i in Fig. 5. The BTFF utilizes features with a scale factor s of 1 to leverage rich spatial context. To align features from different time steps, we fuse events for the current time, i, with events from the target previous $i-1$ or subsequent $i+1$ to obtain the event feature $\mathcal{S}(E)_i^{\{f,b\}}$ for alignment. Note that extracting additional event features for alignment allows us to synthesize the aligned features better since they contain explicit motion information between frames. The forward fused features, $\mathcal{H}(\mathrm{B})_i^f$, considering the previous fused features, are calculated as follows:

$$\begin{aligned} \mathcal{S}(\mathrm{E})_i^f &= \mathrm{Conv}_{3\times3}([\mathcal{F}(\mathrm{E})_{i-1}, \mathcal{F}(\mathrm{E})_i]), \\ \mathcal{H}(\mathrm{B})_i^f &= F_{Res}^f([\tilde{\mathcal{G}}(\mathrm{B})_i, \mathcal{S}(\mathrm{E})_i^f, \mathcal{H}(\mathrm{B})_{i-1}^f]), \end{aligned} \tag{8}$$

where F_{Res}^f denotes "$\mathrm{Conv}_{3\times3} - \mathrm{ReLU} - \mathrm{ResBlocks}$" function for forward. Similarly, backward fused features, $\mathcal{H}(\mathrm{B})_i^b$, are obtained as:

$$\begin{aligned} \mathcal{S}(\mathrm{E})_i^b &= \mathrm{Conv}_{3\times3}([\mathcal{F}(\mathrm{E})_i, \mathcal{F}(\mathrm{E})_{i+1}]), \\ \mathcal{H}(\mathrm{B})_i^b &= F_{Res}^b([\tilde{\mathcal{G}}(\mathrm{B})_i, \mathcal{S}(\mathrm{E})_i^b, \mathcal{H}(\mathrm{B})_{i+1}^b]), \end{aligned} \tag{9}$$

and we aggregate the forward and backward fused features:

$$\tilde{\mathcal{H}}(\mathrm{B})_i = \mathrm{Conv}_{3\times3}([\mathcal{H}(\mathrm{B})_i^b, \mathcal{H}(\mathrm{B})_i^f]) \tag{10}$$

Finally, we obtain the sharp latent frames as:

$$\mathrm{S}_i = \mathrm{Conv}_{3\times3}(U(\tilde{\mathcal{H}}(\mathrm{B})_i)) + \mathrm{B}_i,\ i = \{1, ..., N\} \tag{11}$$

where U denotes the transposed 2D convolution block to upsample features. Finally, we obtain the estimated sharp video frames $\{\mathrm{S}_i\}_{i=1}^T$.

Table 1. Comparison of our REVD dataset with publicly available REBlur dataset [42], recorded by DAVIS346 camera.

	REVD (Ours)	REBlur [42]
№ total frames	**6.3 k**	1.5 k
Image Resolution	**1024 × 768**	346 × 260
№ frames in Seq. (Min)	299	6
Blur type	Real	Real
Event type	Real	Real
Color	✓	✗

4. Real-world Event Video Deblurring Dataset

Several event-based deblurring works have employed sequential sharp frames to synthesize blurred images using GoPro [21, 24, 44] or DAVIS [21]. In the case of synthetic deblurring datasets, we often observe the unrealistic blur frames (*e.g.*, shutter effects) due to a discrete averaging process using consecutive sharp images. Due to these discrepancies between synthetic and real-world blur frames, deep learning models trained on the synthetic deblurring dataset show limited generalization ability to real-world blurry video frames. Therefore, we need real-world deblurring datasets with synchronized real-events for the generalization ability on the real-world blurry videos.

Recently, Sun *et al*. [42] introduced an event-based real-world deblurring dataset by capturing blur frames, corresponding sharp frames, and actual events with a DAVIS-346 camera. However, as shown in Tab. 1, REBlur dataset [42] has a limited number of frames in the sequence, so it is not suitable for an event-based video deblurring research since we can not evaluate the model by leveraging long-range temporal dependency of videos. Furthermore, DAVIS-346 camera has low image quality and resolution. Therefore, high resolution and quality event-based real-world video deblurring datasets are essential for evaluating event-based video deblurring research communities.

To this end, we design a hybrid camera system where two same FLIR BlackFly cameras and one Prophesee Gen4 event camera are co-axis aligned utilizing a two-way 50:50 beam-splitter system as illustrated in Fig. 6. The three cameras are hardware-level temporally synchronized with a microcontroller system. Then, we set a longer exposure time for the capturing blurry video frames (32ms) compared to the sharp video frames (4ms). We adjust the irradiance intensity of the camera for capturing blurry videos to be 1/8 of that of the camera for sharp videos in order to maintain a consistent irradiance intensity. To account for the additional half reduction in irradiance intensity as it passes through the beam-splitter, a 25% neutral density filter is physically inserted in front of the camera for capturing blurry videos.

5. Experiments

5.1. Datasets

GoPro Dataset. We evaluate the proposed method on the GoPro dataset [27], which has been widely used in existing

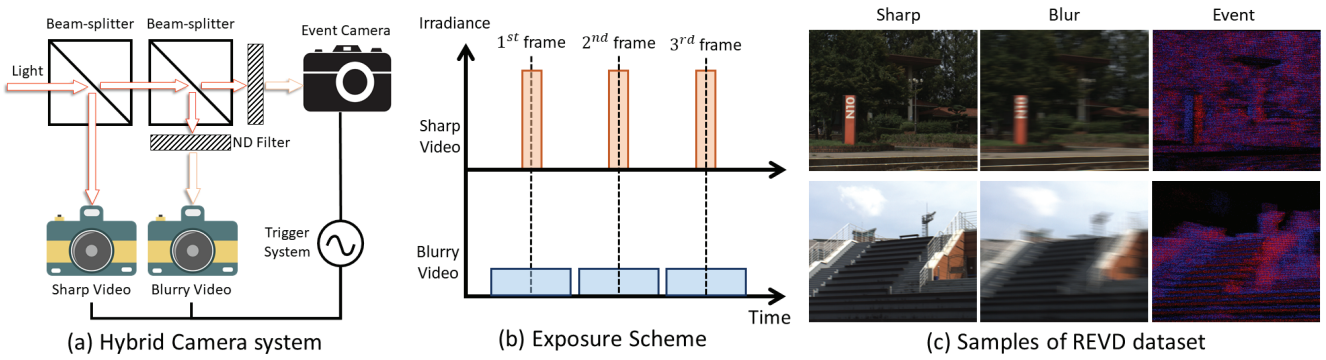

Figure 6. Camera system for acquiring the Real-world Event Video Deblurring (REVD) dataset. As shown in (b), to equalize the total illumination received by cameras due to different exposure times, we install a neutral density filter in front of the camera with a longer exposure time, thus reducing the irradiance.

Table 2. Quantitative evaluations on the GoPro dataset. † represents the results we obtained from training the networks using the same event raw data and representation as ours. All other results are obtained from the original paper.

Frame-based	IFRNN [28]	ESTRNN [64]	CDVD-TSP [33]	MMP-RNN [51]	STDAN [60]	ERDN [19]	DSTNet [34]
PSNRs	29.97	31.07	31.67	32.64	32.29	32.48	34.16
SSIMs	0.895	0.902	0.928	0.936	0.931	0.933	0.968
Event-based	eSL-Net† [48]	REDNet† [56]	D^{2}Nets [38]	UEVD† [21]	EFNet [42]	EFNet† [42]	Ours
PSNRs	28.69	35.07	31.60	35.93	35.46	35.87	**36.70**
SSIMs	0.908	0.971	0.940	0.975	0.972	0.974	**0.978**

event-based deblurring researches [21, 38, 42]. We used the event simulator (ESIM) [37] for generating synthetic events and performing comparisons with other methods.

REVD Dataset provides 21 sequences, including dynamic scenes and extreme blurs. We use 13 sequences for training and 8 sequences for the test set. The image and event resolution is 1024×768. We capture videos of typical urban scenes encompassing diverse motion modes, including ego-motion, object motion, and a combination of both.

5.2. Implementation Details

We set the batch size of 8 and bin size of event voxels as 16. We randomly crop the images and event voxels to 256×256 for the same locations in training time. All networks are trained by utilizing AdamW [25] optimizer with an initial learning rate of 1×10^{-4} using the charbonnier loss [5].

5.3. Comparison on Synthetic Datasets

Table 2 presents the quantitative results on the GoPro dataset. Since the performance results for some existing works [21, 48, 56] on GoPro were not publicly available, we train those models from scratch ourselves, denoted by †, using the same event representation as ours for a fair comparison. The majority of event-based methods outperform frame-based single and video-based methods in terms of performance. Among them, our method achieves the highest performance, leveraging the ability to utilize multi-frame information, which results in a difference of 0.83∼8.01 dB compared to existing event-based methods.

Table 3. Quantitative evaluations on the REVD dataset.

Methods	Event	PSNR	SSIM	Time (ms)	FLOPs (T)
STDAN [60]	✗	29.84	0.8893	929	57.32
ESTRNN [64]	✗	30.51	0.9049	177	6.38
RNN-MBP [66]	✗	31.77	0.9212	2238	153.5
eSL-Net [48]	✓	26.99	0.7868	6	0.25
EFNet [42]	✓	31.75	0.9208	244	13.00
REDNet [56]	✓	31.90	0.9207	479	19.07
UEVD [21]	✓	31.97	0.9211	598	39.03
Ours	✓	**32.99**	**0.9326**	484	30.27

5.4. Comparison on Real-blur Datasets

As shown in Table 3, we compare our method with other frame- and event-based methods on the REVD dataset. We provide the inference times for each model running on an NVIDIA RTX A6000 GPU for a single frame at a resolution of 1024×768, along with the Floating Point Operations (FLOPs) for processing five video clips. Event-based deblurring methods generally outperform frame-based video deblurring methods. However, RNN-MBP [66] shows comparable performance to event-based methods, *i.e.*, 31.77 dB, which is 0.02 dB higher than EFNet [42]. Due to factors such as the inherent noise in events being much more severe in real-world datasets compared to synthetic data, and the presence of saturated pixels that cannot be restored by blur kernels within a single frame, these degradations make it challenging to restore even with the use of events. However, RNN-MBP is computationally inefficient due to the con-

Figure 7. Qualitative comparison of our methods with other methods on the REVD dataset. Please zoom in for a better visualization.

siderable operations required for accurate alignment. On the other hand, our method effectively leverages the event modality, especially in the frequency domain, to restore fine structures. With the help of two temporal alignments based on the properties of the video, our approach achieves superior performance, with a PSNR 1.02 dB higher than the second-best method, UEVD [21]. Furthermore, our approach features an efficient design leveraging events, resulting in time and FLOP requirements that are significantly lower than the best frame-based method, RNN-MBP, and even more efficient than the second-best model, UEVD. Figure 7 illustrates highly challenging blur scenarios. The results of our method are more realistic and clean than other methods.

6. Ablation Study

Table 4 presents the ablation study for each module. Starting from the baseline, excluding all modules, we conduct an ablation study by incrementally adding each module. For FCFE, 3D convolutions with concatenation are used as a replacement, while other modules that can be freely removed.

Frequency-aware Cross-modal Feature Enhancement. FCFE performs feature fusion between the two modalities in the frequency domain. This effect is distinctly evident when compared to the baseline. The baseline performs fusion between events and images solely in the spatial domain using a 3D convolution block but exhibits significantly lower performance. In contrast, adding FCFE module to this baseline results in a 2.25 dB improvement in PSNR. We also note an improvement in performance ranging from 0.23 to 0.52 dB compared to models utilizing alignment techniques like ELTP, BTFF, or a combination of both.

Event-guided Temporal Feature Alignment. We perform temporal feature alignment through the ELTP and BTFF modules. Examining the results of models with and without these alignments highlights the importance of temporal feature alignment in video deblurring. Compared to the baseline, ELTP and BTFF yield improvements of 4.22 dB and 4.94 dB in PSNR, respectively. While BTFF may have a performance advantage over ELTP due to its ability to utilize rich context information in the spatial domain at higher resolutions, the significance of ELTP lies in its capability to perform alignment in both the spatial and frequency domains, even at the lowest resolution, enabling efficient restoration of fine details. Hence, when utilized together, their performance exceeds that of BTFF alone by an improvement of 0.37 dB. Finally, when both cross-modal feature enhancement and temporal feature alignment are present, they collectively deliver the best performance, thereby validating the efficacy of each module.

Table 4. Ablation study on the GoPro dataset.

FCFE		✓		✓			✓	✓
ELTP			✓	✓		✓		✓
BTFF					✓	✓	✓	✓
PSNR	31.14	33.39	35.36	35.59	36.08	36.45	<u>36.60</u>	**36.70**

7. Conclusion

We propose a framework for event-based video deblurring, including cross-modal feature enhancement and event-guided temporal feature alignment. Specifically, we have developed modules in which we can effectively leverage the advantages of events and frames in the frequency domain. Our approach significantly outperforms existing state-of-the-art methods. Additionally, we provide, for the first time, a real-world dataset for the event-based video deblurring communities, acquired using an RGB-Event hybrid camera system.

Acknowledgements This research was supported by National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NRF2022R1A2B5B03002636) and the Challengeable Future Defense Technology Research and Development Program through the Agency For Defense Development (ADD) funded by the Defense Acquisition Program Administration (DAPA) in 2024 (No.912768601).

References

[1] Dawit Mureja Argaw, Junsik Kim, Francois Rameau, and In So Kweon. Motion-blurred video interpolation and extrapolation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 901–910, 2021. 1

[2] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 4

[3] Stephan Brehm, Sebastian Scherer, and Rainer Lienhart. High-resolution dual-stage multi-level feature aggregation for single image and video deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2020. 2

[4] Jiahang Cao, Xu Zheng, Yuanhuiyi Lyu, Jiaxu Wang, Renjing Xu, and Lin Wang. Chasing day and night: Towards robust and efficient all-day object detection guided by an event camera. *arXiv preprint arXiv:2309.09297*, 2023. 1

[5] Pierre Charbonnier, Laure Blanc-Feraud, Gilles Aubert, and Michel Barlaud. Two deterministic half-quadratic regularization algorithms for computed imaging. In *Proceedings of 1st international conference on image processing*, pages 168–172. IEEE, 1994. 7

[6] Huaijin Chen, Jinwei Gu, Orazio Gallo, Ming-Yu Liu, Ashok Veeraraghavan, and Jan Kautz. Reblur2deblur: Deblurring videos via self-supervised learning. In *2018 IEEE International Conference on Computational Photography (ICCP)*, pages 1–9. IEEE, 2018. 2

[7] Liangyu Chen, Xin Lu, Jie Zhang, Xiaojie Chu, and Chengpeng Chen. Hinet: Half instance normalization network for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 182–192, 2021. 2

[8] Lu Chi, Borui Jiang, and Yadong Mu. Fast fourier convolution. *Advances in Neural Information Processing Systems*, 33:4479–4488, 2020. 2

[9] Hoonhee Cho and Kuk-Jin Yoon. Event-image fusion stereo using cross-modality feature propagation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 454–462, 2022. 1

[10] Hoonhee Cho and Kuk-Jin Yoon. Selection and cross similarity for event-image deep stereo. In *European Conference on Computer Vision*, pages 470–486. Springer, 2022. 1

[11] Hoonhee Cho, Yuhwan Jeong, Taewoo Kim, and Kuk-Jin Yoon. Non-coaxial event-guided motion deblurring with spatial alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12492–12503, 2023. 2

[12] Sung-Jin Cho, Seo-Won Ji, Jun-Pyo Hong, Seung-Won Jung, and Sung-Jea Ko. Rethinking coarse-to-fine approach in single image deblurring. *arXiv preprint arXiv:2108.05054*, 2021. 2

[13] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017. 4

[14] Shengyang Dai and Ying Wu. Motion from blur. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008. 2

[15] Jiangxin Dong, Jinshan Pan, Zhongbao Yang, and Jinhui Tang. Multi-scale residual low-pass filter network for image deblurring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12345–12354, 2023. 2

[16] Hongyun Gao, Xin Tao, Xiaoyong Shen, and Jiaya Jia. Dynamic scene deblurring with parameter selective sharing and nested skip connections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1

[17] Chen Haoyu, Teng Minggui, Shi Boxin, Wang YIzhou, and Huang Tiejun. Learning to deblur and generate high frame rate video with an event camera. *arXiv preprint arXiv:2003.00847*, 2020. 2

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[19] Bangrui Jiang, Zhihuai Xie, Zhen Xia, Songnan Li, and Shan Liu. Erdn: Equivalent receptive field deformable network for video deblurring. In *European Conference on Computer Vision*, pages 663–678. Springer, 2022. 2, 7

[20] Zhe Jiang, Yu Zhang, Dongqing Zou, Jimmy Ren, Jiancheng Lv, and Yebin Liu. Learning event-based motion deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3320–3329, 2020. 1, 2

[21] Taewoo Kim, Jeongmin Lee, Lin Wang, and Kuk-Jin Yoon. Event-guided deblurring of unknown exposure time videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 1, 2, 6, 7, 8

[22] Taewoo Kim, Yujeong Chae, Hyun-Kurl Jang, and Kuk-Jin Yoon. Event-based video frame interpolation with cross-modal asymmetric bidirectional motion fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18032–18042, 2023. 4

[23] Lingshun Kong, Jiangxin Dong, Jianjun Ge, Mingqiang Li, and Jinshan Pan. Efficient frequency domain-based transformers for high-quality image deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5886–5895, 2023. 2

[24] Songnan Lin, Jiawei Zhang, Jinshan Pan, Zhe Jiang, Dongqing Zou, Yongtian Wang, Jing Chen, and Jimmy SJ Ren. Learning event-driven video deblurring and interpolation. In *ECCV (8)*, pages 695–710, 2020. 2, 6

[25] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 7

[26] Xintian Mao, Yiming Liu, Fengze Liu, Qingli Li, Wei Shen, and Yan Wang. Intriguing findings of frequency selection for image deblurring. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1905–1913, 2023. 2

[27] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3883–3891, 2017. 1, 2, 6

[28] Seungjun Nah, Sanghyun Son, and Kyoung Mu Lee. Recurrent neural networks with intra-frame iterations for video deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8102–8111, 2019. 2, 7

[29] Mehdi Noroozi, Paramanand Chandramouli, and Paolo Favaro. Motion deblurring in the wild. In *GCPR*, pages 65–77. Springer, 2017. 2

[30] Henri J Nussbaumer and Henri J Nussbaumer. *The fast Fourier transform*. Springer, 1982. 4

[31] Alan V Oppenheim, Alan S Willsky, Syed Hamid Nawab, and Jian-Jiun Ding. *Signals and systems*. Prentice hall Upper Saddle River, NJ, 1997. 3

[32] Jinshan Pan, Haoran Bai, and Jinhui Tang. Cascaded deep video deblurring using temporal sharpness prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[33] Jinshan Pan, Haoran Bai, and Jinhui Tang. Cascaded deep video deblurring using temporal sharpness prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3043–3051, 2020. 7

[34] Jinshan Pan, Boming Xu, Jiangxin Dong, Jianjun Ge, and Jinhui Tang. Deep discriminative spatial and temporal network for efficient video deblurring. In *The IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2023. 2, 7

[35] Liyuan Pan, Cedric Scheerlinck, Xin Yu, Richard Hartley, Miaomiao Liu, and Yuchao Dai. Bringing a blurry frame alive at high frame-rate with an event camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6820–6829, 2019. 2

[36] Liyuan Pan, Richard Hartley, Cedric Scheerlinck, Miaomiao Liu, Xin Yu, and Yuchao Dai. High frame rate video reconstruction based on an event camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 2

[37] Henri Rebecq, Daniel Gehrig, and Davide Scaramuzza. Esim: an open event camera simulator. In *Conference on Robot Learning*, pages 969–982. PMLR, 2018. 7

[38] Wei Shang, Dongwei Ren, Dongqing Zou, Jimmy S. Ren, Ping Luo, and Wangmeng Zuo. Bringing events into video deblurring with non-consecutively blurry frames. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4531–4540, 2021. 2, 7

[39] Noam Shazeer. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020. 4

[40] Wang Shen, Wenbo Bao, Guangtao Zhai, Li Chen, Xiongkuo Min, and Zhiyong Gao. Blurry video frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5114–5123, 2020. 1

[41] Maitreya Suin, Kuldeep Purohit, and AN Rajagopalan. Spatially-attentive patch-hierarchical network for adaptive motion deblurring. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3606–3615, 2020. 2

[42] Lei Sun, Christos Sakaridis, Jingyun Liang, Qi Jiang, Kailun Yang, Peng Sun, Yaozu Ye, Kaiwei Wang, and Luc Van Gool. Event-based fusion for motion deblurring with cross-modal attention. In *European Conference on Computer Vision (ECCV)*, 2022. 1, 2, 4, 6, 7

[43] Lei Sun, Christos Sakaridis, Jingyun Liang, Peng Sun, Jiezhang Cao, Kai Zhang, Qi Jiang, Kaiwei Wang, and Luc Van Gool. Event-based frame interpolation with ad-hoc deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18043–18052, 2023. 1, 2

[44] Minggui Teng, Chu Zhou, Hanyue Lou, and Boxin Shi. Nest: Neural event stack for event-based image enhancement. In *European Conference on Computer Vision*, pages 660–676. Springer, 2022. 6

[45] Fu-Jen Tsai, Yan-Tsung Peng, Chung-Chi Tsai, Yen-Yu Lin, and Chia-Wen Lin. Banet: a blur-aware attention network for dynamic scene deblurring. *IEEE Transactions on Image Processing*, 31:6789–6799, 2022. 2

[46] Stepan Tulyakov, Daniel Gehrig, Stamatios Georgoulis, Julius Erbach, Mathias Gehrig, Yuanyou Li, and Davide Scaramuzza. TimeLens: Event-based video frame interpolation. *IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 4

[47] Stepan Tulyakov, Alfredo Bochicchio, Daniel Gehrig, Stamatios Georgoulis, Yuanyou Li, and Davide Scaramuzza. Time Lens++: Event-based frame interpolation with nonlinear parametric flow and multi-scale fusion. *IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 4

[48] Bishan Wang, Jingwei He, Lei Yu, Gui-Song Xia, and Wen Yang. Event enhanced high-quality image recovery. In *European Conference on Computer Vision*. Springer, 2020. 7

[49] Bishan Wang, Jingwei He, Lei Yu, Gui-Song Xia, and Wen Yang. Event enhanced high-quality image recovery. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, pages 155–171. Springer, 2020. 2

[50] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019. 4

[51] Yusheng Wang, Yunfan Lu, Ye Gao, Lin Wang, Zhihang Zhong, Yinqiang Zheng, and Atsushi Yamashita. Efficient video deblurring guided by motion magnitude. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 4, 7

[52] Patrick Wieschollek, Michael Hirsch, Bernhard Scholkopf, and Hendrik Lensch. Learning blind motion deblurring. In *ICCV*, pages 231–240, 2017. 2

[53] Ruihao Xia, Chaoqiang Zhao, Meng Zheng, Ziyan Wu, Qiyu Sun, and Yang Tang. Cmda: Cross-modality domain adaptation for nighttime semantic segmentation. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 21515–21524, 2023. 1

[54] Fengze Liu Qingli Li Wei Shen Xintian Mao, Yiming Liu and Yan Wang. Intriguing findings of frequency selection for image deblurring. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence*, 2023. 2

[55] Fang Xu, Lei Yu, Bishan Wang, Wen Yang, Gui-Song Xia, Xu Jia, Zhendong Qiao, and Jianzhuang Liu. Motion deblurring with real events. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2583–2592, 2021. 2

[56] Fang Xu, Lei Yu, Bishan Wang, Wen Yang, Gui-Song Xia, Xu Jia, Zhendong Qiao, and Jianzhuang Liu. Motion deblurring with real events. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2583–2592, 2021. 2, 7

[57] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Multi-stage progressive image restoration. *arXiv preprint arXiv:2102.02808*, 2021. 2

[58] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5728–5739, 2022. 2, 4

[59] Hongguang Zhang, Yuchao Dai, Hongdong Li, and Piotr Koniusz. Deep stacked hierarchical multi-patch network for image deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5978–5986, 2019. 1, 2

[60] Huicong Zhang, Haozhe Xie, and Hongxun Yao. Spatio-temporal deformable attention network for video deblurring. In *ECCV*, 2022. 2, 7

[61] Xiang Zhang and Lei Yu. Unifying motion deblurring and frame interpolation with events. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17765–17774, 2022. 2

[62] Xiang Zhang, Lei Yu, Wen Yang, Jianzhuang Liu, and Gui-Song Xia. Generalizing event-based motion deblurring in real-world scenarios. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10734–10744, 2023. 1, 2

[63] Youjian Zhang, Chaoyue Wang, and Dacheng Tao. Video frame interpolation without temporal priors. *Advances in Neural Information Processing Systems*, 33, 2020. 1

[64] Zhihang Zhong, Ye Gao, Yinqiang Zheng, and Bo Zheng. Efficient spatio-temporal recurrent neural network for video deblurring. In *European Conference on Computer Vision*, pages 191–207. Springer, 2020. 7

[65] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Unsupervised event-based learning of optical flow, depth, and egomotion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 989–997, 2019. 2

[66] Chao Zhu, Hang Dong, Jinshan Pan, Boyang Liang, Yuhao Huang, Lean Fu, and Fei Wang. Deep recurrent neural network with multi-scale bi-directional propagation for video deblurring. In *Proceedings of the AAAI conference on artificial intelligence*, pages 3598–3607, 2022. 2, 4, 7

[67] Qi Zhu, Man Zhou, Naishan Zheng, Chongyi Li, Jie Huang, and Feng Zhao. Exploring temporal frequency spectrum in deep video deblurring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12428–12437, 2023. 2