Event-Enhanced Blurry Video Super-Resolution

Dachun Kai[1], **Yueyi Zhang**[1,2*], **Jin Wang**[1], **Zeyu Xiao**[3], **Zhiwei Xiong**[1,2], **Xiaoyan Sun**[1,2]

[1]MoE Key Laboratory of Brain-inspired Intelligent Perception and Cognition,
University of Science and Technology of China
[2]Institute of Artificial Intelligence, Hefei Comprehensive National Science Center
[3]National University of Singapore
{dachunkai, jin01wang}@mail.ustc.edu.cn, zeyuxiao@nus.edu.sg, {zhyuey, zwxiong, sunxiaoyan}@ustc.edu.cn

Abstract

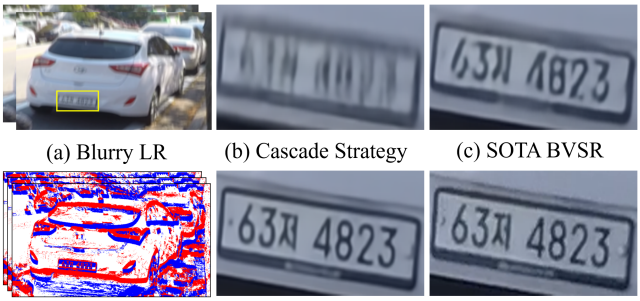
In this paper, we tackle the task of blurry video super-resolution (BVSR), aiming to generate high-resolution (HR) videos from low-resolution (LR) and blurry inputs. Current BVSR methods often fail to restore sharp details at high resolutions, resulting in noticeable artifacts and jitter due to insufficient motion information for deconvolution and the lack of high-frequency details in LR frames. To address these challenges, we introduce event signals into BVSR and propose a novel event-enhanced network, Ev-DeblurVSR. To effectively fuse information from frames and events for feature deblurring, we introduce a reciprocal feature deblurring module that leverages motion information from intra-frame events to deblur frame features while reciprocally using global scene context from the frames to enhance event features. Furthermore, to enhance temporal consistency, we propose a hybrid deformable alignment module that fully exploits the complementary motion information from inter-frame events and optical flow to improve motion estimation in the deformable alignment process. Extensive evaluations demonstrate that Ev-DeblurVSR establishes a new state-of-the-art performance on both synthetic and real-world datasets. Notably, on real data, our method is +2.59dB more accurate and 7.28× faster than the recent best BVSR baseline FMA-Net.

Code — https://github.com/DachunKai/Ev-DeblurVSR

Introduction

Video super-resolution (VSR) aims to recover a high-resolution (HR) video from its low-resolution (LR) counterpart. While existing methods (Xu et al. 2024; Zhou et al. 2024) get good results for general videos, they struggle with hard cases involving severe motion blur. Yet, such a setting is very common in practical VSR applications, like sports broadcasting (Liu et al. 2021) and video surveillance (Shamsolmoali et al. 2019). For example, in sports videos, fast-moving objects often cause unwanted motion blur.

To achieve VSR from a blurry video, *i.e.*, blurry VSR (BVSR), a straightforward approach is to perform video deblurring, followed by VSR methods, which we refer to as the *cascade* strategy. However, this approach has a drawback in that the pixel errors introduced in the deblurring stage are propagated and amplified in the subsequent VSR step, thus degrading the overall performance. To address this, some works (Fang and Zhan 2022; Youk, Oh, and Kim 2024) have proposed joint learning methods of VSR and deblurring. For instance, FMA-Net (Youk, Oh, and Kim 2024) proposes jointly estimating the degradation and restoration kernels through sophisticated representation learning. However, as shown in Fig. 1, these methods still suffer from blurry artifacts, jitter effects, and temporal aliasing.



(a) Blurry LR (b) Cascade Strategy (c) SOTA BVSR
(d) Events **(e) Ours** (f) GT

Figure 1: An example (a) from a challenging blurry video enhanced by (b) SOTA methods in video deblurring (Zhang, Xie, and Yao 2024) + VSR (Xu et al. 2024); (c) a SOTA BVSR method (Youk, Oh, and Kim 2024); and (e) our event-enhanced approach. Our method can restore the license plate with finer details, sharper edges, and less aliasing.

Relying solely on blurry LR frames to restore high-quality HR videos is a highly ill-posed problem. This is due to the inherent lack of motion information needed to deconvolve blurred images and the lack of high-frequency details in LR frames. Recently, event signals captured by event cameras have been used for image deblurring (Xu et al. 2021; Yang et al. 2023). Compared to standard cameras, event cameras have very high temporal resolution, high dynamic range (Gallego et al. 2020), and rich "moving edge" information (Mitrokhin et al. 2020). These characteristics enable events to provide complementary motion information, as well as high-frequency details, for BVSR. Motivated by these advantages, we propose including event signals as auxiliary information to enhance BVSR performance.

In this paper, we present Ev-DeblurVSR, a novel event-enhanced network for BVSR. To effectively fuse informa-

*Corresponding author.

tion from frames and events, we first categorize events into intra-frame and inter-frame events. Intra-frame events provide valuable motion and high-frequency information during the frames' exposure time, aiding in deblurring frame features. Frames, in turn, offer global scene context, further enhancing event features. This synergy motivates our Reciprocal Feature Deblurring (RFD) module. Inter-frame events capture continuous motion between frames, which is crucial for temporal consistency. For this purpose, we propose a Hybrid Deformable Alignment (HDA) module that combines inter-frame event information with optical flow for superior motion estimation in the deformable alignment process. Experimental results on three datasets demonstrate the effectiveness of our proposed Ev-DeblurVSR. Our Ev-DeblurVSR significantly outperforms existing methods in both spatial recovery and temporal consistency. To summarize, our main contributions are:

- We present Ev-DeblurVSR, the **first** event-enhanced scheme for BVSR. Our Ev-DeblurVSR leverages motion information and high-frequency details from both intra-frame and inter-frame events for BVSR.
- We propose the RFD module, which effectively utilizes mutual assistance between frames and intra-frame events to facilitate feature deblurring.
- We propose the HDA module, which fully exploits complementary motion information from optical flow and inter-frame events to improve temporal alignment.
- Ev-DeblurVSR achieves state-of-the-art performance on three datasets, including synthetic and real-world data.

Related Work

Video Super-Resolution. With the rapid development of deep learning, there has been significant progress in VSR (Wang et al. 2019; Xiao et al. 2020, 2021; Liu et al. 2022; Xia et al. 2023; Li et al. 2024a). Compared to single-image super-resolution, VSR focuses more on modeling temporal relationships and aligning frames. For example, BasicVSR++ (Chan et al. 2022) introduced second-order grid propagation and flow-guided deformable alignment to explore long-term information across misaligned frames. However, these methods often perform poorly in challenging cases, such as videos with severe motion blur (Li et al. 2024b). To address this issue, Fang and Zhan (2022) proposed the first deep learning-based BVSR network that uses a parallel-fusion module to combine features from SR and deblurring branches. Recently, Youk, Oh, and Kim (2024) presented FMA-Net, a method for joint learning of spatiotemporally variant degradation and restoration kernels through complex motion representation learning. However, these methods often fail when there are large pixel displacements, resulting in severe temporal inconsistency.

Video Deblurring. Video deblurring aims to recover sharp videos from blurry inputs, where exploring temporal information is crucial (Li et al. 2021; Jiang et al. 2022; Zhu et al. 2022; Cao et al. 2022; Lin et al. 2022; Wang et al. 2023; Li et al. 2024b; Liang et al. 2024). To efficiently transfer useful information from neighboring frames, Zhong et al. (2020) proposed using a global spatiotemporal attention module within a recurrent framework to propagate information from non-local frames. However, incorrect estimation of non-local frames can lead to error propagation through the recurrent process. To address this, Pan et al. (2023) introduced a deep discriminative spatial and temporal network with a channel-wise gated dynamic module to adaptively explore useful information from non-local frames for better video restoration. More recently, Zhang, Xie, and Yao (2024) proposed a Blur-aware Spatio-Temporal Sparse Transformer Network (BSSTNet) for video deblurring. BSSTNet uses a blur map to convert dense attention into a sparse form, allowing for more extensive information utilization throughout the entire video sequence. This approach has shown significant performance improvements in the area of video deblurring.

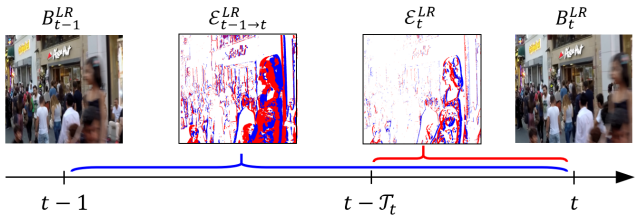


Figure 2: Proposed event processing for BVSR. The exposure time of B_t^{LR} is $\mathcal{T}_t$. Intra-frame events $\mathcal{E}_t^{LR}$ capture motion within the exposure, which is used to deblur frame features. Inter-frame events $\mathcal{E}_{t-1\to t}^{LR}$ capture motion between frames, which helps enhance temporal alignment in VSR.

Event-based Vision. Event cameras, also known as dynamic vision sensors (Lichtsteiner, Posch, and Delbruck 2008), are new bio-inspired vision sensors that measure pixel-wise brightness changes asynchronously and output events. They offer super high temporal resolution (about 1μs) and high dynamic range (140 dB) (Gallego et al. 2020; Chen et al. 2021). With these advantages, event signals have been widely applied in optical flow estimation (Shiba, Aoki, and Gallego 2022; Luo et al. 2024) and video frame interpolation (Xiao et al. 2022; Kim et al. 2023; Liu et al. 2024). With their high temporal resolution, event data can provide rich motion information (Xiao et al. 2024a,b) during the frame's exposure time, which helps deconvolve blurred images (Yang et al. 2022, 2024a,b,c; Kim, Cho, and Yoon 2024; Yu et al. 2024). Sun et al. (2022) devised an event-image fusion module to adaptively integrate event features with image features, alongside a symmetric cumulative event voxel representation for event-based frame deblurring.

Recent studies (Jing et al. 2021; Kai, Zhang, and Sun 2023; Lu et al. 2023; Kai et al. 2024; Xiao et al. 2024d,c) have proposed combining an event camera with an RGB camera to improve VSR performance. Typically, Jing et al. (2021) proposed E-VSR, which first utilizes events to reconstruct intermediate frames. The high-frame-rate video is then encoded into a VSR module to recover HR videos. However, these methods generally assume that the input frames are sharp. In VSR, training with events and blurry frames remains a challenging problem.

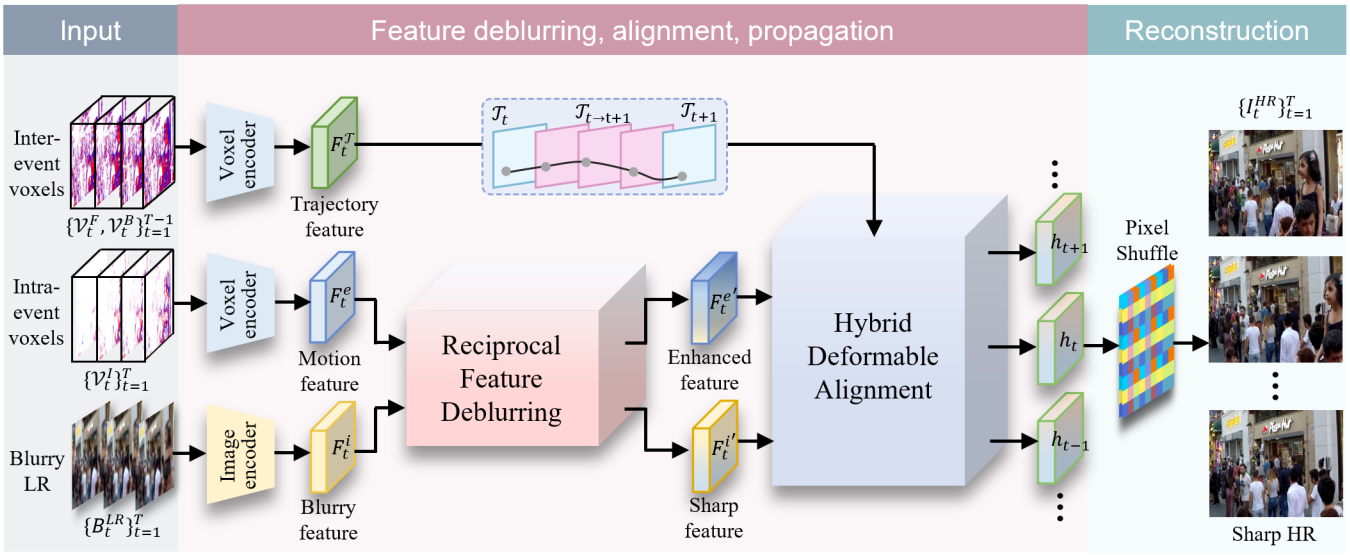


Figure 3: Overview of Ev-DeblurVSR. Intra-frame voxels are fused with blurry frames in the RFD module to deblur frame features and enhance event features with scene context. Inter-frame voxels are integrated into the HDA module, using continuous motion trajectories to guide deformable alignment. Finally, the aligned features are upsampled to reconstruct sharp HR frames.

Method

Event Processing for BVSR

Previous event-based VSR methods are insufficient for BVSR as they only use inter-frame events for alignment. However, temporal misalignment between the timestamps of blurry frames and the inter-frame events hinders the modeling of inherent motion within the blurry frames.

To address this, we propose categorizing events into intra-frame and inter-frame events for BVSR. As shown in Fig. 2, given two blurry LR frames, B_{t-1}^{LR} and B_t^{LR}, and the event stream $\mathcal{E}$, with B_t^{LR} having an exposure time of $\mathcal{T}_t$, our goal is to deblur and upscale these frames to sharp HR frames, I_{t-1}^{HR} and I_t^{HR}. Intra-frame events $\mathcal{E}_t$ capture motion information within the exposure time of each blurry frame and are used for deblurring frame features. Inter-frame events $\mathcal{E}_{t-1\to t}$ capture continuous motion trajectories between frames and are used for feature alignment in VSR.

We represent events as a grid-like event voxel grid $\mathcal{V}$ as in (Zhu et al. 2019). In our experiments, we set the number of bins to 5, consistent with the earlier study (Weng, Zhang, and Xiong 2021). We can thus obtain intra-frame voxels $\mathcal{V}_t^I$. For inter-frame events, since our model uses a bidirectional recurrent network as in BasicVSR (Chan et al. 2021), we generate forward voxels $\mathcal{V}_t^F$ and backward voxels $\mathcal{V}_t^B$.

Ev-DeblurVSR Network

Framework Overview. The architecture of our proposed Ev-DeblurVSR is shown in Fig. 3. The network's input includes a blurry LR sequence consisting of T frames, denoted as $\{B_t^{LR}\}_{t=1}^T$, along with their intra-frame voxels $\{\mathcal{V}_t^I\}_{t=1}^T$, and the $T-1$ intervals' inter-frame voxels, including forward voxels $\{\mathcal{V}_t^F\}_{t=1}^{T-1}$ and backward voxels $\{\mathcal{V}_t^B\}_{t=1}^{T-1}$. The output is a sharp HR sequence $\{I_t^{HR}\}_{t=1}^T$.

The proposed Ev-DeblurVSR comprises two key components: the RFD module and the HDA module. Firstly, the input voxels and frames are passed through their respective feature extractors, comprising five residual blocks as used in (Wang et al. 2018), yielding trajectory, motion, and blurry frame features. In the RFD module, we leverage motion information from intra-frame event features to deblur frame features. Reciprocally, we also enhance event features with global scene context from frame features. In the HDA module, we utilize motion trajectory information from inter-frame events and optical flow to collaboratively enhance motion estimation for the deformable alignment process in VSR. Finally, the aligned features are processed through pixel shuffle (Shi et al. 2016) layers and added with bicubic upsampled results to reconstruct sharp HR frames.

Reciprocal Feature Deblurring. To address the limitations of events in feature deblurring due to sparsity and limited scene context (Messikommer et al. 2020), we propose the RFD module. This module not only utilizes events for effective deblurring but also integrates frames to enhance event features. As shown in Fig. 4, at timestamp t, the RFD module receives the blurry frame feature F_t^i and the intra-frame event feature F_t^e as inputs. They are processed through two pathways, the event and frame pathways, each including a multi-head Channel Attention Block (CAB). The frame pathway captures global scene context, producing F_{CA}^i, while the event pathway learns motion information, resulting in F_{CA}^e. The operation is as follows:

$$F_{CA}^e = \mathbf{CAB}(F_t^e),\ F_{CA}^i = \mathbf{CAB}(F_t^i). \tag{1}$$

The frame feature F_{CA}^i is then fed into the event pathway to enhance event features with scene details, resulting in $F_{CA}^{e'}$. This output is then used to deblur F_{CA}^i, producing $F_{CA}^{i'}$. The above operations are performed using a QKV-based multi-

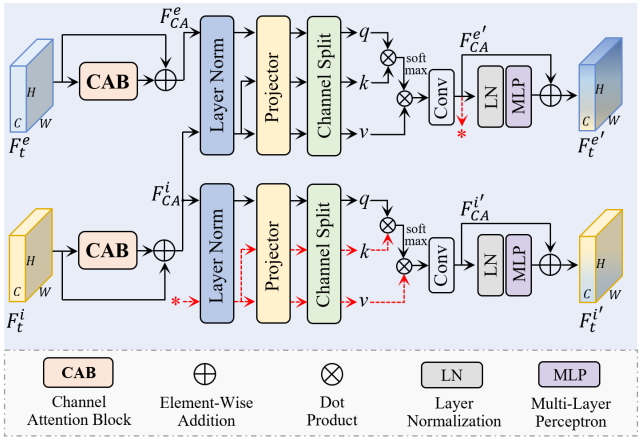


Figure 4: The structure of the RFD module.

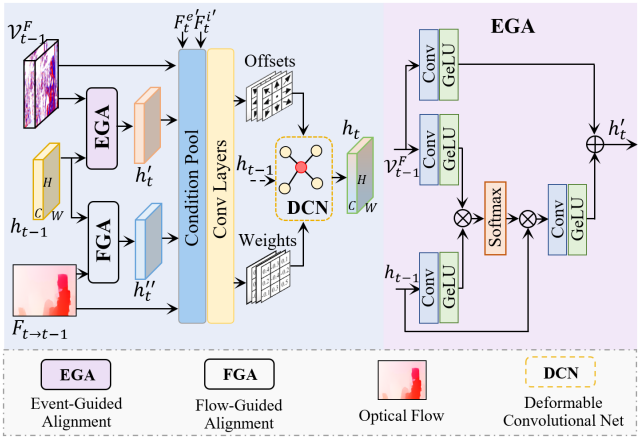


Figure 5: The structure of the HDA module.

head cross-modal attention mechanism as follows:

$$F_{CA}^{e'} = F_{CA}^{e} + \mathbf{V}_i \,\mathrm{softmax}\left(\frac{\mathbf{Q}_e^T \mathbf{K}_i}{\sqrt{C}}\right), \quad (2)$$

$$F_{CA}^{i'} = F_{CA}^{i} + \mathbf{V}_e \,\mathrm{softmax}\left(\frac{\mathbf{Q}_i^T \mathbf{K}_e}{\sqrt{C}}\right), \quad (3)$$

where we use a 1×1 convolutional layer to create attention maps. After that, we apply layer normalization and MLP layers to aggregate information. This results in scene-enhanced event feature $F_t^{e'}$ and sharper frame feature $F_t^{i'}$.

Hybrid Deformable Alignment. Events and optical flow can both represent motion, but they have different characteristics (Wan, Dai, and Mao 2022). Events provide continuous motion but are spatially sparse. Optical flow offers rich spatial information but lacks temporal continuity. To leverage this complementarity, we propose integrating optical flow and events to improve motion estimation in deformable alignment used in VSR (Shi et al. 2022).

We introduce the HDA module, with its structure shown in Fig. 5. We use the feature propagation process from $t-1$ to t as an example to illustrate the alignment process. The HDA module adopts a two-branch structure: the Event-Guided Alignment (EGA) branch and the Flow-Guided Alignment (FGA) branch. In the EGA branch, we use the inter-frame voxel $\mathcal{V}_{t-1}^F$ to align h_{t-1} to h_t'. The FGA branch employs well-established SpyNet (Ranjan and Black 2017) to estimate optical flow $F_{t\to t-1}$. This flow is then used to backward warp h_{t-1}, generating the flow-based alignment feature h''. The process is as follows:

$$h_t' = \mathbf{EGA}(h_{t-1},\, \mathcal{V}_{t-1}^F),\; h_t'' = \mathbf{FGA}(h_{t-1},\, F_{t\to t-1}). \quad (4)$$

In our EGA, we first apply a convolutional layer to $\mathcal{V}_{t-1}^F$ and h_{t-1} to match their channel dimensions. They are then element-wise multiplied, followed by a softmax operation to compute channel-wise similarity scores. The similarity information is used to modulate h_{t-1}, which incorporates the event information into the alignment process. The modulated feature is then combined with the processed $\mathcal{V}_{t-1}^F$ to produce the event-based alignment feature h'.

$$h_t = \mathbf{DCN}(h_{t-1}; h_t', h_t'', F_t^{e'}, F_t^{i'}, F_{t\to t-1}). \quad (5)$$

Finally, h_t' and h_t'', along with $F_{t\to t-1}$, are concatenated with $F_t^{e'}$ and $F_t^{i'}$. As in Eq. (5), these features form our condition pool and are fed into a stack of convolutional layers to predict the motion offsets and modulation weights for the Deformable Convolutional Network (DCN). The learned offsets and weights are used to deform and align h_{t-1} to h_t.

Loss function

Previous VSR studies (Chan et al. 2021, 2022) typically use MSE loss, denoted as $\mathcal{L}_r$, for supervision, calculated between the ground-truth (GT) and super-resolved HR clips. However, this loss treats all pixels equally, regardless of high-frequency and low-frequency regions. It also averages the errors of all pixels, leading to over-smooth results (Xie et al. 2023). To address this, we propose an edge-enhanced loss $\mathcal{L}_e$ that utilizes high-frequency event information to selectively weight pixel reconstruction errors:

$$\mathcal{L}_e = \frac{1}{T}\sum_{t=1}^{T} |\mathcal{V}_t^{HR}| \cdot \sqrt{\left\|I_t^{GT} - I_t^{HR}\right\|^2 + \eta^2}. \quad (6)$$

Here, $\mathcal{V}_t^{HR} \in \mathbb{R}^{sH\times sW\times 3}$ represents the edge-related mask derived from HR voxels within exposure $\mathcal{T}_t$, ranging from $[-1.0, +1.0]$, where s is the upsampling factor. And η is a small smoothing factor to avoid numerical instability. In our experiments, we set $\eta = 1\times 10^{-8}$. Our final loss function is a combination of $\mathcal{L}_r$ and $\mathcal{L}_e$, *i.e.*, $\mathcal{L} = \mathcal{L}_r + \mathcal{L}_e$.

Experiments

Datasets

Synthetic datasets. We use two widely-used datasets for training: **GoPro** (Nah, Hyun Kim, and Mu Lee 2017) and **BSD** (Zhong et al. 2020). Then, we follow the strategy used in previous VSR studies by applying bicubic downsampling to the videos in the datasets to create blurry LR and sharp HR pairs. The GoPro dataset was recorded using a GoPro camera at 240 fps with a resolution of 1280×720. It contains 22 videos for training and 11 for testing. The blurry frames in this dataset are created by averaging several sharp frames. The BSD dataset, on the other hand, consists of real

Method Type		Method	GoPro PSNR↑	GoPro SSIM↑	GoPro LPIPS↓	#Params (M)	FLOPs (G / frame)	Runtime (ms / frame)
RGB-based	Deblur + VSR	DSTNet + BasicVSR++	24.43	0.7471	0.3816	7.45 + 7.32	44.9 + 405.6	7.0 + 64.4
		DSTNet + IART	24.43	0.7467	0.3842	7.45 + 13.41	44.9 + 1972.7	7.0 + 1321.2
		BSSTNet + MIA-VSR	26.40	0.8192	0.3161	48.18 + 16.60	314.7 + 1267.5	67.8 + 831.0
		BSSTNet + IART	26.40	0.8189	0.3148	48.18 + 13.41	314.7 + 1972.7	67.8 + 1321.2
	BVSR	BasicVSR++*	30.79	<u>0.9077</u>	<u>0.2287</u>	7.32	405.6	64.4
		MIA-VSR*	27.91	0.8481	0.2901	16.60	1267.5	831.0
		IART*	27.69	0.8372	0.3050	13.41	1972.7	1321.2
		FMA-Net	29.24	0.8720	0.2682	9.62	1365.0	579.8
Event-based	Deblur + VSR	EFNet + EGVSR	23.53	0.7276	0.4155	8.47 + 2.58	94.9 + 159.6	11.7 + 118.1
		EFNet† + EGVSR	23.80	0.7422	0.3963	9.91 + 2.58	114.5 + 159.6	15.4 + 118.1
		REFID + EvTexture	23.72	0.7448	0.4019	15.92 + 8.90	89.1 + 805.4	16.2 + 100.8
		REFID† + EvTexture	24.28	0.7738	0.3402	17.36 + 8.90	108.7 + 805.4	19.9 + 100.8
	BVSR	eSL-Net++	26.29	0.7959	0.3377	**1.41**	434.4	**59.4**
		eSL-Net++†	26.43	0.8293	0.3052	2.85	454.0	63.1
		EGVSR*	27.79	0.8331	0.3037	2.58	**159.6**	118.1
		EvTexture*	<u>31.00</u>	0.9065	0.2355	8.90	805.4	100.8
		Ev-DeblurVSR	**32.51**	**0.9314**	**0.2041**	8.28	459.5	79.6

Table 1: Quantitative comparison on GoPro for $4\times$ BVSR. **All methods are retrained on the same dataset.** All results are calculated on the RGB channel. **Bold** and <u>underlined</u> numbers represent the best and second-best performance. FLOPs and runtime are computed on one 180×320 LR frame. * **denotes the model initially proposed for sharp VSR, and we retrain it on blurry LR inputs.** † **indicates the single-image model, and we include optical flow from SpyNet to refine it.**

Method	BSD PSNR/ SSIM/ LPIPS	NCER PSNR/ SSIM/ LPIPS
BasicVSR++*	<u>31.12</u> / <u>.9050</u> / <u>.2580</u>	27.05 / <u>.8255</u> / <u>.1975</u>
MIA-VSR*	29.24 / .8643 / .3074	24.55 / .7307 / .3251
IART*	29.47 / .8689 / .2977	25.16 / .7499 / .2908
FMA-Net	30.14 / .8805 / .2887	26.01 / .7779 / .2538
EGVSR*	29.32 / .8665 / .3145	24.26 / .7218 / .3276
EvTexture*	31.06 / .8956 / .2746	<u>27.23</u> / .8136 / .2241
Ev-DeblurVSR	**33.02 / .9304 / .2281**	**28.60 / .8516 / .1712**

Table 2: Comparison on BSD and NCER for $4\times$ BVSR.

GoPro	IART*	FMA-Net	EGVSR*	EvTexture*	**Ours**
tOF↓	2.94	2.30	2.78	1.73	**1.43**
TCC↑$_{\times 10}$	3.73	4.36	3.70	4.90	**5.38**
NCER	IART*	FMA-Net	EGVSR*	EvTexture*	**Ours**
tOF↓	1.61	0.96	1.39	0.72	**0.54**
TCC↑$_{\times 10}$	2.85	3.50	2.62	3.96	**4.73**

Table 3: Temporal consistency on GoPro and NCER.

blurry-sharp video pairs captured using a beam splitter system. These videos have a resolution of 640×480 and a frame rate of 15 fps, which contain severe motion blur. The dataset includes 60 sequences for training and 20 for testing. Since the GoPro and BSD datasets do not have real event data, we use the commonly used event simulator Vid2E (Gehrig et al. 2020) to generate event data from the video clips.

Real-world datasets. We also train and test our method on real-world event data. For this, we use the recently published event-based motion deblurring dataset **NCER** (Cho et al. 2023), which includes 27 videos for training (a total of 2,583 frames) and 16 videos for testing (1,454 frames). The dataset is recorded with a high-frame-rate (522 fps) RGB camera and a 640×480 DVXplorer event camera, covering various scenes and textures suitable for BVSR.

Implementation Details

We follow the previous study (Chan et al. 2022); when training, we use 15 frames as inputs, set the mini-batch size to 8, and center-crop the input frames size and event voxels size as 64×64. We use random horizontal and vertical flips to augment the data. On the three datasets mentioned above, we first train the model on GoPro for 300K iterations using the Adam optimizer and Cosine Annealing scheduler. For the experiments on BSD, we fine-tune the model trained on GoPro with an initial learning rate of 1×10^{-4} for 200K iterations. Then, similar to NCER, we fine-tune the model trained on BSD with the same hyperparameter settings. The entire training process runs on 8 NVIDIA RTX4090 GPUs and takes about four days per dataset to converge.

Comparisons with State-of-the-Art Methods

Baselines. We compare two types of SOTA methods: RGB-based and event-based. Each of these types is further divided into two strategies: the cascade strategy, *i.e.*, deblur + VSR, and BVSR. For RGB-based VSR, we compare our method with three recent methods: BasicVSR++ (Chan et al. 2022), MIA-VSR (Zhou et al. 2024), and IART (Xu et al. 2024). For event-based VSR, we compare two methods: EGVSR (Lu et al. 2023) and EvTexture (Kai et al. 2024). Also, we include two recent video deblurring methods: DSTNet (Pan et al. 2023) and BSSTNet (Zhang, Xie, and Yao 2024), as well as two event-based motion deblurring methods: EFNet (Sun et al. 2022) and REFID (Sun et al. 2023).

Figure 6: Qualitative comparison on BSD. Our method can restore clear road signs and text with sharp edges.

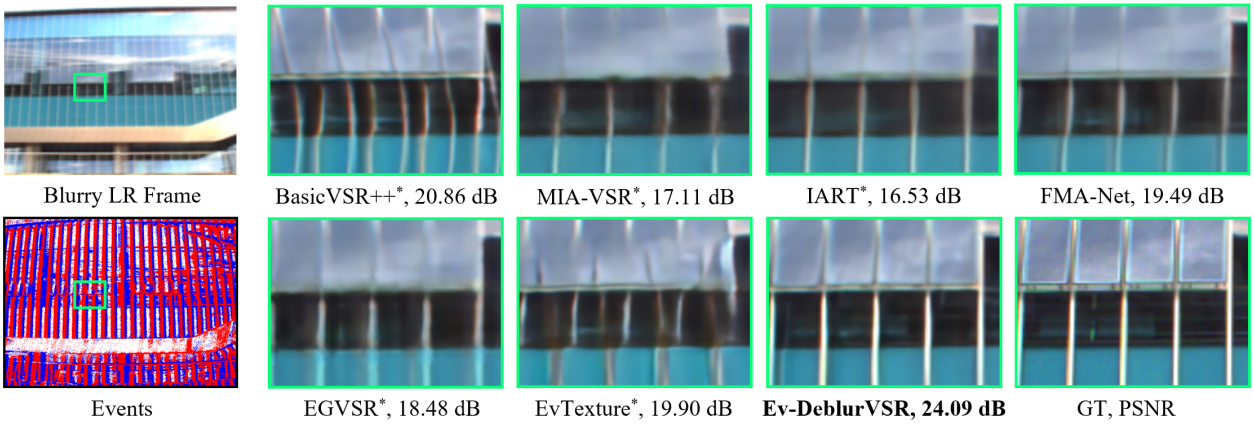


Figure 7: Qualitative comparison on NCER. Our method can restore blurred and distorted window lines to sharp ones.

Additionally, we compare with two BVSR methods: FMA-Net (Youk, Oh, and Kim 2024) and eSL-Net++ (Yu et al. 2023). It should be noted that **we retrain all baselines using the same datasets as ours for fair comparisons.**

Quantitative results. Tabs. 1, 2 and 3 present the comparison results against the baselines mentioned above. From the data, it is evident that our method consistently achieves superior spatial recovery in terms of PSNR, SSIM, and LPIPS (Zhang et al. 2018), as well as temporal consistency metrics, such as tOF (Chu et al. 2020) and TCC (Chi et al. 2020). Notably, our method significantly improves over the recent BVSR method FMA-Net, surpassing it by **3.27** dB, **2.88** dB, and **2.59** dB on the GoPro, BSD, and NCER datasets. Additionally, our Ev-DeblurVSR has fewer parameters, requires only 33.67% of the FLOPs, and is **7.28**× faster than FMA-Net. Furthermore, our method makes better use of event data than other event-based VSR methods. In most cases, EGVSR and EvTexture do not perform better than the image-only method BasicVSR++. However, our method significantly outperforms BasicVSR++ by at least **1.55** dB across all three datasets.

Qualitative results. We also show visual comparisons in Figs. 6 and 7. It is evident that our method can effectively restore clear road signs and window lines, producing sharp, well-defined edges. In contrast, other methods fail to recover fine details, resulting in blurry artifacts and indistinct boundaries. This highlights the superiority of our approach in handling blurry inputs and recovering high-quality HR frames.

	Method	GoPro PSNR↑	GoPro SSIM↑	tOF↓	#Params (M)
Events	(a) w/o inter-	31.32	0.9072	1.67	7.94
	(b) w/o intra-	31.51	0.9180	1.54	8.03
RFD	(c) w/o CAB	31.55	0.9176	1.53	8.25
	(d) w/o CM	31.36	0.9162	1.58	8.25
	(e) w/o $i \rightarrow e$	31.81	0.9226	1.50	8.27
	(f) $e \rightarrow i$, $i \rightarrow e$	32.23	0.9269	1.48	8.28
HDA	(g) w/o EGA	31.72	0.9163	1.60	7.98
	(h) w/o FGA	31.53	0.9082	1.62	6.55
Loss	(i) w/o $\mathcal{L}_r$	32.37	0.9288	1.47	8.28
	(j) w/o $\mathcal{L}_e$	32.21	0.9268	1.49	8.28
	(k) Ours	**32.51**	**0.9314**	**1.43**	8.28

Table 4: Ablation studies of the components.

Ablation Study

Event utilization. Tab. 4(a-b, k) shows that using only intra-frame events for both feature deblurring and alignment re-

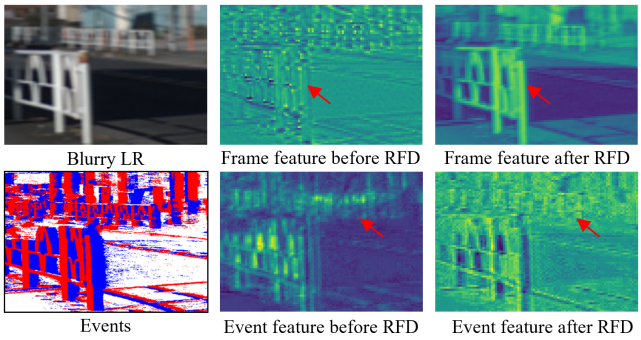


Figure 8: Analysis of the RFD module.

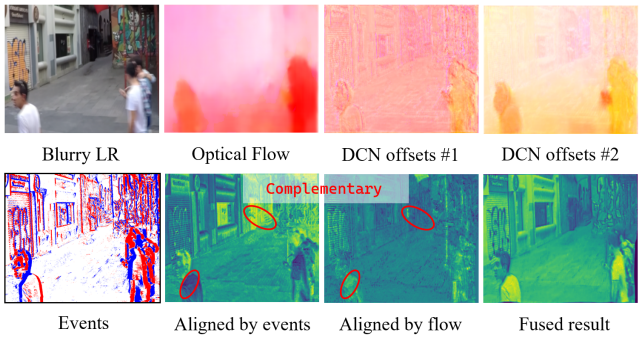


Figure 9: Analysis of the HDA module.

sults in a 1.19 dB drop. This is because the timestamps of intra-frame events are not well-aligned with the nearby frames. Similarly, using only inter-frame events also causes a performance drop. Our method, which combines both intra-frame and inter-frame events, better meets the needs of BVSR, leading to a significant improvement.

The RFD module. Tab. 4(c-f, k) shows the importance of each component in our RFD module. Removing the CAB, which captures global features from event-image modalities, leads to a 0.96 dB drop. Cross-modal (CM) interaction is also critical, as its removal causes a 1.15 dB drop. In Tab. 4(e-f), $i \rightarrow e$ refers to refining event features with image features, while $e \rightarrow i$ represents using event features to deblur image features. While $e \rightarrow i$ is a standard process in event-based motion deblurring, $i \rightarrow e$ is rarely explored. Excluding $i \rightarrow e$ results in a 0.70 dB drop. Our method employs a sequential $i \rightarrow e$ followed by $e \rightarrow i$, which outperforms reversing the order, where performance drops by 0.28 dB.

Fig. 8 illustrates the deblurring process: in blurry areas such as railings, the RFD sharpens frame features and enriches event features with contextual scene information.

The HDA module. Tab. 4(g-h, k) shows that our full model, which combines EGA and FGA alignment methods, achieves significant improvements. Fig. 9 visualizes the learned motion vectors and aligned features. It demonstrates that DCN offsets, similar to optical flow, capture moving objects but provide more diversity. Moreover, event-aligned features can capture background areas affected by camera movement, where flow-aligned features may fail. In contrast, flow-aligned features enhance details in regions with significant motion. These two alignment methods complement each other, and the fused features exhibit sharp edges and detailed scene representations, validating the effectiveness of our hybrid alignment approach.

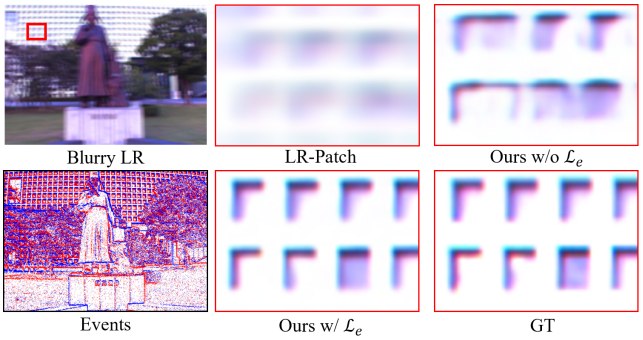


Figure 10: Analysis of the edge-enhanced loss.

Metrics	MIA-VSR	IART	FMA-Net	EvTexture	**Ours**
PSNR↑	34.14	33.95	33.00	34.55	**34.99**
SSIM↑	0.9449	0.9430	0.9315	0.9491	**0.9534**
LPIPS↓	0.1695	0.1719	0.1826	0.1642	**0.1551**
tOF↓$_{\times 10}$	6.71	6.86	7.22	6.37	**5.98**
TCC↑$_{\times 10}$	6.03	5.99	5.79	6.14	**6.26**

Table 5: Comparisons of sharp VSR methods on GoPro.

Edge-enhanced loss. Tab. 4(i-j, k) demonstrates the effectiveness of our edge-enhanced loss. Fig. 10 shows that the model trained with $\mathcal{L}_e$ more accurately restores windows, eliminating blur and producing sharper edges.

Performance on sharp videos. Although our primary focus is on blurry inputs, we compare our method with several recent SOTA VSR methods on sharp inputs using the GoPro dataset. As shown in Tab. 5, our method *consistently* achieves the best performance on sharp inputs, both in spatial recovery and temporal consistency, demonstrating the effectiveness and versatility of our approach.

Limitation. In our setting, we assume that the frame exposure time is known and fixed. However, in real-world scenarios, especially when auto-exposure is enabled, the exposure time can vary dynamically depending on the lighting conditions, making it unknown (Kim et al. 2022). Thus, the problem of handling BVSR under unknown exposure times remains an open and worthwhile area for further research.

Conclusion

This paper presents Ev-DeblurVSR, a novel event-enhanced network for BVSR that leverages high-temporal-resolution and high-frequency event signals. To effectively fuse frame and event information for BVSR, we categorize events into intra-frame and inter-frame types. The RFD module is then introduced, utilizing intra-frame events to deblur frame features while reciprocally enhancing event features with global scene context from frames. Additionally, we propose the HDA module, which combines the complementary motion information from inter-frame events and optical flow to improve motion estimation and temporal alignment. Extensive experiments on both synthetic and real-world datasets demonstrate the effectiveness of our Ev-DeblurVSR.

Acknowledgments

We acknowledge funding from the National Natural Science Foundation of China under Grants 62472399 and 62021001.

References

Cao, M.; Fan, Y.; Zhang, Y.; Wang, J.; and Yang, Y. 2022. Vdtr: Video deblurring with transformer. *IEEE TCSVT*.

Chan, K. C.; Wang, X.; Yu, K.; Dong, C.; and Loy, C. C. 2021. BasicVSR: The search for essential components in video super-resolution and beyond. In *CVPR*.

Chan, K. C.; Zhou, S.; Xu, X.; and Loy, C. C. 2022. BasicVSR++: Improving video super-resolution with enhanced propagation and alignment. In *CVPR*.

Chen, Z.; Zheng, Q.; Niu, P.; Tang, H.; and Pan, G. 2021. Indoor lighting estimation using an event camera. In *CVPR*.

Chi, Z.; Mohammadi Nasiri, R.; Liu, Z.; Lu, J.; Tang, J.; and Plataniotis, K. N. 2020. All at once: Temporally adaptive multi-frame interpolation with advanced motion modeling. In *ECCV*.

Cho, H.; Jeong, Y.; Kim, T.; and Yoon, K.-J. 2023. Non-Coaxial Event-guided Motion Deblurring with Spatial Alignment. In *ICCV*.

Chu, M.; Xie, Y.; Mayer, J.; Leal-Taixé, L.; and Thuerey, N. 2020. Learning temporal coherence via self-supervision for GAN-based video generation. *ACM TOG*.

Fang, N.; and Zhan, Z. 2022. High-resolution optical flow and frame-recurrent network for video super-resolution and deblurring. *Neurocomputing*.

Gallego, G.; Delbrück, T.; Orchard, G.; Bartolozzi, C.; Taba, B.; Censi, A.; Leutenegger, S.; Davison, A. J.; Conradt, J.; Daniilidis, K.; et al. 2020. Event-based vision: A survey. *IEEE TPAMI*.

Gehrig, D.; Gehrig, M.; Hidalgo-Carrió, J.; and Scaramuzza, D. 2020. Video to events: Recycling video datasets for event cameras. In *CVPR*.

Jiang, B.; Xie, Z.; Xia, Z.; Li, S.; and Liu, S. 2022. Erdn: Equivalent receptive field deformable network for video deblurring. In *ECCV*.

Jing, Y.; Yang, Y.; Wang, X.; Song, M.; and Tao, D. 2021. Turning frequency to resolution: Video super-resolution via event cameras. In *CVPR*.

Kai, D.; Lu, J.; Zhang, Y.; and Sun, X. 2024. EvTexture: Event-driven Texture Enhancement for Video Super-Resolution. In *ICML*.

Kai, D.; Zhang, Y.; and Sun, X. 2023. Video Super-Resolution Via Event-Driven Temporal Alignment. In *ICIP*.

Kim, T.; Chae, Y.; Jang, H.-K.; and Yoon, K.-J. 2023. Event-based video frame interpolation with cross-modal asymmetric bidirectional motion fields. In *CVPR*.

Kim, T.; Cho, H.; and Yoon, K.-J. 2024. Frequency-aware Event-based Video Deblurring for Real-World Motion Blur. In *CVPR*.

Kim, T.; Lee, J.; Wang, L.; and Yoon, K.-J. 2022. Event-guided deblurring of unknown exposure time videos. In *ECCV*.

Li, D.; Xu, C.; Zhang, K.; Yu, X.; Zhong, Y.; Ren, W.; Suominen, H.; and Li, H. 2021. Arvo: Learning all-range volumetric correspondence for video deblurring. In *CVPR*.

Li, Z.; Liu, H.; Shang, F.; Liu, Y.; Wan, L.; and Feng, W. 2024a. SAVSR: Arbitrary-Scale Video Super-Resolution via a Learned Scale-Adaptive Network. In *AAAI*.

Li, Z.; Yuan, Z.; Li, L.; Liu, D.; Tang, X.; and Wu, F. 2024b. Object Segmentation-Assisted Inter Prediction for Versatile Video Coding. *IEEE Transactions on Broadcasting*.

Liang, J.; Cao, J.; Fan, Y.; Zhang, K.; Ranjan, R.; Li, Y.; Timofte, R.; and Van Gool, L. 2024. VRT: A Video Restoration Transformer. *IEEE TIP*.

Lichtsteiner, P.; Posch, C.; and Delbruck, T. 2008. A 128×128 120 dB 15 μs latency asynchronous temporal contrast vision sensor. *IEEE journal of solid-state circuits*.

Lin, J.; Cai, Y.; Hu, X.; Wang, H.; Yan, Y.; Zou, X.; Ding, H.; Zhang, Y.; Timofte, R.; and Van Gool, L. 2022. Flow-Guided Sparse Transformer for Video Deblurring. In *ICML*.

Liu, C.; Yang, H.; Fu, J.; and Qian, X. 2022. Learning trajectory-aware transformer for video super-resolution. In *CVPR*.

Liu, H.; Zhao, P.; Ruan, Z.; Shang, F.; and Liu, Y. 2021. Large motion video super-resolution with dual subnet and multi-stage communicated upsampling. In *AAAI*.

Liu, Y.; Deng, Y.; Chen, H.; and Yang, Z. 2024. Video Frame Interpolation via Direct Synthesis with the Event-based Reference. In *CVPR*.

Lu, Y.; Wang, Z.; Liu, M.; Wang, H.; and Wang, L. 2023. Learning Spatial-Temporal Implicit Neural Representations for Event-Guided Video Super-Resolution. In *CVPR*.

Luo, X.; Luo, A.; Wang, Z.; Lin, C.; Zeng, B.; and Liu, S. 2024. Efficient Meshflow and Optical Flow Estimation from Event Cameras. In *CVPR*.

Messikommer, N.; Gehrig, D.; Loquercio, A.; and Scaramuzza, D. 2020. Event-based asynchronous sparse convolutional networks. In *ECCV*.

Mitrokhin, A.; Hua, Z.; Fermuller, C.; and Aloimonos, Y. 2020. Learning visual motion segmentation using event surfaces. In *CVPR*.

Nah, S.; Hyun Kim, T.; and Mu Lee, K. 2017. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *CVPR*.

Pan, J.; Xu, B.; Dong, J.; Ge, J.; and Tang, J. 2023. Deep Discriminative Spatial and Temporal Network for Efficient Video Deblurring. In *CVPR*.

Ranjan, A.; and Black, M. J. 2017. Optical flow estimation using a spatial pyramid network. In *CVPR*.

Shamsolmoali, P.; Zareapoor, M.; Jain, D. K.; Jain, V. K.; and Yang, J. 2019. Deep convolution network for surveillance records super-resolution. *Multimedia Tools and Applications*.

Shi, S.; Gu, J.; Xie, L.; Wang, X.; Yang, Y.; and Dong, C. 2022. Rethinking alignment in video super-resolution transformers. *NeurIPS*.

Shi, W.; Caballero, J.; Huszár, F.; Totz, J.; Aitken, A. P.; Bishop, R.; Rueckert, D.; and Wang, Z. 2016. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *CVPR*.

Shiba, S.; Aoki, Y.; and Gallego, G. 2022. Secrets of event-based optical flow. In *ECCV*.

Sun, L.; Sakaridis, C.; Liang, J.; Jiang, Q.; Yang, K.; Sun, P.; Ye, Y.; Wang, K.; and Gool, L. V. 2022. Event-based fusion for motion deblurring with cross-modal attention. In *ECCV*.

Sun, L.; Sakaridis, C.; Liang, J.; Sun, P.; Cao, J.; Zhang, K.; Jiang, Q.; Wang, K.; and Van Gool, L. 2023. Event-based frame interpolation with ad-hoc deblurring. In *CVPR*.

Wan, Z.; Dai, Y.; and Mao, Y. 2022. Learning dense and continuous optical flow from an event camera. *IEEE TIP*.

Wang, J.; Weng, W.; Zhang, Y.; and Xiong, Z. 2023. Unsupervised Video Deraining with An Event Camera. In *ICCV*.

Wang, X.; Chan, K. C.; Yu, K.; Dong, C.; and Change Loy, C. 2019. EDVR: Video restoration with enhanced deformable convolutional networks. In *CVPRW*.

Wang, X.; Yu, K.; Wu, S.; Gu, J.; Liu, Y.; Dong, C.; Qiao, Y.; and Change Loy, C. 2018. ESRGAN: Enhanced super-resolution generative adversarial networks. In *ECCVW*.

Weng, W.; Zhang, Y.; and Xiong, Z. 2021. Event-based video reconstruction using transformer. In *ICCV*.

Xia, B.; He, J.; Zhang, Y.; Wang, Y.; Tian, Y.; Yang, W.; and Van Gool, L. 2023. Structured sparsity learning for efficient video super-resolution. In *CVPR*.

Xiao, P.; Zhang, Y.; Kai, D.; Peng, Y.; Zhang, Z.; and Sun, X. 2024a. ESTME: Event-driven Spatio-temporal Motion Enhancement for Micro-Expression Recognition. In *ICME*.

Xiao, P.; Zhang, Y.; Kai, D.; Peng, Y.; Zhang, Z.; and Sun, X. 2024b. A Micro-Expression Recognition System with Event Cameras. In *ICMEW*.

Xiao, Z.; Fu, X.; Huang, J.; Cheng, Z.; and Xiong, Z. 2021. Space-time distillation for video super-resolution. In *CVPR*.

Xiao, Z.; Kai, D.; Zhang, Y.; Sun, X.; and Xiong, Z. 2024c. Asymmetric Event-Guided Video Super-Resolution. In *ACM MM*.

Xiao, Z.; Kai, D.; Zhang, Y.; Zha, Z.-J.; Sun, X.; and Xiong, Z. 2024d. Event-Adapted Video Super-Resolution. In *ECCV*.

Xiao, Z.; Weng, W.; Zhang, Y.; and Xiong, Z. 2022. EVA^2: Event-Assisted Video Frame Interpolation via Cross-Modal Alignment and Aggregation. *IEEE TCI*.

Xiao, Z.; Xiong, Z.; Fu, X.; Liu, D.; and Zha, Z.-J. 2020. Space-time video super-resolution using temporal profiles. In *ACM MM*.

Xie, L.; Wang, X.; Shi, S.; Gu, J.; Dong, C.; and Shan, Y. 2023. Mitigating artifacts in real-world video super-resolution models. In *AAAI*.

Xu, F.; Yu, L.; Wang, B.; Yang, W.; Xia, G.-S.; Jia, X.; Qiao, Z.; and Liu, J. 2021. Motion deblurring with real events. In *ICCV*.

Xu, K.; Yu, Z.; Wang, X.; Mi, M. B.; and Yao, A. 2024. Enhancing Video Super-Resolution via Implicit Resampling-based Alignment. In *CVPR*.

Yang, W.; Wu, J.; Li, L.; Dong, W.; and Shi, G. 2023. Event-based Motion Deblurring with Modality-Aware Decomposition and Recomposition. In *ACM MM*.

Yang, W.; Wu, J.; Ma, J.; Li, L.; Dong, W.; and Shi, G. 2022. Learning for motion deblurring with hybrid frames and events. In *ACM MM*.

Yang, W.; Wu, J.; Ma, J.; Li, L.; Dong, W.; and Shi, G. 2024a. Learning Frame-Event Fusion for Motion Deblurring. *IEEE TIP*.

Yang, W.; Wu, J.; Ma, J.; Li, L.; and Shi, G. 2024b. Motion Deblurring via Spatial-Temporal Collaboration of Frames and Events. In *AAAI*.

Yang, Y.; Liang, J.; Yu, B.; Chen, Y.; Ren, J. S.; and Shi, B. 2024c. Latency Correction for Event-guided Deblurring and Frame Interpolation. In *CVPR*.

Youk, G.; Oh, J.; and Kim, M. 2024. FMA-Net: Flow-Guided Dynamic Filtering and Iterative Feature Refinement with Multi-Attention for Joint Video Super-Resolution and Deblurring. In *CVPR*.

Yu, L.; Wang, B.; Zhang, X.; Zhang, H.; Yang, W.; Liu, J.; and Xia, G.-S. 2023. Learning to super-resolve blurry images with events. *IEEE TPAMI*.

Yu, W.; Li, J.; Zhang, S.; and Ji, X. 2024. Learning Scale-Aware Spatio-temporal Implicit Representation for Event-based Motion Deblurring. In *ICML*.

Zhang, H.; Xie, H.; and Yao, H. 2024. Blur-aware Spatio-temporal Sparse Transformer for Video Deblurring. In *CVPR*.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*.

Zhong, Z.; Gao, Y.; Zheng, Y.; and Zheng, B. 2020. Efficient spatio-temporal recurrent neural network for video deblurring. In *ECCV*.

Zhou, X.; Zhang, L.; Zhao, X.; Wang, K.; Li, L.; and Gu, S. 2024. Video Super-Resolution Transformer with Masked Inter&Intra-Frame Attention. In *CVPR*.

Zhu, A. Z.; Yuan, L.; Chaney, K.; and Daniilidis, K. 2019. Unsupervised event-based learning of optical flow, depth, and egomotion. In *CVPR*.

Zhu, C.; Dong, H.; Pan, J.; Liang, B.; Huang, Y.; Fu, L.; and Wang, F. 2022. Deep recurrent neural network with multi-scale bi-directional propagation for video deblurring. In *AAAI*.