Exploiting Blurry Representations for Event-guided Video Super-Resolution

Zeyu Xiao, and Xinchao Wang*

National University of Singapore
zeyuxiao@nus.edu.sg, xinchao@nus.edu.sg

Abstract

Blurry video super-resolution (BVSR) remains fundamentally ill-posed due to the simultaneous loss of high-frequency spatial details and reliable motion cues in blurry low-resolution frames. While cascade-based and joint BVSR methods struggle under severe blur, existing event-guided VSR approaches essentially assume clean inputs and are ineffective against complex motion degradation. These methods fail to model blurry representations or leverage event signals for blur-aware motion cues, leading to sub-optimal performance. We propose *BluR-EVSR*, a unified framework that implicitly models Blurry Representations and leverages Event cameras to jointly address both blur and resolution degradation for VSR. The framework begins with a self-supervised degradation learning strategy guided by event streams and neighboring frames, enabling adaptive blur representation without explicit supervision. A dynamic routing mechanism encodes spatially varying degradations, while a motion-saliency degradation-aware attention module injects motion saliency priors to facilitate efficient RGB-event fusion. Integrated into a bidirectional recurrent framework, *BluR-EVSR* enables temporally consistent and detail-preserving restoration with low computational cost. Extensive experiments across multiple benchmarks show that our method significantly outperforms prior BVSR and event-based approaches.

1 Introduction

Video super-resolution (VSR) aims to reconstruct high-resolution (HR) video frames from low-resolution (LR) inputs and has broad applications in areas such as medical imaging (Peng et al. 2020), remote sensing (Xiao et al. 2021a), and surveillance (Farooq et al. 2021; Zhang et al. 2024). In real-world scenarios such as sports broadcasting, aerial surveillance, and autonomous driving, video frames often suffer from both resolution degradation and motion blur due to rapid motion or camera shake. To better characterize such degradations, the formation of the j-th blurry and LR observation frame $\boldsymbol{y}_j$ ($j = i-N, i-N+1, ..., i+N$), that is, the degraded measurement from its underlying HR reference $\boldsymbol{x}_i$, can be described by a physics-based model (Liu and Sun 2013; Lee, Choi, and Lee 2021; Pan et al. 2021; Jeelani et al. 2023; Xiao et al. 2023; Bai and Pan 2024)

$$\boldsymbol{y}_j = \boldsymbol{S}\boldsymbol{K}_j\boldsymbol{F}_{i\to j}\boldsymbol{x}_i + \boldsymbol{n}, \tag{1}$$

where $\boldsymbol{F}_{i\to j}$ denotes motion warping function from reference frame $\boldsymbol{x}_i$ to target j-th frame frame, $\boldsymbol{K}_j$ is a blurring matrix (w.r.t K_j), $\boldsymbol{S}$ is the downsampling operator (w.r.t S), and $\boldsymbol{n}$ is additive noise. This formulation jointly accounts for resolution loss and motion-induced blur. It reflects the inherent ill-posedness of blurry VSR (BVSR), where the latent HR frame $\boldsymbol{x}_i$, blur kernel $\boldsymbol{K}_j$, and motion field $\boldsymbol{F}_{i\to j}$ are unknown and entangled in the degradation process.

Existing VSR approaches typically focus on spatial upsampling under simplified assumptions, namely clean inputs without noise $\boldsymbol{n}$, no explicit modeling of the blur kernel $\boldsymbol{K}_j$, and a fixed downsampling operator $\boldsymbol{S}$ typically assumed to be bicubic (Chan et al. 2021; Wang et al. 2019). In blurry scenarios, cascade pipelines, whether performing deblurring before super-resolution or vice versa, tend to propagate and amplify artifacts across stages and fail to leverage temporal redundancy effectively. Joint frameworks (Fang and Zhan 2022; Youk, Oh, and Kim 2024) that estimate motion and blur in an end-to-end manner have shown promise but still struggle with residual blur and jitter, particularly under complex motion. A core limitation of these methods lies in the use of RGB-only inputs, which *lack the temporal granularity necessary for reliable motion modeling and ambiguity resolution*. When both spatial detail and motion cues are degraded, RGB-only models are prone to produce temporally inconsistent and visually blurred results. To overcome these challenges, integrating neuromorphic sensors (Zhao et al. 2021; Dong et al. 2024; Zhao et al. 2023b, 2024b, 2023a; Xiao et al. 2022; Xiao, Lu, and Wang 2024) such as event cameras into VSR has recently attracted growing interest, owing to their inherent robustness to motion blur and high temporal resolution.

While event-guided VSR has attracted increasing attention due to the inherent advantages of event cameras, works specifically targeting the BVSR setting remain scarce. To the best of our knowledge, only one recent work (Kai et al. 2025) explicitly addresses this scenario. However, it suffers from two key limitations: (1) it lacks explicit modeling of blur and instead relies on implicit exposure information, which *may not capture the spatio-temporal variation of blur severity*; and (2) it adopts a parallel fusion strategy

*Corresponding Author

that processes frames and event streams independently before fusing them at the feature level. This design limits the model's ability to exploit blur-aware motion cues, making it *less effective at disambiguating motion and restoring fine details*. Motivated by these limitations, we adopt a principled degradation modeling strategy that directly integrates event data into the observation formation process, enabling accurate and flexible blur representation learning. To alleviate the need for explicit motion estimation, we reformulate the degradation process into an event-aware model

$$\boldsymbol{y}_j = \boldsymbol{S}\tilde{\boldsymbol{K}}_j\boldsymbol{x}_j + \boldsymbol{n}, \text{where } \tilde{\boldsymbol{K}}_j = \Psi(\boldsymbol{E}_j, \{\boldsymbol{y}_k\}_{k\in\mathcal{T}_j}), \quad (2)$$

$\boldsymbol{E}_j$ denotes the event stream captured during the exposure of the j-th frame, and $\mathcal{T}_j$ denotes a temporal neighborhood of frame j, typically defined as $\mathcal{T}_j = \{i-N, i-N+1, ..., i+N\}$. The set $\{\boldsymbol{y}_k\}_{k\in\mathcal{T}_j}$ contains neighboring blurry frames that provide additional context to guide blur kernel estimation. The function $\Psi(\cdot)$ is implemented via deep networks that implicitly model blur kernels by fusing event-guided motion cues and degradation context, without requiring explicit motion or blur supervision. In doing so, it enables spatially adaptive and temporally consistent restoration without relying on rigid fusion schemes, offering improved modeling flexibility and interpretability compared to previous designs (Chen et al. 2019; Zhang, Gool, and Timofte 2020; Tang et al. 2023).

To operationalize this event-aware formulation, we propose *BluR-EVSR*, an end-to-end Event-guided VSR framework tailored for blurry LR inputs via Blurry Representations. *BluR-EVSR* begins by estimating implicit blur representations between blurry LR frames and their latent HR counterparts, guided by the high-temporal-resolution cues from event streams. A dynamic routing mechanism adaptively regulates the generation of degradation features across multiple spatial scales, enabling the model to account for spatially varying blur patterns under diverse motion conditions. These features are then used in a degradation regularization process to modulate reconstruction, effectively bridging the gap between degraded and clean representations in a fully self-supervised manner, without requiring explicit blur annotations. To further enhance the modeling of fine-grained spatial structures and temporally coherent motion cues, we propose a Motion-Saliency Degradation-aware Attention (MoSDA) module. MoSDA leverages the event stream and the learned blur representations as motion-aware priors to dynamically modulate attention weights, allowing the network to focus on motion-relevant regions with degraded structures. By integrating these components into a bidirectional recurrent framework (Chan et al. 2021, 2022a), *BluR-EVSR* achieves temporally consistent and detail-preserving reconstruction under severe motion blur, while maintaining high efficiency and interpretability.

Extensive experiments conducted on three benchmark datasets demonstrate that *BluR-EVSR* consistently outperforms state-of-the-art approaches, achieving superior spatial reconstruction quality and temporal consistency. Our contributions can be summarized as follows: (1) We introduce a physics-based formulation of BVSR and extend it to the event-guided setting, where blur is explicitly modeled as an event-guided degradation process. (2) We propose *BluR-EVSR*, a unified framework that combines the dynamic degradation routing and the MoSDA module to jointly model blur and motion. (3) Our method achieves advanced performance in spatial detail recovery and temporal consistency, validated across multiple benchmarks.

2 Related Work

Video super-resolution. VSR enhances LR frames by exploiting temporal information via sliding-window or recurrent structures. Deep learning-based methods (Mao et al. 2025; Xiao, Li, and Jia 2025; Xiao and Wang 2025b; Xiao and Xiong 2025; Xiao et al. 2024b; Bai et al. 2024, 2025; Zhao et al. 2023c; Xiao et al. 2020; Li et al. 2024a, 2025a, 2024b, 2025e,d; Zhang et al. 2025a,b, 2023; Li et al. 2023a, 2025b,c) have become the dominant approach. Sliding-window approaches align neighboring frames to a reference using optical flow (Caballero et al. 2017; Tao et al. 2017; Xiao et al. 2021b; Lu et al. 2023b; Hu et al. 2022; Zhao et al. 2022, 2024a; Ding et al. 2022), dynamic filters (Jo et al. 2018), deformable convolutions (Tian et al. 2020; Wang et al. 2019), or attention (Isobe et al. 2020; Li et al. 2020; Cao et al. 2021; Liang et al. 2024), but often fail to capture long-range dependencies. Recurrent methods such as BasicVSR (Chan et al. 2021), BasicVSR++ (Chan et al. 2022a), and PSRT (Shi et al. 2022) improve feature propagation yet degrade under severe blur (Kai et al. 2025). Joint deblurring and SR frameworks (Fang and Zhan 2022; Youk, Oh, and Kim 2024) address motion degradation but remain vulnerable to large displacements and temporal artifacts, motivating BVSR. To overcome these limitations, we introduce event cameras into BVSR. By leveraging their microsecond-level temporal resolution and robustness to motion blur, our method effectively complements RGB inputs with fine-grained motion cues, enabling sharper and more temporally consistent VSR.

Video deblurring. Video deblurring leverages spatio-temporal cues to improve quality (Xiao and Wang 2025a; Pan, Bai, and Tang 2020; Li et al. 2021; Zhou et al. 2022; Zhong et al. 2022; Zhang, Xie, and Yao 2022; Wang et al. 2022; Suin and Rajagopalan 2021; Chao et al. 2022; Ji and Yao 2022; Jiang et al. 2022; Nah, Son, and Lee 2019; Su et al. 2017; Liang et al. 2024; Rao et al. 2024; Zhang, Xie, and Yao 2024). Recurrent models (Kim et al. 2017; Wieschollek et al. 2017; Nah, Son, and Lee 2019; Zhong et al. 2022; Wang et al. 2022; Chao et al. 2022; Liang et al. 2022b; Li et al. 2023b) propagate features sequentially, while others adopt dynamic filters (Zhou et al. 2019) or blur-invariant flow (Son et al. 2021). Bidirectional propagation (Chan et al. 2022b; Zhu et al. 2022; Ji and Yao 2022; Lin et al. 2022; Zhang, Xie, and Yao 2022) and transformers (Liang et al. 2024, 2022b,a) capture long-range dependencies, with sparse transformer designs improving efficiency (Zhang, Xie, and Yao 2024). In this paper, we propose a unified event-guided framework for BVSR that jointly addresses motion blur and resolution degradation in an end-to-end manner.

Event-guided video super-resolution. Event-guided VSR

leverages event streams for high-temporal-resolution guidance. Early work (Jing et al. 2021) uses events for LR frame interpolation, while bidirectional (Kai, Zhang, and Sun 2023) and implicit representation methods (Lu et al. 2023a) jointly model RGB and events. EvTexture (Kai et al. 2024) exploits high-frequency priors, and asymmetric fusion (Xiao et al. 2024a,b) improves efficiency. Kai *et al.* (Kai et al. 2025) first explore BVSR with events. Building on this, we propose *BluR-EVSR*, an event-guided BVSR framework.

3 Method

3.1 Overview

Given a blurry LR input sequence $\{\boldsymbol{y}_t\}_{t=i-N}^{t=i+N}$ ($\boldsymbol{y}_t \in \mathbb{R}^{H\times W\times 3}$) and the corresponding forward and backward event streams $\{\boldsymbol{E}_t^f\}_{t=i-N}^{t=i+N}$ and $\{\boldsymbol{E}_t^b\}_{t=i-N}^{t=i+N}$, the goal of *BluR-EVSR* is to reconstruct the corresponding HR output sequence $\{\hat{\boldsymbol{y}}_t\}_{t=i-N}^{t=i+N}$ ($\hat{\boldsymbol{y}}_t \in \mathbb{R}^{\alpha H\times \alpha W\times 3}$) that closely approximates the ground-truth HR sequence $\{\boldsymbol{x}_t\}_{t=i-N}^{t=i+N}$. H, W, and α are the frame height, frame width, and the upscaling factor respectively. Because event streams are not directly convenient for observation and processing by convolutional neural networks, we follow the common practice of converting event streams into voxel grids $\{\boldsymbol{V}_t^f\}_{t=i-N}^{t=i+N}$ ($\boldsymbol{V}_t^f \in \mathbb{R}^{H\times W\times bin}$) and $\{\boldsymbol{V}_t^b\}_{t=i-N}^{t=i+N}$ ($\boldsymbol{V}_t^b \in \mathbb{R}^{H\times W\times bin}$) for subsequent processing (Kai et al. 2024; Xiao et al. 2024a). bin is the number of time bins, and is set to 5 in this paper. For simplicity, we omit the time subscript and express the *BluR-EVSR* in its compact form as

$$\hat{\boldsymbol{y}} = \textit{BluR-EVSR}(\boldsymbol{y}, \boldsymbol{V}^f, \boldsymbol{V}^b). \tag{3}$$

Figure 1 illustrates the architecture of *BluR-EVSR*. We take two adjacent blurry frames $\boldsymbol{y}_t$, $\boldsymbol{y}_{t+1}$ and their corresponding event voxel grids $\boldsymbol{V}_t^f$, $\boldsymbol{V}_t^b$ as an example. Frame and event features are first extracted using dedicated encoders $Enc_F(\cdot)$ and $Enc_E(\cdot)$ with n_1 residual blocks

$$\begin{aligned} F_t^{BLR} = Enc_F(\boldsymbol{y}_t), \quad F_{t+1}^{BLR} = Enc_F(\boldsymbol{y}_{t+1}), \\ F_t^f = Enc_E(\boldsymbol{V}_t^f), \quad F_t^b = Enc_E(\boldsymbol{V}_t^b). \end{aligned} \tag{4}$$

The features are fused via a multi-level fusion module

$$F_t^{Fused} = MLF(F_t^{BLR}, F_t^f, F_t^b), \tag{5}$$

and processed through a residual block stack $\Phi(\cdot)$ to obtain the blur representation F_t^B via

$$F_t^B = \Phi(F_t^{Fused}). \tag{6}$$

This blur representation F_t^B serves as both a prior for *degradation-aware routing* and a guidance feature in *bidirectional propagation*. In the backward propagation branch, the concatenated features of $\boldsymbol{y}_t$, F_t^{BLR}, and F_{t+1}^{BLR} yield intermediate representation $F_t^{b'}$, which is further fused with the event information to generate an event-aware degradation feature $F_t^{Deg,b}$

$$F_t^{b'} = \Phi([F_t^{BLR}, F_{t+1}^{BLR}, \boldsymbol{y}_t]), F_t^{Deg,b} = \Phi([F_t^{b'}, F_t^b, V_t^b]). \tag{7}$$

These are passed through n_2 MoSDA blocks to yield the final backward branch feature F_t^{Back}. The forward counterpart F_t^{For} is obtained symmetrically. Finally, the fused temporal features F_t^{Back} and F_t^{For} are decoded using n_3 residual blocks and a pixel-shuffle layer to reconstruct the clear HR output $\hat{\boldsymbol{y}}_t$. A residual connection incorporating the bicubic-upsampled input $\boldsymbol{y}_{t\uparrow}$ is introduced to stabilize the training process and further enhance the performance. The detailed encoder and decoder architectures are provided in the supplementary material.

3.2 Blurry Representations Generation and Degradation Regularization

Multi-level fusion. In BVSR, degraded frames often exhibit scale-dependent blur patterns, where motion or defocus affects coarse structures and fine textures differently. This indicates the presence of intrinsic structural dependencies across scales. We propose a multi-scale fusion strategy to effectively capture these complementary cues that aligns and integrates RGB and event features at multiple resolutions, enabling finer texture recovery and robust motion perception across scales.

To capture scale-dependent blur and motion patterns in degraded frames, we construct multi-scale features from the fused blurry frame F_t^{BLR} and the event representation $[F_t^f, F_t^b]$, denoted as F_{t,s_i}^{BLR} and $[F_{t,s_i}^F, F_{t,s_i}^B]$ for $s_i \in \{1,2,3\}$. At each scale, features are fused via

$$F_{s_i} = \Phi(\text{Conv}([F_{t,s_i}^{BLR}, F_{t,s_i}^f, F_{t,s_i}^b, \hat{F}_{s_i})), \tag{8}$$

where $\hat{F}_{s_i}$ denotes the upsampled output from the previous scale (omitted when $i = 1$). The final enhanced representation is computed by aggregating all scales

$$F_t^{fused} = F_{t,s_i}^{BLR} + \text{Conv}([F_{s_1}, F_{s_2}\uparrow_2, F_{s_3}\uparrow_4]), \tag{9}$$

where $\uparrow_2$ and $\uparrow_4$ denote upsampling by a factor of 2 and 4, respectively. This progressive fusion strategy facilitates coarse-to-fine restoration by adaptively injecting event-guided priors at multiple spatial resolutions, enabling the blurry representation generator to better *resolve motion ambiguity and recover fine-grained textures under severe degradation*. The detailed architecture of the multi-level fusion module is provided in the supplementary material.

Blurry representations generation. Inspired by the self-supervised degradation learning framework (Cao et al. 2023), we incorporate a lightweight degradation representation branch to model the blur and noise characteristics of the current frame, which facilitates adaptive modulation of restoration features. Given the initial blur representation F_t^B (aggregated from blurry frames and event voxel features), we adopt a blurry degradation encoder Enc_K to extract latent degradation representation $\boldsymbol{D}$, where $\boldsymbol{D} = Enc_K(F_t^B)$.

Next, inspired by the mixture-of-experts paradigm (Shazeer et al. 2017; Wang et al. 2025), we construct a lightweight router to dynamically allocate the fused blurry representation F_t^B to *scale-aware degradation kernel generators*. The routing mechanism adaptively modulates the restoration process by capturing the underlying

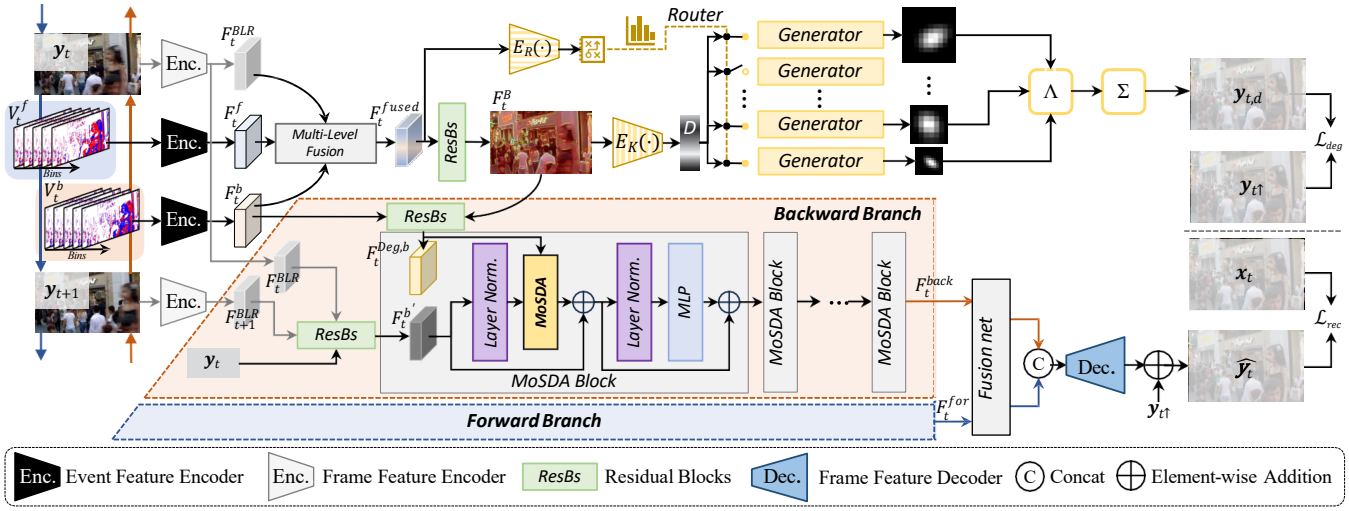

Figure 1: Overview of our proposed *BluR-EVSR*. Given blurry LR frames and their neighbors with event voxels, *BluR-EVSR* extracts spatio-temporal features, estimates event-guided degradation representations, and performs bi-directional reconstruction to generate high-quality HR outputs.

degradation characteristics, such as motion blur, at different resolutions. The learned router $\mathcal{R}$ is defined as

$$\boldsymbol{\mathcal{R}} = \sigma(\text{top}K(Enc_R(F_t^B))), \tag{10}$$

where σ denotes the softmax operation and Enc_R is the routing encoder that transforms F_t^B into a latent relevance vector. Based on the routing scores, the top$K$ operation selects the k most relevant degradation kernel generators out of g candidates. This adaptive routing strategy effectively handles scale-dependent blur and motion by activating kernel experts with varying receptive fields. It enables the network to disentangle structural and textural degradation, supporting fine-grained, content-aware restoration with improved efficiency and interpretability.

Blurry representations regularization. Each selected degradation kernel generator $f_g^{(2i+1)}$ generates a convolutional kernel $k_i \in \mathbb{R}^{(2i+1)\times(2i+1)}$ conditioned on the blurry degradation representation $\boldsymbol{D}$ and routing weights $\boldsymbol{\mathcal{R}}$

$$k_i = f_g^{(2i+1)}(\boldsymbol{\mathcal{R}}, \boldsymbol{D}), \quad i \in \{1,2,3\}. \tag{11}$$

The resulting kernel set $\mathbb{K} = \{k_1, k_2, k_3\}$ is used to convolve the fused input features F_t^{fused} in a residual path, yielding the degradation-aware representation

$$F_t^{deg} = \sum_{i=1}^{3} \Lambda(k_i, F_t^{fused}), \tag{12}$$

where $\Lambda(\cdot)$ denotes depthwise convolution. This branch enables adaptive modulation based on estimated blur representations and encourages the model to become aware of degradation characteristics during restoration. To supervise the learning of degradation representations, we apply an L_1 loss between the synthesized degraded frame $\boldsymbol{y}_{t,d}$ and the upsampled reference $\boldsymbol{y}_{t\uparrow}$

$$\mathcal{L}_{deg} = \|\boldsymbol{y}_{t\uparrow} - \boldsymbol{y}_{t,d}\|_1 . \tag{13}$$

This constraint guides the network to approximate realistic degradation patterns without requiring explicit degradation labels. This component is used only during training and does not introduce any additional computational overhead during the inference stage.

3.3 Motion-Saliency Degradation-aware Attention

To address the challenges of VSR under severe motion blur and complex degradations, it is essential to enhance motion-affected regions with fine granularity. Traditional attention mechanisms often fail to distinguish spatially variant degradation, leading to sub-optimal restoration. With their high temporal resolution and motion sensitivity, event cameras provide valuable cues for identifying motion-salient and blur-prone areas. We propose MoSDA, which modulates feature interactions using fused event-blurry representations to exploit this. Structured as stackable blocks, MoSDA forms a modular backbone that progressively refines spatio-temporal features across propagation stages.

Motion saliency prior generation. To guide the attention toward blurry degraded regions, we introduce a *motion saliency prior* derived from both spatial layout and the fused blurry-event representation. Given an input frame of size $h \times w$, it is divided into HW patches indexed by (i, j), where H and W denote the number of patches along height and width, respectively.

We compute a degradation-aware map by fusing event and blurry features, followed by average pooling within each patch to obtain degradation scores z_{ij}. Based on these scores, we define a motion-sensitive distance matrix

$$D_{ij,i'j'} = |z_{ij} - z_{i'j'}|, \tag{14}$$

where $D \in \mathbb{R}^{HW\times HW}$ quantifies degradation inconsistency, highlighting motion-salient and blur-heavy patch

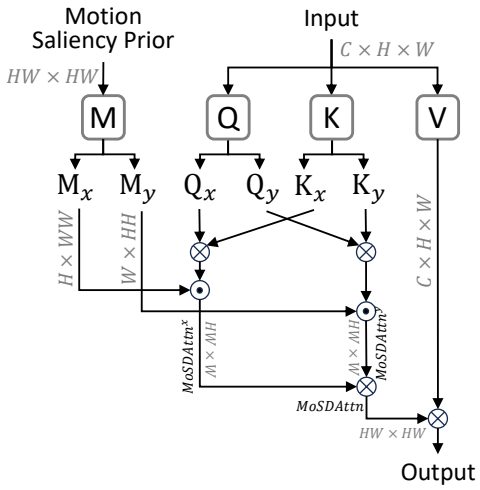

Figure 2: Detailed illustration of the proposed motion-saliency degradation-aware attention mechanism.

pairs. To complement this with structural locality, we compute the Manhattan distance between patch positions

$$S_{ij,i'j'} = |i - i'| + |j - j'|, \tag{15}$$

yielding a spatial matrix $S \in \mathbb{R}^{HW \times HW}$ that models layout proximity.

We fuse D and S into a unified motion saliency prior $M \in \mathbb{R}^{HW \times HW}$

$$M = \theta \cdot D + (1 - \theta) \cdot S, \quad \theta \in [0, 1], \tag{16}$$

where θ is a learnable weight. This fusion encodes both degradation contrast and spatial structure, offering rich geometric guidance for attention.

Motion-saliency degradation-aware attention mechanism. Given a feature map $x \in \mathbb{R}^{HW \times C}$, the standard self-attention in each head is formulated as:

$$\mathrm{SelfAtt}(Q, K, V) = \mathrm{Softmax}(QK^\top)V, \tag{17}$$

where Q, K, and V are the query, key, and value matrices derived from linear projections of x. While effective for content-based interaction, this formulation lacks spatial or degradation-awareness, limiting its ability to focus on motion-degraded regions.

To address this, we incorporate a *motion-saliency prior* $M \in \mathbb{R}^{HW \times HW}$, derived from both spatial layout and fused event-blurry features. This prior encodes relative degradation sensitivity and spatial proximity between patch pairs. Inspired by decay-based positional encoding strategies (Fan et al. 2024), we modulate the attention weights using a learnable exponential decay:

$$\mathrm{MoSDAttn}(Q, K, V, M) = (\mathrm{Softmax}(QK^\top) \odot \beta^M)V, \tag{18}$$

where $\beta \in (0, 1)$ is a learnable scalar and $\beta^M = [\beta^{m_{ij}}]$ applies suppression based on motion discrepancy m_{ij} between patch (i, j) pairs. This selectively enhances attention to motion-salient regions while suppressing irrelevant or ambiguous patches, thus improving feature propagation.

To ensure scalability for high-resolution inputs, we follow sparse attention principles and adopt a decomposed formulation that computes attention along vertical and horizontal axes independently. Specifically, we decompose M into $M^x \in \mathbb{R}^{HW \times W}$ and $M^y \in \mathbb{R}^{HW \times H}$, representing horizontal and vertical motion-aware priors, respectively. The directional attention is then computed as:

$$\mathrm{MoSDAttn}^y = \left(\mathrm{Softmax}(Q^y (K^y)^\top) \odot \beta^{M^y}\right), \tag{19}$$

$$\mathrm{MoSDAttn}^x = \left(\mathrm{Softmax}(Q^x (K^x)^\top) \odot \beta^{M^x}\right), \tag{20}$$

$$\mathrm{MoSDAttn} = \mathrm{MoSDAttn}^y (\mathrm{MoSDAttn}^x V)^\top. \tag{21}$$

Unlike conventional attention that treats all patch interactions equally, MoSDA leverages motion-saliency priors to enhance attention focus on regions affected by motion blur and degradation. By integrating fused event-blurry cues into the attention weights, it enables localized reasoning over dynamic scenes. The decomposed formulation ensures efficient computation while preserving the ability to capture structural alignment and blur-sensitive details.

Motion-saliency degradation-aware attention block. We integrate the proposed MoSDA into a standard Transformer-style backbone to construct a stackable attention unit, referred to as the MoSDA Block, as shown in the figure. Specifically, given the fused backward frame feature $F_t^{b'}$ and the degradation-guided feature F_t^{Deg}, we first apply layer normalization and feed the features into the MoSDAttn module. Motion-saliency priors guide the attention to emphasize blur-sensitive and motion-dominant regions. A residual connection is applied to retain the original context. Following this, another normalization layer and a lightweight MLP are used for further transformation, accompanied by a second skip connection. This design enables effective spatio-temporal modulation under severe motion blur and degradation, while maintaining high efficiency and modularity. By stacking multiple MoSDA blocks, we construct a powerful backbone capable of progressive feature refinement across propagation stages.

4 Experiments

4.1 Experimental Settings

Dataset setup. We follow the dataset protocol established by Kai *et al.* (Kai et al. 2025) to ensure consistency and fair comparisons. Our study utilizes three datasets encompassing both synthetic and real-world scenarios. We adopt two widely used datasets for training: GoPro (Nah, Hyun Kim, and Mu Lee 2017) and BSD (Zhong et al. 2020). Following standard VSR practices, we generate blurry LR and sharp HR pairs by applying the bicubic downsampling operation to video frames. The GoPro dataset is captured at 240 fps with a resolution of 1280×720, containing 22 sequences for training and 11 for testing. Blurry frames are synthesized by averaging consecutive sharp frames to simulate motion blur. BSD provides real blurry-sharp video pairs captured using a beam splitter system, recorded at 15 fps and 640×480 resolution. It includes 60 training and 20 testing sequences and features naturally occurring motion blur. As

Method Type		Method	GoPro			#Params (M)	FLOPs (G / frame)	Runtime (ms / frame)
			PSNR↑	SSIM↑	LPIPS↓			
Frame-only	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	0.9077	0.2287	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*	31.00	0.9065	0.2355	8.90	805.4	100.8
		Ev-DeblurVSR	<u>32.51</u>	<u>0.9314</u>	<u>0.2041</u>	8.28	459.5	79.6
		BluR-EVSR (Ours)	**32.63**	**0.9393**	**0.0967**	8.76	597.9	125.0

Table 1: Quantitative comparison on the GoPro dataset for $4\times$ BVSR. All methods are retrained on the same dataset, and evaluation is performed on the RGB channel. **Bold** and <u>underlined</u> values indicate the best and second-best performance, respectively. FLOPs and runtime are measured per 320×180 LR frame. $*$ denotes models originally designed for standard VSR, retrained here on blurry LR inputs. † indicates single-image models enhanced with optical flow refinement via SpyNet.

neither GoPro nor BSD includes real event data, we employ the Vid2E simulator (Gehrig et al. 2020) to generate synthetic event streams by converting temporal intensity variations into asynchronous event sequences. To evaluate our method under real-world settings, we further use the NCER dataset (Cho et al. 2023), which offers event-based motion deblurring data. NCER contains 27 training videos (2,583 frames) and 16 testing videos (1,454 frames), recorded with a high-speed RGB camera (522 fps) and a 640×480 DVXplorer event camera. It covers diverse scenes and challenging motion patterns, making it well-suited for benchmarking event-guided BVSR models.

Implementation details. Following (Chan et al. 2022a; Kai et al. 2025), we use 15 input frames per clip, a batch size of 8, and center-crop all input frames and event voxels to 64×64. We set $n_1 = 5$, $n_2 = 5$ and $n_3 = 50$. α is set to 4. Supervision is provided by two loss functions (Lai et al. 2017) via $\mathcal{L} = \mathcal{L}_{deg} + \mathcal{L}_{rec} = \mathcal{L}_{deg} + \sqrt{\|\hat{\boldsymbol{y}}_t - \boldsymbol{x}_t\|^2 + \varepsilon^2}$, where ε is set to 1×10^{-3} in our experiments. Data augmentation includes random horizontal and vertical flips. We first train the model on GoPro for 300K iterations using the Adam optimizer with a cosine annealing learning rate schedule. We fine-tune the GoPro-pretrained model for 200K iterations with an initial learning rate of 1×10^{-4} for BSD. The model is then further fine-tuned on NCER using the same settings. Experiments are conducted on two NVIDIA RTX 4090 GPUs, requiring approximately ten days to converge.

4.2 Quantitative and Qualitative Results

We compare our method against several typical advanced approaches across both *RGB-based* and *event-based* paradigms. Each paradigm includes two typical strategies: (1) a cascaded pipeline (*i.e.*, deblurring followed by VSR), and (2) an end-to-end BVSR approach. For RGB-based VSR, we include three recent advanced models: BasicVSR++ (Chan et al. 2022a), MIA-VSR (Zhou et al. 2024), and IART (Xu et al. 2024). We evaluate event-based VSR against two representative methods: EGVSR (Lu et al. 2023a) and EvTexture (Kai et al. 2024). To assess cascaded baselines, we further incorporate two recent leading video deblurring methods: DSTNet (Pan et al. 2023) and BSSTNet (Zhang, Xie, and Yao 2024), along with two event-guided deblurring approaches: EFNet (Sun et al. 2022) and REFID (Sun et al. 2023). Moreover, we compare against two recent BVSR methods that integrate restoration and super-resolution jointly: FMA-Net (Youk, Oh, and Kim 2024) and eSL-Net++ (Yu et al. 2023). Since our implementation strictly follows the protocol and evaluation setting of Kai *et al.* (Kai et al. 2025), we regard Ev-DeblurVSR as the most relevant and competitive baseline for direct comparison.

Method	BSD			NCER		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
BasicVSR++*	31.12	0.9050	0.2580	27.05	0.8255	0.1975
MIA-VSR*	29.24	0.8643	0.3074	24.55	0.7307	0.3251
IART*	29.47	0.8689	0.2977	25.16	0.7499	0.2908
FMA-Net	30.14	0.8805	0.2887	26.01	0.7779	0.2538
EGVSR*	29.32	0.8665	0.3145	24.26	0.7218	0.3276
EvTexture*	31.06	0.8956	0.2746	27.23	0.8136	0.2241
Ev-DeblurVSR	<u>33.02</u>	<u>0.9304</u>	<u>0.2281</u>	<u>28.60</u>	<u>0.8516</u>	<u>0.1712</u>
BluR-EVSR (Ours)	**33.30**	**0.9307**	**0.0876**	**28.85**	**0.8613**	**0.1104**

Table 2: Quantitative comparison on the BSD and NCER datasets for $4\times$ BVSR. $*$ denotes models originally designed for standard VSR, retrained on blurry LR inputs. All methods are retrained and evaluated under consistent settings.

Quantitative results. Table 1 and Table 2 present the comparison results against the baselines mentioned above. The

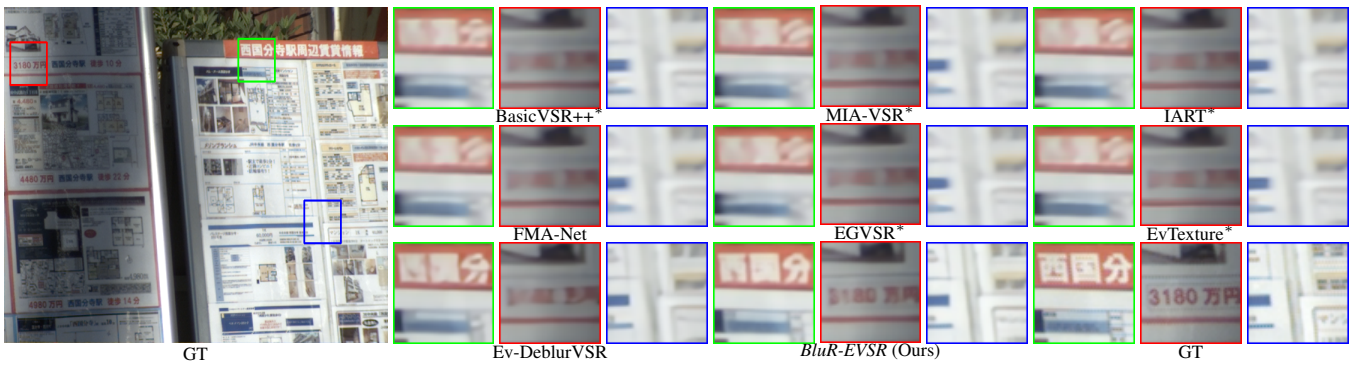

Figure 3: Visual comparisons for $4\times$ BVSR on the BSD dataset and the NCER dataset. Please zoom in to get a better view. *More visual results can be found in the supplementary material.*

Method	Core Components		GoPro	
	Deg. Reg.	MoSDA	PSNR↑	SSIM↑
(a)	✗	✗	31.98	0.9167
(b)	✓	✗	32.24	0.9281
(c)	✗	✓	32.32	0.9298
(d)	✓	✓	32.63	0.9393

Table 3: Ablation study of two core designs in *BluR-EVSR* on GoPro. We replace the removed components with residual blocks to ensure parameter consistency. "Deg. Reg." denotes degradation regularization.

data shows that our method consistently achieves superior spatial recovery in terms of PSNR, SSIM, and LPIPS. Our method, *BluR-EVSR*, consistently outperforms all baselines across PSNR, SSIM, and LPIPS metrics. On BSD, *BluR-EVSR* achieves the highest PSNR of 33.30 dB, surpassing the best-performing frame-based baseline (BasicVSR++: 31.12 dB) by a margin of 2.18 dB, and the strongest event-based competitor (Ev-DeblurVSR: 33.02 dB) by 0.28 dB. Similarly, it obtains the best SSIM of 0.9307 and the lowest LPIPS of 0.0876, indicating better perceptual and structural reconstruction. On NCER, *BluR-EVSR* again delivers the highest PSNR (28.85 dB) and SSIM (0.8613), outperforming the next-best method (Ev-DeblurVSR) by 0.23 dB PSNR and 0.0022 SSIM. It also achieves the lowest LPIPS score of 0.1104, significantly improving perceptual fidelity over prior works.

Qualitative results. We also show visual comparisons in Figure 3, where *BluR-EVSR* produces noticeably sharper edges and finer textures compared to both frame-based and event-based baselines. While competing methods struggle with over-smoothing or residual blur, particularly in regions with fast motion or fine-grained structures, our method successfully restores spatial detail.

4.3 Ablation Study

We conduct experiments on GoPro in terms of PSNR/SSIM. Main results are in Table 3. *Due to the space limitations, please refer to the supplementary material for detailed experiments and analysis.*

Table 3 reports an ablation on the GoPro dataset to assess the contributions of degradation regularization (Deg. Reg.) and the MoSDA module. The baseline (a), which lacks both components, achieves the lowest performance (31.98 dB PSNR, 0.9167 SSIM). Introducing Deg. Reg. alone in (b) improves PSNR by 0.26 dB and SSIM by 0.0111, indicating its effectiveness in modeling spatially varying degradations. Similarly, incorporating only MoSDA in (c) yields comparable gains (+0.34 dB PSNR, +0.0131 SSIM), validating the benefit of motion-aware attention. The full model (d) with both components achieves the highest performance (32.63 dB PSNR, 0.9393 SSIM), demonstrating their strong complementary effect in enhancing both spatial fidelity and temporal consistency.

5 Conclusion

We propose *BluR-EVSR*, a unified framework for BVSR that jointly models blur and motion degradation using event-guided cues. By reformulating BVSR as an event-aware degradation process, we introduce a self-supervised strategy to learn implicit blur representations and a dynamic routing mechanism to handle spatially varying degradations. Our MoSDA module further enhances motion-sensitive attention by fusing event streams and learned degradation priors. Integrated into a recurrent architecture, *BluR-EVSR* enables temporally consistent, detail-preserving restoration under severe blur. Extensive experiments demonstrate the proposed *BluR-EVSR* can achieve superior performance.

Broader impact. Leveraging blurry representations in event-guided BVSR effectively turns motion blur into a useful prior, enabling robust reconstruction in challenging fast-motion and low-light scenarios. This benefits edge applications such as autonomous driving and surveillance, while also supporting lightweight deployment. However, potential misuse for surveillance or synthetic content generation highlights the need for transparency, responsible deployment, and strong privacy safeguards.

Acknowledgments

This project is supported by the National Research Foundation, Singapore, under its Medium Sized Center for Advanced Robotics Technology Innovation.

References

Bai, H.; and Pan, J. 2024. Self-supervised deep blind video super-resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(7): 4641–4653.

Bai, H.; Zhang, J.; Zhao, Z.; Wu, Y.; Deng, L.; Cui, Y.; Feng, T.; and Xu, S. 2025. Task-driven Image Fusion with Learnable Fusion Loss. In *CVPR*.

Bai, H.; Zhao, Z.; Zhang, J.; Wu, Y.; Deng, L.; Cui, Y.; Jiang, B.; and Xu, S. 2024. ReFusion: Learning Image Fusion from Reconstruction with Learnable Loss Via Meta-Learning. *International Journal of Computer Vision*, 1–21.

Caballero, J.; Ledig, C.; Aitken, A.; Acosta, A.; Totz, J.; Wang, Z.; and Shi, W. 2017. Real-time video super-resolution with spatio-temporal networks and motion compensation. In *CVPR*.

Cao, B.; Sun, Y.; Zhu, P.; and Hu, Q. 2023. Multi-modal gated mixture of local-to-global experts for dynamic image fusion. In *ICCV*.

Cao, J.; Li, Y.; Zhang, K.; and Van Gool, L. 2021. Video Super-Resolution Transformer. *arXiv preprint arXiv:2106.06847*.

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. 2022a. BasicVSR++: Improving video super-resolution with enhanced propagation and alignment. In *CVPR*.

Chan, K. C. K.; Zhou, S.; Xu, X.; and Loy, C. C. 2022b. BasicVSR++: Improving Video Super-Resolution with Enhanced Propagation and Alignment. In *CVPR*.

Chao, Z.; Hang, D.; Jinshan, P.; Boyang, L.; Yuhao, H.; Lean, F.; and Fei, W. 2022. Deep Recurrent Neural Network with Multi-scale Bi-directional Propagation for Video Deblurring. In *AAAI*.

Chen, C.; Xiong, Z.; Tian, X.; Zha, Z.-J.; and Wu, F. 2019. Camera lens super-resolution. In *CVPR*.

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

Ding, Z.; Zhao, R.; Zhang, J.; Gao, T.; Xiong, R.; Yu, Z.; and Huang, T. 2022. Spatio-temporal recurrent networks for event-based optical flow estimation. In *AAAI*.

Dong, Y.; Xiong, R.; Zhang, J.; Yu, Z.; Fan, X.; Zhu, S.; and Huang, T. 2024. Super-resolution reconstruction from Bayer-pattern spike streams. In *CVPR*.

Fan, Q.; Huang, H.; Chen, M.; Liu, H.; and He, R. 2024. Rmt: Retentive networks meet vision transformers. In *CVPR*.

Fang, N.; and Zhan, Z. 2022. High-resolution optical flow and frame-recurrent network for video super-resolution and deblurring. *Neurocomputing*, 489: 128–138.

Farooq, M.; Dailey, M. N.; Mahmood, A.; Moonrinta, J.; and Ekpanyapong, M. 2021. Human face super-resolution on poor quality surveillance video footage. *Neural Computing and Applications*, 33: 13505–13523.

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

Hu, L.; Zhao, R.; Ding, Z.; Ma, L.; Shi, B.; Xiong, R.; and Huang, T. 2022. Optical flow estimation for spiking camera. In *CVPR*.

Isobe, T.; Jia, X.; Gu, S.; Li, S.; Wang, S.; and Tian, Q. 2020. Video super-resolution with recurrent structure-detail network. In *ECCV*.

Jeelani, M.; Cheema, N.; Illgner-Fehns, K.; Slusallek, P.; Jaiswal, S.; et al. 2023. Expanding synthetic real-world degradations for blind video super resolution. In *CVPR*.

Ji, B.; and Yao, A. 2022. Multi-Scale Memory-Based Video Deblurring. 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*.

Jo, Y.; Oh, S. W.; Kang, J.; and Kim, S. J. 2018. Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation. 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*.

Kai, D.; Zhang, Y.; Wang, J.; Xiao, Z.; Xiong, Z.; and Sun, X. 2025. Event-Enhanced Blurry Video Super-Resolution. In *AAAI*.

Kim, T. H.; Lee, K. M.; Schölkopf, B.; and Hirsch, M. 2017. Online Video Deblurring via Dynamic Temporal Blending Network. In *ICCV*.

Lai, W.-S.; Huang, J.-B.; Ahuja, N.; and Yang, M.-H. 2017. Deep laplacian pyramid networks for fast and accurate super-resolution. In *CVPR*.

Lee, S.; Choi, M.; and Lee, K. M. 2021. DynaVSR: Dynamic adaptive blind video super-resolution. In *WACV*.

Li, B.; Hu, Y.; Liu, S.; and Wang, X. 2025a. Control and Realism: Best of Both Worlds in Layout-to-Image without Training. In *ICML*.

Li, B.; Hu, Y.; Nie, X.; Han, C.; Jiang, X.; Guo, T.; and Liu, L. 2023a. Dropkey for vision transformer. In *CVPR*.

Li, B.; Zhang, Z.; Nie, X.; Han, C.; Hu, Y.; Qiu, X.; and Guo, T. 2025b. Styo: Stylize your face in only one-shot. In *AAAI*.

Li, B.; Zhang, Z.; Yang, X.; and Wang, X. 2025c. CoSER: Towards Consistent Dense Multiview Text-to-Image Generator for 3D Creation. In *CVPR*.

Li, D.; Shi, X.; Zhang, Y.; Cheung, K. C.; See, S.; Wang, X.; Qin, H.; and Li, H. 2023b. A Simple Baseline for Video Restoration With Grouped Spatial-Temporal Shift. In *CVPR*.

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, W.; Tao, X.; Guo, T.; Qi, L.; Lu, J.; and Jia, J. 2020. Mucan: Multi-correspondence aggregation network for video super-resolution. In *ECCV*.

Li, Y.; Zhang, H.; Li, L.; and Liu, D. 2025d. Learned Image Compression with Hierarchical Progressive Context Modeling. In *ICCV*.

Li, Z.; Li, J.; Li, Y.; Li, L.; Liu, D.; and Wu, F. 2024a. Inloop filtering via trained look-up tables. In *VCIP*.

Li, Z.; Liao, J.; Tang, C.; Zhang, H.; Li, Y.; Bian, Y.; Sheng, X.; Feng, X.; Li, Y.; Gao, C.; et al. 2025e. USTC-TD: A test dataset and benchmark for image and video coding in 2020s. *IEEE Transactions on Multimedia*.

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*.

Liang, J.; Fan, Y.; Xiang, X.; Ranjan, R.; Ilg, E.; Green, S.; Cao, J.; Zhang, K.; Timofte, R.; and Gool, L. V. 2022a. Recurrent Video Restoration Transformer with Guided Deformable Attention.

Liang, J.; Fan, Y.; Xiang, X.; Ranjan, R.; Ilg, E.; Green, S.; Cao, J.; Zhang, K.; Timofte, R.; and Van Gool, L. 2022b. Recurrent Video Restoration Transformer with Guided Deformable Attention.

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

Liu, C.; and Sun, D. 2013. On Bayesian adaptive video super resolution. *IEEE transactions on pattern analysis and machine intelligence*, 36(2): 346–360.

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

Lu, Z.; Xiao, Z.; Bai, J.; Xiong, Z.; and Wang, X. 2023b. Can SAM Boost Video Super-Resolution? *arXiv preprint arXiv:2305.06524*.

Mao, Y.; Xiao, Z.; An, P.; Liu, D.; and Shan, C. 2025. Deep Sparse-to-Dense Inbetweening for Multi-View Light Fields. *IEEE Transactions on Image Processing*.

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

Nah, S.; Son, S.; and Lee, K. M. 2019. Recurrent neural networks with intra-frame iterations for video deblurring. In *CVPR*.

Pan, J.; Bai, H.; Dong, J.; Zhang, J.; and Tang, J. 2021. Deep blind video super-resolution. In *ICCV*.

Pan, J.; Bai, H.; and Tang, J. 2020. Cascaded Deep Video Deblurring Using Temporal Sharpness Prior. 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*.

Peng, C.; Lin, W.-A.; Liao, H.; Chellappa, R.; and Zhou, S. K. 2020. Saint: spatially aware interpolation network for medical slice synthesis. In *CVPR*.

Rao, C.; Li, G.; Lan, Z.; Sun, J.; Luan, J.; Xing, W.; Zhao, L.; Lin, H.; Dong, J.; and Zhang, D. 2024. Rethinking Video Deblurring with Wavelet-Aware Dynamic Transformer and Diffusion Model. In *ECCV*.

Shazeer, N.; Mirhoseini, A.; Maziarz, K.; Davis, A.; Le, Q.; Hinton, G.; and Dean, J. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*.

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

Son, H.; Lee, J.; Lee, J.; Cho, S.; and Lee, S. 2021. Recurrent Video Deblurring with Blur-Invariant Motion Estimation and Pixel Volumes. *IEEE TIP*, 40(5): 185:1–185:18.

Su, S.; Delbracio, M.; Wang, J.; Sapiro, G.; Heidrich, W.; and Wang, O. 2017. Deep Video Deblurring for Hand-held Cameras. In *CVPR*.

Suin, M.; and Rajagopalan, A. N. 2021. Gated Spatio-Temporal Attention-Guided Video Deblurring. In *CVPR*.

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*.

Tang, X.; Zhao, X.; Liu, J.; Wang, J.; Miao, Y.; and Zeng, T. 2023. Uncertainty-aware unsupervised image deblurring with deep residual prior. In *CVPR*.

Tao, X.; Gao, H.; Liao, R.; Wang, J.; and Jia, J. 2017. Detail-revealing deep video super-resolution. In *ICCV*.

Tian, Y.; Zhang, Y.; Fu, Y.; and Xu, C. 2020. Tdan: Temporally-deformable alignment network for video super-resolution. In *CVPR*.

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, Y.; Lu, Y.; Gao, Y.; Wang, L.; Zhong, Z.; Zheng, Y.; and Yamashita, A. 2022. Efficient Video Deblurring Guided by Motion Magnitude. In *ECCV*.

Wang, Z.; Yan, Z.; Pan, J.; Gao, G.; Zhang, K.; and Yang, J. 2025. DORNet: A Degradation Oriented and Regularized Network for Blind Depth Super-Resolution. In *CVPR*.

Wieschollek, P.; Hirsch, M.; Schölkopf, B.; and Lensch, H. P. A. 2017. Learning Blind Motion Deblurring. In *ICCV*.

Xiao, Y.; Su, X.; Yuan, Q.; Liu, D.; Shen, H.; and Zhang, L. 2021a. Satellite video super-resolution via multiscale deformable convolution alignment and temporal grouping projection. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–19.

Xiao, Y.; Yuan, Q.; Zhang, Q.; and Zhang, L. 2023. Deep blind super-resolution for satellite video. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–16.

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

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

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

Xiao, Z.; Li, Z.; and Jia, W. 2025. Occlusion-Embedded Hybrid Transformer for Light Field Super-Resolution. In *AAAI*.

Xiao, Z.; Lu, Z.; and Wang, X. 2024. P-bic: Ultra-high-definition image moiré patterns removal via patch bilateral compensation. In *ACM MM*.

Xiao, Z.; and Wang, X. 2025a. Asymmetric Dual-Lens Video Deblurring. *NeurIPS*.

Xiao, Z.; and Wang, X. 2025b. Event-based Video Super-Resolution via State Space Models. In *CVPR*.

Xiao, Z.; Weng, W.; Zhang, Y.; and Xiong, Z. 2022. EVA2: Event-Assisted Video Frame Interpolation via Cross-Modal Alignment and Aggregation. *IEEE Transactions on Computational Imaging*, 8: 1145–1158.

Xiao, Z.; and Xiong, Z. 2025. Incorporating degradation estimation in light field spatial super-resolution. *Computer Vision and Image Understanding*, 252: 104295.

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

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

Zhang, C.; Lin, M.; Zhang, X.; Jiang, C.; and Yu, L. 2024. Super-Resolving Blurry Images with Events. *arXiv preprint arXiv:2405.06918*.

Zhang, H.; Xie, H.; and Yao, H. 2022. Spatio-Temporal Deformable Attention Network for Video Deblurring. In *ECCV*.

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

Zhang, K.; Gool, L. V.; and Timofte, R. 2020. Deep unfolding network for image super-resolution. In *CVPR*.

Zhang, X.; Cai, N.; Zhang, H.; Zhang, Y.; Di, J.; and Lin, W. 2023. AFD-Former: A Hybrid Transformer With Asymmetric Flow Division for Synthesized View Quality Enhancement. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(8): 3786–3798.

Zhang, X.; Ma, J.; Wang, G.; Zhang, Q.; Zhang, H.; and Zhang, L. 2025a. Perceive-IR: Learning to Perceive Degradation Better for All-in-One Image Restoration. *IEEE Transactions on Image Processing*, 1–1.

Zhang, X.; Zhang, H.; Wang, G.; Zhang, Q.; Zhang, L.; and Du, B. 2025b. UniUIR: Considering Underwater Image Restoration as an All-in-One Learner. *IEEE Transactions on Image Processing*, 34: 6963–6977.

Zhao, J.; Xie, J.; Xiong, R.; Zhang, J.; Yu, Z.; and Huang, T. 2021. Super resolve dynamic scene from continuous spike streams. In *ICCV*.

Zhao, J.; Xiong, R.; Zhang, J.; Zhao, R.; Liu, H.; and Huang, T. 2023a. Learning to super-resolve dynamic scenes for neuromorphic spike camera. In *AAAI*.

Zhao, R.; Xiong, R.; Zhang, J.; Yu, Z.; Zhu, S.; Ma, L.; and Huang, T. 2023b. Spike camera image reconstruction using deep spiking neural networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(6): 5207–5212.

Zhao, R.; Xiong, R.; Zhang, J.; Zhang, X.; Yu, Z.; and Huang, T. 2024a. Optical Flow for Spike Camera with Hierarchical Spatial-Temporal Spike Fusion. In *AAAI*.

Zhao, R.; Xiong, R.; Zhao, J.; Yu, Z.; Fan, X.; and Huang, T. 2022. Learning optical flow from continuous spike streams. In *NeurIPS*.

Zhao, R.; Xiong, R.; Zhao, J.; Zhang, J.; Fan, X.; Yu, Z.; and Huang, T. 2024b. Boosting spike camera image reconstruction from a perspective of dealing with spike fluctuations. In *CVPR*.

Zhao, Z.; Bai, H.; Zhang, J.; Zhang, Y.; Xu, S.; Lin, Z.; Timofte, R.; and Van Gool, L. 2023c. CDDFuse: Correlation-Driven Dual-Branch Feature Decomposition for Multi-Modality Image Fusion. In *CVPR*.

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

Zhong, Z.; Gao, Y.; Zheng, Y.; Zheng, B.; and Sato, I. 2022. Real-World Video Deblurring: A Benchmark Dataset and an Efficient Recurrent Neural Network. *IJCV*.

Zhou, K.; Li, W.; Lu, L.; Han, X.; and Lu, J. 2022. Revisiting Temporal Alignment for Video Restoration. In *CVPR*.

Zhou, S.; Zhang, J.; Pan, J.; Zuo, W.; Xie, H.; and Ren, J. S. J. 2019. Spatio-Temporal Filter Adaptive Network for Video Deblurring. In *ICCV*.

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, 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*.