Time-Specialized Event-Image Alignment for Blur-to-Video Decomposition

Zhijing Sun, Senyan Xu, Ruixuan Jiang, Kean Liu, Runze Tian, Xueyang Fu, Zheng-Jun Zha*
University of Science and Technology of China, China
{sunzhijing, syxu, ruixuanjiang, rickyliu, trz220765}@mail.ustc.edu.cn
{xyfu, zhazj}@ustc.edu.cn

Abstract

Motion blur is a common degradation in dynamic imaging. Recent studies have moved beyond restoring a single sharp image from a blurred input and instead target blur decomposition: recovering a temporally continuous sharp video sequence from one motion-blurred image. Event cameras, with their microsecond temporal resolution, can effectively alleviate motion ambiguity. However, existing event-based methods often fail to explicitly model time-aligned event–image features. How to accurately exploit event data to reconstruct frames at different time instants remains largely underexplored. In this paper, we propose TSANet, an event-based blur-to-video decomposition method that time-specializes both event features and image features for alignment. Specifically, we introduce a Relative Time-Encoded Attention module that steers event features toward motion information relevant to a given target time, and a Timesurface Dynamic Warping module that warps image features into the spatial configuration corresponding to that time. With time-specialized motion and image features explicitly aligned at arbitrary query times, our framework can decompose a single blurred image into a high–frame-rate sharp video sequence. In addition, we collect a new dataset containing real events and high-quality color videos, and synthesize blurred inputs by averaging sharp frames to evaluate our method. Experiments on multiple datasets with both synthetic and real events demonstrate that our approach consistently outperforms previous state-of-the-art methods on the blur decomposition task. Code at https://github.com/ZhijingS/TSANet.

1. Introduction

Relative motion between camera and captured object during photography often results in motion blur, a prevalent image degradation phenomenon. Previous researchers [4, 15, 25, 36, 40] have done extensive studies on reconstructing a sharp image from a blurred one. Recently, the research direction has been extended to a more challenging task of motion decomposition, which aims to recover a sharp video sequence from a single blurred image [12]. The problem is inherently ill-posed because different motion trajectories can integrate over the exposure to produce the same blurred image. This phenomenon is known as the motion ambiguity of motion-blurred frames. Figure 1 provides a straightforward example of this ambiguity.

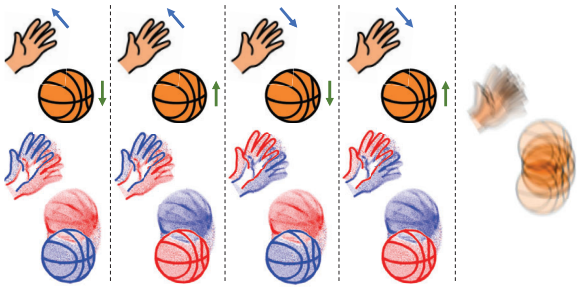

Figure 1. Illustration of motion ambiguity in a toy hand–ball example. The top row in (a) shows four possible motion patterns during exposure: hand moving up while ball moves down, both moving up, both moving down, and hand moving down while ball moves up. After temporal averaging, all of them produce the same blurred image in (b). Relying only on the blurred image therefore leads to intrinsic motion ambiguity. The bottom row in (a) shows the corresponding event data, which encodes the underlying motion process and provides strong cues to disambiguate the blur.

Prior image-only methods attempt to resolve the inherent motion ambiguity in a single blurry image through various strategies, such as imposing temporal consistency losses [12, 29], using multi-frame inputs [26, 30, 46, 47], or leveraging information from rolling shutter images [10]. However, these approaches often fail when faced with large, complex motion, as the crucial temporal information is irrecoverably lost in the blur. In contrast, event cameras [1, 5, 6, 20, 33] offer a powerful alternative. With their microsecond-level temporal resolution, they asynchronously capture pixel-level brightness changes, thereby providing a precise and robust record of the very motion trajectories lost to the conventional sensor. Researchers

have leveraged events as an auxiliary input to great effect [8, 22, 28, 42], though existing methods exhibit clear limitations. Pan et al. [27], Wang et al. [37] relied on physics-based models to define the relationship between the latent sharp frames, the blurry image, and the events, but these idealized models are often sensitive to the noise and imperfections of real-world event data. Other works, like [22], adopted a two-stage pipeline that decomposes the problem into cascaded deblurring and interpolation tasks, but this design is prone to error accumulation. More recent learning-based methods, such as [45] generate time-specific event representations to guide reconstruction, yet they lack mechanisms for explicit feature-level alignment between the two modalities, which fundamentally limits their performance.

Synthesizing these limitations reveals a common underlying challenge: to truly unlock the potential of event data for motion decomposition, a model must have the ability to explicitly align the motion-rich features from the event stream with the texture-rich features from the image to an arbitrary target time t. Based on this insight, we propose TSANet, an event-based blur decomposition framework centered on the principle of Time Specialized Alignment. To be specific, we argue that **Time-Specialization** of information from both modalities is necessary before feature fusion. We direct event features to focus on motion relevant to time t, while warping image features to their corresponding spatial configuration at time t.

For the former, we propose a Relative Time-Encoded Attention (RTEA) module that generates an attention score bias based on the relative temporal distance between the given time t and each event frame, guiding the aggregation of event features to produce a time-specialized motion feature. For the latter, we design a Timesurface Dynamic Warping (TDW) module, which leverages event timesurface representations enriched with motion priors to guide the warping of average-exposed image features into the spatial configuration corresponding to the target time t. With the support of these two modules, we obtain time-specialized image and motion features alignable at any time. Through a simple yet efficient gating fusion mechanism, we can reconstruct a high-frame-rate sharp video sequence.

Furthermore, to supplement the research community with a high-quality real-world event dataset featuring well-aligned modalities and to validate the efficacy of our method in handling real-world blur decomposition, we have collected a dataset named EBD comprising sensor-level hardware-aligned RGB frames and corresponding event data. Following prior works, we generated blurred images by averaging sharp video sequences.

Overall, our contributions can be summarized as follows:

- We present TSANet, a novel event-based blur decomposition framework centered on the principle of **T**ime **S**pecialized **A**lignment.
- We propose Relative Time-Encoded Attention (RTEA) encoding the relative temporal distance to extract motion cues tightly focused on the target time.
- We design Timesurface Dynamic Warping (TDW) to geometrically align image features to time-specialized spatial configuration with timesurface guidance.

Our TSANet achieves state-of-the-art performance across various datasets, yielding PSNR gains of +0.64dB on GoPro, +1.06dB on HighREV, and +1.94dB on EBD over previous methods.

2. Related Work

2.1. Image-based Blur Decomposition

The blur decomposition task seeks to "unroll" a single motion-blurred frame into a temporally ordered sequence of sharp frames, despite exposure-time integration destroying temporal order and masking motion direction. Jin et al. [12] first exploit and cast the problem as predicting frame pairs from a blurred input and use order-invariant loss to neutralize forward/backward ambiguity, establishing a practical training paradigm. Purohit et al. [30] pretrain a convolutional recurrent autoencoder on sharp videos to learn a motion representation, then transfer it to encode motion from a single blurred image, guiding sequential frame synthesis.

More recent methods advance along three complementary axes: (i) stronger sequence modeling, where the BiT [47] leverages multi-scale transformer with two-sided temporal supervision and symmetry aggregation to better recover latent temporal correlations; (ii) explicit ambiguity resolution, where Zhong et al. [46] condition a two-stage decomposition network on discretized motion directions to deliver physically plausible, multi-modal reconstructions, and Pham et al. [29] replaces order-invariant losses with a self-supervised temporal ordering scheme that explicitly separates forward and reverse motion; and (iii) external temporal priors, where Ji et al. [10] exploits rolling-shutter scan timing as an extrinsic ordering cue and couples it with global shutter branches to improve detail fidelity and identifiability in real scenes. Overall, researchers try to reduce the ill-posedness of motion ambiguity in blur-to-video recovery; nevertheless, under extreme displacements and compound camera motion, purely image-based methods still face direction ambiguity and performance degradation.

2.2. Event-based Blur Decomposition

Event cameras provide microsecond-level temporal cues, which have been widely applied in low-level vision [7, 13, 14, 18, 19, 23, 43, 44] and high-level vision [2, 3, 17, 24, 41, 48]. The advantages of event streams are critical for resolving the motion ambiguity inherent in blur decomposition. Early works, such as the seminal physics-guided model EDI [27], coupled the blur formation process with

asynchronous events to analytically recover a high-frame-rate video. However, these methods are often sensitive to sensor noise [42], limiting their flexibility in complex, real-world scenarios.

To overcome these limitations, learning-based methods have become the dominant paradigm. Methods like EVDI [45] and [22] demonstrated that end-to-end models could reconstruct the sharp video sequence. However, a fundamental limitation persists in these and subsequent approaches, including those exploring continuous representations like E-CIR [32]. They typically treat events as a holistic motion descriptor for the entire exposure window, lacking mechanisms for explicit, feature-level alignment to an arbitrary query time t. For instance, EVDI [45] conditions on t only during input preprocessing, which is inefficient for dense video generation and fails to align features dynamically within the network. Consequently, the alignment of time-specific motion cues from events with rich but time-agnostic texture from the image remains a largely implicit and unresolved challenge.

We argue that the key lies in explicit time-specialized alignment. Our approach is built on a simple but critical principle: for any target time t, the network must first align both modalities to that instant—filtering event features to represent instantaneous motion and warping image features to the corresponding spatial configuration. Fusing these properly aligned features is the cornerstone for reconstructing a high-quality, temporally coherent video sequence.

3. Method

3.1. Architecture Overview

Figure 2 illustrates the overall workflow of our proposed method. Given a single blurry image B and the corresponding events $\mathcal{E}$ captured during its exposure time, our goal is to reconstruct a high-frame-rate, sharp video sequence by restoring the latent sharp frame S_t at arbitrary time t, where $t \in [0, 1]$ with 0 and 1 denoting the start and end of the exposure period, respectively. This process can be formally expressed as:

$$S_t = \phi(B, \mathcal{E}, t), \tag{1}$$

where ϕ denotes our proposed event-based blur decomposition network. First, we convert the event data into event voxels E [35] and event timesurfaces TS [16]. The event voxels are fed into an event branch consisting of Conv blocks and Swin Transformer blocks, which capture the spatiotemporal motion dynamics throughout the entire exposure, detailed structure in $Supp$. In the image branch, we use SFHBlocks [11] to extract the global texture features. The crux of our method lies in the subsequent Time Specialization stage. Here, we introduce two novel modules to align these global features to the target time t. The Relative Time-Encoded Attention (RTEA) module aggregates the event motion features to instant t, and the Timesurface Dynamic Warping (TDW) module uses the timesurfaces to geometrically transform the image features to the corresponding spatial configuration specialized at time t. Finally, these time-specialized motion and texture features are fused via a lightweight Event Guide Gating Fusion (EGGF) module and passed to a decoder to reconstruct the final sharp frame S_t.

3.2. Relative Time-Encoded Attention

Accurately extracting motion information for a specific time t from the high-temporal-resolution event stream is a cornerstone of event-based blur decomposition. However, most existing methods treat events as a holistic motion descriptor over the entire exposure, lacking a precise mechanism to specialize features for a target time. Although recent works such as EVDI [45] address this by generating a different event representation for each t in preprocessing. This approach introduces a significant computational burden during both training and inference, making it inefficient for generating dense video sequences. To overcome these limitations, we propose the Relative Time-Encoded Attention (RTEA) module, guiding the model to focus on motion information temporally close to the target t by explicitly encoding their relative temporal distance into the network.

As illustrated in Figure 2(a), the RTEA module dynamically re-weights a sequence of event feature maps based on their relative temporal distance to the query time t. Given an input event feature sequence $F_E \in \mathbb{R}^{N\times C\times H\times W}$, where N is the number of temporal bins, the process unfolds as follows. First, we establish a content-based relevance score using a standard query-key attention mechanism. We compute a global feature representation for each event frame by spatially averaging F_E, yielding a tensor $\hat{F_E} \in \mathbb{R}^{N\times C}$. This is then linearly projected to form the key tensor $K \in \mathbb{R}^{N\times D}$. Concurrently, the query time t is passed through a Fourier positional embedding and a small MLP to produce the query vector $q \in \mathbb{R}^{D\times 1}$. The initial content-based attention logits W are then computed via scaled dot-product:

$$\begin{aligned} K &= MLP(AvgPool(F_E)), \\ q &= MLP(Embed(t)), \\ W &= K \cdot q. \end{aligned} \tag{2}$$

Crucially, to inject a strong temporal prior, we introduce a relative temporal position bias T_{bias} to refine these content-based scores, which is the core of RTEA module. We first compute the normalized relative distance d_n, between each frame index n and the target time t:

$$d_n = \frac{n-p}{N-1}, \quad p = t \times (N-1). \tag{3}$$

This simple yet effective metric d_n quantifies how far each event frame is from the query time t. We then generate the

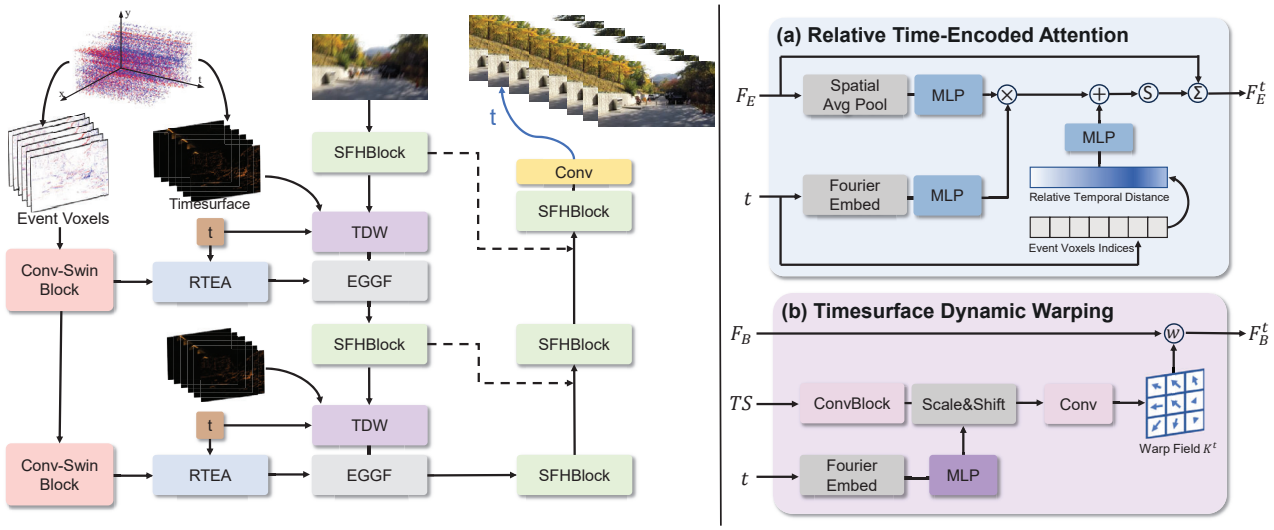

Figure 2. Overall framework of the proposed method. The RTEA module adds relative-time biases to attention so the model focuses on motion near the queried time. The TDW module uses the timesurface as a motion prior to predict a warp field that moves time-averaged image features into the correct spatial position at that time. Details of other modules are in *Supp*.

bias T_{bias} by passing d_n and its square d_n^2 through an MLP. The bias is then added to the initial logits W to get the final attention weights η, explicitly encouraging the model to favor frames that are temporally closer to the target t. After the weighted sum over input event feature F_E and η, we get the time-specialized motion feature $F_E^t \in \mathbb{R}^{C\times H\times W}$:

$$\begin{aligned} T_{bias} &= MLP([d_n, d_n^2]), \\ \eta &= Softmax(W + T_{bias}), \\ F_E^t &= \eta \cdot F_E, \end{aligned} \tag{4}$$

where F_E^t represents an aggregation of event features that is precisely focused on the motion characteristics around the desired instant. To better balance the global and local motion during the blurring process, we employ the RTEA module in both the global spatial scale and the window spatial scale. Using a learnable weight α, we adaptively fuse the time-specialized features aggregated from both scales.

3.3. Timesurface Dynamic Warping

Due to the temporal integration inherent in the image capture process, features extracted from a blurry image represent time-averaged texture. This creates a fundamental spatial misalignment between these average features and the true scene configuration at a specific time t. We argue that aligning these texture-rich features to the correct spatial state at time t is a critical prerequisite for effective fusion with our time-specialized motion features. To achieve this, we propose the Timesurface Dynamic Warping (TDW) module, which leverages the rich motion history encoded in the event timesurface to guide the geometric transformation of the blurry image features. The event timesurface is uniquely suited for this task. By encoding the timestamp of the most recent event at each pixel, it forms a map that implicitly traces the motion trajectories across the time window. It therefore contains the precise local motion priors required to predict the spatial warp field needed to transform the averaged features to a specific position.

The TDW module takes the global blurry image features F_B, the event timesurface TS and the target time t as input. The core of the module is to generate a time-conditioned warp field K from TS and use it to transform F_B.

First, we process the timesurface TS through a convolutional block to extract its latent motion patterns. To make this process conditional on query time t, we employ a Scale and Shift layer. The embedded time t is passed through a small MLP to generate channel-wise scaling parameters γ and shifting parameters β. It can be formulated as:

$$\begin{aligned} \gamma, \beta &= MLP(Embed(t)), \\ M^t &= \gamma \cdot ConvBlock(TS) + \beta, \end{aligned} \tag{5}$$

where M^t denotes the time-conditioned motion representation. These parameters modulate the intermediate timesurface features, steering the feature extraction to be most relevant to the state at instant t.

Subsequently, a final convolutional head maps M^t to a 2-channel warp field $K^t \in \mathbb{R}^{2\times H\times W}$, which represents the predicted pixel-wise displacement $(\Delta x, \Delta y)$ required for alignment. The final time-specialized image feature F_B^t is obtained by the process that can be formulated as:

$$\begin{aligned} K^t &= Conv(M^t), \\ F_B^t &= Warp(F_B, K^t). \end{aligned} \tag{6}$$

where $Warp(\cdot)$ represents the warping operation. The original feature map F_B is warped using the field K^t via differentiable bilinear sampling.

3.4. Event Guide Gating Fusion

With both the event and image features now aligned to the target time t, the final step is to fuse these complementary features. Because our preceding time specialization operations have already handled the complex task of spatio-temporal alignment, we can bypass computationally expensive fusion mechanisms, such as elaborate cross-attention transformers. We therefore propose a lightweight yet effective Event Guide Gating Fusion (EGGF) module designed for precise information integration.

Initially, we refine the time-specialized event feature F_E^t with the dense motion representation M^t from the TDW module by scale and shift. This produces an enhanced feature $\hat{F_E^t}$, which holds a comprehensive summary of motion at time t. Subsequently, this refined feature is processed by another convolutional layer and a RELU activation to generate a spatial gating map G. This gate then selectively scales the time-specialized image features F_B^t through element-wise multiplication, emphasizing texture details in regions where significant motion was detected. Finally, a residual connection adds the original image feature back, ensuring the module primarily injects event-guided details without overwriting the base texture information. The workflow of the EGGF module can be summarized as:

$$\begin{aligned} \gamma_m, \beta_m &= Chunk(ReLU(Conv(M^t))), \\ \hat{F_E^t} &= \gamma_m \cdot F_E^t + \beta_m, \\ G &= ReLU(Conv(\hat{F_E^t})), \\ F_{fused} &= G \odot F_B^t + F_B^t, \end{aligned} \tag{7}$$

where F_{fused} represents the resulting fused feature, which combines the spatial precision of the warped image features with the motion-specific details from the event stream. The fused feature also provides a precise input for the final reconstruction decoder.

4. Experiments

4.1. Dataset Preparation

EBD dataset. We collect a new event–RGB dataset using a DVSync event camera, comprising 29 color video sequences with a total of 25,608 sharp frames and their corresponding events. The sequences cover diverse scenes (indoor, playground, road, architecture) and motion types (camera motion and object motion). We split the data into 23 training sequences (19,305 frames) and 6 test sequences (6,303 frames). All videos are captured at a resolution of 640×1120. The EBD dataset is also suitable for event-based interpolation and deblurring tasks. Additional details are provided in *Supp.*

Training Preparation. We extensively evaluate our method on both synthetic-event and real-event datasets. For the synthetic setting, we adopt the GoPro dataset and use the GoPro_Large_all split provided by [25], which contains 240 fps sharp videos captured with a GoPro Hero4 Black camera. All sharp frames are downsampled to a resolution of 640×360. To establish a one-to-one correspondence between target time stamps and ground-truth frames, we average 11 consecutive sharp frames to synthesize a blurry image, and assign the normalized time labels $\{0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$ to the 11 constituent sharp frames. In parallel, we use the v2e [9] event simulator to generate event streams from these 11 sharp frames associated with each blurry image.

With the procedure, we obtain 2058 blurry–event pairs for training. During training, we randomly sample one target time t from the 11 candidates and use the corresponding sharp frame as supervision. For testing, we obtain 1089 blurry–event pairs. In the "$\times 5$" setting, we decompose each blurry image into five sharp frames at time stamps $\{0.0, 0.3, 0.5, 0.7, 1.0\}$ and report the average PSNR and SSIM over all reconstructed frames and their ground truths.

For real-event evaluation, we use the HighREV [34] dataset and our collected EBD dataset, both of which provide high-quality color videos and synchronized real events. We synthesize blurry inputs in the same way, by averaging 11 consecutive sharp frames, to assess our method under diverse real-world scenarios. For all public datasets, we follow the official training/testing splits.

4.2. Comparison with SOTA Method

We compare our approach with both image-based and event-based methods for blur decomposition. On the image side, we include two recent blur-to-video methods, LEVS [12] and BiT [47], as well as the interpolation method DeMFI [26] and BimVFI [31]. On the event side, we compare against deblurring methods that can reconstruct sharp sequences (RED [42], E-CIR [32]) and event-guided interpolation methods (LEDVDI [21], EVDI [45], EBFI [39], REFID [34], EvEnhancer [38]). To fairly assess the capability of all methods on the blur decomposition task, we retrain every model on the three datasets used in our experiments.

Tab. 1 reports quantitative results, where "$\times 5$" denotes that the reconstructed video contains 5 frames. Our method consistently outperforms all competing approaches. In particular, compared with image-based methods that rely on multiple blurry inputs, our approach achieves at least 1.14 dB, 4.6 dB, and 3.4 dB PSNR gains on GoPro[25], HighREV[34], and EBD, respectively. This indicates that events captured during the exposure provide a more precise and reliable description of the underlying motion than what

Table 1. Quantitative results on GoPro, HighREV and EBD datasets. The best and second-best results are boldfaced and underlined, respectively.

Methods	Events	GoPro×5		HighREV×5		EBD×5		Params
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	(M)
LEVS[12]	✗	24.29	0.831	27.22	0.893	24.79	0.856	10.8
BiT[47]	✗	26.99	0.868	31.01	0.926	25.62	0.873	11.3
DeMFI[26]	✗	27.15	0.881	31.89	0.951	24.42	0.879	7.4
BimVFI[31]	✗	27.26	<u>0.885</u>	32.24	0.954	24.38	0.876	15.4
LEDVDI[21]	✓	25.50	0.871	31.37	0.837	23.54	0.845	16.3
RED[42]	✓	25.25	0.793	30.79	0.829	23.38	0.848	8.5
E-CIR[32]	✓	24.78	0.861	30.16	0.845	24.58	0.867	2.1
EVDI[45]	✓	24.61	0.861	32.55	0.915	26.36	0.882	0.4
EBFI[39]	✓	25.25	0.830	32.72	<u>0.957</u>	25.77	0.842	13.2
REFID[34]	✓	27.25	0.873	35.67	0.947	<u>27.84</u>	<u>0.890</u>	15.9
EvEnhancer[38]	✓	<u>27.76</u>	0.878	<u>35.78</u>	0.951	27.62	0.878	6.6
Ours	✓	**28.40**	**0.908**	**36.84**	**0.974**	**29.02**	**0.916**	6.3

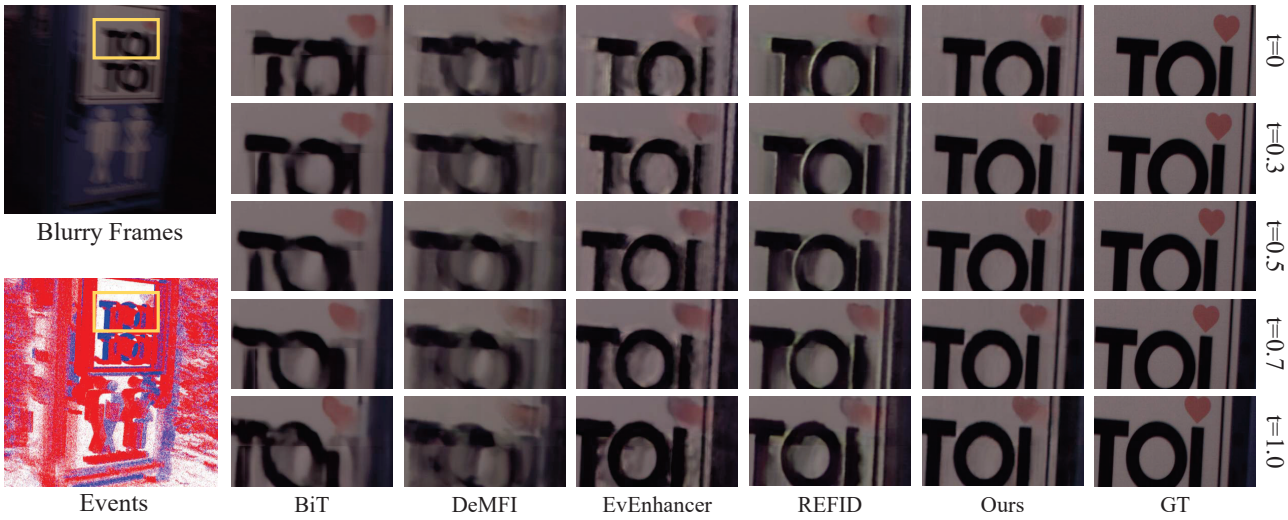

Figure 3. Qualitative comparisons on the HighREV dataset. We visualize five reconstructed frames at selected times from a single blurred input. Methods that infer motion only from neighboring RGB frames (BiT, DeMFI) exhibit noticeably more artifacts in challenging scenes compared with event-aided approaches. Among all methods, ours delivers the most faithful texture recovery and the most consistent motion reconstruction. Zoom in for better view.

can be inferred solely from multiple blurred frames. Compared with event-based baselines, the performance gains further highlight that simply introducing events is insufficient; accurate time-specific alignment between motion features and image features is crucial for high-quality blur decomposition. Our RTEA module extracts motion features that are tightly focused on the target time, while TDW warps time-averaged texture features toward the correct spatial configuration at that instant. Under this time-specialized dual-modality fusion, our method can produce sharp videos and achieves at least 0.64dB, 1.06dB, and 1.4dB improvements over prior event-based methods on the three datasets.

We also show the longer video generation results in Tab. 2. Qualitative comparisons in Figures 3 to 5 further show that our approach recovers sharper local details and more temporally consistent motion trajectories than existing methods.

To better demonstrate the temporal coherence of our reconstructed videos, we further compare spatio-temporal slices on the EBD dataset in Figure 6. Each slice is constructed by stacking a narrow band of pixels from the same spatial location across time. As shown in Figure 6, competing methods produce blurred, duplicated, or jittery patterns that reveal temporal inconsistency across frames, whereas our approach yields sharp, continuous motion traces that

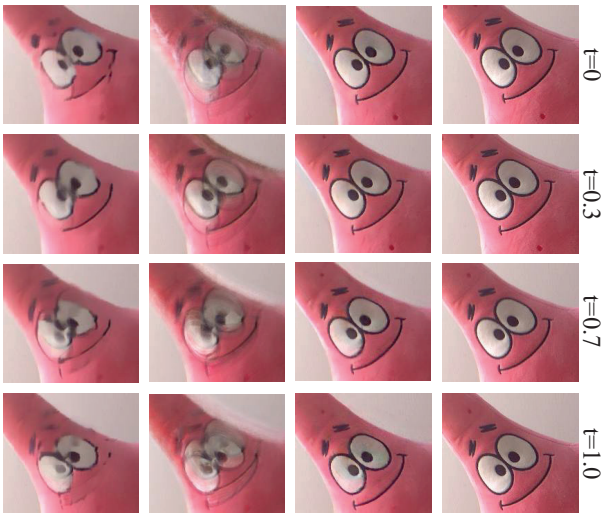

Figure 4. Qualitative comparisons on the EBD dataset. Zoom in for better view.

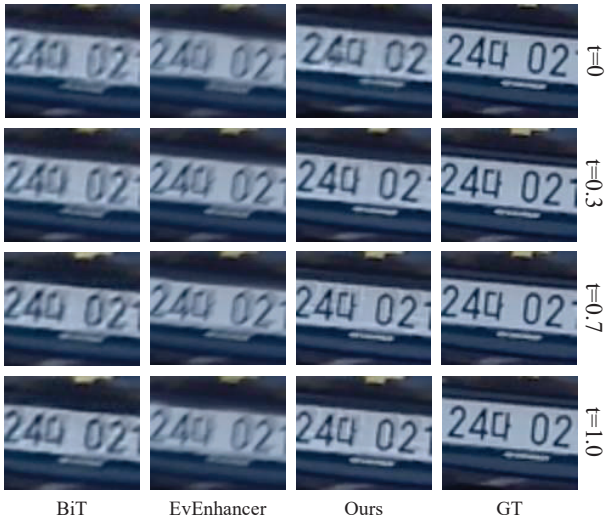

Figure 5. Qualitative comparisons on the GoPro dataset. Zoom in for better view.

closely follow the ground-truth trajectories.

4.3. Ablation studies

Adequate ablation experiments were conducted on the HighREV [34] dataset to validate the effectiveness of our module designs.

Effects of the RTEA Module. As shown in Tab. 3, we constructed the baseline model (Case 1) by replacing the RTEA module with a simple module that computes attention scores via global average pooling followed by an MLP, while directly removing the TDW module. By incorporating the RTEA module in Case 2, we achieved improvements of 1.5dB in PSNR and 0.007 in SSIM. This demonstrates that the RTEA module endows the model with the ability to focus on motion states at arbitrary time instants, which serves as the cornerstone for fully leveraging event data to decompose blurred images into sharp video sequences. In Figure 7, we present the attention weight generated by the RTEA module, along with visualization results of the model with and without the RTEA module. It can be observed that the RTEA module acts as an adaptive temporal focus lens, concentrating on motion features adjacent to the target time instant. In contrast, the model without the RTEA module tends to recover over-smoothed images with motion trails.

Effects of the TDW Module. Building upon Case 2, we implemented two variants: Case 3, which adds the EDW module that guides image feature warping using event features, and our proposed method (Ours), which employs timesurface features for warping guidance. The performance improvements of both schemes over Case 2 indicate that transforming average texture features to the spatial positions corresponding to the target time instant effectively enhances the quality of the reconstructed video sequence. Compared to Case 3, our method achieves an additional 0.49dB gain, verifying that the motion priors contained in timesurface provide more accurate guidance for image feature transformation and highlighting the importance of utilizing task-aligned characteristics within event data. In Figure 8, we visualize the warp field K^t generated by TDW module and compare the results between the variants with

Table 2. Comparison of long video sequence generation. $\times 9$ denotes decomposing one blurry frame into a 9-frame sequence.

Methods	HighREV$\times 9$		EBD$\times 9$	
	PSNR	SSIM	PSNR	SSIM
BiT	30.74	0.932	25.33	0.861
REFID	35.45	0.941	<u>27.57</u>	<u>0.872</u>
EvEnhancer	<u>35.59</u>	<u>0.943</u>	27.44	0.869
Ours	**36.81**	**0.973**	**28.99**	**0.914**

Table 3. Ablation study for each module on HighREV dataset. EDW means estimating a warp field from event-voxel features and using it to warp the image features.

Case	RTEA	Warp Guidance	Fusion Module	PSNR	SSIM	FLOPs (G)
1	-	-	EGGF	33.92	0.959	94.12
2	✓	-	EGGF	35.42	0.966	101.67
3	✓	EDW	EGGF	36.35	0.971	108.76
4	✓	TDW	Concat	36.31	0.970	115.96
5	✓	TDW	Add	36.33	0.970	106.31
6	✓	TDW	Cross attn.	36.79	0.973	108.45
Ours	✓	TDW	EGGF	**36.84**	**0.974**	107.91

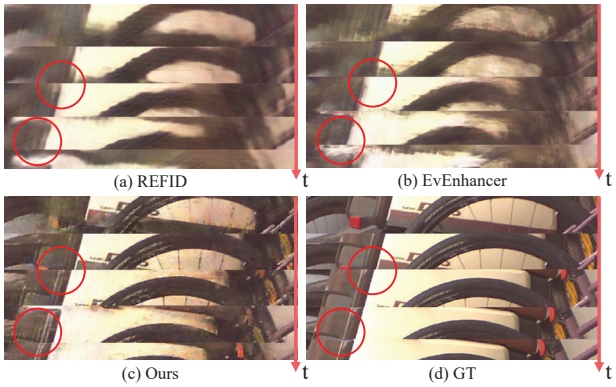

Figure 6. Comparison of the spatio–temporal slices on EBD dataset. We visualize a narrow band of pixels from the same spatial location across time. Our method produces smooth and coherent motion trajectories, while competing methods exhibit broken or jittery patterns.

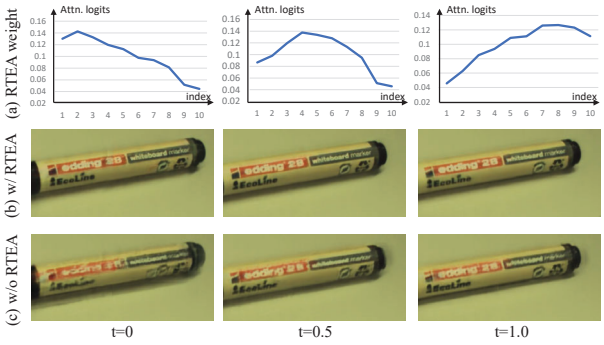

Figure 7. Visual comparison results of w/ and w/o RTEA. (a) shows the RTEA weights for different target times, exhibiting sharp, time-localized distributions. Without RTEA, the images exhibit motion artifacts, whereas RTEA yields sharper results.

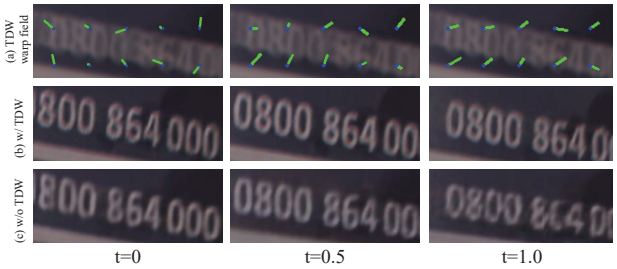

Figure 8. Visual comparison results of w/ and w/o TDW. (a) shows the warp field of TDW module. Without TDW, the images show oversmooth artifacts. Zoom in for better view.

and without TDW. As shown in Figure 8, the object moves from left to right as time t evolves from 0 to 1. The marker in subfigure (a) denotes that the pixel corresponding to the blue dot is warped along the direction of the green line. Accordingly, during the reconstruction process, the blurred features are aggregated toward the left (backward motion) at $t = 0$ and toward the right (forward motion) at $t = 1$. As the query time t traverses the entire exposure interval, the TDW module produces adaptive and time-varying warp fields, which progressively align the aggregated features to the desired spatial configuration at each timestamp. In this way, our TDW module guarantees consistent and physically plausible motion trajectories across the whole reconstructed high-frame-rate video.

Effects of the EGGF Module. As shown in Cases 4, 5, and 6 in Tab. 3, we replaced the EGGF module with concatenation, element-wise addition, and a basic cross-modal attention module, respectively. Benefiting from the time-specialized and aligned bimodal features, our EGGF module efficiently accomplishes cross-modal information fusion, achieving a performance improvement of up to 0.53dB. Meanwhile, the EGGF module reduces network FLOPs by 0.54G compared to the second-best cross-modal attention module.

5. Conclusion

In this paper, we propose TSANet, a novel approach that leverages the high temporal resolution of event cameras to enhance the performance and temporal consistency of motion decomposition. To fully exploit the motion information embedded in event data and mitigate motion ambiguity, TSANet comprises two key components. The Relative Time-Encoded Attention (RTEA) module incorporates relative time encoding biases into the attention mechanism, capturing specialized motion features around the target time instant. The Timesurface Dynamic Warping (TDW) module utilizes motion priors contained in temporal surfaces to guide the specialization of average image texture features toward the spatial positions corresponding to the target time instant. Finally, the Event Guide Gating Fusion (EGGF) module complementarily fuses the time-specialized bimodal information. Experimental results across multiple datasets demonstrate that TSANet achieves state-of-the-art performance.

Despite these promising results, TSANet still has limitations in real-world generalization. In particular, we observe that a model trained on synthetic blur performs suboptimally on real-world sequences because of the domain gap between synthetic and real data. In future work, we plan to alleviate this issue by collecting paired real-blur, sharp, and event data using a beam-splitter setup with an additional RGB camera, which may provide a more reliable foundation for bridging the synthetic-to-real gap.

Acknowledgements. This work was supported by the National Natural Science Foundation of China (NSFC) under Grants 62225207, 62436008, 62422609 and 62276243.

References

[1] Christian Brandli, Raphael Berner, Minhao Yang, Shih-Chii Liu, and Tobi Delbruck. A 240 × 180 130 db 3 μs latency global shutter spatiotemporal vision sensor. *IEEE Journal of Solid-State Circuits*, 49(10):2333–2341, 2014. 1

[2] Chengzhi Cao, Xueyang Fu, Hongjian Liu, Yukun Huang, Kunyu Wang, Jiebo Luo, and Zheng-Jun Zha. Event-guided person re-identification via sparse-dense complementary learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17990–17999, 2023. 2

[3] Zhiwen Chen, Zhiyu Zhu, Yifan Zhang, Junhui Hou, Guangming Shi, and Jinjian Wu. Segment any event streams via weighted adaptation of pivotal tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3890–3900, 2024. 2

[4] Sung-Jin Cho, Seo-Won Ji, Jun-Pyo Hong, Seung-Won Jung, and Sung-Jea Ko. Rethinking coarse-to-fine approach in single image deblurring. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4641–4650, 2021. 1

[5] Thomas Finateu, Atsumi Niwa, Daniel Matolin, Koya Tsuchimoto, Andrea Mascheroni, Etienne Reynaud, Pooria Mostafalu, Frederick Brady, Ludovic Chotard, Florian LeGoff, Hirotsugu Takahashi, Hayato Wakabayashi, Yusuke Oike, and Christoph Posch. 5.10 a 1280×720 back-illuminated stacked temporal contrast event-based vision sensor with 4.86μm pixels, 1.066geps readout, programmable event-rate controller and compressive data-formatting pipeline. In *IEEE International Solid-State Circuits Conference*, pages 112–114, 2020. 1

[6] Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J Davison, Jörg Conradt, Kostas Daniilidis, et al. Event-based vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(1):154–180, 2020. 1

[7] Chengjie Ge, Xueyang Fu, Peng He, Kunyu Wang, Chengzhi Cao, and Zheng-Jun Zha. Neuromorphic event signal-driven network for video de-raining. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1878–1886, 2024. 2

[8] Jin Han, Yixin Yang, Chu Zhou, Chao Xu, and Boxin Shi. Evintsr-net: Event guided multiple latent frames reconstruction and super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4882–4891, 2021. 2

[9] Yuhuang Hu, Shih-Chii Liu, and Tobi Delbruck. v2e: From video frames to realistic dvs events. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1312–1321, 2021. 5

[10] Xiang Ji, Haiyang Jiang, and Yinqiang Zheng. Motion blur decomposition with cross-shutter guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12534–12543, 2024. 1, 2

[11] Xingyu Jiang, Xiuhui Zhang, Ning Gao, and Yue Deng. When fast fourier transform meets transformer for image restoration. In *European Conference on Computer Vision*, pages 381–402. Springer, 2024. 3

[12] Meiguang Jin, Givi Meishvili, and Paolo Favaro. Learning to extract a video sequence from a single motion-blurred image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6334–6342, 2018. 1, 2, 5, 6

[13] Taewoo Kim and Kuk-Jin Yoon. Event-guided unified framework for low-light video enhancement, frame interpolation, and deblurring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8524–8534, 2025. 2

[14] Taewoo Kim, Hoonhee Cho, and Kuk-Jin Yoon. Frequency-aware event-based video deblurring for real-world motion blur. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24966–24976, 2024. 2

[15] Dilip Krishnan, Terence Tay, and Rob Fergus. Blind deconvolution using a normalized sparsity measure. In *CVPR 2011*, pages 233–240. IEEE, 2011. 1

[16] Xavier Lagorce, Garrick Orchard, Francesco Galluppi, Bertram E Shi, and Ryad B Benosman. Hots: a hierarchy of event-based time-surfaces for pattern recognition. *IEEE transactions on pattern analysis and machine intelligence*, 39(7):1346–1359, 2016. 3

[17] Siqi Li, Zhikuan Zhou, Zhou Xue, Yipeng Li, Shaoyi Du, and Yue Gao. 3d feature tracking via event camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18974–18983, 2024. 2

[18] Guoqiang Liang, Kanghao Chen, Hangyu Li, Yunfan Lu, and Lin Wang. Towards robust event-guided low-light image enhancement: A large-scale real-world event-image dataset and novel approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23–33, 2024. 2

[19] Jinxiu Liang, Yixin Yang, Boyu Li, Peiqi Duan, Yong Xu, and Boxin Shi. Coherent event guided low-light video enhancement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10615–10625, 2023. 2

[20] Patrick Lichtsteiner, Christoph Posch, and Tobi Delbruck. A 128 x 128 120db 30mw asynchronous vision sensor that responds to relative intensity change. In *2006 IEEE International Solid State Circuits Conference-Digest of Technical Papers*, pages 2060–2069. IEEE, 2006. 1

[21] Songnan Lin, Jiawei Zhang, Jinshan Pan, Zhe Jiang, Dongqing Zou, Yongtian Wang, Jing Chen, and Jimmy Ren. Learning event-driven video deblurring and interpolation. In *European Conference on Computer Vision*, pages 695–710. Springer, 2020. 5, 6

[22] Songnan Lin, Jiawei Zhang, Jinshan Pan, Zhe Jiang, Dongqing Zou, Yongtian Wang, Jing Chen, and Jimmy Ren. Learning event-driven video deblurring and interpolation. In *European Conference on Computer Vision*, pages 695–710. Springer, 2020. 2, 3

[23] Haoyue Liu, Shihan Peng, Lin Zhu, Yi Chang, Hanyu Zhou, and Luxin Yan. Seeing motion at nighttime with an event camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25648–25658, 2024. 2

[24] Nico Messikommer, Carter Fang, Mathias Gehrig, and Davide Scaramuzza. Data-driven feature tracking for event cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5642–5651, 2023. 2

[25] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *CVPR*, 2017. 1, 5

[26] Jihyong Oh and Munchurl Kim. Demfi: deep joint deblurring and multi-frame interpolation with flow-guided attentive correlation and recursive boosting. In *European Conference on Computer Vision*, pages 198–215. Springer, 2022. 1, 5, 6

[27] Liyuan Pan, Cedric Scheerlinck, Xin Yu, Richard Hartley, Miaomiao Liu, and Yuchao Dai. Bringing a blurry frame alive at high frame-rate with an event camera. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6820–6829, 2019. 2

[28] Liyuan Pan, Richard Hartley, Cedric Scheerlinck, Miaomiao Liu, Xin Yu, and Yuchao Dai. High frame rate video reconstruction based on an event camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2519–2533, 2020. 2

[29] Bang-Dang Pham, Phong Tran, Anh Tran, Cuong Pham, Rang Nguyen, and Minh Hoai. Hypercut: Video sequence from a single blurry image using unsupervised ordering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9843–9852, 2023. 1, 2

[30] Kuldeep Purohit, Anshul Shah, and AN Rajagopalan. Bringing alive blurred moments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6830–6839, 2019. 1, 2

[31] Wonyong Seo, Jihyong Oh, and Munchurl Kim. Bim-vfi: Bidirectional motion field-guided frame interpolation for video with non-uniform motions. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7244–7253, 2025. 5, 6

[32] Chen Song, Qixing Huang, and Chandrajit Bajaj. E-cir: Event-enhanced continuous intensity recovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7803–7812, 2022. 3, 5, 6

[33] Yunjae Suh, Seungnam Choi, Masamichi Ito, Jeongseok Kim, Youngho Lee, Jongseok Seo, Heejae Jung, Dong-Hee Yeo, Seol Namgung, Jongwoo Bong, Sehoon Yoo, Seung-Hun Shin, Doowon Kwon, Pilkyu Kang, Seokho Kim, Hoonjoo Na, Kihyun Hwang, Changwoo Shin, Jun-Seok Kim, Paul K. J. Park, Joonseok Kim, Hyunsurk Ryu, and Yongin Park. A 1280×960 dynamic vision sensor with a 4.95-μm pixel pitch and motion artifact minimization. In *IEEE International Symposium on Circuits and Systems*, pages 1–5, 2020. 1

[34] Lei Sun, Christos Sakaridis, Jingyun Liang, Peng Sun, Jiezhang Cao, Kai Zhang, Qi Jiang, Kaiwei Wang, and Luc Van Gool. Event-based frame interpolation with ad-hoc deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18043–18052, 2023. 5, 6, 7

[35] Zhijing Sun, Xueyang Fu, Longzhuo Huang, Aiping Liu, and Zheng-Jun Zha. Motion aware event representation-driven image deblurring. In *European Conference on Computer Vision*, pages 418–435. Springer, 2024. 3

[36] Xin Tao, Hongyun Gao, Xiaoyong Shen, Jue Wang, and Jiaya Jia. Scale-recurrent network for deep image deblurring. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8174–8182, 2018. 1

[37] Bishan Wang, Jingwei He, Lei Yu, Gui-Song Xia, and Wen Yang. Event enhanced high-quality image recovery. In *European Conference on Computer Vision*, pages 155–171. Springer, 2020. 2

[38] Shuoyan Wei, Feng Li, Shengeng Tang, Yao Zhao, and Huihui Bai. Evenhancer: Empowering effectiveness, efficiency and generalizability for continuous space-time video super-resolution with events. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 17755–17766, 2025. 5, 6

[39] Wenming Weng, Yueyi Zhang, and Zhiwei Xiong. Event-based blurry frame interpolation under blind exposure. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1588–1598, 2023. 5, 6

[40] Jie Xiao, Xueyang Fu, Yurui Zhu, and Zheng-Jun Zha. Bayesian window transformer for image restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(3):2431–2444, 2026. 1

[41] Shiao Wang Yuan Chen Zhe Wu Bo Jiang Yonghong Tian Jin Tang Xiao Wang, Yao Rong. Unleashing the power of cnn and transformer for balanced rgb-event video recognition. *Machine Intelligence Research*, 22:1031–1047, 2025. 2

[42] Fang Xu, Lei Yu, Bishan Wang, Wen Yang, Gui-Song Xia, Xu Jia, Zhendong Qiao, and Jianzhuang Liu. Motion deblurring with real events. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2583–2592, 2021. 2, 3, 5, 6

[43] Senyan Xu, Zhijing Sun, Mingchen Zhong, Chengzhi Cao, Yidi Liu, Xueyang Fu, and Yan Chen. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8942–8950, 2025. 2

[44] Yixin Yang, Jinxiu Liang, Bohan Yu, Yan Chen, Jimmy S Ren, and Boxin Shi. Latency correction for event-guided deblurring and frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24977–24986, 2024. 2

[45] Xiang Zhang and Lei Yu. Unifying motion deblurring and frame interpolation with events. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17765–17774, 2022. 2, 3, 5, 6

[46] Zhihang Zhong, Xiao Sun, Zhirong Wu, Yinqiang Zheng, Stephen Lin, and Imari Sato. Animation from blur: Multi-modal blur decomposition with motion guidance. In *Eu-*

ropean Conference on Computer Vision, pages 599–615. Springer, 2022. 1, 2

[47] Zhihang Zhong, Mingdeng Cao, Xiang Ji, Yinqiang Zheng, and Imari Sato. Blur interpolation transformer for real-world motion from blur. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5713–5723, 2023. 1, 2, 5, 6

[48] Nikola Zubic, Mathias Gehrig, and Davide Scaramuzza. State space models for event cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5819–5828, 2024. 2