

EventAid: Benchmarking Event-aided Image/Video Enhancement Algorithms with Real-captured Hybrid Dataset

Peiqi Duan[†], Boyu Li[†], Yixin Yang, Hanyue Lou, Minggui Teng, Xinyu Zhou, Yi Ma,
Boxin Shi[‡], Senior Member, IEEE

Abstract—Event cameras are emerging imaging technology that offer advantages over conventional frame-based imaging sensors in dynamic range and sensing speed. Complementing the rich texture and color perception of traditional image frames, the hybrid camera system of event and frame-based cameras enables high-performance imaging. With the assistance of event cameras, high-quality image/video enhancement methods make it possible to break the limits of traditional frame-based cameras, especially exposure time, resolution, dynamic range, and frame rate limits. This paper focuses on five event-aided image and video enhancement tasks (*i.e.*, event-based video reconstruction, event-aided high frame rate video reconstruction, image deblurring, image super-resolution, and high dynamic range image reconstruction), provides an analysis of the effects of different event properties, a real-captured and ground truth labeled benchmark dataset, a unified benchmarking of state-of-the-art methods, and an evaluation for two mainstream event simulators. In detail, this paper collects a real-captured evaluation dataset EVENTAID for five event-aided image/video enhancement tasks, by using “Event-RGB” multi-camera hybrid system, taking into account scene diversity and spatiotemporal synchronization. We further perform quantitative and visual comparisons for state-of-the-art algorithms, provide a controlled experiment to analyze the performance limit of event-aided image deblurring methods, and discuss open problems to inspire future research.

Index Terms—Event camera, image/video enhancement, benchmark dataset, simulated-to-real gap

1 INTRODUCTION

EVENT cameras, also known as Dynamic Vision Sensors (DVS) [1], [2], draw on the perception mechanism of the human retina [3] to sense brightness changes in the scene in the form of “event” signals [1], [2], [4]. Each pixel of the event camera compares the current and last light intensity state on a logarithmic scale and triggers a binary form event when the intensity variation exceeds the preset threshold [1], [4]. Such a trigger mechanism enables event cameras to high-speed ($\sim 10\mu s$) perceive dynamic visual scenarios with a high dynamic range (HDR) ($\sim 120dB$) while lacking the absolute radiance intensity recording and static sensing [1]. With the superior properties of HDR, low latency, and low redundancy [1], event cameras make it possible to break through the bottlenecks of computer vision and robotic technologies based on traditional frame-based cameras.

Thus far, event cameras have shown promising capability in solving classical as well as new computer vision and robotics tasks, including low-level tasks such as high frame-rate (HFR) video synthesis [5] and HDR image reconstruction [6], middle-level tasks such as optical flow [7] and scene depth estimation [8], and high-level tasks such as 3D scene

reconstruction [9], object tracking [10], object detection [11], SLAM [12], and autonomous wheel steering [13] tasks.

Due to the special triggering mechanism and much shorter research times compared to frame-based cameras, the signal quality of event cameras degrades when the scenes are relatively static, and suffers from severe noises and poor color perception [14]. By contrast, traditional frame-based RGB cameras are the mainstream sensors of computer vision and robotic technologies that feature rich color, texture, and semantic information as well as lower noise. Witnessing and experiencing the success of frame-based RGB sensors and corresponding algorithms over the past decades, researchers have built large-scale datasets [15], various well-designed network architectures [16], and even foundation models [17] for frame-based vision. Such sensory motivates researchers to leverage the complementary advantages of both ends through an “Event-RGB” hybrid multi-camera fusion [18], [19] and to use the existing achievements of frame-based cameras for accelerating the research of event cameras. This fusion has been extensively explored in the field of high-quality imaging. To take advantage of the high speed and HDR features of event cameras, break through the traditional imaging bottlenecks, and meet the image quality requirements of human and machine vision, researchers have bridged the event and image modality in recent years [18]–[25]. We classify such tasks as **event-aided** (also known as event-guided [26]–[28]) **image/video enhancement** methods, and we focus on five event-aided tasks in this paper (as shown in Fig. 1).

To benchmark event-aided image/video enhancement

• [†] Contributed equally to this work as first authors

• [‡] Corresponding author: shiboxin@pku.edu.cn

• P. Duan, B. Li, Y. Yang, H. Lou, M. Teng, Y. Ma, and B. Shi are with National Key Laboratory for Multimedia Information Processing and National Engineering Research Center of Visual Technology, School of Computer Science, Peking University.

• X. Zhou is with the National Key Lab of General Artificial Intelligence, School of Intelligence Science and Technology, Peking University.

• Project page: <https://sites.google.com/view/eventaid-benchmark>

Tasks	Event-based video reconstruction	Event-aided high frame rate video reconstruction	Event-aided image deblurring	Event-aided image super resolution	Event-aided high dynamic range image reconstruction
Illustrations/ Inputs					
Event characteristics	High temporal resolution, Low latency			High dynamic range	
Outputs					
Formulations	$I_{t_1} = \mathcal{F}_{\text{passion}}(\mathcal{F}_{\text{warp}}(E_{t_0:t_1}))$ (1)	$L_{t_1} = L_{t_0} + c \cdot \int E_{t_0:t_1}$ (3)	$L^{\text{clear}} = L^{\text{blur}} - c \cdot \iint E_{t_0:t_1}$ (5)	$I_{t_0}^{\text{SR}} = \mathcal{F}_{\text{mix}}(I_{t_0}^{\text{LR}}, \mathcal{F}_{\text{rec}}(E_{t_0-w:t_0-w+\epsilon}), \dots, \mathcal{F}_{\text{rec}}(E_{t_0+w-c:t_0+w}))$ (7)	$I_{t_1}^{\text{HDR}} = \mathcal{F}_{\text{fusion}}(I_{t_1}^{\text{LDR}}, \mathcal{F}_{\text{rec}}(E_{t_0:t_1}))$ (8)
	$I_{t_1} = \mathcal{F}_{\text{rec}}(E_{t_0:t_1})$ (2)	$I_{t_1} = \mathcal{F}_{\text{fusion}}(\mathcal{F}_{\text{syn}}(I_{t_0}, E_{t_0:t_1}), \mathcal{F}_{\text{warp}}(I_{t_0}, E_{t_0:t_1}))$ (4)	$I^{\text{clear}} = \mathcal{F}_{\text{fusion}}(\mathcal{F}_{\text{syn}}(I^{\text{blur}}, E_{t_0:t_1}), \mathcal{F}_{\text{warp}}(I^{\text{blur}}, E_{t_0:t_1}))$ (6)	$I_{t_1}^{\text{HDR}} = \mathcal{F}_{\text{fusion}}(I_{t_1}^{\text{LDR}}, E_{t_0:t_1})$ (9)	

Figure 1: 1st row: Five event-aided image/video enhancement tasks. 2nd row: The illustrations of different tasks, when the event camera and the RGB camera shoot rotating fan blades at the same time and in the same field of view, the event camera asynchronously triggers positive (blue dot) and negative (red dot) events with high temporal resolution, while the RGB camera outputs images frame by frame. 3rd row: The Event characteristics that each task used to break through the performance bottlenecks limited by frames. 4th row: Outputs of different tasks. 5th row: Formulations of each task.

methods with real data, a real “Event-RGB” hybrid multi-camera system with high-precision spatiotemporal synchronization is necessary. Currently, mainly three types of “Event-RGB” hybrid multi-camera systems are in use: (1) As early attempts, ATIS [29] and DAVIS [2], [30] cameras embed intensity-recording subpixels or conventional Active Pixel Sensors (APS) with event sensors to “Event-RGB” simultaneous imaging. This kind of design is the optimal way to achieve multi-camera spatiotemporal synchronization, while there are bottlenecks that lead to severe noise and low resolution [14] (e.g., the resolution of the widely-used DAVIS346 [2] is only 346×260). (2) To match high-quality frame cameras, building an event camera and a frame camera into a dual-camera system becomes another option [19]. Nevertheless, the different homography matrices of different scene depths make it difficult for this dual-camera system to avoid spatial matching errors. (3) Wang *et al.* [26] and Han *et al.* [24] first propose a hybrid camera system that physically co-located an event camera and a frame camera via a beam splitter, with two cameras sharing a common field of view. Although there are still problems of non-portability and unstable matching accuracy, such a system can easily replace cameras while avoiding field-of-view misalignment caused by binocular disparity. Based on the above three hybrid systems, many evaluation datasets for event-aided image/video enhancement methods have been proposed (as listed in Table 1). To avoid the imperfect conditions of the above real systems, using event simulators to generate events and prepare datasets is also a popular choice [19], [31], while the unavoidable gap between real-captured and simulated data [14] (real-sim gap) makes the benchmark results less convincing to reveal the real performance of algorithms. As event-aided image/video enhancement methods are continuously springing up, a high-

quality real-captured dataset for comprehensive benchmarking them on a unified scale is urgently needed.

In this paper, we propose a real-world and comprehensive evaluation dataset, named EVENTAID, for evaluating five mainstream event-aided image/video enhancement tasks, taking into account geometric and photometric alignment, temporal synchronization of sensors, and scene diversity. All data, including input events and frames and the ground truth, are real-captured by beam-splitter-mounted hybrid camera systems. The sub-datasets corresponding to each task are: EVENTAID-R for event-based video reconstruction, EVENTAID-F for HFR video reconstruction, EVENTAID-B for image deblurring, EVENTAID-S for image super-resolution (SR), and EVENTAID-D for HDR image reconstruction. We benchmark 20 state-of-the-art event-based algorithms and 12 state-of-the-art single-image-based methods. To evaluate the real-sim gap of event simulators, we generate events with different simulators and execute the benchmarking again. To our best knowledge, this is the first high-quality benchmark dataset for event-aided image/video enhancement tasks with real-captured data, and the first comprehensive benchmarking of existing methods in diverse scenes with unified evaluation protocols.

This paper makes the following contributions:

- We categorize five mainstream event-aided image/video enhancement tasks using unified formulations, as well as summarize and compare the existing evaluation datasets.
- We collect the first real-captured dataset EVENTAID to evaluate the performance of existing methods of five tasks, with high accuracy of spatiotemporal synchronization between two sensors as well as great scene diversity (61 scenes in total).
- We benchmark a total of 32 state-of-the-art methods

and statistically analyze their performance. We further compare and evaluate the real-sim gap of two widely used event simulators by referring to the real data benchmark results.

- Based on our benchmark evaluation, open problems from different aspects of these five tasks, such as evaluation metric, feature extraction, artifacts suppression, color restoration, and so on, are discussed to inspire future research.

2 EVENT-AIDED IMAGING CATEGORIZATION

This section first presents the mathematical form of the event trigger model and categorizes and formulates existing event-aided image/video enhancement methods into five tasks with unified notations. Then, we list, organize, and summarize existing evaluation datasets for these five tasks.

2.1 Event-aided imaging model formulation

We first formulate the event trigger model and build the relationship to the image-based counterpart. Consider a latent spatiotemporal volume in which an intensity field is sampled by an ideal frame-based camera that can output blur-free, high-resolution, and HDR intensity images I_t at any moment. The event output at t_0 can be described as:

$$E_{t_0} = \Gamma\left\{ \log\left(\frac{I_{t_0} + b}{I_{t_0-1} + b}\right), c + n_{\text{event}} \right\}, \quad (10)$$

where $\Gamma\{\theta, c\}$ represents the conversion function from log-intensity to events, and b is an offset value to prevent $\log(0)$. $\Gamma\{\theta, c\} = 1$ when $\theta \geq c$, indicating a positive event; $\Gamma\{\theta, c\} = -1$ when $\theta \leq -c$, indicating a negative event; and $\Gamma\{\theta, c\} = 0$ when $|\theta| < c$, indicating that no event has been fired. n_{event} represents the perturbation noise pivoted at the firing threshold c . The duration from $t_0 - 1$ to t_0 corresponds to the minimum response delay of the event camera. The dead pixels can be interpreted as c being significantly low or high. In event-aided image/video enhancement tasks, I_t is taken as the ground truth.

Here we model five tasks and compare their relationships and differences. Figure 1 shows the illustrations of input and process for each task as well as the equations to be solved for each task.

Event-based video reconstruction. The basic task of bridging events and images that directly reconstructs images from pure event signals, which can be formulated as a process of $E_{t_0:t_1} \rightarrow I_{t_1}$, where the $E_{t_0:t_1}$ denotes the event stream triggered between t_0 and t_1 . This is an ill-posed problem because the event signals only record the intensity change but not the absolute intensity in the scene, so it is difficult to accurately measure the light intensity via events. To solve this problem, Barua *et al.* [49] propose to use the optical flow consistency hypothesis and motion compensation to obtain the gradient of images from events, and then employ the Poisson reconstruction method to restore the image, *i.e.*, $I_{t_1} = \mathcal{F}_{\text{passion}}(\mathcal{F}_{\text{warp}}(E_{t_0:t_1}))$ (*c.f.*, Eq. 1 in Fig. 1). Deep learning-based methods [5], [20], [50], [51] directly learn the mapping model from events to images by $I_{t_1} = \mathcal{F}_{\text{rec}}(E_{t_0:t_1})$ (*c.f.*, Eq. 2). It is worth noting that this task cannot reconstruct textures in the static scene.

Event-aided HFR video reconstruction. This task aims to interpolate new frames, *i.e.*, reconstruct latent frames, between two adjacent frames with the assistance of events, which is formulated as $I_{t_0} \& E_{t_0:t_1} \rightarrow I_{t_1}$. Since events record the logarithmic changes of I_{t_0} over $t_0 : t_1$ with high time accuracy, the I_{t_1} in the logarithmic domain (*i.e.*, L_{t_1}) can be easily obtained by the events synthesis model $L_{t_1} = L_{t_0} + c \cdot \int E_{t_0:t_1}$ (*c.f.*, Eq. 3), despite the interference of event noise [52], [53]. In order to improve performance, deep learning-based methods generally employ the events synthesis model to constraint intensity values and event-based optical flow estimation to constraint motion trajectories in reconstructed videos, then use a fusion model to fuse the two branches and output the final result [19], [23], [54], formulated as $I_{t_1} = \mathcal{F}_{\text{fusion}}(\mathcal{F}_{\text{syn}}(I_{t_0}, E_{t_0:t_1}), \mathcal{F}_{\text{warp}}(I_{t_0}, E_{t_0:t_1}))$ (*c.f.*, Eq. 4).

Event-aided image deblurring. This task aims to restore a clear image from the long-exposure image suffering from motion blur, formulated as $I^{\text{blur}} \& E_{t_0:t_1} \rightarrow I^{\text{clear}}$, where $t_0 : t_1$ corresponds to the exposure time period. Pan *et al.* [22] find the event-based double integral model to bridge I^{blur} and I^{clear} via events and reconstruct the clear image via $L^{\text{clear}} = L^{\text{blur}} - c \cdot \iint E_{t_0:t_1}$ (*c.f.*, Eq. 5). Learning-based methods [43], [55]–[57] continuously improve the deblurring performance by upgrading the network model. Optical estimation is also introduced to improve performance [42], *i.e.*, $I^{\text{clear}} = \mathcal{F}_{\text{fusion}}(\mathcal{F}_{\text{syn}}(I^{\text{blur}}, E_{t_0:t_1}), \mathcal{F}_{\text{warp}}(I^{\text{blur}}, E_{t_0:t_1}))$ (*c.f.*, Eq. 6). Due to the high temporal resolution of events, most methods achieve intra-frame interpolation as well.

Event-aided image super-resolution. This task aims to reconstruct a high-resolution image from a low-resolution image by converting event-recorded motion information into sub-pixel shifts, *i.e.*, $I_{t_0}^{\text{LR}} \& E_{t_0-w:t_0+w} \rightarrow I_{t_0}^{\text{SR}}$, where w adjust time window length. EvIntSR [47] and E2SRI [21], [46] generally convert event data to multiple latent intensity frames and learn to mix the frame sequence to achieve super-resolution, expressed as $I_{t_0}^{\text{SR}} = \mathcal{F}_{\text{mix}}(I_{t_0}^{\text{LR}}, \mathcal{F}_{\text{rec}}(E_{t_0-w:t_0-w+\epsilon}), \dots, \mathcal{F}_{\text{rec}}(E_{t_0+w-\epsilon:t_0+w}))$ (*c.f.*, Eq. 7), ϵ is the time length of events to convert each latent frames. EventSR [58] can also achieve image SR, while it mainly learns the mapping from LR images generated by events to HR images through GAN-based methods [59].

Event-aided HDR image reconstruction. This task aims to recover an HDR image from a low dynamic range (LDR) image by extracting texture features of over-/under-exposed areas from events in dynamic scenes, *i.e.*, $I_{t_1}^{\text{LDR}} \& E_{t_0:t_1} \rightarrow I_{t_1}^{\text{HDR}}$. Han *et al.* [24], [48] first explore this task and propose to reconstruct an intensity frame from events before fusing it with input LDR images via a refinement network module, expressed as $I_{t_1}^{\text{HDR}} = \mathcal{F}_{\text{fusion}}(I_{t_1}^{\text{LDR}}, \mathcal{F}_{\text{rec}}(E_{t_0:t_1}))$ (*c.f.*, Eq. 8). Yang *et al.* [28] eliminate the step of reconstructing the image from events and recover an HDR image by $I_{t_1}^{\text{HDR}} = \mathcal{F}_{\text{fusion}}(I_{t_1}^{\text{LDR}}, E_{t_0:t_1})$ (*c.f.*, Eq. 9).

2.2 Evaluation datasets for event-aided imaging

The field of conventional image/video enhancement already has a large amount of research on benchmark evaluation. For example, Köhler *et al.* [60] propose a real-captured dataset to benchmark image SR task and Rim *et al.* [61]

Table 1: The summary of existing evaluation datasets of five event-aided image/video enhancement tasks. The following four characteristics as we marked in the 1st row are compared: 1. Real-captured data: the input images and events, and ground truth are real-captured or simulated. 2. Spatiotemporal synchronization of two sensors. 3. Event/Frame-based sensor: the spatial resolution, color imaging type, and frame rate parameters of cameras. 4. Scene diversity. (“-” represents a “not applied” attribute. The resolution and frame rate of some datasets are not completely consistent, we use “~” and “<” to represent the approximate value distribution. “*” indicates indirect estimation due to the dataset are not public and whose frame rates are unspecified in the paper.

	Simulated or real dataset	Dataset name	Real-captured data			Spatiotemporal synchronization of two sensors		Event sensor		Frame-based sensor			Scene diversity				
			Input image	Input events	Ground truth	Spatial matching	Temporal synchronization	Camera model	Spatial resolution	Spatial resolution	Color / gray	Frame rate	Indoor+ Outdoor	Ego+ Local motion	Slow+ Fast motion	High texture	
Event-based video reconstruction	Real	Simulation	EventNFS [32]	-	✓	✗	Display+camera calibration	Mark points matching	DAVIS346 mono	222×124	222×124	color	-	✓	✓	✓	✓
		IJRR [33]	-	✓	✓	✓	Frame+event sensor	Chip synchronization	DAVIS240	240×180	240×180	gray	<24 FPS	✓	✓	✓	✓
		HQF [34]	-	✓	✓	✓	Frame+event sensor	Chip synchronization	DAVIS240	240×180	240×180	gray	17-25 FPS	✓	✓	✓	✓
		DVS-Dark [35]	-	✓	✓	✓	Frame+event sensor	Chip synchronization	DAVIS240	240×180	240×180	gray	~10 FPS*	✓	✗	✗	✗
		MVSEC [36]	-	✓	✓	✓	Frame+event sensor	Chip synchronization	DAVIS346 mono	346×260	346×260	gray	50 FPS	✓	✓	✓	✗
		CED [37]	-	✓	✓	✓	Frame+event sensor	Chip synchronization	DAVIS346 color	346×260	346×260	color	~33 FPS	✓	✓	✗	✗
		EVENTAID-R	-	✓	✓	✓	Beam splitter	External clock triggering	Prophesee	~ 954 × 636	~ 954 × 636	color	20-150 FPS	✓	✓	✓	✓
Event-aided high frame rate video reconstruction	Real	Simulation	Tulyakov <i>et al.</i> [19]	✓	✗	✓	-	-	Simulation	1280×720	1280×720	color	-	✓	✓	✓	✓
		GoPro+ESIM [38]	✓	✗	✓	✓	-	-	Simulation	1280×720	1280×720	color	240 FPS	✓	✓	✓	✓
		SloMo-DVS [39]	✓	✓	✓	✓	Frame+event sensor	Chip synchronization	DAVIS240	240×180	240×180	gray	<30 FPS	✓	✓	✓	✓
		GEP [18]	✓	✓	✓	✓	Beam splitter	Mark points matching	DAVIS240	190×180	1520×1440	color	20 FPS	✓	✓	✗	✓
		HS-ERGB [19]	✓	✓	✓	✓	Dual camera setup	External clock triggering	Prophesee	~ 900 × 800	~ 900 × 800	color	150-163 FPS	✓	✓	✓	✓
		BS-ERGB [23]	✓	✓	✓	✓	Beam splitter	External clock triggering	Prophesee	970×625	970×625	color	28 FPS	✓	✓	✓	✓
		ERF-X170FPS [40]	✓	✓	✓	✓	Beam splitter	External clock triggering	Prophesee	1440×975	1440×975	color	170 FPS	✗	✓	✓	✓
Event-aided image deblurring	Real	Simulation	RBE [31]	✓	✓	✓	Frame+event sensor	Chip synchronization	DAVIS240	240×180	240×180	gray	<9 FPS*	✗	✓	✗	✓
		Kim <i>et al.</i> [27]	✗	✗	✓	✓	Frame+event sensor	Chip synchronization	DAVIS346 color	346×260	346×260	color	~17 FPS	✓	✗	✗	✗
		REBlur [43]	✓	✓	✓	✓	Repetitive motion scenes	Mark points matching	DAVIS346 mono	320×260	320×260	gray	-	✗	✗	✓	✗
		REVD [44]	✓	✓	✓	✓	Beam splitter	External clock triggering	Prophesee	1024×768	1024×768	color	-	✗	✓	✓	✓
		EVRB [45]	✓	✓	✓	✓	Beam splitter	External clock triggering	Prophesee	960×640	960×640	color	-	✗	✓	✓	✓
		EVENTAID-B	✓	✓	✓	✓	Beam splitter	External clock triggering	Prophesee	~ 835 × 620	~ 835 × 620	color	30 FPS	✓	✓	✓	✓
		ESC [46]	✗	✗	✓	✓	-	-	Simulation	128 × 128	512 × 512	gray	-	✓	✓	✓	✓
Event-aided image super resolution	Simulation	GoPro+V2E [47]	✗	✗	✓	✓	-	-	Simulation	320 × 180	1280 × 720	color	-	✓	✓	✓	✓
		GoPro+ESIM [38]	✗	✗	✓	✓	-	-	Simulation	320 × 180	1280 × 720	gray	-	✓	✓	✓	✓
		EVENTAID-S	✗	✓	✓	✓	Beam splitter	External clock triggering	Prophesee	1270 × 710	2540 × 1420	color	30 FPS	✓	✓	✓	✓
Event-aided high dynamic range image reconstruction	Real	Simulation	Yang <i>et al.</i> [28]	✗	✗	✓	-	-	Simulation	256×256	256×256	color	-	✓	✗	✗	✓
		Han <i>et al.</i> [24]	✓	✓	✓	✗	Beam splitter	Mark points matching	DAVIS240	240×180	1520×1440	color	20 FPS	✓	✓	✗	✗
		HES-HDR [48]	✓	✓	✓	✗	Beam splitter	Mark points matching	DAVIS346 mono	329×237	2032×1446	color	20 FPS	✓	✓	✗	✓
		Yang <i>et al.</i> [28]	✓	✓	✓	✗	Beam splitter	Mark points matching	DAVIS346 color	346×260	346×260	color	~10 FPS	✗	✗	✗	✓
		EVENTAID-D	✓	✓	✓	Reference	Beam splitter	External clock triggering	Prophesee	~ 800 × 500	~ 800 × 500	color	20-30 FPS	✓	✓	✓	✓

propose a real-captured dataset to benchmark image/video deblurring task. In contrast, due to the lack of comprehensive real-captured datasets and quantitative benchmarks, the performance of event-aided image/video enhancement methods on real-captured data is still largely unexplored. In Table 1, we summarize the widely used evaluation datasets for five event-aided image/video enhancement tasks and compare their characteristics¹. EVENTAID is also added to this comparison to highlight its advantage.

We focus on the following properties to evaluate the characteristics of these datasets: (1) Whether the triplet (input frames, input events, and ground truth) are real-captured data: Compared to using simulated data as evaluation datasets, real-captured data enables benchmarking the enhancement performance of the algorithm in real-world scenarios. However, since it is extremely challenging to collect real-captured data simultaneously, existing datasets often complete the triplet data by simulating events or synthesizing blur images, LDR images, *etc.*, which will introduce a real-sim gap and make the benchmark results

less convincing. (2) Spatiotemporal synchronization manner of event sensor and frame-based sensor²: To capture real data, there are four spatial matching patterns, where “frame+event sensor” is the best solution for synchronization but lacks the flexibility of switching cameras, “dual camera setup” can easily replace different frame cameras but disparity prevents pixels from being accurately aligned, “repetitive motion scenes” is difficult to collect diverse scenes. In contrast, “beam splitter” can effectively avoid the above problems. There are three temporal synchronization patterns, where “chip synchronization” is the best choice, and “external clock triggering” can also achieve microsecond level synchronization error. (3) Performance parameters of event sensor and frame-based sensor: Higher resolution and frame rates help to collect high-quality data. (4) The diversity of the captured scenes: Datasets covering diverse scenarios help evaluate the robustness of algorithms.

Event-based video reconstruction. For this task, evaluation data should contain input events and ground truth frames. EventNFS [32] develops a display-camera system to observe

1. Tables on page 3 of the supplementary material analyzes EVENTAID-F/-B/-D in terms of data diversity compared to existing datasets.

2. The illustrations of different spatiotemporal synchronization ways are shown in Sec. 0 of the supplementary material.

real-scenario event data via playback of $240FPS$ 720p videos on the display, while the real-sim gap of events still exists due to the relatively low refresh rate and dynamic range display. IJRR [33], HQF [34], DVS-Dark [35], MVSEC [36] and CED [37] all capture event data with the DAVIS series cameras [2], [30] and use APS as the ground truth frames. However, these cameras have low resolution and frame rate, and APS suffers from severe noise. In contrast, EVENTAID-R collects real-captured data at an average 954×636 resolution and $150FPS$, enabling the benchmark results to meet the requirements of real-world application scenarios.

Event-aided HFR video reconstruction. For this task, evaluation data should contain the triplet: input events, input LFR frame sequence, and ground truth HFR frames. The simulated datasets [38], [62] take in HFR video datasets as ground truth and pass them into simulators to generate event signals. SloMo-DVS [39] and GEF [18], [26], while real-captured, also suffer from low resolution and low frame rates. HS-ERGB [19] first collects HFR and HR videos as the ground truth, while the baseline existing in the dual camera system introduces inevitable errors. BS-ERGB [23] selects an LFR frame-based camera and the two cameras mount different lenses. ERF-X170FPS [40] and ERDS [41] lack the data on indoor scenes. The proposed EVENTAID-F avoids the above shortcomings and provides ground truth videos of $150FPS$ for the algorithms to be evaluated.

Event-aided image deblurring. This task requires evaluation data containing the triplet: input events, input blur images, and ground truth blur-free images. Different from the above two tasks, simultaneous capturing of blur and blur-free images greatly increases the difficulty of data collection. Therefore, most datasets simulate blurry images by averaging multi-frames [27], [31], [38], [42]. To achieve real-captured data collection, REBlur [43] performs controlled experiments indoors to collect the triplet data by repeating the same motion scenario multiple times, which can only capture indoor scenes with nondiversity. REVd [44] lacks the data on indoor scenes. To collect diversity and real-world datasets, the proposed EVENTAID-B first captures all real-captured triplet data by synchronizing two frame cameras and an event camera via beam splitters.

Event-aided image super-resolution. This task requires evaluation data containing input events, input LR images, and ground truth HR images. Since this task is still in its initial exploration stage, the existing evaluation datasets are all simulated. The proposed EVENTAID-S dataset is the first real-captured evaluation dataset, which captures $2 \times$ HR frames as the ground truth and $1 \times$ events as the LR inputs. The input $1 \times$ LR frames are downsampled from HR ones following the process in single image SR tasks.

Event-aided HDR image reconstruction. For this task, evaluation data contains input events, input LDR images, and ground truth HDR images. Existing datasets, *i.e.*, Han *et al.* [24], HES-HDR [48], and Yang *et al.* [28] only contain input events and LDR images. They mainly evaluate the quality of the reconstructed HDR images through no-reference quality assessment. EVENTAID-D uses an alternating-exposure camera to cyclically get short-/middle-/long-exposure LDR

images as the diverse input data, and mix multi-exposure images to restore HDR images as the reference.

3 EVENTAID DATASET COLLECTING

This section introduces the collection process of EVENTAID. Figure 2 shows the equipment setup we used.

3.1 Sensor and optics configuration

To collect datasets with high imaging quality, spatiotemporal synchronization, and unified scale, we use one Prophesee EVK4 HD (1280×720) event camera to capture event signals, two Hikvision MV-CA050-12UC RGB cameras (2448×2048 , $60FPS$) to simultaneously capture short-/low-exposure and long-/high-exposure frames for deblur/HDR task, one Hikvision MV-CA050-12UC RGB cameras (2448×2048 , $60FPS$) to capture HR images, one Hikvision MV-CA016-10UC RGB camera (1440×1080 , $165FPS$) to capture HFR frames, and one Basler acA800-510uc RGB camera (800×600 , $510FPS$) to capture alternating-exposure frame for HDR reconstruction task. We mount the same lenses ($16mm$ or $50mm$, $F = 1 : 2.8$, C-mount, fixed focus) for each task to avoid the influence of focal length and distortion differences. During the capturing process, we balance image quality and depth of field to determine aperture parameters and keep them consistent across all lenses. For scenes with multiple objects at different depths, we adjust focus rings to ensure the objects at the image center are in focus. We use Thorlabs CCM1-BS013 beam splitters ($50 : 50$ Split Ratio) to share light input for multiple cameras. In addition, when collecting EVENTAID-B, we set 25% , 10% , or 2% transmission ND filters to ensure luminosity consistency for short-exposure and long-exposure cameras.

Since event-aided image deblurring has been widely studied, and the algorithm performance is related to the degree of blur caused by motion, we execute a controllable experiment to evaluate the limits of blur levels that existing methods can withstand. Figure 3 shows the equipment setup and the experiment site layout of the controlled experiment. We use a servo steering gear to control a rigid rod to swing periodically at an opening angle of 120° , and fix a flat plate with a high-definition photo $1m$ away from the center of rotation as the main shooting target. The equipment setup is placed about $1m$ in front of the photo. Thus, when the exposure time of the cameras is fixed, we can adjust the motion speed of the photo by adjusting the swing period of the steering gear to obtain frames with different blur degrees. We set the exposure time of the short exposure camera to $2ms$ and the long exposure camera to $8ms$. The swing period is sequentially sampled at intervals of $0.25s$ from $1.5s$ to $4s$. We set up a DC fill light behind the scenes to ensure clear frames are less affected by noise imaging.

3.2 Geometric and photometric alignment

We use two types of camera hybrid system setups to collect data, as shown in Fig. 2 (a) and (b). The first setup used to collect the EVENTAID-B and EVENTAID-D datasets contains three cameras and we use three $50 : 50$ split ratio beam splitters docked and mounted in front of the lenses to ensure the input light is evenly split across all three

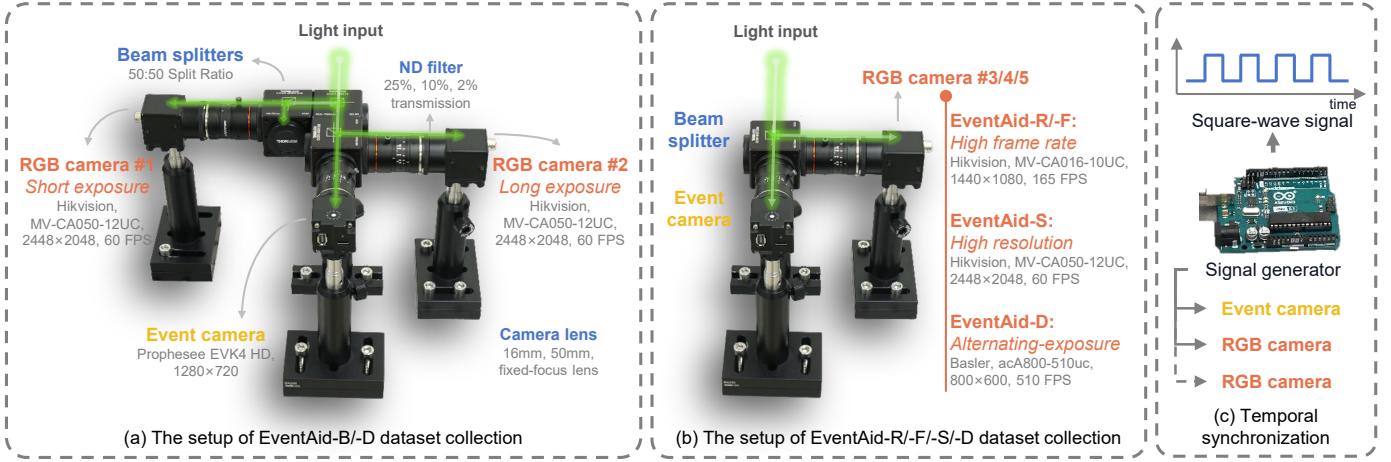


Figure 2: The equipment setup we used to collect the proposed dataset. (a) For the EVENTAID-B, we use one RGB camera to capture long-exposure blur images as the input, and another RGB camera to capture short-exposure clear images as the ground truth, the corresponding events are captured by one event camera. This setup can also collect EVENTAID-D where two different exposure images can be merged into one HDR image. (b) For the EVENTAID-R/-F/-S/-D dataset, we collocate an event camera and an RGB camera by mounting a 50 : 50 split ratio beam splitter in front of them. For each task, an RGB camera with corresponding attributes is selected to ensure that effective ground truths are captured. (c) We use a signal generator to simultaneously send square-wave signals to all cameras to achieve synchronized shooting.

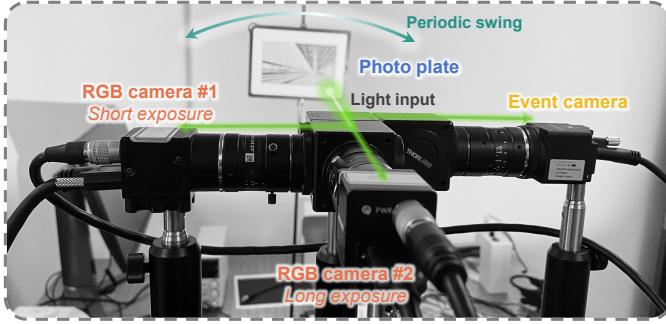


Figure 3: The equipment setup we used for the controlled experiment of event-aided image deblurring task.

cameras, *i.e.*, each camera receives an equal 25% of the input light. The three beam splitters are tightly connected through two SM1 external thread couplers, each lens is tightly connected with a beam splitter through a M27-to-SM1 thread adapter. The second setup used to collect the EVENTAID-R/-F/-S/-D contains two cameras and we use a 50 : 50 split ratio beam splitter to connect them. All cameras are fixed to a breadboard via poles to ensure they remain stable during severe shaking. Although the above-mentioned tight connection can make the fields of view of the three cameras well overlapped, it is still difficult to avoid pixel-level misalignment. Similar to Wang *et al.* [26], we use a 13.9" monitor for an offline geometric register for three cameras. We calculate the homography among three views. Then we use the homography matrix to transfer the views of two RGB cameras to the view of the event camera to ensure that the three views are aligned. Before dataset collection, we synchronize RGB cameras' exposure and capture video and event data simultaneously in optimal lighting. We iteratively refine the camera setup to minimize misalignment. For RGB-event matching, we compare residuals from adjacent RGB frames with corresponding event frames and minimize their misalignment.

The first setup should ensure the photometric alignment for two RGB cameras. For color consistency, we turn off the auto white balance and calibrate two cameras with a ColorChecker, adjust the same parameters of R/G/B channels in the camera SDK, and also use image editing software to fine-tune and ensure color consistency. For brightness consistency, exposure time ratios are set to $100 : N$ between the N%-transmission-ND-filtered and unfiltered cameras to equalize light intake. Before data capture, we use a grayscale board to calibrate the brightness of the two cameras to make them consistent.

In contrast to the relay-lens-based geometric alignment strategy [83], our setup achieves a similar FOV while avoiding the significant aberrations introduced by relay lenses. The DSLR-lens-based strategy [84] positions a DSLR lens in front of beam splitters to enable multiple cameras to share one lens. In comparison, our setup makes it easy to switch cameras or filters. We use a DSLR-lens-based setup to capture 4 groups in EVENTAID-R. The results show similar geometric alignment accuracy between the two strategies.

When collecting EVENTAID-D, we adopt the Zou *et al.* [85] method and use the first setup to capture EVENTAID-D-Dynamic sub-dataset with HDR video reference. Where two RGB cameras take overexposed and underexposed LDR images to merge HDR reference. We also use Han *et al.* [48] method to collect EVENTAID-D-Static sub-dataset with high-quality static reference with an alternate exposure camera.

3.3 Temporal synchronization

We use an Arduino Uno Rev3 microcontroller board as the signal generator to simultaneously send 5-volt square-wave signals to all cameras for achieving synchronized capturing. The corresponding interface of the signal generator and the GPIO ports of the cameras are connected through cables. The RGB camera takes a frame on each rising edge of the square-wave signal. The Prophesee event camera starts to trigger events when it receives the first rising edge

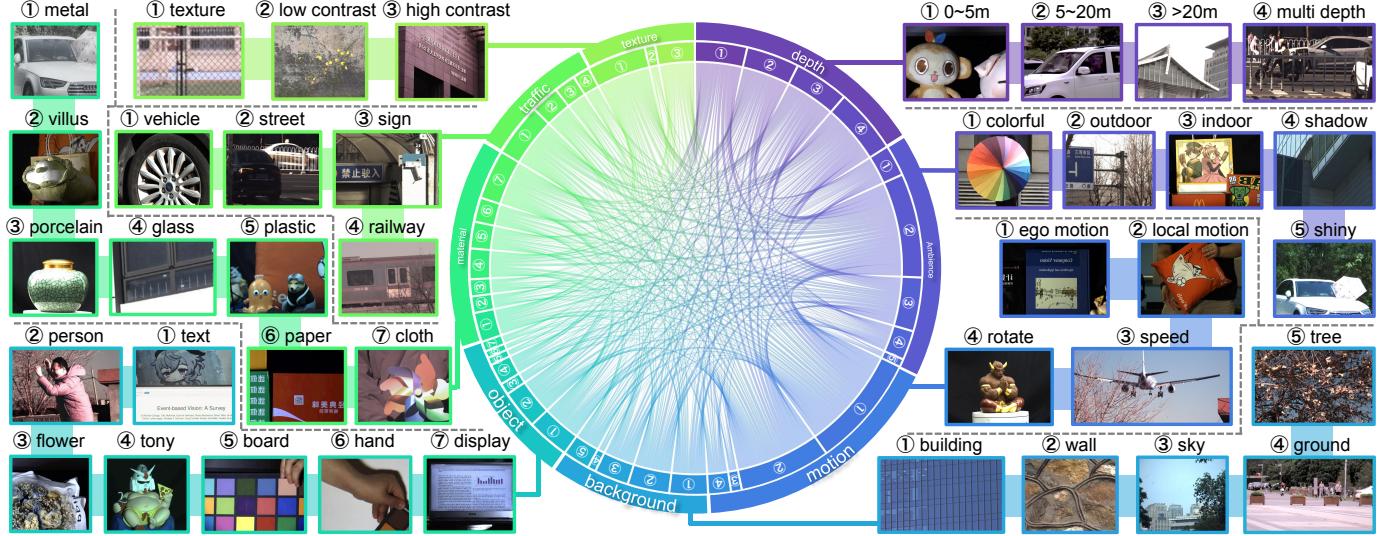


Figure 4: Overview of the real-world scenes on the proposed datasets. The frequency of associations between 8 types of scene attributes is visualized in the chord diagram. Each attribute is subdivided into detailed categories, and one sample scene of each category is showcased on the left or right sides. The colored bands below the sample scenes serve to indicate the positions of categories under the same attribute.

and marks each rising edge and falling edge with special marking events. By shooting a high-precision LED timer, we confirm that the time error is within 10ms. For EVENTAID-B, the trigger times of short exposure images are set at the center of the exposure time of long exposure images.

3.4 Scene diversity and dataset size

We consider the scenario diversity of all sub-datasets. As shown in Fig. 4, all sub-datasets include indoor and outdoor, global and local motion, slow and fast motion, and various texture scenes³. To analyze the performance for challenging scenes, we collect fast, non-linear motion, and smooth texture scenes in EVENTAID-R/-F/-B, complex texture scenes in EVENTAID-S, and wide dynamic range scenes in EVENTAID-D. We collect a large amount of test data for each sub-dataset, with EVENTAID-F containing 10 groups totaling 44,496 frames, EVENTAID-R containing 14 groups⁴ totaling 45,270 frames, EVENTAID-B containing 14 groups totaling 4,088 frames, EVENTAID-S containing 10 groups totaling 10,986 frames, and EVENTAID-D containing 13 groups totaling 5,173 frames.

We also collect two simulated datasets, *i.e.*, EVENTAID-V2E/-VM, to evaluate and compare the real-sim gap of two widely used event simulators, *i.e.*, V2E [86] and DVS-Voltmeter [87] on event-based video reconstruction, event-aided HFR video reconstruction, image deblurring, and image SR reconstruction tasks. To simulate the corresponding events, the ground truth videos of EVENTAID-R/-F/-B/-S are input into two simulators to generate event data. All parameters are set according to the author's suggestions.

3. More scenes, as well as completed scene numbers and names are recorded in the supplementary material.

4. 10 groups are shared with EVENTAID-F and 4 groups are collected by a SilkyEvCam BothView camera.

4 EXPERIMENTS AND BENCHMARK ANALYSIS

4.1 Methods and evaluation metrics

We use the sub-datasets of EVENTAID to evaluate representative methods of five event-aided image/video enhancement tasks respectively. (1) For event-based video reconstruction task, we choose E2VID (CVPR19 [20], TPAMI20 [5]), FireNet (WACV20 [63]), ET-Net (ICCV21 [64]), SPADE-E2VID (TIP21 [51]), SSL-E2VID (ICCV21 [65]), and EVSNN (CVPR22 [66]). (2) For the event-aided HFR video reconstruction task, we choose TimeLens (CVPR21 [19]), E-VFIA (ICRA23 [67]), and CBMNet (CVPR23 [40]). (3) For the event-aided image deblurring task, we choose EDI (CVPR19 [22], TPAMI20 [71]), RED-Net (ICCV21 [72]), D2Net (ICCV21 [73]), EVDI (CVPR22 [74]), EFNet (ECCV22 [75]), NEST (ECCV22 [76] and REFID (CVPR23 [55])). (4) For the event-aided image SR reconstruction task, we choose E2SRI (CVPR20 [21], TPAMI22 [46]) and EvIntSR (ICCV21 [47]). (5) For the event-aided HDR image reconstruction task, we choose HDRev (CVPR23 [28]) and NeuImg-HDR (CVPR20 [24], TPAMI23 [48]). For algorithms with both conference and journal papers, we chose their latest version for evaluation. We use the original code and pre-trained model of each method released from their project websites.

For each of the four event-aided image/video enhancement tasks (*i.e.*, event-aided HFR reconstruction, image deblurring, image SR, and HDR restoration), we also benchmark three state-of-the-art single-image-based methods. (1) For HFR video reconstruction, we choose FLAVR (WACV23 [68]), RIFE (ECCV22 [69]), and VFIFformer (CVPR22 [70]). (2) For image deblurring, we choose FFTFormer (CVPR23 [77]), NAFNet (ECCV22 [78]), and Restormer (CVPR22 [79]). (3) For image SR, we choose ATD (CVPR24 [80]), CAMixer (CVPR24 [82]), and BFSR (CVPR24 [81]). (4) For HDR restoration, we choose CEVR (WACV23 [88]), KUNet (IJCAI22 [89]), and SingleHDR (CVPR20 [90]).

Table 2: The benchmark results for four tasks on real-captured EVENTAID and simulated EVENTAID-V2E/-VM dataset. The top three values for each group are highlighted in color block, with redder shades indicating higher rankings.

Event-based video reconstruction											
	Methods	EventAid-R			EventAid-R-V2E			EventAid-R-VM			
		PSNR (\uparrow)	SSIM (\uparrow)	LPIPS (\downarrow)	PSNR (\uparrow)	SSIM (\uparrow)	LPIPS (\downarrow)	PSNR (\uparrow)	SSIM (\uparrow)	LPIPS (\downarrow)	
Event-based methods	E2VID [5], [20]	6.912	0.398	0.582	6.295	0.390	0.624	10.464	0.532	0.500	
	FireNet [63]	8.461	0.405	0.571	8.890	0.397	0.617	10.536	0.501	0.528	
	ET-Net [64]	13.757	0.483	0.548	14.225	0.560	0.566	13.139	0.473	0.536	
	SPADE-E2VID [51]	9.570	0.393	0.592	9.018	0.371	0.630	10.118	0.425	0.567	
	SSL-E2VID [65]	9.282	0.409	0.601	10.537	0.448	0.614	9.340	0.433	0.576	
	EVSNN [66]	9.660	0.396	0.606	9.441	0.370	0.657	9.184	0.420	0.582	
Event-aided HFR video reconstruction											
	Methods	EventAid-F			EventAid-F-V2E			EventAid-F-VM			
		PSNR (\uparrow)	SSIM (\uparrow)	LPIPS (\downarrow)	PSNR (\uparrow)	SSIM (\uparrow)	LPIPS (\downarrow)	PSNR (\uparrow)	SSIM (\uparrow)	LPIPS (\downarrow)	
Event-based methods	TimeLens [19]	31.634	0.916	0.122	28.169	0.861	0.159	28.547	0.871	0.154	
	E-VFIA [67]	29.282	0.872	0.204	27.227	0.833	0.234	27.490	0.839	0.228	
	CBMNet [40]	32.436	0.926	0.132	29.951	0.895	0.153	30.212	0.899	0.151	
Image-based methods	FLAVR [68]	31.500	0.903	0.179	/	/	/	/	/	/	
	RIFE [69]	32.198	0.906	0.117	/	/	/	/	/	/	
	VFIFormer [70]	31.366	0.908	0.152	/	/	/	/	/	/	
Event-aided image deblurring											
	Methods	EventAid-B			EventAid-B-V2E			EventAid-B-VM			
		PSNR (\uparrow)	SSIM (\uparrow)	LPIPS (\downarrow)	PSNR (\uparrow)	SSIM (\uparrow)	LPIPS (\downarrow)	PSNR (\uparrow)	SSIM (\uparrow)	LPIPS (\downarrow)	
Event-based methods	EDI [71]	27.243	0.824	0.308	25.987	0.826	0.298	25.285	0.796	0.336	
	RED-Net [72]	23.134	0.847	0.263	24.316	0.808	0.318	23.129	0.789	0.327	
	D2Net [73]	26.205	0.862	0.299	25.912	0.855	0.306	25.822	0.853	0.308	
	EVDI [74]	27.528	0.865	0.261	25.329	0.829	0.306	23.834	0.801	0.317	
	EFNet [75]	28.522	0.884	0.262	25.852	0.843	0.318	25.191	0.830	0.324	
	NEST [76]	28.628	0.887	0.224	24.675	0.828	0.310	25.239	0.812	0.305	
	REFID [55]	28.504	0.894	0.235	25.189	0.838	0.309	24.266	0.823	0.313	
Image-based methods	FFTFormer [77]	23.691	0.831	0.324	/	/	/	/	/	/	
	NAFNet [78]	26.291	0.860	0.316	/	/	/	/	/	/	
	Restormer [79]	26.716	0.855	0.260	/	/	/	/	/	/	
Event-aided image super-resolution											
	Methods	EventAid-S			EventAid-S-V2E			EventAid-S-VM			
		PSNR (\uparrow)	SSIM (\uparrow)	LPIPS (\downarrow)	PSNR (\uparrow)	SSIM (\uparrow)	LPIPS (\downarrow)	PSNR (\uparrow)	SSIM (\uparrow)	LPIPS (\downarrow)	
Event-based methods	E2SRI [46]	8.813	0.497	0.606	8.912	0.504	0.611	10.052	0.539	0.577	
	EvIntSR [47]	18.292	0.777	0.303	18.287	0.780	0.307	18.064	0.764	0.308	
Image-based methods	ATD [80]	45.180	0.982	0.115	/	/	/	/	/	/	
	BFSR [81]	46.047	0.986	0.095	/	/	/	/	/	/	
	CAMixer [82]	45.255	0.984	0.111	/	/	/	/	/	/	

The output results can be evaluated with the full reference evaluation metrics except for EVENTAID-D. We adopt PSNR to approximate estimate the human perception of reconstruction quality, and SSIM to evaluate the similarity of two images from the luminance, contrast, and structure components. We further use LPIPS [91], which better models the humane judgment by extracting the features from the pre-trained classification network, to evaluate the perceptual similarity between results and the ground truth.

4.2 Benchmarking for event-aided methods

We show the quantitative results on both real and simulated data for 4 tasks in Table 2. Representative qualitative results are presented in Fig. 5, Fig. 6, Fig. 7, and Fig. 8.⁵

5. In the supplementary material, Sec. 1 to Sec. 5 of the document file respectively shows more comparison results for five tasks on the real-captured EVENTAID dataset and the simulated EVENTAID-V2E/-VM datasets, Sec. 6 shows the distribution of quantitative results across all frames by boxplots, and the video results from different comparison methods are provided in the video file.

4.2.1 Event-based video reconstruction

We feed the input events of EVENTAID-R into the methods to be benchmarked and each of them reconstructs a video with a frame rate of 150FPS with timestamps matching the ground truth video. The quantitative comparison result in Table 2 shows that ET-Net [64] achieves more promising performance than other methods, the quantitative result distribution⁶ also shows that ET-Net [64] performs optimally in most groups. Nevertheless, the qualitative results of E2VID [5], [20] and FireNet [63] in Fig. 5 seem more natural, retaining more detail, while other methods tend to reconstruct images with sharper edges and high contrast. It is consistent with the results in their original papers.

Inspiration: There are two main goals in the current research for event-based video reconstruction, one is to faithfully restore details and contrasts of natural images (e.g., E2VID [5] and FireNet [63]), and the other one is to enhance

6. The quantitative result distribution is shown in Fig. S6-1 of the supplementary material.

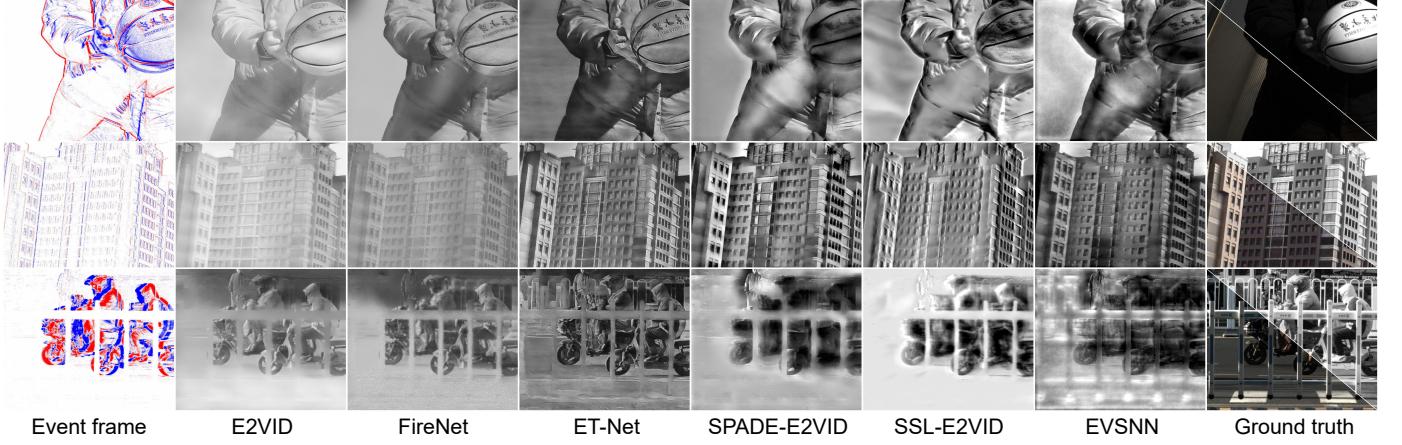


Figure 5: Event-based video reconstruction result examples from the EVENTAID-R dataset. We show the results output from E2VID [5], [20], FireNet [63], ET-Net [64], SPADE-E2VID [51], SSL-E2VID [65], and EVSNN [66]. In each square of the ground truth column, lower-left shows the RGB images, and upper-right shows the corresponding gray channel to facilitate comparison with the grayscale images produced from evaluated methods.

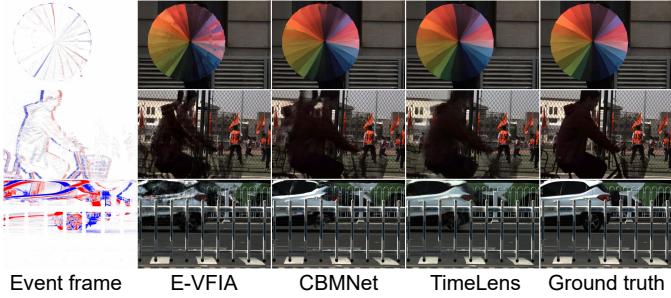


Figure 6: Event-aided HFR video reconstruction result examples from the EVENTAID-F. We show the results produced from E-VFIA [67], CBMNet [40], and TimeLens [19].

the contrast and sharpness as well as suppress noises of images (*e.g.*, ET-Net [64] and SSL-E2VID [65]) to highlight the main objects of the scene. Current metrics for evaluating reconstruction quality such as PSNR only focus on global pixel value similarity, and ignore evaluating the image contrast enhancement and noise suppression quality. To comprehensively evaluate the reconstruction performance of different methods, developing new metrics for balancing detail recovery and noise suppression in event-based video reconstruction is necessary for future research. In addition, current research focuses on video reconstruction of event data from DAVIS240 [30] or DAVIS346 [2], and event cameras with higher resolution (*e.g.*, Prophesee EVK4 HD) also become popular. The comparison show that some methods perform much better on the DAVIS-captured event data in the original papers (*e.g.*, Fig. 6 in EVSNN [66]) than on the Prophesee-captured event data (*i.e.*, EVENTAID-R). How to reconstruct high-quality video given event data with more pixels or introduce image pre-training models to improve the reconstruction performance is worth further exploration.

4.2.2 Event-aided HFR video reconstruction

We extract frames from the HFR ground truth videos in EVENTAID-F by a factor of 1/8 as the LFR input videos, then feed the input events and videos into the methods to be benchmarked to reconstruct 8 \times HFR videos. Note that we only consider inter-frame interpolation and not intra-frame

interpolation. We classify intra-frame interpolation into the image deblurring task. We use the first 700 frames of each group as the training dataset and finetune the pretrained models. The retraining strategies follow the original papers or acquired from the authors. The quantitative comparison results in Table 2 and additional results⁷ show the best frame interpolation performance of CBMNet [40]. Qualitative results show that TimeLens [19] and CBMNet [40] perform significantly better in challenging scenarios such as high-speed motion. Besides, CBMNet [40] tends to reconstruct better performance for intermediate frames.

Inspiration: Introducing events into this task mainly aims to use the high-temporal precision motion information recorded by events to restore the motion trajectory of objects, and events are mostly desired when there are non-linear and complex motions in the scene. However, existing algorithms do not always extract the motion information precisely, resulting in distorted edges and inaccurate color recovery in the interpolated frames. Besides, event degradations such as noise, tailing, and signal loss also affect the accuracy of motion extraction and the quality of image detail recovery. Some algorithms also exhibit significant performance variations in reconstructing skipped frames at different positions. How to model and represent non-linear motion while eliminating the interference caused by event degradations is the main bottleneck encountered in this task. With future progress in the accurate motion extraction of event data, the effect of HFR reconstruction is expected to be further improved.

4.2.3 Event-aided image deblurring

We feed the input blur image sequence and corresponding events within its exposure period into deblurring methods to restore clear images at the input image's exposure period. Since RED-Net [72] can only process grayscale images, we evaluate the output results with ground truth in grayscale space. we randomly select half of the frames as the training dataset and finetune the pretrained model of each method

⁷ More results and results of image-based methods are shown in Sec. 2.1 and Sec. 6.4 of the supplementary material.

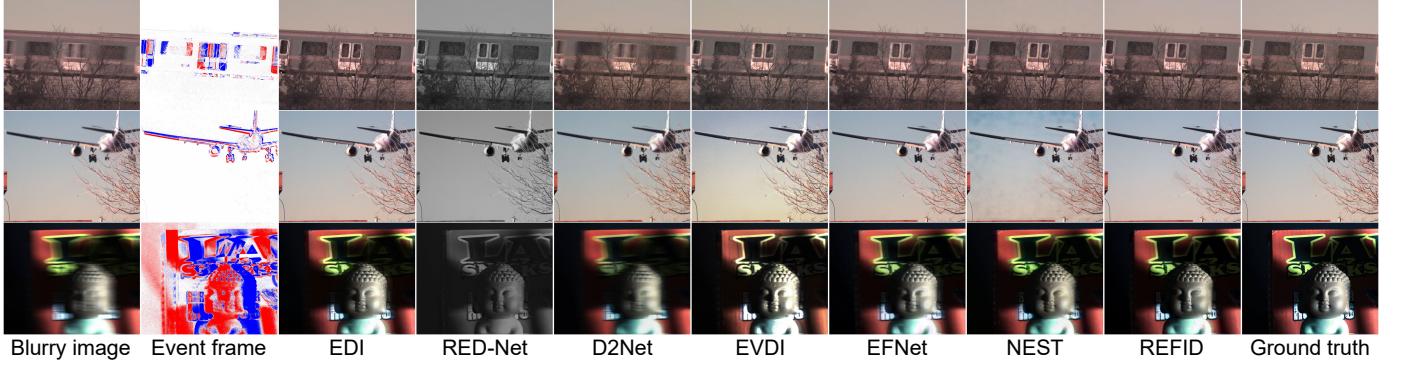


Figure 7: Event-aided image deblurring result examples from the EVENTAID-B dataset. We show the results produced from EDI [22], RED-Net [72], D2Net [73], EVDI [74], EFNet [75], NEST [76], and REFID [55].

accordingly. The retraining strategies are implemented following the descriptions in the original papers or acquired from the authors. The PSNR and SSIM comparison results⁸ in Table 2 show that the performance of EFNet [75], NEST [76] and REFID [55] are all competitive, which benefit from their well-designed network model. The qualitative comparisons in Fig. 7 show that EDI [71], EFNet [75] and REFID [55] can reconstruct clear images. However, the quantitative results for some methods are relatively low, which may be due to a mismatch between the training data and real-captured data distribution, leading to lower pixel values in outputs and subsequently lowering the test metrics.

controlled experiment: We further experiment with event-aided image deblurring to evaluate the limits of blur levels that existing methods can withstand. Table 3 records the quantitative results from the pretrained models. We perform gamma correction on the results of RED-Net [72] and NEST [76] to alleviate the brightness issues and mark the refined results with “*”. The average PSNR values of D2Net [73], EVDI [74], EFNet [75], REFID [55] and refined RED-Net [72] are comparable and better than other methods including image-based ones, while EFNet [75] and REFID [55] are significantly superior to other methods in LPIPS.

The changing trend of the colored blocks in Table 3 reveals the algorithm’s tolerance and compatibility with different degrees of blur⁹. It is interesting that the PSNR values of most algorithms decrease as the degree of blur diminishes. This counterintuitive result may be due to a smoothing effect of the PSNR metric, where the algorithm smooths the reconstruction of blurred edges in images with higher blur, leading to a higher PSNR value. In contrast, the trend in LPIPS more reasonably reflects the performance improvement of the algorithms as the blur decreases.

Inspiration: From the comparison, we can find that two issues prevent existing methods from being further improved: The modal differences between events and images and the difficulty in calibrating event trigger thresholds lead to additional artifacts introduced by event signals. Designing an event representation model that is more effective for image fusion, and proposing a robust online threshold estimation method might help in conquering these bottlenecks. Besides,

Table 3: The image deblurring quantitative results on the dataset collected from the controlled experiment (Fig. 3). We calculate PSNR and SSIM for 7 methods that perform the deblurring process on input images with 11 blur levels. The blur levels correspond to the swing periods, a shorter period indicates a higher degree of blur. “Ave.” indicates the average values, redder blocks represent better performance with a higher PSNR value or lower LPIPS value. The values in yellow/green blocks represent the differences from average values, greener blocks represent better performance.

Methods	Ave.	PSNR									
		1.50s	1.75s	2.00s	2.25s	2.50s	2.75s	3.00s	3.25s	3.50s	3.75s
EDI [71]	32.80	+0.08	+0.16	+0.32	-0.07	+0.07	+0.11	+0.00	-0.18	-0.11	-0.22
RED-Net [72]	14.17	-0.02	+0.05	+0.28	-0.24	-0.23	-0.07	+0.05	+0.19	-0.03	-0.04
RED-Net*	33.15	-0.51	-0.22	-0.07	-0.17	-0.04	+0.25	+0.40	+0.09	+0.10	-0.05
D2Net [73]	33.21	+0.74	+0.41	+0.36	+0.05	+0.04	+0.05	-0.16	-0.37	-0.42	-0.51
EVDI [74]	33.26	+0.09	+0.19	+0.36	+0.01	+0.18	+0.25	-0.05	-0.22	-0.23	-0.35
EFNet [75]	33.22	+0.62	+0.52	+0.60	+0.07	+0.06	+0.08	-0.22	-0.40	-0.39	-0.51
NEST [76]	20.10	-0.01	+0.02	+0.03	-0.05	+0.03	-0.04	-0.01	-0.01	+0.03	+0.02
NEST*	27.46	-0.09	+0.04	-0.34	+0.10	+0.45	-0.06	-0.01	+0.05	-0.13	-0.12
REFID [55]	33.52	+0.82	+0.60	+0.57	+0.16	+0.13	+0.11	-0.31	-0.46	-0.51	-0.62
FFTFormer [77]	32.55	+0.08	+0.02	+0.15	-0.07	+0.18	+0.11	-0.02	-0.14	-0.06	-0.20
NAFNet [78]	32.12	+0.48	+0.39	+0.29	+0.16	+0.20	+0.05	-0.14	-0.32	-0.31	-0.47
Restormer [79]	32.85	+0.49	+0.40	+0.40	+0.01	+0.03	+0.05	-0.16	-0.29	-0.28	-0.37

Methods	Ave.	LPIPS									
		1.50s	1.75s	2.00s	2.25s	2.50s	2.75s	3.00s	3.25s	3.50s	3.75s
EDI [71]	0.293	0.010	0.010	0.008	0.014	0.013	0.012	-0.009	-0.007	-0.014	-0.015
RED-Net [72]	0.327	0.013	0.018	0.013	0.015	0.004	0.000	-0.012	-0.012	-0.009	-0.007
RED-Net*	0.246	-0.005	0.007	0.005	0.012	0.004	-0.002	-0.005	0.001	-0.001	0.000
D2Net [73]	0.230	-0.017	-0.002	-0.001	0.004	0.003	-0.003	0.001	0.008	0.008	0.006
EVDI [74]	0.231	0.015	0.014	0.005	0.010	0.000	-0.006	-0.008	-0.006	-0.005	-0.004
EFNet [75]	0.207	-0.018	-0.005	-0.008	0.003	0.004	0.001	0.004	0.009	0.008	0.006
NEST [76]	0.437	0.008	0.002	0.005	0.002	-0.005	-0.005	-0.006	0.001	0.002	0.004
NEST*	0.451	0.010	0.004	0.004	0.002	-0.004	-0.007	-0.006	0.001	0.000	0.003
REFID [55]	0.206	-0.019	-0.007	-0.006	0.000	-0.001	-0.001	0.004	0.008	0.013	0.011
FFTFormer [77]	0.229	-0.001	0.009	0.004	0.011	0.001	-0.004	-0.002	0.001	-0.001	-0.003
NAFNet [78]	0.240	-0.020	-0.010	-0.005	0.003	-0.003	-0.003	0.006	0.009	0.013	0.011
Restormer [79]	0.216	-0.009	0.000	-0.003	0.006	0.001	-0.003	0.003	0.004	0.005	0.003

similar to the HFR reconstruction task, the difficulty of extracting the non-linear motion accurately makes it hard to restore sharp textures correctly. So precise motion extractions are also desired here. In addition, event cameras perceive in grayscale space, so it is difficult to assist in recovering the color information of blurry areas. Introducing color event cameras or image colorization models is a possible way of improvement. Finally, due to the large spatial resolution gap between event cameras and frame

8. More results are shown in Sec. 3 of the supplementary material.

9. Visual results are shown in Fig. S3-71 to Fig. S3-74.

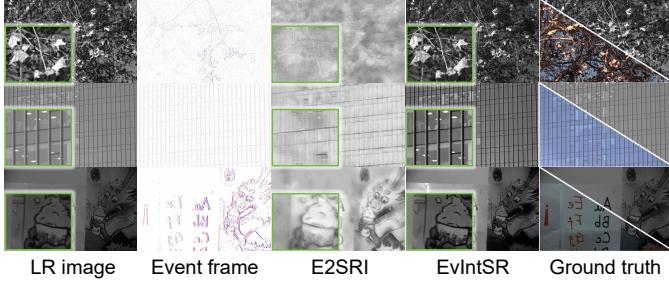


Figure 8: Event-aided image super-resolution result examples from the EVENTAID-S dataset. We show the results produced from E2SRI [21], [46], and EvIntSR [47]. Green boxes show the closed-up views.

cameras, how to use LR event signals to deblur HR images with much higher resolution (like 20 times larger) is of practical value for enhancing the photography experience in future camera phones. The performance advantages over image-based methods in dealing with severe blur validate the necessity of introducing event cameras.

4.2.4 Event-aided image super-resolution

We generate input $1\times$ LR images by downsampling HR images of EVENTAID-S at a factor of $1/2$ following the process in single image SR tasks, then feed the $1\times$ LR image sequences and events into selected methods to reconstruct $2\times$ SR frames. Note that E2SRI [5], [21] executes the SR process directly from pure event data, so we only feed $1\times$ LR events into it. We evaluate the output results with ground truth in grayscale space since E2SRI [5], [21] and EvIntSR [47] can only process grayscale images. EventZoom [32] and ESR [92] are not compared here because these algorithms output high-resolution event signals rather than directly output high-resolution images. The quantitative and qualitative comparison results¹⁰ in Table 2 and Fig. 8 show the best frame SR performance of EvIntSR [47]. However, the comparison can not validate that the performance of EvIntSR [47] is better than the other method because the input of the two methods is inconsistent. For E2SRI [21] with only input events, the spatial resolution cannot be amplified by events when the scene is static.

Inspiration: The core principle of this task is to use the high temporal motion information recorded by events to convert sub-pixel displacements in the spatial domain, thereby achieving spatial upsampling. The results show that the performance of the single-image methods is significantly better than event-based methods because the two tested event-based methods are trained on low-resolution data, while the higher resolution of EVENTAID-S challenges the generalization of them. In addition, $2\times$ SR is easy for single-image methods, while event noise affects the performance of event-based methods. Besides, pure image-based SR has been studied for decades, and some learning-based methods can even achieve $16\times$ upsampling (e.g., ABPN [93]), while event-aided methods can only achieve lower-factor upsampling. New explorations should make better use of events

10. More results are shown in Sec. 4.1 and Sec. 6.10 of the supplementary material.

to reconstruct SR images of higher quality than pure image-based methods to further demonstrate the practical value and research significance of this event-aided task.

4.2.5 Event-aided HDR image reconstruction

We feed the LDR images and corresponding events into selected HDRev [28] and NeuImg-HDR [24], [48] to restore HDR images. For EVENTAID-D-Static, we obtain LDR images captured through short-/middle-/long-exposure by alternating exposure and use them as input to fully test the robustness of methods. To obtain HDR reference, we first hold the scene still while shooting the data and capture 11 multi-exposure frames to synthesize the reference image byDebevec *et al.* [94]¹¹. For EVENTAID-D-Dynamic, overexposed and underexposed LDR videos are merged to an HDR video reference. It can be seen from the comparison that both algorithms show a trend that the recovery performance of the under-exposed area is better than that of the over-exposed area, perhaps because the white background color of the over-exposed area makes it easier to highlight the reconstructed artifacts. The color of the image reconstructed by HDRev [28] is more in line with human vision, while the texture details restored by NeuImg-HDR [24], [48] are clearer. The experiment also shows that the event-based methods have obvious advantages over the single image methods in the area where the image is over/under-exposed but the events perceive the texture information effectively.

Inspiration: Event-aided HDR image reconstruction methods use the texture motion of over-/under-exposed areas perceived by events to recover the lost information of these areas and fuse them with LDR images. Accurately restoring texture and color are two major attributes that the methods of this task need to own. The challenge of reconstructing realistic textures in over-/under-exposed areas is similar to that of the event-based video reconstruction task. In addition, current methods for Event-aided HDR image restoration [24] require shaking the event camera when capturing data to make the event camera sufficiently perceive the texture in the scene. Breaking through this limitation will enable the algorithm to be more conveniently and broadly used. Moreover, it is difficult for existing methods to correctly restore the color of over-/under-exposed areas because neither the event nor the image provides color priors. Therefore, how to recover the color of HDR areas is also a challenge that needs to be explored. Using color restoration strategies in algorithms such as image colorization, image inpainting, and semantic-based image restoration, or using existing color recovery pre-trained models might be helpful to solve the problem faced by this task.

4.3 Comparison with event simulators

Event simulators are useful for generating large-scale training datasets, but the real-sim gap makes it difficult for trained models to work efficiently on real-captured data, which has been verified by NeuroZoom [14]. We execute the above benchmark processing again on the simulated EVENTAID-V2E/-VM datasets to compare the construction results. The quantitative results on simulated data are

11. Results are shown in Sec. 5 of the supplementary material.

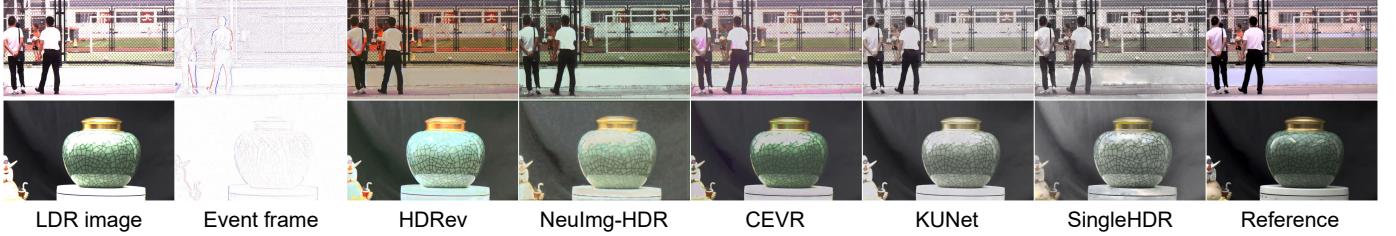


Figure 9: Event-aided HDR image reconstruction results from EVENTAID-D. We show results of event-based methods HDRev [28] and NeuImg-HDR [24], [48], and single image-based methods CEVR [88], KUNet [89], and SingleHDR [90].

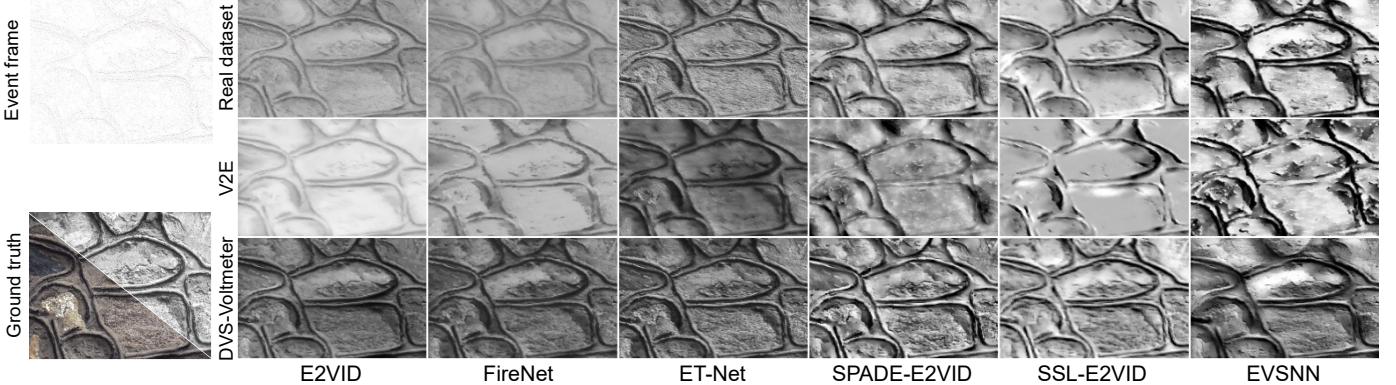


Figure 10: Event-based video reconstruction result examples from real-captured EVENTAID-R, V2E [86] simulated dataset EVENTAID-R-V2E, and DVS-Voltmeter [87] simulated dataset EVENTAID-R-VM.

recorded in Table 2. One qualitative result sample on event-based video reconstruction task is presented in Fig. 10¹².

The quantitative results show that the performance ranking on simulated data is close to the ranking on the real-captured dataset EVENTAID. This indirectly proves that the spatiotemporal synchronization error of our EVENTAID dataset is not significant, and the comparisons on real-captured EVENTAID are convincing. The results in Table 2 show that the performances on V2E-simulated datasets are lower than on real data, while some methods perform against this trend, such as ET-Net [64] in the video reconstruction task. This may be because the pre-trained models of these methods are trained on V2E-simulated datasets. In contrast, the performances on EVENTAID-VM datasets are similar or even higher than the corresponding performance on real data, especially the video reconstruction task. Because the DVS-Voltmeter [87] more realistically models the triggering process of event signals, the simulated event distribution model matches the real data. In contrast, the quality of V2E [86] depends on the frame rate of the input video. When the frame rate is low, the generated events are difficult to simulate the continuous distribution of real events in the time domain. The comparison of qualitative results also shows that EVENTAID-VM dataset results in better visual effects. Note that in the event-based video reconstruction, the image reconstructed on the EVENTAID-VM has a greater contrast than the result on real-captured EVENTAID, which may be related to the inaccurate setting of the trigger threshold of the simulator. Whether the event simulator can accurately simulate the trigger mechanism of the real event sensor, correctly model the degradation process such as noise, trailing, and signal loss, and solve the

problem of discontinuous event distribution in time dimension when converting low frame rate video into events will determine whether simulators can provide effective training and evaluation data for the study of event algorithms.

5 CONCLUSIONS

We propose the first evaluation dataset for event-aided image/video enhancement tasks with real-captured data that allow quantitative and qualitative evaluations. All data are real-captured by beam-splitter-mounted hybrid camera systems. We benchmark 20 event-based methods and 12 image-based methods for both five tasks and analyze their performances, and also benchmark 2 widely used event simulators. Finally, we discuss the performance of existing methods and propose several open problems for future researchers.

Limitations: Some published methods have not been benchmarked in this paper because the codes are unavailable or they have just been published. we will release a benchmark website that allows researchers to update their methods to continuously facilitate research on event-aided image/video enhancement tasks after the acceptance of this paper.

ACKNOWLEDGMENT

This work was supported by National Natural Science Foundation of China (Grant No. 62088102, 62402014, 62136001), Beijing Natural Science Foundation (Grant No. L233024), Beijing Municipal Science & Technology Commission, Administrative Commission of Zhongguancun Science Park (Grant No. Z241100003524012). Peiqi Duan was also supported by China National Postdoctoral Program for Innovative Talents (Grant No. BX20230010) and China Postdoctoral Science Foundation (Grant No. 2023M740076).

12. More results are included in the supplementary material, Sec. 6 shows the quantitative result distribution across all frames by boxplots.

REFERENCES

- [1] P. Lichtsteiner, C. Posch, and T. Delbrück, "A 128×128 120 db 15 μ s latency asynchronous temporal contrast vision sensor," *IEEE Journal of Solid-State Circuits*, vol. 43, no. 2, 2008.
- [2] G. Taverni, D. P. Moeyns, C. Li, C. Cavaco, V. Motsnyi, D. S. S. Bello, and T. Delbrück, "Front and back illuminated dynamic and active pixel vision sensors comparison," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 65, no. 5, pp. 677–681, 2018.
- [3] C. D. Schuman, S. R. Kulkarni, M. Parsa, J. P. Mitchell, P. Date, and B. Kay, "Opportunities for neuromorphic computing algorithms and applications," *Nature Computational Science*, vol. 2, no. 1, pp. 10–19, 2022.
- [4] G. Gallego, T. Delbrück, G. Orchard, C. Bartolozzi, B. Taba, A. Censi, S. Leutenegger, A. J. Davison, J. Conradt, K. Daniilidis, and D. Scaramuzza, "Event-based vision: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 1, pp. 154–180, 2022.
- [5] H. Rebecq, R. Ranftl, V. Koltun, and D. Scaramuzza, "High speed and high dynamic range video with an event camera," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 6, pp. 1964–1980, 2021.
- [6] R. Christian, G. Gottfried, and P. Thomas, "Real-time intensity-image reconstruction for event cameras using manifold regularisation," in *Proc. of British Machine Vision Conference (BMVC)*, 2016.
- [7] C. Lee, A. Kosta, A. Z. Zhu, K. Chaney, K. Daniilidis, and K. Roy, "Spike-FlowNet: Event-based optical flow estimation with energy-efficient hybrid neural networks," in *Proc. of European Conference on Computer Vision (ECCV)*, 2020.
- [8] A. Z. Zhu, L. Yuan, K. Chaney, and K. Daniilidis, "Unsupervised event-based learning of optical flow, depth, and egomotion," in *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pp. 989–997, 2019.
- [9] A. Baudron, Z. W. Wang, O. Cossairt, and A. K. Katsaggelos, "E3d: Event-based 3d shape reconstruction," *arXiv*, vol. abs/2012.05214, 2020.
- [10] X. Wang, J. Li, L. Zhu, Z. Zhang, Z. Chen, X. Li, Y. Wang, Y. Tian, and F. Wu, "VisEvent: Reliable object tracking via collaboration of frame and event flows," *ArXiv*, vol. abs/2108.05015, 2021.
- [11] B. Ramesh and H. Yang, "Boosted kernelized correlation filters for event-based face detection," in *Proc. of Winter Conference on Applications of Computer Vision Workshops (WACVW)*, 2020.
- [12] A. R. Vidal, H. Rebecq, T. Horstschafer, and D. Scaramuzza, "Ultimate SLAM? combining events, images, and IMU for robust visual SLAM in HDR and high-speed scenarios," *IEEE Robotics and Automation Letters*, vol. 3, no. 2, pp. 994–1001, 2018.
- [13] J. Li, S. Dong, Z. Yu, Y. Tian, and T. Huang, "Event-based vision enhanced: A joint detection framework in autonomous driving," in *Proc. of IEEE International Conference on Multimedia and Expo (ICME)*, 2019.
- [14] P. Duan, Y. Ma, X. Zhou, X. Shi, Z. W. Wang, T. Huang, and B. Shi, "NeuroZoom: Denoising and super resolving neuromorphic events and spikes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–14, 2023.
- [15] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. of European Conference on Computer Vision (ECCV)*, pp. 740–755, 2014.
- [16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. of Advances in Neural Information Processing Systems (NeurIPS)*, pp. 5998–6008, 2017.
- [17] D. Ko, J. Choi, H. K. Choi, K.-W. On, B. Roh, and H. J. Kim, "MELTR: Meta loss transformer for learning to fine-tune video foundation models," in *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pp. 20105–20115, 2023.
- [18] P. Duan, Z. Wang, B. Shi, O. Cossairt, T. Huang, and A. Katsaggelos, "Guided Event Filtering: Synergy between intensity images and neuromorphic events for high performance imaging," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 11, pp. 8261–8275, 2021.
- [19] S. Tulyakov, D. Gehrig, S. Georgoulis, J. Erbach, M. Gehrig, Y. Li, and D. Scaramuzza, "Time Lens: Event-based video frame interpolation," in *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2021.
- [20] H. Rebecq, R. Ranftl, V. Koltun, and D. Scaramuzza, "Events-to-video: Bringing modern computer vision to event cameras," in *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [21] S. M. Mostafavi I., J. Choi, and K.-J. Yoon, "Learning to super resolve intensity images from events," in *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [22] L. Pan, C. Scheerlinck, X. Yu, R. Hartley, M. Liu, and Y. Dai, "Bringing a blurry frame alive at high frame-rate with an event camera," in *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [23] S. Tulyakov, A. Bochicchio, D. Gehrig, S. Georgoulis, Y. Li, and D. Scaramuzza, "Time lens++: Event-based frame interpolation with parametric non-linear flow and multi-scale fusion," in *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2022.
- [24] J. Han, C. Zhou, P. Duan, Y. Tang, C. Xu, C. Xu, T. Huang, and B. Shi, "Neuromorphic camera guided high dynamic range imaging," in *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [25] X. Zhou, P. Duan, Y. Ma, and B. Shi, "EvUnroll: Neuromorphic events based rolling shutter image correction," in *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2022.
- [26] Z. W. Wang, P. Duan, O. Cossairt, A. Katsaggelos, T. Huang, and B. Shi, "Joint filtering of intensity images and neuromorphic events for high-resolution noise-robust imaging," in *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [27] T. Kim, J. Lee, L. Wang, and K.-J. Yoon, "Event-guided deblurring of unknown exposure time videos," *Proc. of European Conference on Computer Vision (ECCV)*, 2022.
- [28] Y. Yang, J. Han, J. Liang, I. Sato, and B. Shi, "Learning event guided high dynamic range video reconstruction," in *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2023.
- [29] C. Posch, D. Matolin, and R. Wohlgemant, "A QVGA 143 db dynamic range frame-free PWM image sensor with lossless pixel-level video compression and time-domain CDS," *IEEE Journal of Solid-State Circuits*, vol. 46, no. 1, 2010.
- [30] C. Brandli, R. Berner, M. Yang, S.-C. Liu, and T. Delbrück, "A 240×180 130 db 3 μ s latency global shutter spatiotemporal vision sensor," *IEEE Journal of Solid-State Circuits*, vol. 49, no. 10, pp. 2333–2341, 2014.
- [31] L. Pan, C. Scheerlinck, X. Yu, R. Hartley, M. Liu, and Y. Dai, "Bringing a blurry frame alive at high frame-rate with an event camera," in *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [32] P. Duan, Z. Wang, X. Zhou, Y. Ma, and B. Shi, "EventZoom: Learning to denoise and super resolve neuromorphic events," in *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2021.
- [33] E. Mueggler, H. Rebecq, G. Gallego, T. Delbrück, and D. Scaramuzza, "The event-camera dataset and simulator: Event-based data for pose estimation, visual odometry, and SLAM," *The International Journal of Robotics Research*, vol. 36, no. 2, pp. 142–149, 2017.
- [34] T. Stoffregen, C. Scheerlinck, D. Scaramuzza, T. Drummond, N. Barnes, L. Kleeman, and R. Mahony, "Reducing the sim-to-real gap for event cameras," in *Proc. of European Conference on Computer Vision (ECCV)*, 2020.
- [35] S. Zhang, Y. Zhang, Z. Jiang, D. Zou, J. Ren, and B. Zhou, "Learning to see in the dark with events," in *Proc. of European Conference on Computer Vision (ECCV)*, 2020.
- [36] A. Z. Zhu, D. Thakur, T. Özaslan, B. Pfommer, V. Kumar, and K. Daniilidis, "The multivehicle stereo event camera dataset: An event camera dataset for 3D perception," *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 2032–2039, 2018.
- [37] C. Scheerlinck, H. Rebecq, T. Stoffregen, N. Barnes, R. Mahony, and D. Scaramuzza, "CED: Color event camera dataset," in *Proc. of Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1684–1693, 2019.
- [38] B. Wang, J. He, L. Yu, G.-S. Xia, and W. Yang, "Event enhanced high-quality image recovery," in *Proc. of European Conference on Computer Vision (ECCV)*, 2020.
- [39] Z. Yu, Y. Zhang, D. Liu, D. Zou, X. Chen, Y. Liu, and J. S. Ren, "Training weakly supervised video frame interpolation with events," in *Proc. of International Conference on Computer Vision (ICCV)*, 2021.
- [40] T. Kim, Y. Chae, H.-K. Jang, and K.-J. Yoon, "Event-based video frame interpolation with cross-modal asymmetric bidirectional motion fields," in *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pp. 18032–18042, June 2023.
- [41] H. Cho, T. Kim, Y. Jeong, and K.-J. Yoon, "TTA-EVF: Test-time adaptation for event-based video frame interpolation via reliable pixel and sample estimation," in *Proceedings of the IEEE/CVF*

- Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25701–25711, June 2024.
- [42] Z. Jiang, Y. Zhang, D. Zou, J. Ren, J. Lv, and Y. Liu, “Learning event-based motion deblurring,” in *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pp. 3320–3329, 2020.
- [43] L. Sun, C. Sakaridis, J. Liang, Q. Jiang, K. Yang, P. Sun, Y. Ye, K. Wang, and L. V. Gool, “Event-based fusion for motion deblurring with cross-modal attention,” in *Proc. of European Conference on Computer Vision (ECCV)*, 2021.
- [44] T. Kim, H. Cho, and K.-J. 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 (CVPR)*, pp. 24966–24976, June 2024.
- [45] T. Kim, H. Cho, and K.-J. Yoon, “Cmta: Cross-modal temporal alignment for event-guided video deblurring,” in *Proc. of European Conference on Computer Vision (ECCV)*, 2025.
- [46] M. Mostafavi, Y. Nam, J. Choi, and K.-J. Yoon, “E2SRI: Learning to super-resolve intensity images from events,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, pp. 6890–6909, 2022.
- [47] J. Han, Y. Yang, C. Zhou, C. Xu, and B. Shi, “EvIntSR-Net: Event guided multiple latent frames reconstruction and super-resolution,” in *Proc. of International Conference on Computer Vision (ICCV)*, 2021.
- [48] J. Han, Y. Yang, P. Duan, C. Zhou, L. Ma, C. Xu, T. Huang, I. Sato, and B. Shi, “Hybrid high dynamic range imaging fusing neuromorphic and conventional images,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 7, pp. 8553–8565, 2023.
- [49] S. Barua, Y. Miyatani, and A. Veeraraghavan, “Direct face detection and video reconstruction from event cameras,” in *Proc. of Winter Conference on Applications of Computer Vision (WACV)*, pp. 1–9, 2016.
- [50] C. Scheerlinck, H. Rebecq, D. Gehrig, N. Barnes, R. Mahony, and D. Scaramuzza, “Fast image reconstruction with an event camera,” in *Proc. of Winter Conference on Applications of Computer Vision (WACV)*, pp. 156–163, 2020.
- [51] P. R. G. Cadena, Y. Qian, C. Wang, and M. Yang, “SPADE-E2VID: Spatially-adaptive denormalization for event-based video reconstruction,” *IEEE Transactions on Image Processing*, vol. 30, pp. 2488–2500, 2021.
- [52] C. Brandli, L. Muller, and T. Delbrück, “Real-time, high-speed video decompression using a frame- and event-based DAVIS sensor,” in *Proc. of International Symposium on Circuits and Systems (ISCAS)*, pp. 686–689, 2014.
- [53] Z. Wang, Y. Ng, C. Scheerlinck, and R. Mahony, “An asynchronous kalman filter for hybrid event cameras,” in *Proc. of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.
- [54] Y. Gao, S. Li, Y. Li, Y. Guo, and Q. Dai, “SuperFast: 200x video frame interpolation via event camera,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 6, pp. 7764–7780, 2023.
- [55] L. Sun, C. Sakaridis, J. Liang, P. Sun, J. Cao, K. Zhang, Q. Jiang, K. Wang, and L. Van Gool, “Event-based frame interpolation with ad-hoc deblurring,” in *Proc. of Computer Vision and Pattern Recognition (CVPR)*, June 2023.
- [56] H. Chen, M. Teng, B. Shi, Y. Wang, and T. Huang, “A residual learning approach to deblur and generate high frame rate video with an event camera,” *IEEE Transactions on Multimedia*, pp. 1–14, 2022.
- [57] S. Lin, J. Zhang, J. Pan, Z. Jiang, D. Zou, Y. Wang, J. Chen, and J. Ren, “Learning event-driven video deblurring and interpolation,” in *Proc. of European Conference on Computer Vision (ECCV)*, 2020.
- [58] L. Wang, T.-K. Kim, and K.-J. Yoon, “EventSR: From asynchronous events to image reconstruction, restoration, and super-resolution via end-to-end adversarial learning,” in *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [59] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. C. Loy, “ESRGAN: Enhanced super-resolution generative adversarial networks,” in *Proc. of European Conference on Computer Vision Workshops (ECCVW)*, 2018.
- [60] T. Köhler, M. Bätz, F. Naderi, A. Kaup, A. Maier, and C. Riess, “Toward bridging the simulated-to-real gap: Benchmarking super-resolution on real data,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 11, pp. 2944–2959, 2020.
- [61] J. Rim, H. Lee, J. Won, and S. Cho, “Real-world blur dataset for learning and benchmarking deblurring algorithms,” in *Proc. of European Conference on Computer Vision (ECCV)*, 2020.
- [62] S. Tulyakov, F. Fleuret, M. Kiefel, P. Gehler, and M. Hirsh, “Learning an event sequence embedding for dense event-based deep stereo,” in *Proc. of International Conference on Computer Vision (ICCV)*, pp. 1527–1537, 2019.
- [63] C. Scheerlinck, H. Rebecq, D. Gehrig, N. Barnes, R. E. Mahony, and D. Scaramuzza, “Fast image reconstruction with an event camera,” in *Proc. of Winter Conference on Applications of Computer Vision (WACV)*, pp. 156–163, 2020.
- [64] W. Weng, Y. Zhang, and Z. Xiong, “Event-based video reconstruction using transformer,” in *Proc. of International Conference on Computer Vision (ICCV)*, pp. 2563–2572, 2021.
- [65] F. Paredes-Vallés and G. C. H. E. de Croon, “Back to event basics: Self-supervised learning of image reconstruction for event cameras via photometric constancy,” in *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2021.
- [66] L. Zhu, X. Wang, Y. Chang, J. Li, T. Huang, and Y. Tian, “Event-based video reconstruction via potential-assisted spiking neural network,” in *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pp. 3594–3604, 2022.
- [67] O. S. Kılıç, A. Akman, and A. A. Alatan, “E-VFIA: Event-based video frame interpolation with attention,” in *Proc. of IEEE International Conference on Robotics and Automation (ICRA)*, 2023.
- [68] T. Kalluri, D. Pathak, M. Chandraker, and D. Tran, “FLAVR: Flow-agnostic video representations for fast frame interpolation,” in *WACV*, 2023.
- [69] Z. Huang, T. Zhang, W. Heng, B. Shi, and S. Zhou, “Real-time intermediate flow estimation for video frame interpolation,” in *Proc. of European Conference on Computer Vision (ECCV)*, 2022.
- [70] L. Lu, R. Wu, H. Lin, J. Lu, and J. Jia, “Video frame interpolation with transformer,” in *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2022.
- [71] L. Pan, R. Hartley, C. Scheerlinck, M. Liu, X. Yu, and Y. Dai, “High frame rate video reconstruction based on an event camera,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 5, pp. 2519–2533, 2022.
- [72] F. Xu, L. Yu, B. Wang, W. Yang, G.-S. Xia, X. Jia, Z. Qiao, and J. Liu, “Motion deblurring with real events,” in *Proc. of International Conference on Computer Vision (ICCV)*, pp. 2583–2592, 2021.
- [73] W. Shang, D. Ren, D. Zou, J. S. Ren, P. Luo, and W. Zuo, “Bringing events into video deblurring with non-consecutively blurry frames,” in *Proc. of International Conference on Computer Vision (ICCV)*, pp. 4531–4540, 2021.
- [74] X. Zhang and L. Yu, “Unifying motion deblurring and frame interpolation with events,” in *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pp. 17765–17774, 2022.
- [75] L. Sun, C. Sakaridis, J. Liang, Q. Jiang, K. Yang, P. Sun, Y. Ye, K. Wang, and L. Van Gool, “Event-based fusion for motion deblurring with cross-modal attention,” in *Proc. of European Conference on Computer Vision (ECCV)*, 2022.
- [76] M. Teng, C. Zhou, H. Lou, and B. Shi, “NEST: Neural event stack for event-based image enhancement,” in *Proc. of European Conference on Computer Vision (ECCV)*, 2022.
- [77] L. Kong, J. Dong, J. Ge, M. Li, and J. Pan, “Efficient frequency domain-based transformers for high-quality image deblurring,” in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.
- [78] L. Chen, X. Chu, X. Zhang, and J. Sun, “Simple baselines for image restoration,” in *Proc. of European Conference on Computer Vision (ECCV)*, 2022.
- [79] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, “Restormer: Efficient transformer for high-resolution image restoration,” in *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2022.
- [80] L. Zhang, Y. Li, X. Zhou, X. Zhao, and S. Gu, “Transcending the limit of local window: Advanced super-resolution transformer with adaptive token dictionary,” in *Proc. of Computer Vision and Pattern Recognition (CVPR)*, June 2024.
- [81] L.-Y. Tsao, Y.-C. Lo, C.-C. Chang, H.-W. Chen, R. Tseng, C. Feng, and C.-Y. Lee, “Boosting flow-based generative super-resolution models via learned prior,” in *Proc. of Computer Vision and Pattern Recognition (CVPR)*, June 2024.
- [82] Y. Wang, Y. Liu, S. Zhao, J. Li, and L. Zhang, “CAMixerSR: Only details need more “attention”,” in *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2024.
- [83] Z. Zhong, Y. Zheng, and I. Sato, “Towards rolling shutter correction and deblurring in dynamic scenes,” in *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pp. 9219–9228, 2021.

- [84] J. Rim, G. Kim, J. Kim, J. Lee, S. Lee, and S. Cho, "Realistic blur synthesis for learning image deblurring," in *Proc. of European Conference on Computer Vision (ECCV)*, 2022.
- [85] Y. Zou, Y. Zheng, T. Takatani, and Y. Fu, "Learning to reconstruct high speed and high dynamic range videos from events," in *Proc. of Computer Vision and Pattern Recognition (CVPR)*, June 2021.
- [86] Y. Hu, S.-C. Liu, and T. Delbruck, "V2E: From video frames to realistic dvs events," in *Proc. of Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2021.
- [87] S. Lin, Y. Ma, Z. Guo, and B. Wen, "DVS-Voltmeter: Stochastic process-based event simulator for dynamic vision sensors," in *Proc. of European Conference on Computer Vision (ECCV)*, 2022.
- [88] P.-H. Le, Q. Le, R. Nguyen, and B.-S. Hua, "Single-image hdr reconstruction by multi-exposure generation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, January 2023.
- [89] H. Wang, M. Ye, X. Zhu, S. Li, C. Zhu, and X. Li, "KUNet: Imaging knowledge-inspired single hdr image reconstruction," in *Proceedings of the International Joint Conferences on Artificial Intelligence (IJCAI-ECAI)*, 2022.
- [90] Y.-L. Liu, W.-S. Lai, Y.-S. Chen, Y.-L. Kao, M.-H. Yang, Y.-Y. Chuang, and J.-B. Huang, "Single-image hdr reconstruction by learning to reverse the camera pipeline," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.
- [91] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [92] W. Weng, Y. Zhang, and Z. Xiong, "Boosting event stream super-resolution with a recurrent neural network," in *Proc. of European Conference on Computer Vision (ECCV)*, 2022.
- [93] Z.-S. Liu, L.-W. Wang, C.-T. Li, and W.-C. Siu, "Image super-resolution via attention based back projection networks," in *Proc. of International Conference on Computer Vision Workshops (ICCVW)*, 2019.
- [94] P. E. Debevec and J. Malik, "Recovering high dynamic range radiance maps from photographs," in *ACM SIGGRAPH*, 1997.



Peiqi Duan is currently a Boya Postdoctoral Fellow at the School of Computer Science, Peking University. He received his PhD degree from Peking University in 2023. His research interests span event-based imaging and vision, single-image super-resolution, and HDR image reconstruction. He has served as a reviewer/program committee member for TPAMI, IJCV, TCSV, CVPR, ICCV, ECCV, NeurIPS, etc.



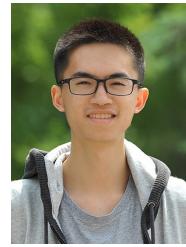
Boyu Li is currently a Ph.D. student in the School of Computer Science of Peking University. He received the B.S. degree from Peking University in 2023. His research interests include event-based vision and related topics.



Xixin Yang is a Ph.D. student in the School of Computer Science, Peking University. Her research interests span event-based imaging and vision, hybrid-camera super-resolution and HDR reconstruction.



Hanyue Lou received the B.S. degree summa cum laude from Peking University, Beijing, China, in 2023. She is currently working toward the Ph.D. degree with the National Engineering Research Center of Video Technology, School of Computer Science, Peking University. Her research interests are focused on applications of neuromorphic cameras.



Minggui Teng received the B.S. degree from Peking University, Beijing, China, in 2021. He is currently working toward the Ph.D. degree with the National Engineering Research Center of Video Technology, School of Computer Science, Peking University. His research interests are focused on neuromorphic camera and image enhancement. He has served as a reviewer for CVPR, ICCV, ECCV, etc.



Xinyu Zhou is a Ph.D. student in the School of Intelligence Science and Technology of Peking University. He received the B.S. degree from Peking University in 2022. His research interests include computational photography and event-based vision.



Yi Ma received the B.S. degree and M.E. degree from Peking University in 2021 and 2024. He is currently a research engineer at Chang Guang Satellite. His research interests are centered around event-based vision signal processing.



Boxin Shi received the BE degree from the Beijing University of Posts and Telecommunications, the ME degree from Peking University, and the PhD degree from the University of Tokyo, in 2007, 2010, and 2013. He is currently a Boya Young Fellow Assistant Professor and Research Professor at Peking University, where he leads the Camera Intelligence Lab. Before joining PKU, he did research with MIT Media Lab, Singapore University of Technology and Design, Nanyang Technological University, National Institute of Advanced Industrial Science and Technology, from 2013 to 2017. His papers were awarded as Best Paper Runner-Up at ICCP 2015 and selected as Best Papers from ICCV 2015 for IJCV Special Issue. He has served as an associate editor of TPAMI/IJCV and an area chair of CVPR/ICCV. He is a senior member of IEEE.