Asymmetric Event-Guided Video Super-Resolution

Zeyu Xiao*
MoE Key Laboratory of Brain-inspired Intelligent Perception and Cognition, University of Science and Technology of China
Hefei, China
zeyuxiao@mail.ustc.edu.cn

Dachun Kai*
MoE Key Laboratory of Brain-inspired Intelligent Perception and Cognition, University of Science and Technology of China
Hefei, China
dachunkai@mail.ustc.edu.cn

Yueyi Zhang†
MoE Key Laboratory of Brain-inspired Intelligent Perception and Cognition, University of Science and Technology of China
Hefei, China
zhyuey@ustc.edu.cn

Xiaoyan Sun
MoE Key Laboratory of Brain-inspired Intelligent Perception and Cognition, University of Science and Technology of China
Hefei, China
sunxiaoyan@ustc.edu.cn

Zhiwei Xiong
MoE Key Laboratory of Brain-inspired Intelligent Perception and Cognition, University of Science and Technology of China
Hefei, China
zwxiong@ustc.edu.cn

Abstract

Event cameras are novel bio-inspired cameras that record asynchronous events with high temporal resolution and dynamic range. Leveraging the auxiliary temporal information recorded by event cameras holds great promise for the task of video super-resolution (VSR). However, existing event-guided VSR methods assume that the event and RGB cameras are strictly calibrated (*e.g.*, pixel-level sensor designs in DAVIS 240/346). This assumption proves limiting in emerging high-resolution devices, such as dual-lens smartphones and unmanned aerial vehicles, where such precise calibration is typically unavailable. To unlock more event-guided application scenarios, we perform the task of asymmetric event-guided VSR for the first time, and we propose an Asymmetric Event-guided VSR Network (AsEVSRN) for this new task. AsEVSRN incorporates two specialized designs for leveraging the asymmetric event stream in VSR. Firstly, the content hallucination module dynamically enhances event and RGB information by exploiting their complementary nature, thereby adaptively boosting representational capacity. Secondly, the event-enhanced bidirectional recurrent cells align and propagate temporal features fused with features from content-hallucinated frames. Within the bidirectional recurrent cells, event-enhanced flow is employed to simultaneously utilize and fuse temporal information at both the feature and pixel levels. Comprehensive experimental results affirm that our method consistently generates superior quantitative and qualitative results. The code is publicly available at: https://github.com/zeyuxiao1997/AsEVSRN.

*Zeyu and Dachun contribute equally to this work.
†Corresponding author.

MM '24, October 28-November 1, 2024, Melbourne, VIC, Australia

ACM ISBN 979-8-4007-0686-8/24/10
https://doi.org/10.1145/3664647.3681357

CCS Concepts

• **Computing methodologies** → **Reconstruction**.

Keywords

Video Super-Resolution, Event Camera, Stereo Images

ACM Reference Format:
Zeyu Xiao, Dachun Kai, Yueyi Zhang, Xiaoyan Sun, and Zhiwei Xiong. 2024. Asymmetric Event-Guided Video Super-Resolution. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24), October 28-November 1, 2024, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3664647.3681357

1 Introduction

High-resolution (HR) videos are attracting increasing attention in both academia and industry, and have been already widely used in modern society, especially for the multimedia field. Video super-resolution (VSR) stands as a foundational task within the domains of computer vision to generate HR videos. The primary goal of VSR is to enhance visual quality by reconstructing an HR video from a low-resolution (LR) observation. VSR has garnered substantial attention and popularity due to its diverse applications, encompassing areas such as video surveillance [1, 77], high-definition television [12], and satellite imagery [8, 38, 62, 64]. In contrast to single-image super-resolution, which primarily addresses spatial dimensions, VSR uniquely exploits both spatial and temporal dependencies. Advanced VSR methods focus on harnessing temporal information through various techniques such as sliding windows [16, 24, 29, 30, 37, 57, 60, 74] and recurrent structures [5–7, 27, 28, 48, 73]. Recently, with the rapid development of Transformers in computer vision, several attempts have been made to exploit Transformers for better recovering missing details in LR sequences [31, 34, 43, 51]. However, effectively modeling and harnessing temporal relationships continues to pose an open and formidable challenge in the task of VSR.

Event cameras represent an innovative class of bio-inspired sensors capable of asynchronously detecting intensity changes in individual pixels at the microsecond level [47]. These sensors can

generate asynchronous event data, amounting to millions of events per second while maintaining robustness in HDR lighting conditions. Therefore, recent event-guided VSR methods have been proposed [19, 21, 22, 36, 67] to leverage the advantages of the events for VSR, comparing favorably to RGB-only methods. In practice, however, these event-guided VSR methods assume *strict alignment between images and events*. To substantiate this assumption, [19] leverages the CED dataset [49], which comprises aligned images and events sourced from DAVIS346. Similarly, [36] introduces the well-aligned ALPIX-VSR dataset for event-guided VSR.

In practice, however, in our daily imaging systems, especially in edge devices like dual-lens (or more) smartphones and unmanned aerial vehicles [11], deploying strictly aligned event and RGB cameras is challenging, let alone leveraging data collected from aligned cameras for downstream tasks. It is crucial to develop an asymmetric VSR algorithm based on asymmetric stereo events and RGB cameras to address this issue. However, utilizing information from different modalities and dealing with non-aligned data pose significant challenges for this task.

In this paper, to solve this new task, we propose the Asymmetric Event-guided VSR Network (AsEVSRN). AsEVSRN is the first end-to-end learning-based network that can generally be applied to super-resolve an LR video using an asymmetric event camera. Our proposed AsEVSRN introduces two key components to leverage asymmetric event streams for VSR. Firstly, our proposed content hallucination (CH) module dynamically enhances both event and RGB information by exploiting their complementary characteristics, thereby adaptively boosting representational capacity. Specifically, we adopt a dual-branch architecture to fuse event and RGB information adaptively and employ a dynamic convolution for dynamic enhancement of representational capacity. Secondly, drawing inspiration from successful practices in existing VSR methods [6], we design the event-enhanced bidirectional recurrent cells. The event-enhanced bidirectional recurrent cells align and propagate temporal features, integrating them with features extracted from content-hallucinated frames. Due to misalignment between the event stream and RGB frames, direct utilization for temporal fusion and propagation is not feasible. Therefore, our proposed event-enhanced bidirectional recurrent cells first pre-align event information with RGB views using a deformable convolution, enabling simultaneous utilization and fusion of temporal information at both the feature and pixel levels. We conduct experiments using event-RGB stereo data. Through extensive experimentation, we have quantitatively and qualitatively demonstrated the effectiveness of AsEVSRN.

In summary, our contributions can be summarized as follows:

(1) We propose AsEVSRN for super-resolving RGB frames with the guidance of the asymmetric event data. To the best of our knowledge, this is the first event-guided VSR method based on asymmetric event and RGB cameras.

(2) We propose the CH module dynamically to enhance event and RGB information by leveraging their complementary nature, thereby adaptively boosting representational capacity.

(3) We propose the event-enhanced bidirectional recurrent cells to align and propagate temporal features fused with features from content-hallucinated frames. Within these recurrent cells, event-enhanced flow facilitates simultaneous utilization and fusion of temporal information at both feature and pixel levels.

(4) Extensive experiments demonstrate that our proposed AsEVSRN is superior to the existing advanced potential methods.

2 Related Work

Video Super-Resolution. Existing RGB-only VSR methods enhance LR frames using temporal information via sliding windows [63, 65, 66, 69] and recurrent structures [6, 7]. Sliding-window techniques like 3DSRNet [24], TDAN [57], and EDVR [60] recover HR frames by predicting dynamic offsets and sampling convolution kernels from adjacent LR frames. They employ methods such as 3D convolution [24], optical flow estimation [25, 55], and deformable convolution [57, 60] to align temporal features. However, capturing long-range temporal features remains challenging for these approaches. To address this, recurrent structures-based methods [6, 13, 15, 55, 73] have been developed to model long-range temporal dependencies by utilizing hidden states to connect video frames. For example, BasicVSR [6] uses a bidirectional recurrent structure that merges forward and backward propagation features, resulting in significant performance gains. Vision Transformer-based methods [3, 34, 51, 75] have also achieved remarkable success in VSR. This paper focuses on event-guided VSR, which integrates event cameras to enhance VSR performance.

Event-Guided VSR. Event-guided VSR emerges as a pivotal application, leveraging the high-frame-rate motion details captured by event cameras. In event-guided VSR, consecutive frames and event data are utilized to generate HR frames. Various approaches have emerged in this domain. Jing *et al.* [19] propose a two-stage method that leverages events to interpolate the LR video, resulting in a high-frequency video that is then used to reconstruct HR frames. Kai *et al.* [22] introduce a bidirectional VSR framework, which harnesses nonlinear motion information from events to aid temporal alignment and incorporates a bidirectional cross-modal synthesis module to enhance the model's robustness to lighting variations. Lu *et al.* [36] present a joint framework that learns implicit neural representations from both RGB frames and events, enabling arbitrary-scale VSR. However, these methods presume strict alignment between the event stream and RGB images, posing practical challenges in real-world applications. For instance, in edge devices like dual-lens smartphones and unmanned aerial vehicles [11], acquiring strictly aligned event and RGB cameras can be challenging. To tackle this issue, we propose the first framework for asymmetric event-guided VSR.

Event-Guided Video Restoration. Event cameras have the unique capability to measure intensity changes at each pixel independently with microsecond accuracy, making them valuable for various video restoration tasks. One of the notable advantages of event cameras is their ability to provide motion information within the exposure time, which serves as a natural motion cue for deblurring [18, 50, 52, 53, 76]. In the context of video frame interpolation, the integration of event cameras has sparked innovations, such as TimeLens [59]. Subsequently, there has been a growing emphasis on designing interaction modules facilitating the exchange of information between event data and RGB frames, ultimately enhancing the performance of event-based video frame interpolation [42, 58, 68]. Additionally, event cameras have proven useful

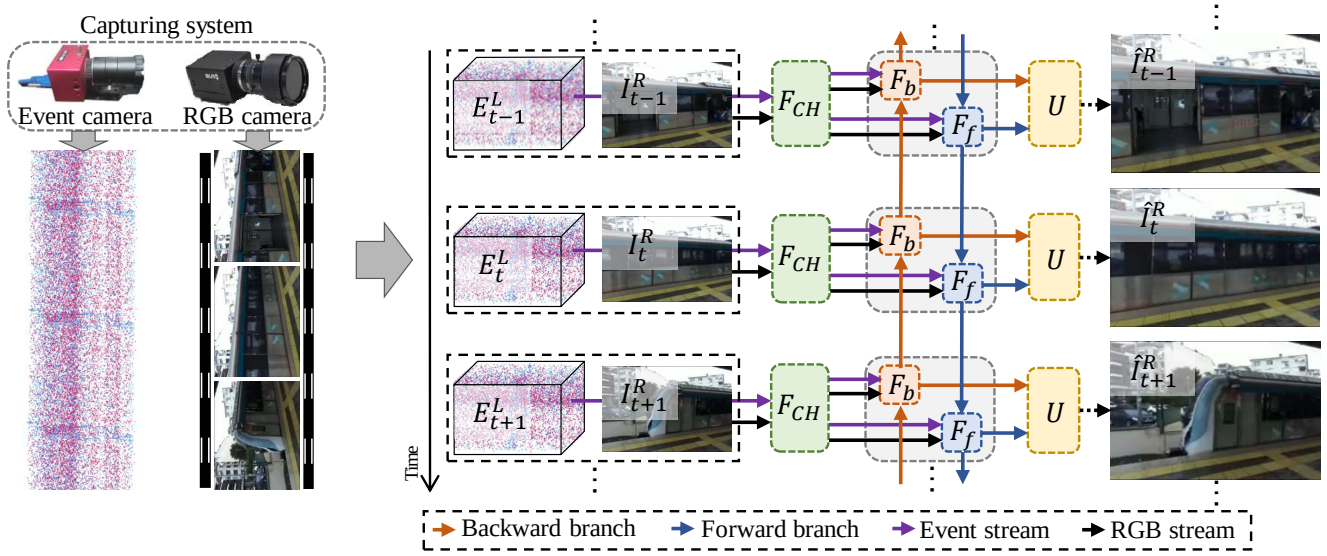

Figure 1: Left: Illustration of the asymmetric Event and RGB cameras System. Right: Overview of the proposed AsEVSRN. The information from the left event stream and the right RGB frame stream are fed to content hallucination module, aiming to highlight valuable information while mitigating interference from misaligned data from different modalities. Then the hallucinated event and image features are fed to bidirectional recurrent cells for further alignment and aggregation. Finally, the bidirectional features are fed to the upsampling module to generate the super-resolved results. For simplicity, the upsampling residual operations are omitted in the figure.

in correcting rolling shutter artifacts in consecutive global shutter frames [10, 35, 61, 78]. They have also been applied to tasks such as high-dynamic-range imaging [40, 46, 72, 80], deraining [53], and low-illumination enhancement [17], showcasing their versatility across various domains. In this paper, to the best of our knowledge, we propose the first event-guided VSR method based on asymmetric event and RGB cameras.

3 Method

3.1 Overview

Given a multi-modal stereo camera capturing system, where the left camera is an event camera and the right one is a normal LR RGB camera, our goal is to reconstruct consecutive HR clear right frames $\hat{\mathcal{I}}^R = \{\dots, \hat{I}^R_{t-1}, \hat{I}^R_t, \hat{I}^R_{t+1}, \dots\}$ ($\hat{\mathcal{I}}^R \in \mathbb{R}^{T\times sH\times sW\times 3}$) using the captured right LR frames $\mathcal{I}^R = \{\dots, I^R_{t-1}, I^R_t, I^R_{t+1}, \dots\}$ ($\mathcal{I}^R \in \mathbb{R}^{T\times H\times W\times 3}$) and the corresponding left event streams $\mathcal{E}^L = \{\dots, E^L_{t-1}, E^L_t, E^L_{t+1}, \dots\}$ triggered within T. E^L_t denotes the left event stream at the time stamp t. $\hat{\mathcal{I}}^R$ should be close to the ground truth $\mathcal{I}^{R,GT} = \{\dots, I^{R,GT}_{t-1}, I^{R,GT}_t, I^{R,GT}_{t+1}, \dots\}$ ($\mathcal{I}^{R,GT} \in \mathbb{R}^{T\times sH\times sW\times 3}$). T, H, and W are the frame number, height, and width, respectively. s is the upscaling factor and $T = \{\dots, t-1, t, t+1, \dots\}$. Note that, the camera system can consist of an event camera on the left and an RGB camera on the right, or vice versa with the RGB camera on the left and the event camera on the right.

Due to the fact that the event streams are not convenient for observation and processing by convolutional neural networks because of their sparse, irregular and unstructured properties, we convert event streams $\mathcal{E}^L$ into voxel grids $\mathcal{V}^L \in \mathbb{R}^{T\times H\times W\times B}$ using temporal bilinear transformation, suitable for convolutional neural networks [36, 68]

$$\mathcal{V}(k) = \sum_i p_i \max(0, 1 - |k - \frac{t_i - t_0}{t_{N_e} - t_0}(B-1)|), \tag{1}$$

where t_0 and t_{N_e} denote the start time and the end time of the event stream, and N_e denotes the number of the event data. t_k is the event firing timestamp, and p_k is the polarity indicating the sign of illumination changes, respectively. The index k spans from 0 to $B-1$, with B set as 5 in our experiments.

Figure 1 shows an overview of the proposed AsEVSRN. AsEVSRN employs bidirectional recurrent cells F_f and F_b akin to the scheme proposed in [6]. However, it introduces novel elements such as extra inputs and specialized modules to harness event streams, setting it apart from prior approaches. The left event streams and the right LR frames are first converted into the feature domain using the feature encoders (f^V_{En} and f^I_{En}), following which they are directed into the CH module, denoted as $f_{CH}(\cdot)$. This step aims to accentuate valuable information while simultaneously mitigating interference originating from misaligned data across distinct modalities. The above process can be denoted as

$$\boldsymbol{F}^E = f^V_{En}(\mathcal{V}^L), \boldsymbol{F}^I = f^I_{En}(\mathcal{I}^R), \tag{2}$$

$$\boldsymbol{F}^{E,CH} = f_{CH}(\boldsymbol{F}^E, \boldsymbol{F}^I), \boldsymbol{F}^{I,CH} = f_{CH}(\boldsymbol{F}^I, \boldsymbol{F}^E), \tag{3}$$

Then the hallucinated event and image features ($\boldsymbol{F}^{E,CH}$ and $\boldsymbol{F}^{I,CH}$) are fed to recurrent cells, and for a time step t, each recurrent cell F_f or F_b not only takes the hallucinated event and image features at the current time step, but also the corresponding features (h^f_{t-1} and h^b_{t+1}) propagated from its neighbors. Moreover, each recurrent

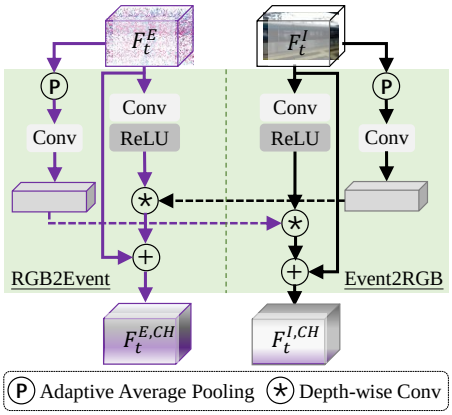

Figure 2: The structure of the content hallucination module.

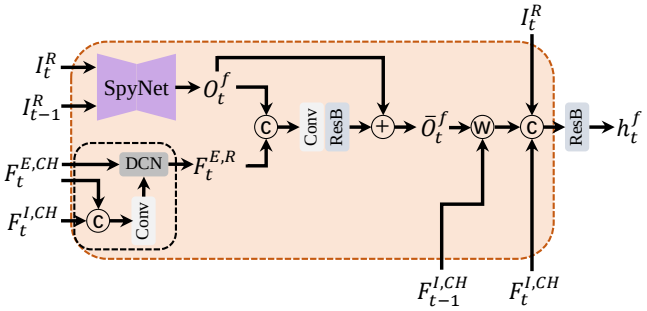

Figure 3: The structure of the event-enhanced forward recurrent cell. The event-enhanced backward recurrent cell one can be obtained in a similar way.

cell propagates the resulting features $h_t^{\{f,b\}}$ to the next cell. The above process can be denoted as

$$\begin{aligned} h_t^f &= F_f(F_t^{E,CH},\, F_t^{I,CH},\, h_{t-1}^f), \\ h_t^b &= F_b(F_t^{E,CH},\, F_t^{I,CH},\, h_{t+1}^b). \end{aligned} \tag{4}$$

To generate a super-resolved output $\hat{I}_t^R$ at timestamp t, the upsampling module U incorporates multiple convolutional layers along with pixel-shuffle operations. This module takes intermediate features $h_t^{f,b}$ and the LR frame I_t^R as inputs, yielding the final super-resolved frame $\hat{I}_t^R$

$$\hat{I}_t^R = U([h_t^f, h_t^b, I_t^R]) + (I_t^R)\uparrow_s, \tag{5}$$

where $[\cdot,\cdot]$ is the concatenation operation, and $(\cdot)\uparrow_s$ is the bilinear upsampling operation with a scaling factor of s.

3.2 Content Hallucination

In scenarios involving LR scenes, both event streams and LR images inherently contain noise in the form of missing details and artifacts, respectively. To tackle this challenge and leverage the complementary nature of information across modalities, we propose the CH module (see Figure 2). Specifically, the CH module adopts a dual-branch structure (*i.e.*, RGB2Event, and Event2RGB branches), enabling the simultaneous hallucination of representations for two modal features.

Given the left event feature F_t^E and the right frame feature F_t^I, the proposed CH module initially employs an adaptive estimation process to derive dynamic filters of high-level contextual information independently for each modality branch. Subsequently, these dynamic filters are utilized to enhance the features of the corresponding modality, facilitating cross-modal feature refinement and integration for improved representation learning. We have

$$\begin{aligned} K_t^E &= \phi_3(\mathrm{P}(F_t^E)), K_t^I = \phi_3(\mathrm{P}(F_t^I)), \\ F_t^{E,CH} &= K_t^I \circledast \mathrm{A}(\phi_3(F_t^E)) + F_t^E, \\ F_t^{I,CH} &= K_t^E \circledast \mathrm{A}(\phi_3(F_t^I)) + F_t^I, \end{aligned} \tag{6}$$

where $\phi_3(\cdot)$ denotes the convolution operation, $\mathrm{P}(\cdot)$ denotes the adaptive average pooling operation, $\mathrm{A}(\cdot)$ denotes the ReLU activation operation, and $\circledast$ is the depth-wise convolution. Using a learned dynamic filter from one modality to modulate the feature representation of another, the proposed CH module enhances valuable information while mitigating interference. This module facilitates the refinement and integration of features across modalities, thereby promoting more effective representation learning in the task of event-guided VSR.

3.3 Bidirectional Recurrent Cells

In VSR, exploiting temporal information is crucial, particularly in asymmetric event-guided VSR scenarios. The bidirectional propagation scheme has been widely acknowledged for its effectiveness [6, 7, 54]; therefore, we adopt it in AsEVSRN. Specifically, we draw inspiration from the bidirectional recurrent cells employed in BasicVSR to implement the utility of bidirectional propagation.

In each recurrent cell $F_f(\cdot)$ and $F_b(\cdot)$, a flow estimation network (*e.g.*, SpyNet[44]) is typically employed to estimate the optical flow between the LR frame I_t^R at the current time step and $I_{t\pm1}^R$ at the previous or the next time steps for alignment and further processing. The optical flow estimation can be denoted as

$$O_t^f = \mathrm{SpyNet}(I_t^R, I_{t-1}^R), O_t^b = \mathrm{SpyNet}(I_t^R, I_{t+1}^R). \tag{7}$$

However, under the practical setting of asymmetric event-guided VSR, the conventional approach described above fails to fully exploit the asymmetric information from events. Considering that the event information can effectively assist in aligning and fusing multiple frames [9, 18, 33, 42, 59], we propose the event-enhanced bidirectional recurrent cells.

Although the event stream contains crucial temporal information and cues for alignment, asymmetric events may adversely affect the alignment of RGB frames [4, 26, 27]. Therefore, we first employ deformable convolution to align the event stream with the RGB frames in feature space. Specifically, taking the forward recurrent cell as an example (see Figure 3), the content-hallucinated features $F_t^{E,CH}$ and $F_t^{I,CH}$ are first concatenated, followed by a convolutional layer to generate the RGB-aware offset map ΔP_t^f. ΔP_t^f and $F_t^{E,CH}$ are then fed to the deformable convolution layer, resulting in $F_t^{E,R}$. This procedure can be denoted as

$$\Delta P_t^f = \mathrm{Conv}([F_t^{I,CH}, F_t^{E,CH}]), \tag{8}$$

$$F_t^{E,R} = \mathrm{DConv}(F_t^{E,CH}, \Delta P_t^f), \tag{9}$$

Table 1: Quantitative comparison on the KITTI 2012 and KITTI 2015 datasets for 4× asymmetric event-guided VSR in terms of PSNR and SSIM. The best results are marked in bold, the second ones are marked with underlines, and the third ones are marked with wavy lines. The number of parameters (M) and runtime (ms) are calculated using an NVIDIA GTX 1080Ti GPU for 4× asymmetric event-guided VSR (spatial resolution: $48 \times 48 \rightarrow 192 \times 192$).

Method		#Params (M)	Runtime (ms)	KITTI 2012		KITTI 2015	
				PSNR	SSIM	PSNR	SSIM
SISR	Bicubic	-	-	25.36	0.7530	25.76	0.7613
	SwinIR	11.504	240.99	29.14	0.8618	29.98	0.8700
	SRFormer	10.396	291.91	29.19	0.8623	30.00	0.8697
VSR	VSRNet	0.439	37.63	27.65	0.8253	28.18	0.8308
	DUF	5.822	78.42	29.05	0.8585	29.75	0.8642
	TOF	1.406	104.10	28.74	0.8512	28.07	0.8286
	EDVR	20.699	71.69	30.27	0.8888	30.80	0.8869
	BasicVSR	6.291	21.31	30.69	0.8970	31.17	0.8938
	BasicVSR++	7.323	20.99	30.83	0.8992	31.30	0.8961
	TTVSR	3.450	19.32	30.74	0.8983	31.11	0.8935
	PSRT	13.367	187.09	30.48	0.8955	31.02	0.8936
	IART	13.411	193.97	30.52	0.8980	31.01	0.8951
Event-guided VSR	EGVSR	2.574	44.91	30.41	0.8946	30.90	0.8971
	EBVSR	12.151	46.76	30.89	0.8997	31.31	0.8965
	AsEVSRN (Ours)	9.648	49.37	**31.17**	**0.9052**	**31.91**	**0.9048**

where $[\cdot,\cdot]$, $\mathrm{Conv}(\cdot)$ and $\mathrm{DConv}(\cdot)$ denotes the concatenation operation, the convolution layer and the deformable convolution layer, respectively. The aligned event feature $F_t^{E,R}$ is then employed for flow refinement. On one hand, it is used to mitigate the influence of low-resolution frames on flow estimation. On the other hand, the motion information from events is utilized to further optimize the optical flow O_t^f. To obtain refined optical flow, we directly concatenate $F_t^{E,R}$ and O_t^f and feed them into a convolutional layer and a residual block. We then utilize residual connections to obtain $\overline{O}_t^f$

$$\overline{O}_t^f = \mathrm{ResB}(\mathrm{Conv}([F_t^{E,R}, O_t^f])) + O_t^f, \tag{10}$$

where $ResB(\cdot)$ is the residual block.

To obtain the temporally aggregated feature h_t^f, we utilize the refined optical flow to warp the RGB feature through a warping operation. Then, we concatenate the warped result with I_t^R and feed it into a residual block. This design effectively leverages information from both the feature domain and the pixel domain [68], enhancing the recurrent cell's representation capability. Formally, we have

$$h_t^f = \mathrm{ResB}([I_t^R, F_t^{I,CH}, \mathrm{Warp}(F_{t-1}^{I,CH}, \overline{O}_t^f)]), \tag{11}$$

where $\mathrm{Warp}(\cdot,\cdot)$ denotes the warping operation.

The event-enhanced backward recurrent cell can be obtained using a similar method to obtain h_t^b.

4 Experiments

4.1 Experimental Settings

Datasets. We train and evaluate our proposed AsEVSRN on KITTI 2012 [2] and KITTI 2015 [39] datasets. KITTI 2012 is a real-world dataset with street views from a driving car. It consists of 194 training stereo video clips and 195 testing clips, each with a resolution of 1242×375 pixels and a total of 21 frames per clip. KITTI 2015 is also a real-world dataset that shares the same shooting conditions as KITTI 2012 but with higher quality. It contains 200 training stereo video clips and 200 testing stereo video clips. The resolution and frame number are the same as KITTI 2012. Without loss of generality, we transform the left view of KITTI 2012 and KITTI 2015 into event data, while keeping the right view unchanged. This creates asymmetric event-RGB inputs. We first utilize the pre-trained RIFE interpolation model [14] to generate additional left-view frames at a 4× higher frame rate. Then we use the event camera simulator ESIM [45], to simulate events from the interpolated high-frame-rate left videos.

Training Settings. During the training stage, we follow the division of the training sets of the KITTI 2012 and KITTI 2015 datasets. We utilize bicubic downsampling by a factor of 4 on the left and right view video frames to obtain LR frames. In other words, we set $s = 4$. The proposed AsEVSRN aims to learn the mapping relationship from low-resolution frames to high-resolution frames. Given the ground-truth frame $\mathcal{I}^{R,GT}$ and the super-resolved results $\hat{\mathcal{I}}^R$ generated by our proposed AsEVSRN, we adopt the simple but effective Charbonnier loss [60] to train it from scratch, which can be described as:

$$\mathcal{L} = \sqrt{\left\| \mathcal{I}^{R,GT} - \hat{\mathcal{I}}^R \right\|^2 + \varepsilon^2}, \tag{12}$$

where ε is set to $1e-6$ in our experiments. Following previous works, we use a pre-trained SpyNet to estimate optical flow in the event-enhanced bidirectional recurrent cells. We utilize the Adam optimizer with parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$, and we utilize the Cosine Annealing scheduler for optimization. Each mini-batch consists of 6 samples. The input patch size is set to 64×64. Experiments are conducted using PyTorch [85] on two NVIDIA

3090 GPUs. We fix the weights of the pre-trained SpyNet in the first 5K iterations, and the total number of iterations is 300K.

Inference Settings. During the testing stage, we follow the division of the training sets of the KITTI 2012 and KITTI 2015 datasets. To quantitatively evaluate the reconstructed HR videos, we choose Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) on Y channel as metrics. The temporal consistency can be analyzed by evaluating the estimated optical flow of the reconstructed HR videos and the extracted temporal profiles.

4.2 Quantitative and Qualitative Comparisons

We compare the proposed AsEVSRN with a wide range of potential methods that could be used to address asymmetric event-guided VSR, aiming to explore as many diverse and rich approaches as possible. (1) Single image SR (SISR) methods: Bicubic, SwinIR [32], and SRFormer [79]. Specifically, we process each LR frame sequentially through the SISR network for reconstruction, and then assemble the reconstructed frames to form an HR video. (2) VSR methods: VSRNet [23], DUF [20], TOF [71], EDVR [60], BasicVSR [6], BasicVSR++ [7], PSRT [51], TTVSR [34], and IART [70]. In particular, we exclude the event stream and solely feed the LR video frames into the VSR network for reconstruction, resulting in the final reconstructed video output. (3) Event-guided VSR methods: as this direction is relatively less explored, we compare our AsEVSRN with EGVSR and EBVSR to provide a thorough evaluation. It is important to note that for fair comparison, we retrain all these methods on the KITTI 2012 and KITTI 2015 datasets using their publicly released codes. We refrain from using fine-tuning or directly utilizing pre-trained models on Vimeo90K [71] or REDS [41].

Quantitative Results. Table 1 shows the quantitative comparisons on various testsets in terms of PSNR and SSIM. From the table, we can draw several conclusions. Firstly, VSR methods generally outperform SISR methods in terms of PSNR and SSIM. This indicates that the temporal information provided by video sequences helps in achieving better reconstruction quality compared to SISR methods. For instance, BasicVSR and its variants consistently outperform SwinIR, which is a representative leading SISR method. Secondly, event-guided VSR shows potential for achieving higher performance compared to traditional VSR methods. For instance, methods like EBVSR, which incorporates event information, exhibit higher PSNR and SSIM values than some traditional VSR methods like BasicVSR++ and TTVSR. For example, for EBVSR compared to TTVSR on the KITTI 2012 dataset, there is a PSNR increase of 0.15 dB and a SSIM increase of 0.0014. On the KITTI 2015 dataset, EBVSR shows a PSNR increase of 0.20 dB and an SSIM increase of 0.0030 compared to TTVSR. Lastly, our proposed AsEVSRN demonstrates superior performance compared to both traditional VSR methods and other event-guided VSR methods. Specifically, AsEVSRN achieves the highest PSNR and SSIM values among all methods evaluated on both KITTI 2012 and KITTI 2015 datasets. For instance, AsEVSRN achieves a PSNR of 31.17 dB on KITTI 2012, outperforming all other baseline methods.

Computational Efficiency. We compare AsEVSRN to other methods in terms of the number of parameters and the runtime. Results are listed in Table 1. Comparing the number of parameters and runtime among different methods, we observe that AsEVSRN achieves competitive performance with fewer parameters and comparable runtime. Specifically, AsEVSRN has 9.648M parameters and a runtime of 49.37 ms, while EBVSR, which has a similar performance, requires 12.151M parameters and a runtime of 46.76 ms. Despite having fewer parameters, AsEVSRN achieves higher PSNR and SSIM scores compared to EBVSR. This indicates that AsEVSRN effectively utilizes parameter efficiency to improve reconstruction quality, demonstrating its superiority in terms of performance-complexity trade-off. Therefore, even under similar computational constraints, AsEVSRN outperforms other methods in terms of PSNR and SSIM on both KITTI 2012 and KITTI 2015 datasets.

Qualitative Results. In Figure 4, we present visual comparisons between the results obtained by our proposed AsEVSRN and those of other competing baselines on the KITTI 2012 dataset. It is evident from these visual comparisons that our proposed AsEVSRN method outperforms the baseline methods, yielding superior qualitative results characterized by more accurate details and significantly reduced blurring artifacts. For instance, in the restoration of fine texture details, our AsEVSRN excels in reconstructing the intricate brick patterns on the rooftops, while artifacts are visibly present in the results produced by other methods. Furthermore, AsEVSRN demonstrates a clearer restoration of textural details compared to other methods, which often generate blurry results. This is particularly noticeable in regions with complex textures and high-frequency details, where our method successfully recovers fine structures without introducing unwanted noise or smoothing effects. These qualitative improvements underscore the efficacy of our proposed architecture in enhancing the visual quality of super-resolved video content.

Temporal Consistency. To evaluate the temporal consistency of the super-resolved video clips, we estimate the optical flow on the KITTI 2012 dataset using the advanced RAFT algorithm [56]. Specifically, we compute the optical flow between consecutive frames to assess the motion coherence of the reconstructed sequences. As illustrated in Figure 5, the optical flow estimated from the results produced by our proposed AsEVSRN is remarkably close to that obtained from the original HR frames. This observation substantiates the superior temporal consistency and motion accuracy of our method, demonstrating its effectiveness in generating high-quality, temporally coherent super-resolved video content.

4.3 Ablation Studies

To analyze the effectiveness of the proposed AsEVSRN, we conduct the following experiments on KITTI 2012.

Effectiveness of Two Modules in AsEVSRN. The CH module and the event-enhanced bidirectional recurrent cells are two core modules in AsEVSRN. We analyze the performance of each component by removing different modules and replacing them with residual blocks of equivalent parameter amount. Table 2 presents the effectiveness of the CH module and the event-enhanced bidirectional recurrent cells in terms of PSNR and SSIM. Each method is evaluated with different combinations of these modules. Firstly, when excluding the CH module (AsEVSRN-w/o-F_{CH}), the PSNR is 30.72 dB and SSIM

Figure 4: Visual comparison on the 4× asymmetric event-guided VSR task. Frames are from the KITTI 2012 dataset.

Figure 5: Temporal consistency comparison on the 4× asymmetric event-guided VSR task. We show the estimated optical flow of the results from different methods using the pre-trained RAFT [56].

is 0.8975. This indicates that the absence of the CH module leads to a decrease in performance compared to the full AsEVSRN method. Secondly, omitting the backward recurrent cell (F_b) while including the CH module (AsEVSRN-w/o-F_b) results in a slight improvement in PSNR to 30.95 dB and SSIM to 0.9011 compared to the case without the CH module. Similarly, excluding the forward recurrent cell (F_f) while keeping the CH module (AsEVSRN-w/o-F_f) yields a PSNR of 30.96 dB and SSIM of 0.9014, slightly higher than the previous case. Finally, the full AsEVSRN method, incorporating both the CH module and bidirectional recurrent cells, achieves the highest PSNR of 31.17 dB and SSIM of 0.9052, demonstrating the effectiveness of both modules in enhancing the performance of the proposed method.

Investigation of the CH Module. The CH module aims at leveraging the complementary nature of information across the event and RGB modalities while enhancing the representation ability. To showcase its effectiveness, we design and analyze several variants: (1) CH-w/o-dynamicfilter: we replace the dynamic filter with a simple addition operation. (2) CH-w/o-RGB2Event: in this variant, we directly remove the RGB2Event branch. (3) CH-w/o-Event2RGB: this variant involves the direct removal of the Event2RGB branch.

Table 2: Effectiveness of the CH module and the event-enhanced bidirectional recurrent cells.

Method	F_{CH}	F_{fb}		PSNR	SSIM
		F_b	F_f		
AsEVSRN-w/o-F_{CH}	✗	✓	✓	30.72	0.8975
AsEVSRN-w/o-F_b	✓	✗	✓	30.95	0.9011
AsEVSRN-w/o-F_f	✓	✓	✗	30.96	0.9014
AsEVSRN	✓	✓	✓	31.17	0.9052

Table 3: Effectiveness of the designs in the CH module.

Method	PSNR	SSIM
CH-w/o-dynamicfilter	30.93	0.8995
CH-w/o-RGB2Event	30.99	0.9016
CH-w/o-Event2RGB	31.00	0.9020
CH Module	31.17	0.9052

Results are shown in Table 3. The results show that including both

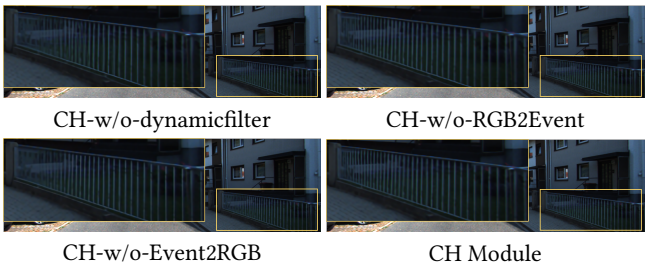

Figure 6: Visual comparison on removing different parts in the CH module. Please zoom in for better viewing.

Table 4: Effectiveness of the designs in the event-enhanced bidirectional recurrent cells.

Method	PSNR	SSIM
Cell-w/o-DCN	30.96	0.9013
Cell-w/o-$F_t^{E,R}$	31.00	0.9021
Cell-w/o-$\overline{O}_t^f$	31.04	0.9033
Cell-w/o-$F_t^{I,CH}$	31.14	0.9036
Cell-w/o-I_t^R	31.12	0.9035
Cell	31.17	0.9052

Table 5: Embedding the components of the AsEVSRN into existing baseline methods, *i.e.*, BasicVSR and EBVSR. † denotes the method with the CH module and event-enhanced bidirectional recurrent cells.

Method	PSNR	SSIM
BasicVSR	30.69	0.8970
BasicVSR†	30.92	0.9001
EBVSR	30.89	0.8997
EBVSR†	31.03	0.9029

RGB-to-Event and Event-to-RGB fusion, along with dynamic filtering in the CH module enhances performance, as evidenced by the highest PSNR and SSIM values compared to the configurations with individual components excluded. In Figure 6, we also visualize the results after removing different parts in the CH module. The removal of dynamic convolution resulted in an overall deterioration in performance. Furthermore, eliminating the RGB2event branch lead to a degradation in the reconstructed details, while removing the event2RGB branch caused the results to become blurrier. These observations align with the findings presented in Table 3.

Investigation of the Event-Enhanced Bidirectional Recurrent Cells. The event-enhanced bidirectional recurrent cells aim at leveraging the event information for RGB feature fusion and propagation. To showcase its effectiveness, we design and analyze several variants: (1) Cell-w/o-DCN: in this variant, we directly remove the G^2DT module. (2) Cell-w/o-$F_t^{E,R}$: we feed the content-hallucinated event feature to the following parts directly. (3) Cell-w/o-$\overline{O}_t^f$: we utilize the optical flow estimated by SpyNet directly. (4) Cell-w/o-$F_t^{I,CH}$: we perform the warping operation at the pixel level. (5) Cell-w/o-I_t^R: we perform the warping operation at the feature level. Table 4 presents the effectiveness of different designs in the event-enhanced bidirectional recurrent cells based on PSNR and SSIM metrics. The results show that each design variation contributes to improving performance, with the complete cell achieving the highest PSNR of 31.17 dB and SSIM of 0.9052 compared to its variants without specific components.

Embedding the Components of the AsEVSRN into Existing Baseline Methods. Table 5 presents the results after integrating two important components into BasicVSR and EBVSR. We observe that incorporating these components leads to improvements in both PSNR and SSIM metrics for both methods. Specifically, BasicVSR with the integrated components achieves a PSNR of 30.92 dB and SSIM of 0.9001, while EBVSR with the integrated components achieves a PSNR of 31.03 dB and SSIM of 0.9029. These results indicate that the integration of the components enhances the performance of both BasicVSR and EBVSR methods, demonstrating the effectiveness of the proposed components.

4.4 Limitation

While our AsEVSRN demonstrates promising results, there are still challenges that need to be addressed. For instance, we encounter difficulties in reconstructing small objects, as elaborated in the supplementary material. These small objects often require finer detail recovery, which can be challenging due to the limited amount of information available at lower resolutions.

5 Conclusion

In this paper, we address the challenge of performing asymmetric event-guided VSR for the first time, introducing AsEVSRN tailored specifically for this novel task. The AsEVSRN leverages two specialized designs: a content hallucination module that dynamically enhances event and RGB information, boosting representational capacity, and event-enhanced bidirectional recurrent cells that align and propagate temporal features fused with content-hallucinated frames. These cells employ event-enhanced flow for the simultaneous utilization and fusion of temporal information at both the feature and pixel levels. The AsEVSRN consistently generates superior results both quantitatively and qualitatively.

Acknowledgments

We acknowledge funding from National Natural Science Foundation of China under Grants 62131003, 62021001 and 61901435.

References

[1] Andreas Aakerberg, Kamal Nasrollahi, and Thomas B Moeslund. 2022. Real-world super-resolution of face-images from surveillance cameras. *IET Image Processing* 16, 2 (2022), 442–452.
[2] Philip Lenz Andreas Geiger and Raquel Urtasun. 2012. Are we ready for autonomous driving? The KITTI vision benchmark suite. In *CVPR*.
[3] Jiezhang Cao, Yawei Li, Kai Zhang, and Luc Van Gool. 2021. Video Super-Resolution Transformer. *arXiv preprint arXiv:2106.06847* (2021).
[4] Jiezhang Cao, Jingyun Liang, Kai Zhang, Yawei Li, Yulun Zhang, Wenguan Wang, and Luc Van Gool. 2022. Reference-based Image Super-Resolution with Deformable Attention Transformer. In *ECCV*.
[5] Jiezhang Cao, Jingyun Liang, Kai Zhang, Wenguan Wang, Qin Wang, Yulun Zhang, Hao Tang, and Luc Van Gool. 2022. Towards interpretable video super-resolution via alternating optimization. In *ECCV*. 393–411.

[6] Kelvin CK Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. 2021. Basicvsr: The search for essential components in video super-resolution and beyond. In *CVPR*. 4947–4956.

[7] Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. 2022. BasicVSR++: Improving video super-resolution with enhanced propagation and alignment. In *CVPR*. 5972–5981.

[8] Michel Deudon, Alfredo Kalaitzis, Israel Goytom, Md Rifat Arefin, Zhichao Lin, Kris Sankaran, Vincent Michalski, Samira E Kahou, Julien Cornebise, and Yoshua Bengio. 2020. Highres-net: Recursive fusion for multi-frame super-resolution of satellite imagery. *arXiv preprint arXiv:2002.06460* (2020).

[9] Xin Ding, Tsuyoshi Takatani, Zhongyuan Wang, Ying Fu, and Yinqiang Zheng. 2022. Event-guided Video Clip Generation from Blurry Images. In *ACM MM*. 2672–2680.

[10] Julius Erbach, Stepan Tulyakov, Patricia Vitoria, Alfredo Bochicchio, and Yuanyou Li. 2023. EvShutter: Transforming Events for Unconstrained Rolling Shutter Correction. In *CVPR*. 13904–13913.

[11] Mathias Gehrig, Willem Aarents, Daniel Gehrig, and Davide Scaramuzza. 2021. Dsec: A stereo event camera dataset for driving scenarios. *IEEE Robotics and Automation Letters* 6, 3 (2021), 4947–4954.

[12] Tomio Goto, Takafumi Fukuoka, Fumiya Nagashima, Satoshi Hirano, and Masaru Sakurai. 2014. Super-resolution System for 4K-HDTV. In *ICCV*. 4453–4458.

[13] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. 2019. Recurrent back-projection network for video super-resolution. In *CVPR*. 3897–3906.

[14] Zhewei Huang, Tianyuan Zhang, Wen Heng, Boxin Shi, and Shuchang Zhou. 2022. Real-time intermediate flow estimation for video frame interpolation. In *ECCV*. 624–642.

[15] Takashi Isobe, Xu Jia, Shuhang Gu, Songjiang Li, Shengjin Wang, and Qi Tian. 2020. Video super-resolution with recurrent structure-detail network. In *ECCV*. 645–660.

[16] Takashi Isobe, Xu Jia, Xin Tao, Changlin Li, Ruihuang Li, Yongjie Shi, Jing Mu, Huchuan Lu, and Yu-Wing Tai. 2022. Look back and forth: Video super-resolution with explicit temporal difference modeling. In *CVPR*. 17411–17420.

[17] Yu Jiang, Yuehang Wang, Siqi Li, Yongji Zhang, Minghao Zhao, and Yue Gao. 2023. Event-based Low-illumination Image Enhancement. *IEEE Transactions on Multimedia* (2023).

[18] Zhe Jiang, Yu Zhang, Dongqing Zou, Jimmy Ren, Jiancheng Lv, and Yebin Liu. 2020. Learning event-based motion deblurring. In *CVPR*. 3320–3329.

[19] Yongcheng Jing, Yiding Yang, Xinchao Wang, Mingli Song, and Dacheng Tao. 2021. Turning frequency to resolution: Video super-resolution via event cameras. In *CVPR*. 7772–7781.

[20] Younghyun Jo, Seoung Wug Oh, Jaeyeon Kang, and Seon Joo Kim. 2018. Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation. In *CVPR*. 3224–3232.

[21] Dachun Kai, Jiayao Lu, Yueyi Zhang, and Xiaoyan Sun. 2024. EvTexture: Event-driven Texture Enhancement for Video Super-Resolution. In *ICML*.

[22] Dachun Kai, Yueyi Zhang, and Xiaoyan Sun. 2023. Video Super-Resolution Via Event-Driven Temporal Alignment. In *ICIP*. 2950–2954.

[23] Armin Kappeler, Seunghwan Yoo, Qiqin Dai, and Aggelos K Katsaggelos. 2016. Video super-resolution with convolutional neural networks. *IEEE transactions on computational imaging* 2, 2 (2016), 109–122.

[24] Soo Ye Kim, Jeongyeon Lim, Taeyoung Na, and Munchurl Kim. 2018. 3DSRnet: Video super-resolution using 3d convolutional neural networks. *arXiv preprint arXiv:1812.09079* (2018).

[25] Tae Hyun Kim, Mehdi SM Sajjadi, Michael Hirsch, and Bernhard Scholkopf. 2018. Spatio-temporal transformer network for video restoration. In *ECCV*. 106–122.

[26] Zeqiang Lai, Ying Fu, and Jun Zhang. 2024. Hyperspectral Image Super Resolution With Real Unaligned RGB Guidance. *IEEE Transactions on Neural Networks and Learning Systems* (2024).

[27] Junyong Lee, Myeonghee Lee, Sunghyun Cho, and Seungyong Lee. 2022. Reference-based video super-resolution using multi-camera video triplets. In *CVPR*. 17824–17833.

[28] Fei Li, Linfeng Zhang, Zikun Liu, Juan Lei, and Zhenbo Li. 2023. Multi-frequency representation enhancement with privilege information for video super-resolution. In *ICCV*. 12814–12825.

[29] Sheng Li, Fengxiang He, Bo Du, Lefei Zhang, Yonghao Xu, and Dacheng Tao. 2019. Fast spatio-temporal residual network for video super-resolution. In *CVPR*. 10522–10531.

[30] Wenyi Lian and Wenjing Lian. 2022. Sliding window recurrent network for efficient video super-resolution. In *ECCV*. 591–601.

[31] Jingyun Liang, Jiezhang Cao, Yuchen Fan, Kai Zhang, Rakesh Ranjan, Yawei Li, Radu Timofte, and Luc Van Gool. 2022. Vrt: A video restoration transformer. *arXiv preprint arXiv:2201.12288* (2022).

[32] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. 2021. Swinir: Image restoration using swin transformer. In *ICCV*. 1833–1844.

[33] Guixu Lin, Jin Han, Mingdeng Cao, Zhihang Zhong, and Yinqiang Zheng. 2023. Event-guided Frame Interpolation and Dynamic Range Expansion of Single Rolling Shutter Image. In *ACM MM*. 3078–3088.

[34] Chengxu Liu, Huan Yang, Jianlong Fu, and Xueming Qian. 2022. Learning Trajectory-Aware Transformer for Video Super-Resolution. In *CVPR*. 5687–5696.

[35] Yunfan Lu, Guoqiang Liang, and Lin Wang. 2023. Self-supervised Learning of Event-guided Video Frame Interpolation for Rolling Shutter Frames. *arXiv preprint arXiv:2306.15507* (2023).

[36] Yunfan Lu, Zipeng Wang, Minjie Liu, Hongjian Wang, and Lin Wang. 2023. Learning Spatial-Temporal Implicit Neural Representations for Event-Guided Video Super-Resolution. In *CVPR*. 1557–1567.

[37] Zhihe Lu, Zeyu Xiao, Jiawang Bai, Zhiwei Xiong, and Xinchao Wang. 2023. Can SAM Boost Video Super-Resolution? *arXiv preprint arXiv:2305.06524* (2023).

[38] Yimin Luo, Liguo Zhou, Shu Wang, and Zhongyuan Wang. 2017. Video satellite imagery super resolution via convolutional neural networks. *IEEE Geoscience and Remote Sensing Letters* 14, 12 (2017), 2398–2402.

[39] Moritz Menze and Andreas Geiger. 2015. Object scene flow for autonomous vehicles. In *CVPR*.

[40] Nico Messikommer, Stamatios Georgoulis, Daniel Gehrig, Stepan Tulyakov, Julius Erbach, Alfredo Bochicchio, Yuanyou Li, and Davide Scaramuzza. 2022. Multi-bracket high dynamic range imaging with event cameras. In *CVPR*. 547–557.

[41] Seungjun Nah, Sungyong Baik, Seokil Hong, Gyeongsik Moon, Sanghyun Son, Radu Timofte, and Kyoung Mu Lee. 2019. NTIRE 2019 challenge on video deblurring and super-resolution: Dataset and study. In *CVPRW*.

[42] Genady Paikin, Yotam Ater, Roy Shaul, and Evgeny Soloveichik. 2021. Efi-net: Video frame interpolation from fusion of events and frames. In *CVPR*. 1291–1301.

[43] Zhongwei Qiu, Huan Yang, Jianlong Fu, and Dongmei Fu. 2022. Learning Spatiotemporal Frequency-Transformer for Compressed Video Super-Resolution. In *ECCV*.

[44] Anurag Ranjan and Michael J Black. 2017. Optical flow estimation using a spatial pyramid network. In *CVPR*. 4161–4170.

[45] Henri Rebecq, Daniel Gehrig, and Davide Scaramuzza. 2018. ESIM: an open event camera simulator. In *Conference on robot learning*. PMLR, 969–982.

[46] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. 2019. High speed and high dynamic range video with an event camera. *IEEE transactions on pattern analysis and machine intelligence* 43, 6 (2019), 1964–1980.

[47] Viktor Rudnev, Vladislav Golyanik, Jiayi Wang, Hans-Peter Seidel, Franziska Mueller, Mohamed Elgharib, and Christian Theobalt. 2021. EventHands: Real-Time Neural 3D Hand Pose Estimation From an Event Stream. In *ICCV*. 12385–12395.

[48] Mehdi SM Sajjadi, Raviteja Vemulapalli, and Matthew Brown. 2018. Frame-recurrent video super-resolution. In *CVPR*. 6626–6634.

[49] Cedric Scheerlinck, Henri Rebecq, Timo Stoffregen, Nick Barnes, Robert Mahony, and Davide Scaramuzza. 2019. CED: Color event camera dataset. In *CVPRW*.

[50] Wei Shang, Dongwei Ren, Dongqing Zou, Jimmy S Ren, Ping Luo, and Wangmeng Zuo. 2021. Bringing events into video deblurring with non-consecutively blurry frames. In *ICCV*. 4531–4540.

[51] Shuwei Shi, Jinjin Gu, Liangbin Xie, Xintao Wang, Yujiu Yang, and Chao Dong. 2022. Rethinking alignment in video super-resolution transformers. *NeurIPS* 35 (2022), 36081–36093.

[52] Lei Sun, Christos Sakaridis, Jingyun Liang, Qi Jiang, Kailun Yang, Peng Sun, Yaozu Ye, Kaiwei Wang, and Luc Van Gool. 2022. Event-based fusion for motion deblurring with cross-modal attention. In *ECCV*. 412–428.

[53] Lei Sun, Christos Sakaridis, Jingyun Liang, Peng Sun, Jiezhang Cao, Kai Zhang, Qi Jiang, Kaiwei Wang, and Luc Van Gool. 2023. Event-Based Frame Interpolation with Ad-hoc Deblurring. In *CVPR*. 18043–18052.

[54] Qi Tang, Yao Zhao, Meiqin Liu, Jian Jin, and Chao Yao. 2024. Semantic Lens: Instance-Centric Semantic Alignment for Video Super-resolution. In *AAAI*, Vol. 38. 5154–5161.

[55] Xin Tao, Hongyun Gao, Renjie Liao, Jue Wang, and Jiaya Jia. 2017. Detail-revealing deep video super-resolution. In *ICCV*. 4472–4480.

[56] Zachary Teed and Jia Deng. 2020. Raft: Recurrent all-pairs field transforms for optical flow. In *ECCV*.

[57] Yapeng Tian, Yulun Zhang, Yun Fu, and Chenliang Xu. 2020. Tdan: Temporally-deformable alignment network for video super-resolution. In *CVPR*. 3360–3369.

[58] Stepan Tulyakov, Alfredo Bochicchio, Daniel Gehrig, Stamatios Georgoulis, Yuanyou Li, and Davide Scaramuzza. 2022. Time lens++: Event-based frame interpolation with parametric non-linear flow and multi-scale fusion. In *CVPR*. 17755–17764.

[59] Stepan Tulyakov, Daniel Gehrig, Stamatios Georgoulis, Julius Erbach, Mathias Gehrig, Yuanyou Li, and Davide Scaramuzza. 2021. Time lens: Event-based video frame interpolation. In *CVPR*. 16155–16164.

[60] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. 2019. EDVR: Video restoration with enhanced deformable convolutional networks. In *CVPRW*.

[61] Yangguang Wang, Xiang Zhang, Mingyuan Lin, Lei Yu, Boxin Shi, Wen Yang, and Gui-Song Xia. 2023. Self-Supervised Scene Dynamic Recovery from Rolling Shutter Images and Events. *arXiv preprint arXiv:2304.06930* (2023).

[62] Yi Xiao, Xin Su, Qiangqiang Yuan, Denghong Liu, Huanfeng Shen, and Liangpei Zhang. 2021. Satellite video super-resolution via multiscale deformable convolution alignment and temporal grouping projection. *IEEE Transactions on*

Geoscience and Remote Sensing 60 (2021), 1–19.

[63] Yi Xiao, Qiangqiang Yuan, Jiang He, Qiang Zhang, Jing Sun, Xin Su, Jialian Wu, and Liangpei Zhang. 2022. Space-time super-resolution for satellite video: A joint framework based on multi-scale spatial-temporal transformer. *International Journal of Applied Earth Observation and Geoinformation* 108 (2022), 102731.

[64] Yi Xiao, Qiangqiang Yuan, Kui Jiang, Xianyu Jin, Jiang He, Liangpei Zhang, and Chia-wen Lin. 2023. Local-Global Temporal Difference Learning for Satellite Video Super-Resolution. *arXiv preprint arXiv:2304.04421* (2023).

[65] Zeyu Xiao, Zhen Cheng, and Zhiwei Xiong. 2023. Space-time super-resolution for light field videos. *IEEE Transactions on Image Processing* (2023).

[66] Zeyu Xiao, Xueyang Fu, Jie Huang, Zhen Cheng, and Zhiwei Xiong. 2021. Space-time distillation for video super-resolution. In *CVPR*. 2113–2122.

[67] Zeyu Xiao, Dachun Kai, Yueyi Zhang, Zheng-Jun Zha, Xiaoyan Sun, and Zhiwei Xiong. 2024. Event-Adapted Video Super-Resolution. In *ECCV*.

[68] Zeyu Xiao, Wenming Weng, Yueyi Zhang, and Zhiwei Xiong. 2022. EVA2: Event-Assisted Video Frame Interpolation via Cross-Modal Alignment and Aggregation. *IEEE Transactions on Computational Imaging* 8 (2022), 1145–1158.

[69] Zeyu Xiao, Zhiwei Xiong, Xueyang Fu, Dong Liu, and Zheng-Jun Zha. 2020. Space-time video super-resolution using temporal profiles. In *ACM MM*. 664–672.

[70] Kai Xu, Ziwei Yu, Xin Wang, Michael Bi Mi, and Angela Yao. 2023. An implicit alignment for video super-resolution. *arXiv preprint arXiv:2305.00163* (2023).

[71] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. 2019. Video enhancement with task-oriented flow. *International Journal of Computer Vision* 127 (2019), 1106–1125.

[72] Yixin Yang, Jin Han, Jinxiu Liang, Imari Sato, and Boxin Shi. 2023. Learning event guided high dynamic range video reconstruction. In *CVPR*. 13924–13934.

[73] Peng Yi, Zhongyuan Wang, Kui Jiang, Junjun Jiang, Tao Lu, Xin Tian, and Jiayi Ma. 2021. Omniscient video super-resolution. In *ICCV*. 4429–4438.

[74] Huanjing Yue, Zhiming Zhang, and Jingyu Yang. 2022. Real-RawVSR: Real-World Raw Video Super-Resolution with a Benchmark Dataset. In *ECCV*. 608–624.

[75] Yanhong Zeng, Huan Yang, Hongyang Chao, Jianbo Wang, and Jianlong Fu. 2021. Improving visual quality of image synthesis by a token-based generator with transformers. *NeurIPS* 34 (2021), 21125–21137.

[76] Limeng Zhang, Hongguang Zhang, Jihua Chen, and Lei Wang. 2020. Hybrid deblur net: Deep non-uniform deblurring with event camera. *IEEE Access* 8 (2020), 148075–148083.

[77] Liangpei Zhang, Hongyan Zhang, Huanfeng Shen, and Pingxiang Li. 2010. A super-resolution reconstruction algorithm for surveillance images. *Signal Processing* 90, 3 (2010), 848–859.

[78] Xinyu Zhou, Peiqi Duan, Yi Ma, and Boxin Shi. 2022. EvUnroll: Neuromorphic events based rolling shutter image correction. In *CVPR*. 17775–17784.

[79] Yupeng Zhou, Zhen Li, Chun-Le Guo, Song Bai, Ming-Ming Cheng, and Qibin Hou. 2023. Srformer: Permuted self-attention for single image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 12780–12791.

[80] Yunhao Zou, Yinqiang Zheng, Tsuyoshi Takatani, and Ying Fu. 2021. Learning to reconstruct high speed and high dynamic range videos from events. In *CVPR*. 2024–2033.