Bilateral Event Mining and Complementary for Event Stream Super-Resolution

Zhilin Huang[1,2*] Quanmin Liang[2,3*] Yijie Yu[1,2] Chujun Qin[4]
Xiawu Zheng[2,5] Kai Huang[3] Zikun Zhou[2†] Wenming Yang[1,2†]
[1] Shenzhen International Graduate School, Tsinghua University [2] Peng Cheng Laboratory
[3] School of Computer Science and Engineering, Sun Yat-Sen University
[4] China Southern Power Grid [5] Xiamen University
{zerinhwang03,chujun.qin}@pku.edu.cn, liangqm5@mail2.sysu.edu.cn, yangelwm@163.com
yyj23@mails.tsinghua.edu.cn, huangk36@mail.sysu.edu.cn, {zhengxw01,zhouzk01}@pcl.ac.cn

Abstract

*Event Stream Super-Resolution (ESR) aims to address the challenge of insufficient spatial resolution in event streams, which holds great significance for the application of event cameras in complex scenarios. Previous works for ESR often process positive and negative events in a mixed paradigm. This paradigm limits their ability to effectively model the unique characteristics of each event and mutually refine each other by considering their correlations. In this paper, we propose a bilateral event mining and complementary network (**BMCNet**) to fully leverage the potential of each event and capture the shared information to complement each other simultaneously. Specifically, we resort to a two-stream network to accomplish comprehensive mining of each type of events individually. To facilitate the exchange of information between two streams, we propose a bilateral information exchange (**BIE**) module. This module is layer-wisely embedded between two streams, enabling the effective propagation of hierarchical global information while alleviating the impact of invalid information brought by inherent characteristics of events. The experimental results demonstrate that our approach outperforms the previous state-of-the-art methods in ESR, achieving performance improvements of over **11%** on both real and synthetic datasets. Moreover, our method significantly enhances the performance of event-based downstream tasks such as object recognition and video reconstruction. Our code is available at https://github.com/Lqm26/BMCNet-ESR.*

1. Introduction

Event cameras are a novel bio-inspired asynchronous sensor [9, 23] with advantages such as high dynamic range, high temporal resolution, and low power consumption [2, 9, 23]. It holds the potential to provide solutions for a wide range of visually challenging scenarios and has already found extensive applications [1, 3, 16, 17, 22, 27, 28, 33, 35, 42, 44]. However, most existing event cameras still exhibit lower spatial resolution (e.g., DAVIS346 with a resolution of 346×260 [33]) and are subject to sensor-induced noise. Increasing the spatial resolution at the hardware level inevitably leads to event loss and increased pixel response latency [11, 19]. Therefore, there is a need to achieve super-resolution and enhancement of event streams, enabling better adaptability to various complex scenes.

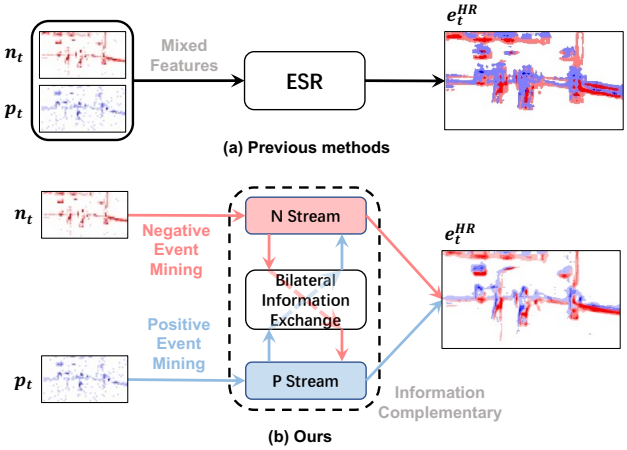

Figure 1. Illustration of various approaches in event stream super resolution for processing positive and negative events.

To address this issue, several methods have been proposed for Event Stream Super-Resolution (ESR). One approach involves preserving the spatiotemporal distribution information of event streams, using non-uniform Poisson processes or Spiking Neural Networks (SNNs) to simultaneously estimate the spatiotemporal distribution of event streams during ESR [18, 19]. Alternatively, it is possi-

*Equal Contribution
†Corresponding Author

ble to employ image frames as auxiliary data, aligning the event stream's resolution with high-resolution images through optical flow estimation [39]. However, these approaches face challenges in accurately estimating the spatiotemporal distribution of the event stream, requiring high-quality images as assistance, and struggling to achieve high-factor resolution enhancement. Another approach is to apply learning-based methods for super-resolution on event streams [7, 41]. By unifying the data format of events with natural images, these methods can seamlessly leverage the experiences and techniques of well-developed image and video super-resolution methods. However, all learning-based methods [7, 41] for ESR ignored explicitly exploring the correlations between two types of events by directly process two different types of events in a mixed way, as shown in Figure 1 (a). Though simple, these approaches require elaborate designs and high capacities in the models, as they need to not only effectively model the distinctive data distribution of each event by addressing inherent characteristics, but also capture the complementary information between two events simultaneously.

In this paper, we propose **three principles** for the model design of ESR based on inherent characteristics of event data: (i). Positive and negative events, which are obtained by decoupling the event stream, have unique distribution characteristics. However, they also demonstrate strong correlations in both spatial and temporal domain, allowing them to complement and enhance each other. Therefore, it is crucial for ESR models to effectively mine the information from each type of events and facilitate their interaction. (ii). Events are always triggered near object edges, resulting in event data that primarily consists of global structures. Therefore, the models are required to have the capability to effectively capture and interact global structures between different events. (iii). Due to the sparsity and noise inherent in events, it is essential to avoid misleading resulting from invalid information during context modeling of event data.

According to these three principles for ESR, we propose a bilateral event mining and complementary network (BMCNET). In addition to effectively modeling the unique characteristics of each event, BMCNET also possesses the capability to efficiently interact and complement global information between different events. Specifically, BMCNET consists of two parallel stream for processing positive and negative events, respectively. And a novel bilateral information exchange (BIE) module we proposed is applied to facilitate the information exchange between two streams. With a single BIE, each channel is treated as a structural representation, and the correlations between two events are modeled across the channel dimension instead of the spatial ones. In this way, the global structures of each event can be efficiently captured and transmitted to another event as a complementary, while the potential misleading effects caused by invalid information in the spatial dimensions can be mitigated. Additionally, a cross-layer interaction representation (CIR) is introduced into BIE for storing useful local and global contexts from previous layers and frames. By layer-wisely stacking the BIE, hierarchical information can be propagated between two streams. Moreover, BIE could be embedded into each stream and serves as a bridge of exchanging spatial and temporal information of each event.

The main contributions of our work are as follows:

- We propose BMCNET, a two-stream network that models the unique characteristics of positive and negative events in event streams while exploiting complementary spatiotemporal contexts to mutually refine each event other, improving overall performance in ESR.
- A novel bilateral information exchange (BIE) module is proposed to facilitate the exchange of complementary global information between the two types of events while mitigating the impact of invalid information such as noise and empty pixels that are inherent in event data.
- Experimental results demonstrate that our method achieve a significant improvement of over **11%** in ESR, and exhibits a visual fidelity that is closer to that of real event streams.
- Our approach also effectively enhances downstream tasks, such as event stream recognition and reconstruction, further validating the effectiveness of our method.

2. Related work

Compared to super-resolution in natural images or videos [4–6, 10, 13, 14, 20, 21, 25, 37], ESR presents unique challenges due to its distinct spatiotemporal characteristics. Initial approaches aimed to preserve the spatiotemporal properties of event streams and directly perform super-resolution. Li et al. [18] introduced an Event Count Map (ECM) to describe the spatial distribution of events, modeling events on each pixel using a non-homogeneous Poisson distribution. However, this method often encountered issues of inaccurate spatiotemporal distribution estimation. Subsequently, based on a mixed imaging strategy, Wang et al. [39] proposed a novel optimization framework called GEF, which utilizes motion-related probability to filter event noise. With the assistance of image frames, GEF achieves event stream super-resolution. However, GEF may degrade significantly when the quality of image frames diminishes. Li et al. [19], leveraging the spatiotemporal properties of Spiking Neural Networks (SNNs), introduced a spatiotemporal constraint learning approach capable of simultaneously learning the temporal and spatial features of event streams. Yet, this method falls short of achieving high-factor super-resolution.

To apply learning-based methods to ESR tasks, researchers proposed projecting event streams onto a 2D plane [26, 45] and then performing super-resolution. Al-

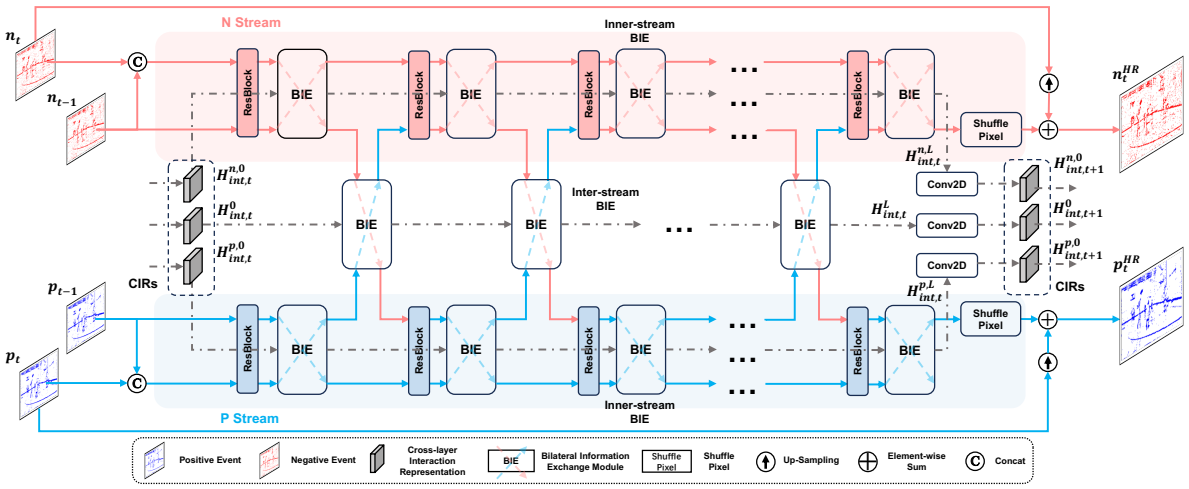

Figure 2. Overall framework of our BMCNET. Blue/red represent events with positive/negative polarity, respectively. BMCNET consists of two parallel streams, each dedicated to mining the information of two events with different polarities. The inter-stream BIE is applied to exchange and complement global structures between two events. Additionally, an inner-stream BIE is embedded within each stream to model the spatio-temporal context of each event. Through the layer-wise introduction of inter- and inner-stream BIE, BMCNET can effectively model hierarchical spatio-temporal contextual correlations between different events, thereby enhancing the performance in ESR.

though this approach compresses the temporal information of event streams, it doesn't impact tasks that require event streams to be converted into a 2D format [12, 17, 34]. Duan et al. [7] were the first to convert event streams into an event stack and introduced a network based on the 3D U-Net for ESR. However, this method faces challenges in terms of memory requirements and training difficulty when handling high-factor super-resolution. Subsequently, Weng et al. [41] presented an event stream super-resolution method based on Recurrent Neural Networks. They effectively addressed the challenges of high-factor super-resolution through the design of the temporal propagation net and spatiotemporal fusion net. Nonetheless, these approaches overlooked the importance of decoupling event data [12, 32] and didn't fully consider its sparsity and structural information. In light of these issues, we propose a bilateral event mining and complementary network capable of independently handling positive and negative events and facilitating mutual improvement between the two event types.

3. Bilateral Event Mining and Complementary

3.1. Preliminary

Applying learning-based methods to Event Stream Super-Resolution typically involves three steps [7, 41]. Firstly, low-resolution event streams are transformed into a 2D representation, compressing the temporal dimension of event streams. Subsequently, a super-resolution network is employed to obtain high-resolution event representations. Finally, recovering the high-resolution event stream through resampling methods. For a set of event streams, we can represent them as $\varepsilon_n = \{\mathbf{e}_k\}_{k=1}^{N}$, where N represents the number of events. Each $\mathbf{e}_i \in \varepsilon_n$ can be represented by a tuple (x_i, y_i, t_i, p_i), where x_i and y_i represent the spatial position of the event, t_i represents the timestamp of the event, and $p_i = \pm 1$ indicates the polarity of the event. Subsequently, we partition ε_n into positive events $\{\mathbf{e}_k\}_{k=1}^{N_p}$ and negative events $\{\mathbf{e}_k\}_{k=1}^{N_n}$ based on their polarity p_i. We transform $\{\mathbf{e}_k\}_{k=1}^{N_p}$ and $\{\mathbf{e}_k\}_{k=1}^{N_n}$ into event count image [41, 45] to describe the spatial distribution of events. Thus, we can build up two-channel event representations from ε_n: positive $\mathbf{p}_k \in \mathbb{R}^{H \times W}$ and negative $\mathbf{n}_k \in \mathbb{R}^{H \times W}$.

3.2. Overall Pipeline

An overview of our proposed bilateral event mining and complementary networks (BMCNET) is depicted in Figure 2. BMCNET comprises two parallel L-layer streams that individually process decoupled positive and negative events. In each layer of BMCNET, the bilateral information exchange (BIE) module are embedded to facilitate the exchange of processed spatial information between the two streams. We refer to this BIE as the inter-stream BIE. Moreover, to fully incorporate temporal dynamics and exploit temporal contexts, we introduce another sub-stream in both P- and N-stream. Specifically, each sub-stream utilizes a residual block consisting of two 3×3 convolution layers to process spatial or temporal information. The spatial and temporal representation are initialized from the current frame and the concatenation of two consecutive frames (the current frame and the previous one), respectively. And the BIE is introduced to layer-wisely facilitate the exchange of information between spatial and temporal contexts. We refer to this BIE as the inner-stream BIE. Both the inter- and inner-stream BIE incorporate the cross-level interac-

tion representations (CIR) that serve different roles. The CIR in the inner-stream BIE provides the interaction between temporal and spatial information, while the CIR in the inter-stream BIE restores the shared global structural information of different types of events. At the time step t, each CIR is updated through an 1×1 convolutional layer and a ReLU activation function:

$$\mathbf{H}_{int,t+1}^{\cdot,0} = \text{ReLU}(\text{Conv}_{1\times1}(\mathbf{H}_{int,t}^{\cdot,L})) \tag{1}$$

where $\cdot$ denotes the type of CIR. $\mathbf{H}_{int,t}^{p,L}$, $\mathbf{H}_{int,t}^{n,L}$ and $\mathbf{H}_{int,t}^{L}$ are CIRs from the last layer of inner-stream BIE in P-, N-stream and the last layer of inter-stream BIE, respectively. And at each time step, both CIRs of inner- and inter-stream BIE are initialized by the updated CIR from previous time step. Particularly, at time step 0, CIRs are initialized as empty representations. This design enables hierarchical information propagation within BMCNET, enhancing its ability to capture spatial-temporal contextual information. After extracting deep-level features, we utilized pixel shuffle [29] to transform the feature information into high-resolution event count image E_t^{SR}. Finally, high-resolution event streams can be obtained by resampling.

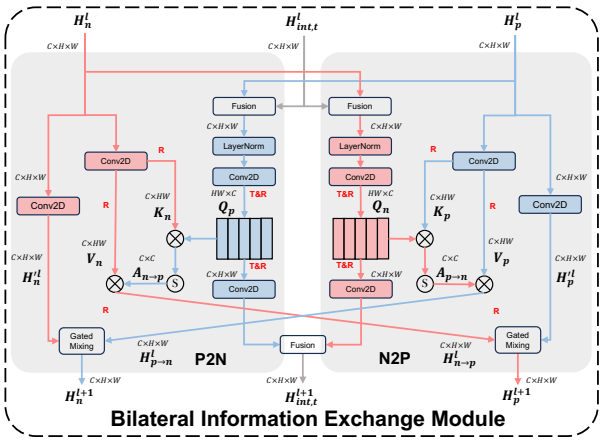

Figure 3. The architecture of the proposed bilateral information exchange (BIE) module. **[Best viewed with zoom-in.]**

3.3. Bilateral Information Exchange Module

Since the positive and negative events exhibit high correlation in corresponding spatial locations and their neighboring regions, the global structures in two types of events can mutually benefit each other by means of providing complementary information. To facilitate the selective integration of global structural information between two types of events and mitigate the potential misleading effects of noises, we propose a bilateral information exchange (BIE) module, as shown in Figure 3. For the event at time step t, we denote the representations of positive and negative events in l-th layer of BMCNET as $\mathbf{H}_p^l \in \mathbb{R}^{C\times H\times W}$ and $\mathbf{H}_n^l \in \mathbb{R}^{C\times H\times W}$, respectively. And a cross-level interaction representation (CIR) $\mathbf{H}_{int,t}^l \in \mathbb{R}^{C\times H\times W}$ is introduced to incorporate the hierarchical contexts into the information exchanging process. In this subsection, we describe the details of the information propagation from the positive to the negative event, and the similar process is applied in the propagation from the negative event to the positive one.

Firstly, the CIR $\mathbf{H}_{int,t}^l$ is updated through the LayerNorm and an 1×1 convolution, and then the updated CIR is fused with the $\mathbf{H}_n^l$ through another 1×1 convolutional layer, obtained the reshaped query $\mathbf{Q}_n^l \in \mathbb{R}^{M\times HW}$. Here M represent the number of global structures we assumed. Then, $\mathbf{H}_p^l$ is clustered into $\mathbf{V}_p^l \in \mathbb{R}^{M\times HW}$ and $\mathbf{K}_p^l \in \mathbb{R}^{M\times HW}$ through the stacked convolutional layers and reshape operation. By calculating the correlation of $\mathbf{Q}_n^l$ and $\mathbf{K}_p^l$ and applying the Softmax function along the horizontal dimension, we can obtain the attention scores matrix $\mathbf{A}_{p\to n} \in \mathbb{R}^{M\times M}$ which reflects the correlations between M semantics in $\mathbf{Q}_n^l$ and $\mathbf{K}_p^l$. Following Scaled Dot-Product Attention [36], the scale coefficient $\sqrt{HW}$ is applied in the calculation process. The attention scores are then utilized to aggregate the global structural representations in $\mathbf{V}_p^l$, obtaining the $\mathbf{H}_{p\to n}^{'l} \in \mathbb{R}^{M\times HW}$. $\mathbf{H}_{p\to n}^{'l}$ is further reshaped and projected into $\mathbf{H}_{p\to n}^l \in \mathbb{R}^{C\times H\times W}$ through an 1×1 convolutional layer. Besides, a residual branch consists of stacked convolutional layers is introduced to process the original information of $\mathbf{H}_n^l$ for the local structural details, obtained the $\mathbf{H}_n^{'l}$. The process is formulated as:

$$\mathbf{H}_{p\to n}^l = \text{Conv}_{1\times1}(\text{Softmax}(\mathbf{Q}_n^l(\mathbf{K}_p^l)^T/\sqrt{HW})\mathbf{V}_p^l) \tag{2}$$

Finally, the $\mathbf{H}_n^l$ is updated by fusing $\mathbf{H}_{p\to n}^{'l}$ and $\mathbf{H}_n^{'l}$ through a gated mixing module:

$$\mathbf{H}_n^{l+1} = \mathbf{Z}_n^l \odot \mathbf{H}_n^{'l} + (1 - \mathbf{Z}_n^l) \odot \mathbf{H}_{p\to n}^l \tag{3}$$

$$\mathbf{Z}_n^l = \sigma(\mathbf{W}_{n1}\mathbf{H}_n^{'l} + \mathbf{W}_{n2}\mathbf{H}_{p\to n}^l) \tag{4}$$

where σ denotes the Sigmoid function whose range lies in $[0, 1]$, $\odot$ denotes element-wise production and $\mathbf{Z}_n^l \in \mathbb{R}$. The similar procedure of information propagation from the negative to the positive event can be written directly:

$$\mathbf{H}_{n\to p}^l = \text{Conv}_{1\times1}(\text{Softmax}(\mathbf{Q}_p^l(\mathbf{K}_n^l)^T/\sqrt{HW})\mathbf{V}_n^l) \tag{5}$$

$$\mathbf{H}_p^{l+1} = \mathbf{Z}_p^l \odot \mathbf{H}_p^{'l} + (1 - \mathbf{Z}_p^l) \odot \mathbf{H}_{n\to p}^l \tag{6}$$

$$\mathbf{Z}_p^l = \sigma(\mathbf{W}_{p1}\mathbf{H}_p^{'l} + \mathbf{W}_{p2}\mathbf{H}_{n\to p}^l) \tag{7}$$

And the CIR $\mathbf{H}_{int,t}^l$ is updated by feeding the concatenation of the reshaped $\mathbf{Q}_p^l$ and $\mathbf{Q}_n^l$ into the convolutional layer:

$$\mathbf{H}_{int,t}^{l+1} = \text{Conv}([\![\mathbf{Q}_p^{l,r}, \mathbf{Q}_n^{l,r}]\!]) + \mathbf{H}_{int}^l \tag{8}$$

where $[\![\cdot]\!]$ denotes concatenation along channel, r denotes the query matrix is reshaped from $\mathbb{R}^{C\times HW}$ into $\mathbb{R}^{C\times H\times W}$. And the pseudo-code of BIE is provided in **Appendix B**.

3.4. Training Objectives

In the training process, to ensure the continuity of the event stream, following the approach of [41], we divide the event stream into segments of length $T(T = 9)$ and calculate the mean square error for each segment using a sliding window:

$$\mathcal{L} = \sum_{t=1}^{T} MSE\left(E_t^{SR},\ E_t^{HR}\right) \tag{9}$$

where E_t^{SR} represents the event count image obtained through BMCNET, E_t^{HR} represents the ground truth event count image, and MSE is the mean square error function.

4. Experiments

4.1. Datasets and Training Settings

In this work, we validate our approach using both real and synthetic data. The real event dataset, including multi-scale LR-HR pairs, is limited due to the difficulty of aligning both temporal and spatial information simultaneously. Event-NFS [7] is the first real dataset involving LR-HR pairs, captured by a DAVIS346 monochromatic camera. However, due to device resolution constraints, the minimum resolution is 55×31, and only $2\times$ and $4\times$ data pairs are available. To obtain high-quality LR-HR pairs with higher multiples, following [41], we use an event simulator [24] on the NFS [15] dataset and RGB-DAVIS dataset [39] to transform them into event streams, resulting in two new synthetic datasets, NFS-syn and RGB-syn.

To ensure a fair comparison, we maintain consistency with the training settings from [41]. We set the batch size to 2, the initial learning rate to 0.001, a decay factor of 0.95, and decay every 4000 iterations. All models were trained for 100,000 iterations, and the entire experimentation was conducted on a Tesla V100 GPU. For more details on dataset processing, ablation studies and experimental results, please refer to the **Appendix A**.

4.2. Comparison with State-of-the-Art Models

In the field of ESR, our primary comparisons are made with two previous learning-based methods, EventZoom [7] and RecEvSR [41]. Comparing with other ESR methods is challenging [18, 19, 39], as they either require real frames as assistance or may fail in complex scenarios, making a fair comparison difficult. EventZoom [7], the first learning-based method in ESR, faced training difficulties with its 3D-Unet architecture for large-factor SR. Following prior practices [41], we ran EventZoom-$2\times$ multiple times to obtain results for large-factor SR. RecEvSR [41] overcomes the challenge of large-scale SR using recurrent neural networks, and we retrained it using the provided code. Additionally, we included two representative methods for video and image super-resolution for comparison: bicubic and SRFBN [21] for image super-resolution and lightweight models RLSP [8] and RSTT [13] for video super-resolution. To fully investigate the effectiveness of the proposed paradigm that process two events independently and mutually refine each other, we introduced BMCNET-plain which is obtained by removing the inner-stream BIE in BMCNET. We utilize the RMSE, model parameters and FLOPs to evaluate each method from three perspectives: the performance, model and computational complexity.

Qualitative Analysis Results. Figure 4 presents the $4\times$ SR results of various methods on both synthetic and real datasets (please refer to **Appendix D.1** for more results). It is evident that bicubic interpolation struggles to effectively contribute to ESR. EventZoom exhibits numerous issues with detail loss, potentially arising from error accumulation in multiple runs of EventZoom-$2\times$. While SRFBN and RLSP can restore the overall structures of event count images, they suffer from significant detail loss and blurred object edges. In comparison, RSTT and RecEvSR perform reasonably in recovering event count image. However, they still fall short in detail repair. Contrasted with these methods, our BMCNET-plain and BMCNET excel in mutually complementing and repairing details by leveraging overall structural information from two events, resulting in richer details and clearer edges.

Quantitative Analysis Results. As shown in Tab. 1, both BMCNET-plain and BMCNET have achieved state-of-the-art (SOTA) performance across all datasets. Compared to RecEvSR, the previous SOTA method in ESR, our BMCNET-plain exhibits an average improvement of 7.5% in super-resolution across all datasets, accompanied by a 40% reduction in parameters. Furthermore, in comparison to the leading Transformer-based video super-resolution method, RSTT, our BMCNET-plain achieves an average improvement of 3.5%, with a reduction in parameters of 73%. Meanwhile, BMCNET, which includes both inter- and inner-stream BIE, demonstrates superior performance. BMCNET boasts an average improvement of 11% over RecEvSR, but with a parameters increase of 50%. In contrast, BMCNET achieves an average improvement of 7% over RSTT, with a reduction in parameters of 31%. Additionally, FLOPs values indicate that our BMCNET-plain exhibits the lowest average complexity, while BMCNET maintains competitive one. These findings collectively underscore the efficiency of our methods.

4.3. Model Analysis

The Validation of Main Components in BMCNET. We conducted a series of ablation experiments on the NFS-syn datasets to investigate the impact of the main components in BMCNET. The experimental results are presented in Tab. 2, **Decoup.** refers to the model processing positive and negative events decoupled from the event stream sep-

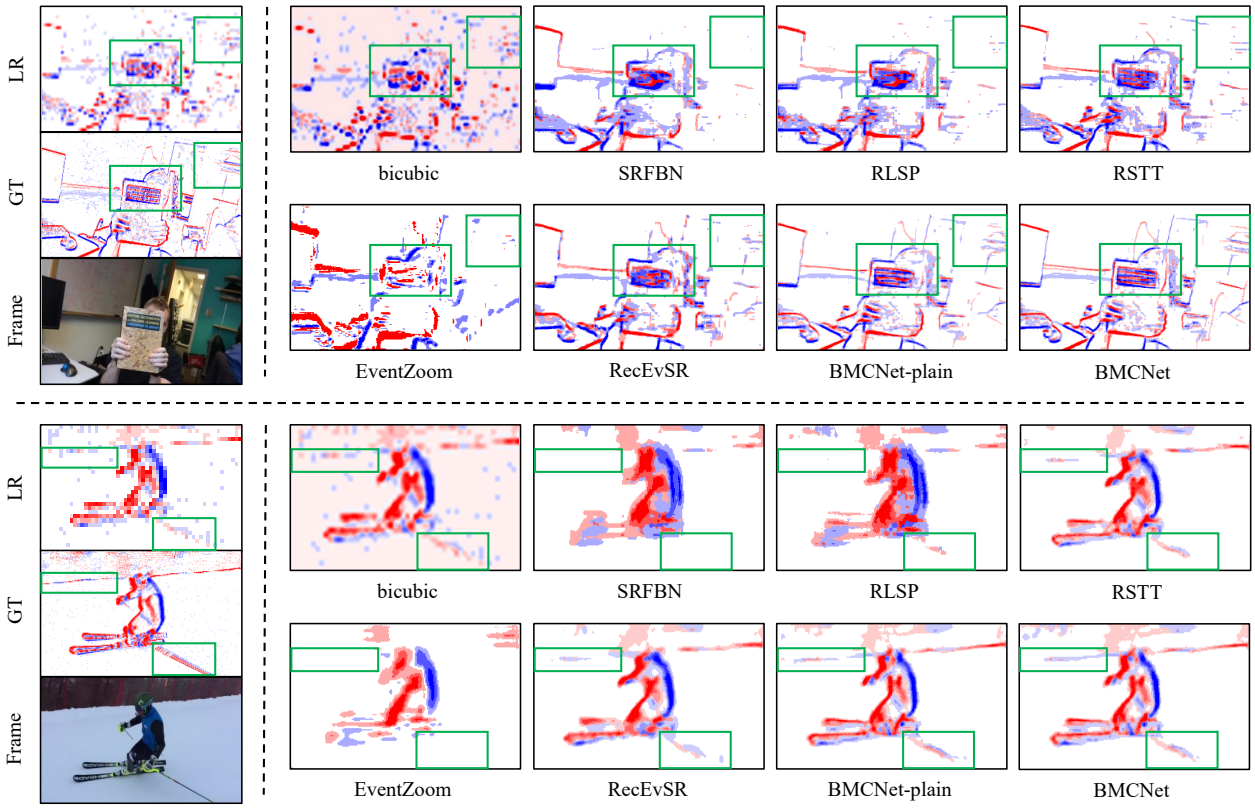

Figure 4. Qualitative analysis comparison on synthetic and real datasets. The upper and lower sections represent the $4\times$ super-resolution results for NFS-syn and EventNFS, respectively. "GT" denotes the $4\times$ ground-truth. Our BMCNET-plain and BMCNET demonstrates superior detail recovery and clearer edges on both datasets. **[Best viewed with zoom-in.]**

Methods	NFS-syn			RGB-syn		EventNFS-real		# Param (M)			# FLOPs (G)		
	2×	4×	8×	2×	4×	2×	4×	2×	4×	8×	2×	4×	8×
bicubic	0.785	0.729	0.738	0.346	0.378	0.872	0.948	-	-	-	-	-	-
SRFBN-esr [21]	0.641	0.628	0.628	0.301	0.300	0.644	0.738	2.1	3.6	7.9	39.5	116.7	984.1
RLSP-esr [8]	0.642	0.624	0.621	0.298	0.294	0.623	0.705	1.2	1.2	1.5	23.1	24.9	32.1
RSTT-esr [13]	0.624	0.605	0.604	0.298	0.292	0.557	0.632	3.8	4.1	4.3	22.3	43.3	61.4
EventZoom [7]	0.898	1.024	1.113	0.479	0.975	0.882	1.117	11.5	11.5	11.5	65.3	81.0	220.3
RecEvSR [41]	0.656	0.607	0.576	0.319	0.296	0.613	0.670	1.8	1.8	1.8	**2.8**	10.7	42.1
BMCNET-plain (Ours)	0.592	0.577	0.579	0.287	0.285	0.541	**0.619**	**0.9**	**1.0**	**1.4**	7.96	**8.16**	**8.96**
BMCNET (Ours)	**0.564**	**0.552**	**0.553**	**0.276**	**0.274**	**0.527**	0.625	2.6	2.7	3.1	35.35	35.65	36.84

Table 1. Quantitative analysis comparison on real and synthetic datasets, and RMSE, model parameters and FLOPs are reported. BMCNET denotes the network additionally equipped with both the inter- and inner-stream BIE for spatial and temporal contexts modeling, and BMCNET-plain denotes only the inter-BIE is equipped as described in Sec. 4.3. The FLOPs is calculated on the LR events of NFS-syn dataset with resolutions of 80×45. Top 2 results are highlighted with **bold text** and underlined text, respectively.

arately, **Rec.** signifies the introduction of a recurrent hidden state into the model for cross-frame modeling, **Inter-BIE** indicates the layer-wise application of inter-stream BIE for global structures exchange between two types of events, and **Inner-BIE** denotes the inner-stream BIE embedded in the model for spatial and temporal contexts modeling.

By comparing Exp0 and Exp1, we observed that the inclusion of a recurrent hidden state improves performance on both synthetic and real event datasets. This improvement can be attributed to the capture of cross-frame correlations and the utilization of useful cues from previous frames for global structure understanding. Additionally, the comparisons between Exp1-4 demonstrate that simply processing decoupled positive and negative events individually can not

model the correlations between two events, resulting in bad performance. Compared to directly process positive and negative events in a mixed paradigm, separately process decoupled events and exchange global structures through BIE can effectively capture inner- and inter-event information, enhancing the performance in super-resolutions. More ablation studies about comparison between mixed and decoupled paradigm please refer to subsections below. Furthermore, the comparison between Exp4 and Exp5 demonstrates that simultaneously modeling the temporal-spatial information correlations and leveraging the complementary nature of the two types of events can fully exploit spatial and temporal contextual information in the event streams. This approach significantly promotes the final performance of the model in event super-resolution.

Exps	Decoup.	Rec.	Inter-BIE	Inner-BIE	NFS-syn	EventNFS-real	# Params.
Exp0	✘	✘	✘	✘	0.605	0.655	**0.38** M
Exp1	✘	✔	✘	✘	0.599	0.648	0.68 M
Exp2	✔	✔	✘	✘	0.637	0.768	0.97 M
Exp3	✔	✘	✔	✘	0.580	0.641	0.94 M
Exp4	✔	✔	✔	✘	0.577	**0.619**	1.00 M
Exp5	✔	✔	✔	✔	**0.552**	0.625	2.72 M

Table 2. Quantitative analysis of branch ablation experiments for $4\times$ SR on RMSE metrics.

Benefit from Decoupling Positive and Negative Events. The core motivation behind our BMCNET is to fully exploit the correlations between positive and negative events, allowing them to benefit each other by providing the complementary information. Through the decoupled paradigm, the model has capability to capture unique characteristics of each event and enable effective interactions between them. And overall performance can be enhanced through leveraging the correlation and complementary nature of two events. To verify this, we conduct experiments that process positive and negative events in both the mixed and decoupled paradigms, respectively. The comparison results are presented in Tab. 2 and Figure 5. Moreover, we observed

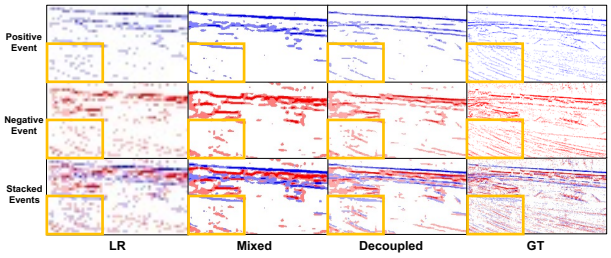

Figure 5. Qualitative comparison on the real dataset EventNFS-real between the mixed and decoupled paradigm.

that simultaneously processing positive and negative events in the mixed paradigms always results in overlapped artifacts. The reason for this is that simultaneous processing of positive and negative events makes the model confused in

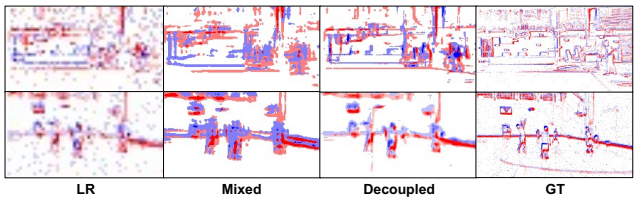

Figure 6. Artifacts of overlapped events appear when two events are processed in a mixed paradigm.

distinguishing between the two types of events. This phenomenon is resolved when we process the positive and negative events in a decoupled paradigm, as shown in Figure 6.

Effectiveness of Bilateral Information Exchange Module. In this section, we use the BMCNET-plain (equipped with inter-stream BIE only) as baseline to conduct ablation studies on NFS-syn and EventNFS-real datasets for investigating the BIE we proposed. The BIE serves as the core operation in our BMCNET, allowing for the effective exchange of global structural information between positive and negative events without being misled by noise. And the introduction of the cross-level interaction representation (CIR) enables the model to exchange information across different levels. To verify the effectiveness of BIE, we replace BIE in BMCNET-plain with three different type operations: (1) concatenation, (2) cross attention (CA) operation, (3) the BIE without CIR. The quantitative comparisons are presented in Tab. 3, which are demonstrated the effectiveness the BIE and CIR on improving the overall performance. Additionally, the qualitative comparisons are shown in Figure 7. The model exchanges information between two events through the concatenation cannot effectively utilize the complementary information, resulting in incomplete structures. The model equipped with CA is easily misled by noises which results in distort structures and artifacts. On the contrary, our BIE has capability to alleviate the impact of invalid information and reconstruct clear edges by treating each channel as a structural representation and exchanging it along the channel dimension.

Metrics	NFS-syn	EventNFS-real	# Params	Complexity
Concat.	0.606	0.664	1.40 M	$\mathcal{O}(C^2)$
CA	0.610	0.643	1.37 M	$\mathcal{O}(CH^2W^2)$
BIE w/o CIR	0.580	0.641	**0.94** M	$\mathcal{O}(C^2HW)$
BIE	**0.577**	**0.619**	1.00 M	$\mathcal{O}(C^2HW)$

Table 3. Effectiveness of BIE and CIR. The experiments are conducted for $4\times$ SR and evaluated on RMSE metrics.

4.4. Enhancing Downstream Applications

Object Recognition. We investigated the performance of bicubic, SRFBN [21], RLSP [8], RSTT [13], EventZoom [7], RecEvSR [41], and our BMCNET-plain and BMCNET

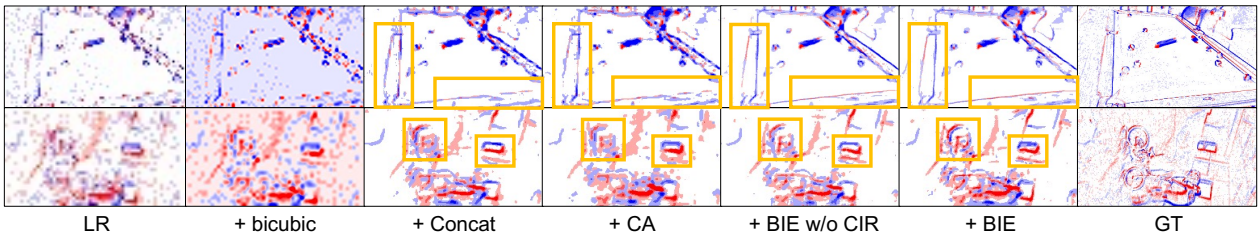

Figure 7. Qualitative analysis comparison between bicubic, Concat, CA, BIE with and without CIR on synthetic and real datasets.

Methods	Object Recognition					
	2x		4x		8x	
	ACC ↑	AUC ↑	ACC ↑	AUC ↑	ACC ↑	AUC ↑
bicubic	56.67	57.43	56.01	56.89	49.95	50.77
SRFBN	61.12	61.94	60.89	61.03	50.02	50.86
RLSP	61.33	62.16	61.19	61.25	50.26	50.97
RSTT	63.51	63.96	63.02	64.29	52.97	54.07
EventZoom	54.68	56.03	49.56	50.45	47.96	48.74
RecEvSR	62.91	63.47	62.37	63.07	53.57	54.48
BMCNet-plain	66.95	67.73	67.31	67.63	54.31	54.93
BMCNet	**68.58**	**69.33**	**68.95**	**69.23**	**57.97**	**58.86**
GT	85.16	84.99	93.44	93.52	94.96	94.81

Methods	Video Reconstruction					
	SSIM ↑	LPIPS ↓	SSIM ↑	LPIPS ↓	SSIM↑	LPIPS ↓
bicubic	0.568	0.395	0.609	0.522	0.598	0.545
SRFBN	0.608	0.389	0.618	0.455	0.612	0.489
RLSP	0.603	0.384	0.620	0.439	0.617	0.476
RSTT	0.627	0.359	0.639	0.424	0.622	0.472
EventZoom	0.542	0.429	0.575	0.488	0.574	0.542
RecEvSR	0.611	0.371	0.637	0.426	0.630	0.466
BMCNet-plain	0.638	0.349	0.651	0.418	0.632	0.468
BMCNet	**0.646**	**0.343**	**0.661**	**0.414**	**0.652**	**0.452**

Table 4. Quantitative analysis results on downstream tasks of object recognition and video reconstruction. For object recognition, the evaluation is conducted on the NCars dataset [30], where AUC and ACC represent accuracy and area under the curve, respectively. GT denotes the reference result obtained by directly recognizing downsampled event streams. Video reconstruction is performed on the NFS-syn dataset. **Bold** and underline indicate the best and the second-best performance.

in the context of object recognition applications. Utilizing the classifier proposed by Gehrig et al. [12], we conducted object recognition on the NCars dataset [30]. Initially, the NCars dataset underwent an $8\times$ downsampling, followed by applying each model trained on the NFS-syn dataset to perform $2(4,8)\times$ super-resolution. The super-resolved data was then used for recognition, and the results were compared. We utilized accuracy (ACC) and area under the curve (AUC) as the metrics for comparison. The GT represents the classification results obtained by directly downsampling the event stream to the same resolution, indicating the upper limit performance. Tab. 4 presents the quantitative analysis results for object recognition, demonstrating that our BMCNET-plain and BMCNET outperform other approaches significantly across $2(4,8)\times$ super-resolution scales, validating the robust detail recovery capability of our methods.

Video Reconstruction. Video reconstruction is a crucial task in event-based vision [28, 31, 40]. We also compared all methods in this task. Initially, we applied each model to $2(4,8)\times$ super-resolution on the NFS-syn dataset, which was downsampled by a factor of 16. Then, we utilized the E2VID [28] for reconstructing images based on the super-resolved event stream. Finally, we evaluated the reconstructed images using the structural similarity (SSIM) [38] and the perceptual similarity (LPIPS) [43] metrics. Please refer to **Appendix D.2** for reconstructed images. Tab. 4 presents quantitative comparisons for reconstructed images, indicating that our BMCNET-plain and BMCNET outperform other approaches significantly at different scales.

5. Conclusion

In this paper, we fully consider the inherent characteristics of the event data and propose a novel bilateral event mining and complementary network (BMCNET) for the ESR task. By simultaneously modeling the distinct data distribution of positive and negative events which are decoupled from the event stream, and capturing the complementary information to mutually refine each other, BMCNET can effectively reconstruct clear structures of each type of event. Extensive empirical studies and analysis experiments conducted on two synthetic and one real datasets demonstrate the effectiveness and superiority of BMCNET. Moreover, BMCNET are evaluated on two downstream tasks, achieving outstanding results and further highlighting its superiority over other approaches.

Acknowledgments

This work was partly supported by the National Natural Science Foundation of China (No.62171251 & 62311530100), the Special Foundations for the Development of Strategic Emerging Industries of Shenzhen (No.JSGG20211108092812020) and the Major Key Research Project of PCL (No.PCL2023A08).

References

[1] Patrick Bardow, Andrew J Davison, and Stefan Leutenegger. Simultaneous optical flow and intensity estimation from an event camera. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 884–892, 2016. 1

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

[3] Marco Cannici, Marco Ciccone, Andrea Romanoni, and Matteo Matteucci. A differentiable recurrent surface for asynchronous event-based data. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, pages 136–152. Springer, 2020. 1

[4] Kelvin CK Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Basicvsr: The search for essential components in video super-resolution and beyond. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4947–4956, 2021. 2

[5] Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5972–5981, 2022.

[6] Xiangyu Chen, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Activating more pixels in image super-resolution transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22367–22377, 2023. 2

[7] Peiqi Duan, Zihao W Wang, Xinyu Zhou, Yi Ma, and Boxin Shi. Eventzoom: Learning to denoise and super resolve neuromorphic events. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12824–12833, 2021. 2, 3, 5, 6, 7, 12, 13, 14, 15, 16, 17

[8] Dario Fuoli, Shuhang Gu, and Radu Timofte. Efficient video super-resolution through recurrent latent space propagation. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 3476–3485. IEEE, 2019. 5, 6, 7, 12, 13, 14, 15, 16, 17

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

[10] Ganzorig Gankhuyag, Kihwan Yoon, Jinman Park, Haeng Seon Son, and Kyoungwon Min. Lightweight real-time image super-resolution network for 4k images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1746–1755, 2023. 2

[11] Daniel Gehrig and Davide Scaramuzza. Are high-resolution event cameras really needed? *arXiv preprint arXiv:2203.14672*, 2022. 1

[12] Daniel Gehrig, Antonio Loquercio, Konstantinos G Derpanis, and Davide Scaramuzza. End-to-end learning of representations for asynchronous event-based data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5633–5643, 2019. 3, 8

[13] Zhicheng Geng, Luming Liang, Tianyu Ding, and Ilya Zharkov. Rstt: Real-time spatial temporal transformer for space-time video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17441–17451, 2022. 2, 5, 6, 7, 12, 13, 14, 15, 16, 17

[14] Takashi Isobe, Xu Jia, Xin Tao, Changlin Li, Ruihuang Li, Yongjie Shi, Jing Mu, Huchuan Lu, and Yu-Wing Tai. Look back and forth: Video super-resolution with explicit temporal difference modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17411–17420, 2022. 2

[15] Hamed Kiani Galoogahi, Ashton Fagg, Chen Huang, Deva Ramanan, and Simon Lucey. Need for speed: A benchmark for higher frame rate object tracking. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1125–1134, 2017. 5, 11

[16] Junho Kim, Jaehyeok Bae, Gangin Park, Dongsu Zhang, and Young Min Kim. N-imagenet: Towards robust, fine-grained object recognition with event cameras. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2146–2156, 2021. 1

[17] Dianze Li, Jianing Li, and Yonghong Tian. Sodformer: Streaming object detection with transformer using events and frames. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 1, 3

[18] Hongmin Li, Guoqi Li, and Luping Shi. Super-resolution of spatiotemporal event-stream image. *Neurocomputing*, 335: 206–214, 2019. 1, 2, 5

[19] Siqi Li, Yutong Feng, Yipeng Li, Yu Jiang, Changqing Zou, and Yue Gao. Event stream super-resolution via spatiotemporal constraint learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4480–4489, 2021. 1, 2, 5

[20] Xiaoming Li, Wangmeng Zuo, and Chen Change Loy. Learning generative structure prior for blind text image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10103–10113, 2023. 2

[21] Zhen Li, Jinglei Yang, Zheng Liu, Xiaomin Yang, Gwanggil Jeon, and Wei Wu. Feedback network for image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3867–3876, 2019. 2, 5, 6, 7, 12, 13, 14, 15, 16, 17

[22] Quanmin Liang, Xiawu Zheng, Kai Huang, Yan Zhang, Jie Chen, and Yonghong Tian. Event-diffusion: Event-based image reconstruction and restoration with diffusion models. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 3837–3846, 2023. 1

[23] Patrick Lichtsteiner, Christoph Posch, and Tobi Delbrück. A 128×128 120 db 15 μs latency asynchronous temporal contrast vision sensor. *IEEE J. Solid State Circuits*, 43(2):566–576, 2008. 1

[24] Songnan Lin, Ye Ma, Zhenhua Guo, and Bihan Wen. Dvs-voltmeter: Stochastic process-based event simulator for dy-

namic vision sensors. In *European Conference on Computer Vision*, pages 578–593. Springer, 2022. 5, 11

[25] Yunfan Lu, Zipeng Wang, Minjie Liu, Hongjian Wang, and Lin Wang. Learning spatial-temporal implicit neural representations for event-guided video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1557–1567, 2023. 2

[26] Ana I Maqueda, Antonio Loquercio, Guillermo Gallego, Narciso García, and Davide Scaramuzza. Event-based vision meets deep learning on steering prediction for self-driving cars. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5419–5427, 2018. 2

[27] Nico Messikommer, Carter Fang, Mathias Gehrig, and Davide Scaramuzza. Data-driven feature tracking for event cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5642–5651, 2023. 1

[28] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. High speed and high dynamic range video with an event camera. *IEEE transactions on pattern analysis and machine intelligence*, 43(6):1964–1980, 2019. 1, 8, 13

[29] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1874–1883, 2016. 4

[30] Amos Sironi, Manuele Brambilla, Nicolas Bourdis, Xavier Lagorce, and Ryad Benosman. Hats: Histograms of averaged time surfaces for robust event-based object classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1731–1740, 2018. 8

[31] Timo Stoffregen, Cedric Scheerlinck, Davide Scaramuzza, Tom Drummond, Nick Barnes, Lindsay Kleeman, and Robert Mahony. Reducing the sim-to-real gap for event cameras. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII 16*, pages 534–549. Springer, 2020. 8

[32] Ganchao Tan, Yang Wang, Han Han, Yang Cao, Feng Wu, and Zheng-Jun Zha. Multi-grained spatio-temporal features perceived network for event-based lip-reading. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20094–20103, 2022. 3

[33] Gemma Taverni, Diederik Paul Moeys, Chenghan Li, Celso Cavaco, Vasyl Motsnyi, David San Segundo Bello, and Tobi Delbruck. Front and back illuminated dynamic and active pixel vision sensors comparison. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 65(5):677–681, 2018. 1

[34] Stepan Tulyakov, Daniel Gehrig, Stamatios Georgoulis, Julius Erbach, Mathias Gehrig, Yuanyou Li, and Davide Scaramuzza. Time lens: Event-based video frame interpolation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16155–16164, 2021. 3

[35] Stepan Tulyakov, Alfredo Bochicchio, Daniel Gehrig, Stamatios Georgoulis, Yuanyou Li, and Davide Scaramuzza. Time lens++: Event-based frame interpolation with parametric non-linear flow and multi-scale fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17755–17764, 2022. 1

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4

[37] Yingwei Wang, Takashi Isobe, Xu Jia, Xin Tao, Huchuan Lu, and Yu-Wing Tai. Compression-aware video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2012–2021, 2023. 2

[38] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4): 600–612, 2004. 8

[39] Zihao W Wang, Peiqi Duan, Oliver Cossairt, Aggelos Katsaggelos, Tiejun Huang, and Boxin Shi. Joint filtering of intensity images and neuromorphic events for high-resolution noise-robust imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1609–1619, 2020. 2, 5, 11

[40] Wenming Weng, Yueyi Zhang, and Zhiwei Xiong. Event-based video reconstruction using transformer. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 2543–2552. IEEE, 2021. 8

[41] Wenming Weng, Yueyi Zhang, and Zhiwei Xiong. Boosting event stream super-resolution with a recurrent neural network. In *European Conference on Computer Vision*, pages 470–488. Springer, 2022. 2, 3, 5, 6, 7, 12, 13, 14, 15, 16, 17

[42] Lan Xu, Weipeng Xu, Vladislav Golyanik, Marc Habermann, Lu Fang, and Christian Theobalt. Eventcap: Monocular 3d capture of high-speed human motions using an event camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4968–4978, 2020. 1

[43] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 586–595. Computer Vision Foundation / IEEE Computer Society, 2018. 8

[44] Yi Zhou, Guillermo Gallego, Henri Rebecq, Laurent Kneip, Hongdong Li, and Davide Scaramuzza. Semi-dense 3d reconstruction with a stereo event camera. In *Proceedings of the European conference on computer vision (ECCV)*, pages 235–251, 2018. 1

[45] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Ev-flownet: Self-supervised optical flow estimation for event-based cameras. *arXiv preprint arXiv:1802.06898*, 2018. 2, 3