Boosting Event Stream Super-Resolution with A Recurrent Neural Network

Wenming Weng, Yueyi Zhang(✉), and Zhiwei Xiong

University of Science and Technology of China, Hefei, China
wmweng@mail.ustc.edu.cn, {zhyuey, zwxiong}@ustc.edu.cn

Abstract. Existing methods for event stream super-resolution (SR) either require high-quality and high-resolution frames or underperform for large factor SR. To address these problems, we propose a recurrent neural network for event SR without frames. First, we design a temporal propagation net for incorporating neighboring and long-range event-aware contexts that facilitates event SR. Second, we build a spatiotemporal fusion net for reliably aggregating the spatiotemporal clues of event stream. These two elaborate components are tightly synergized for achieving satisfying event SR results even for 16× SR. Synthetic and real-world experimental results demonstrate the clear superiority of our method. Furthermore, we evaluate our method on two downstream event-driven applications, i.e., object recognition and video reconstruction, achieving remarkable performance boost over existing methods.

Keywords: Event Camera, Super-Resolution, Recurrent Network

1 Introduction

Event cameras, the novel bio-inspired imaging sensors, have accelerated the innovation in machine vision-enabled systems [4,13,43]. By taking a clue from human vision system, event cameras release a new form of vision capture enabling a promising ability to support diverse operating requirements such as stringent power consumption, demanding memory needs, high speed motion perception, and high dynamic range (HDR) scene imaging [13]. Therefore, the use of event cameras has fast gained wide acceptance recently [1,3,9,14,15,15,20,22,27,30,37–42,44,45,45–47,52–55,59,68–71,73].

In spite of many advantages, the spatial resolution of most commercial event cameras is relatively low due to the physical limitation [43]. Although some high-resolution (HR) devices, such as Prophesee Gen4 CD event cameras, have been developed, they are inevitably limited by low speed and high power consumption. To reconcile this problem, some works have been proposed to approach the event super-resolution (SR) task. Duan et al. [12] and Wang et al. [58] attempted to directly predict a super-resolved event count map via deep learning techniques or optimization frameworks for restoring HR event streams. Li et al. [34] and Li et al. [33] performed event SR by simultaneously estimating the spatiotemporal distribution using either a sparse signal representation [64]

with a non-homogeneous poisson process or spiking neural networks (SNNs) [18]. However, these methods still have two major limitations: **1)** Joint filtering with intensity frames for restoring the HR event stream relies heavily on the quality of HR frames and the accuracy of optical flow estimation, which is computationally expensive [58]. **2)** Super-resolving event streams by large factors i.e., 8×, 16×, is rather comprehensively unexplored due to tough network training [12, 34] or inaccurate spatiotemporal distribution estimation [33].

This work aims to address the above two challenges. Ideally, frame-assisted event SR strongly demands high-quality HR frames. However, the capture of RGB cameras is susceptible to deterioration caused by harsh environments, e.g., high speed motion and HDR scenes, leading to blurry and over(under)-exposure frames that may harm event SR due to the lack of sharp edges. Moreover, further processing of frames for accurate optical flow is typically time-consuming [58]. In this paper, we demonstrate the feasibility of achieving event SR solely with pure events, which is practical for real-world capture without RGB cameras. On the other hand, for large factor event SR, 3D UNet and SNNs used in existing works [12,34] to build an LR-HR projection are intractable to train due to unaccessible memory requirement. It thus calls for multiple times of small-factor SR to obtain large factor SR results, which leads to inevitable performance degradation while sacrificing the running time. In this paper, we make an attempt to use the recurrent neural network for event SR. Without the assistance of frames, we demonstrate that the proposed method is able to achieve the state-of-the-art results in one forward pass, even for 16× SR.

Specifically, we design our recurrent network by constructing two novel components, i.e., a temporal propagation net ($TPNet$) and a spatiotemporal fusion net ($STFNet$). In order to effectively relate and model the sequential temporal correlation of event stream, we devise the $TPNet$ by jointly embedding a local temporal correlation module implemented by the attention mechanism and a global temporal correlation module built upon the recursive state update mechanism. With these two elaborate modules, we can adaptively incorporate the event-aware contexts from both a local range and a global range, favorably promoting event SR. Moreover, for reliably aggregating the local and global temporal correlations captured by the $TPNet$, we build the $STFNet$ that contains a gated temporal fusion module and an adaptive spatiotemporal fusion module, which are tightly collaborated for reliable aggregation.

We summarize our main contributions as four-fold: **1**) We propose an advanced solution using a recurrent neural network to approach event SR, which gets rid of dependence on high-quality and HR frames and suits for large factor SR (even 16×); **2**) We design two novel components, i.e., a temporal propagation net and a spatiotemporal fusion net, for effectively relating and aggregating the sequential spatiotemporal clues of event stream; **3**) Synthetic and real-world experimental results demonstrate the clear superiority of our method over existing methods for event SR; **4**) Superior performance on two downstream event-driven applications, i.e., object recognition and video reconstruction, reconfirms the effectiveness of our method.

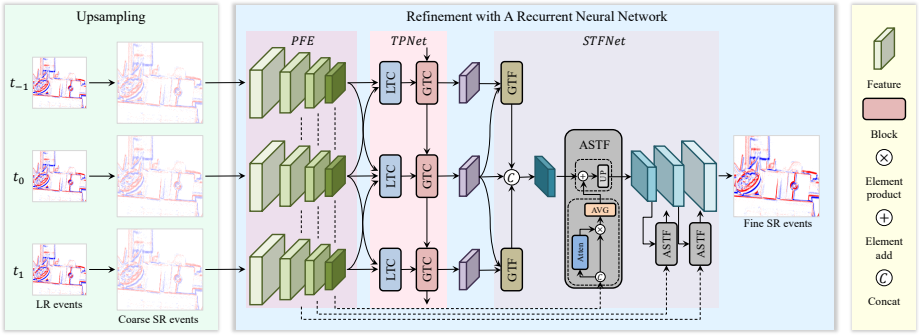


Fig. 1. Overall pipeline. First, the input LR event stream is spatially enlarged by upsampling tools to generate the coarse SR event stream. Then, the recurrent neural network is employed for the fine SR event stream. Three components, i.e., pyramidal feature extractor (PFE), temporal propagation net ($TPNet$) and spatiotemporal fusion net ($STFNet$) are tightly integrated in our recurrent network. The blocks with the same color are sharing weights.

2 Related Work

Frame SR. The frame SR task has been investigated for many years and achieved significant progresses. Classic methods aim to super-resolve images by exploiting the statistical image priors to build a regression function from LR to HR images, which can be achieved by neighbor embedding [2, 7], sparse coding [49, 50, 65, 67] and internal patch recurrence [19, 24]. The recent deep learning based SR methods have shown excellent performance and dominated the field of frame SR. For single image SR (SISR), Dong et al. [11] first applied convolutional neural networks (CNNs) to build an end-to-end LR-HR mapping for SISR. The works with advanced neural network architectures [32, 36, 51] are further reported, significantly promoting performance of SISR. For video SR (VSR), VESCP [5] is the pioneering work for approaching VSR by jointly training optical flow estimation and spatiotemporal networks. Recently, more works [6, 25, 26, 28, 48, 57, 61, 62] have been proposed by investigating more sophisticated components to address the propagation and alignment problems.

Event SR. Compared with natural images, the event stream, captured by sensing the intensity changes, is essentially a kind of spatiotemporal data. To super-resolve the event stream, accurate spatial and temporal distributions need to be estimated. A few works [12,33,34,58] have been developed to approach event SR. Similar to hybrid imaging [23,31,56,66], Wang et al. [58] developed a novel optimization framework termed as GEF, which took advantages of both frame-based and event-based sensing, to achieve the HR and noise-robust event stream. However, GEF is severely deteriorated when the auxiliary frame is visually blurry or the optical flow is inaccurately estimated. Duan et al. [12] proposed a novel network based on 3D U-Net with an event-to-image (E2I) module to learn the correspondence between the LR event stream and the HR event stream. Li et

al. [33] utilized a sparse signal representation method [64] to acquire the spatial distribution of the event stream and then modeled a spatiotemporal filter to generate the temporal rate function. They finally used a non-homogeneous poisson process to simulate the per-pixel events. To build an end-to-end projection in the event domain, Li et al. [34] proposed a novel learning-based method with SNNs to achieve the HR event stream. The spatiotemporal constraint learning enables SNNs to learn the spatial and temporal distribution simultaneously. These works [12, 33, 34, 58] are pioneers in the field of event SR, however, they either strongly depend on extra frames or fail in dealing with large factor event SR. This paper presents a new method to super-resolve the event stream even by a large factor, without the auxiliary of high-quality and HR frames.

3 Method

3.1 Problem Definition

The output of event cameras can be represented as a sparse stream $\mathcal{E} = \{e_k\}_{k=1}^{N_e}$, where N_e is the number of events. Each event $e_k \in \mathcal{E}$ is denoted as a four-element tuple (x_k, y_k, t_k, p_k), representing spatial coordinates, timestamp and polarity respectively. The problem of event SR is to predict HR event stream based on LR event stream. Specifically, we denote LR event stream as $\mathcal{E}^{\mathrm{L}} = \left\{e_k^{\mathrm{L}}(x_k^{\mathrm{L}}, y_k^{\mathrm{L}})\right\}$ (timestamp and polarity are omitted here for brevity), where $x_k^{\mathrm{L}} \in [1, W^L]$, $y_k^{\mathrm{L}} \in [1, H^L]$. The goal of event SR is to obtain HR event stream $\mathcal{E}^{\mathrm{H}} = \left\{e_k^{\mathrm{H}}(x_k^{\mathrm{H}}, y_k^{\mathrm{H}})\right\}$ using LR event stream $\mathcal{E}^{\mathrm{L}}$. The spatial coordinates of e_k^{H} are subject to $x_k^{\mathrm{H}} \in [1, W^H]$, $y_k^{\mathrm{H}} \in [1, H^H]$.

The event stream is essentially sparse and spatiotemporal data that is different from natural images. Usually, a three-stage solution [12, 58] is utilized to super-resolve the event stream. First, the temporal dimension of LR event stream is reduced by counting the event number to get a 2D LR Event Count Map (ECM), which describes the spatial distribution. Then the LR ECM is further processed by the designed algorithm, generating the HR ECM. Finally, the temporal distribution can be restored by randomly or uniformly assigning the timestamps according to the HR ECM, yielding the final HR event stream.

3.2 Overall Pipeline

In this paper, we construct an upsampling-refinement pipeline for event SR. Specifically, we first upsample LR ECM^{L} by counting the event number of LR event stream $\mathcal{E}^{\mathrm{L}}$ to acquire coarse SR ECM: $ECM_{coarse}^{\mathrm{SR}} = \mathrm{Upsample}(ECM^{\mathrm{L}})$. Then the refinement is performed on coarse $ECM_{coarse}^{\mathrm{SR}}$ for producing fine ECM: $ECM_{fine}^{\mathrm{SR}} = \mathrm{Refine}(ECM_{coarse}^{\mathrm{SR}})$, which is then redistributed for HR sparse event stream. We illustrate the overview of upsampling-refinement pipeline in Fig. 1.

The commonly-used upsampling tool in frame-based vision, bicubic, can be a natural choice for the operator Upsample$(\cdot)$. However, bicubic as an interpolation method derives a new value for a new coordinate, inevitably introducing interpolation noise that may be harmful for event SR as shown in Fig. 4.

As an alternative, we develop a simpler yet effective upsampling tool tailored for event data, named coordinate relocation. In contrast to bicubic, coordinate relocation is directly performed in event domain, which is noise-free and preserves the spatiotemporal distribution of the input event stream. Specifically, coordinate relocation is used to convert LR event stream $\mathcal{E}^{\mathrm{L}}$ to SR event stream: $\mathcal{E}_{cr}^{\mathrm{SR}} = \mathrm{CoordinateRelocate}(\mathcal{E}^{\mathrm{L}})$. The spatial coordinates of $\mathcal{E}_{cr}^{\mathrm{SR}}$ can be calculated as: $x_{cr}^{\mathrm{SR}} = \mathrm{Round}\left(\frac{x^{\mathrm{L}}}{W^{\mathrm{L}}} \cdot W^{\mathrm{H}}\right)$, $y_{cr}^{\mathrm{SR}} = \mathrm{Round}\left(\frac{y^{\mathrm{L}}}{H^{\mathrm{L}}} \cdot H^{\mathrm{H}}\right)$. The operator $\mathrm{Round}(\cdot)$ is employed to convert the derived coordinates x_{cr}^{SR}, y_{cr}^{SR} to integer values. We then convert $\mathcal{E}_{cr}^{\mathrm{SR}}$ to the coarse SR $ECM_{coarse}^{\mathrm{SR}}$ by counting the event number, which is further enhanced by the refinement network. We provide experimental comparisons between bicubic and coordinate relocation in Sec. 5.

3.3 Recurrent Neural Network for Event SR

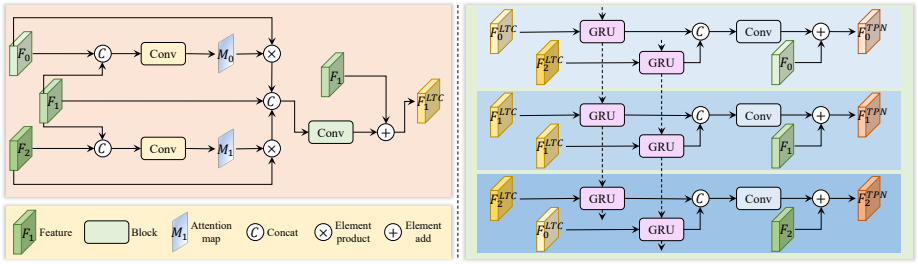
Through upsampling operation, we obtain the coarse SR $ECM_{coarse}^{\mathrm{SR}}$. As shown in Fig. 1, $ECM_{coarse}^{\mathrm{SR}}$ is still severely corrupted, calling for further detail restoration to approach the ground-truth ECM^{GT}. To achieve this, we propose a ***Rec***urrent neural network for ***Ev***ent stream ***S***uper-***R***esolution, termed as ***RecEvSR***, to model the internal spatiotemporal correlation of event stream. We demonstrate the overview network in Fig. 1. Our RecEvSR consists of three elaborately designed components, i.e., pyramidal feature extractor, temporal propagation net and spatiotemporal fusion net. In the following, we elaborate the motivations of designing these network components.

Pyramidal Feature Extractor (PFE). After upsampling the LR event stream $\mathcal{E}^{\mathrm{L}}$, we consider the sequence of $ECM_{coarse}^{\mathrm{SR}}$ as the input to the pyramidal feature extractor. The potential reason of choosing a sequence is related to the balance between expensive computation of event-by-event processing and temporal correlation loss of event-by-count processing. As aforementioned in Sec. 3.2, producing ECM is computationally tractable, yet inevitably introduces temporal correlation loss. Therefore, all $ECM_{coarse}^{\mathrm{SR}}$ in a sequence are utilized as a remedy for partially recovering the lost temporal correlation in a single ECM. The number of ECM in a sequence is 3 for all experiments (as in Fig. 1), which can be increased for better results but with high training cost.

Specifically, our pyramidal feature extractor consists of a head and three stacked convolutional blocks. The head is employed to transform each $ECM_{coarse}^{\mathrm{SR}}$ in a sequence to a high-dimensional feature while keeping the spatial resolution. Subsequently, three stacked convolutional blocks further embed the sequence of high-dimensional features while reducing the spatial resolution by $2\times$ step by step. In such a way, we finally obtain the deep pyramidal features at different timestamps, forming a feature sequence

$$Seq[F_{PFE}] = PFE(Seq[ECM_{coarse}^{\mathrm{SR}}]), \tag{1}$$

where $Seq[\cdot]$ denotes a sequence of specified components. We omit the timestamp for brevity.



(a) Local temporal correlation module (b) Global temporal correlation module

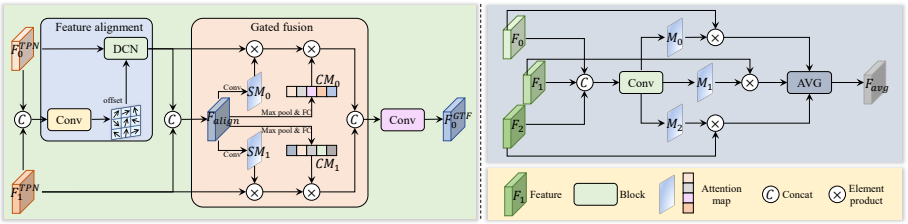
Fig. 2. The details of temporal propagation net, which consists of a local temporal correlation module and a global temporal correlation module. Zoom in for best view.

Notably, the pyramidal feature sequence generated by PFE implies two kinds of hidden clues: the temporal clues encoded by the same-scale features at different timestamps and the spatial clues encoded by the different-scale features at the same timestamp. We need to answer two questions here: **1)** how to excavate the temporal clues? **2)** how to aggregate the spatiotemporal clues?

Temporal Propagation Net (TPNet). At the beginning, let's take the first question into account. As discussed above, PFE is utilized to embed the input sequence of $ECM^{\mathrm{SR}}_{coarse}$ to acquire the deep pyramidal features at each timestamp. However, the temporal correlation among the deep pyramid features over different timestamps are not considered explicitly or implicitly, inevitably resulting in loss of intersected temporal event-aware information that favorably promotes event SR. To solve this problem, we build a temporal propagation net to further recover the lost intersected temporal clues.

Inspired by [60, 63], we design the temporal propagation net by constructing two separated modules: local temporal correlation module (LTC) for locally modeling short-term temporal clues, and global temporal correlation module (GTC) for globally modeling long-term temporal clues. As shown in Fig. 2, for each timestamp, LTC take as input the feature sequence $Seq[F_{PFE}] : \{F_0, F_1, F_2\}$ generated by PFE. For efficient processing, we only utilize the features with the smallest scale. We then use two convolutional blocks to produce two spatial attention maps $\{M_0, M_1\}$, which describe the spatial reliability of $\{F_0, F_2\}$. The rectified boundary features and central feature are concatenated and fused with a residual connection to obtain the final output F_1^{LTC}. For GTC, we choose the bidirectional temporal propagation [60,63] built upon GRU [8], for fully exploring the global temporal information of event streams. First bidirectional local temporal feature sequences: $\{F_0^{LTC}, F_1^{LTC}, F_2^{LTC}\}$ and $\{F_2^{LTC}, F_1^{LTC}, F_0^{LTC}\}$ are generated, which are then fed into GRU. We concatenate the outputs of GRU at each timestamp and then embed them before adding the original central feature sequence $\{F_0, F_1, F_2\}$ to obtain the final output $Seq[F_{TPN}] : \{F_0^{TPN}, F_1^{TPN}, F_2^{TPN}\}$. The forward process of $TPNet$ can be formulated as

$$Seq[F_{TPN}] = GTC(LTC(Seq[F_{PFE}])). \tag{2}$$



(a) Gated temporal fusion module (b) Adaptive spatiotemporal fusion module

Fig. 3. The details of spatiotemporal fusion net, which consists of a gated temporal fusion module and an adaptive spatiotemporal module. Zoom in for best view.

The designed LTC module implemented by attention mechanism is responsible for adaptively incorporating the neighboring event-aware contexts, which facilitates event SR from a local range. Exploiting recursive state update mechanism, GTC module is capable of implicitly embedding spatiotemporal clues of event stream into internal memories of the model for effective propagation in a long-range event sequence, thus favorably boosting event SR from a global range. The resultant F_{TPN} preserves the local and global temporal correlation simultaneously, capable of recovering fine-grained event-aware details.

Spatiotemporal Fusion Net (STFNet). We then answer the second question. As aforementioned, the feature sequence $Seq[F_{TPN}]$ generated by $TPNet$ captures intersected temporal clues both locally and globally. Nevertheless, each F_{PFE} in $Seq[F_{PFE}]$ maintains the unique spatiotemporal context details from currently fired events that are not included in other F_{PFE}. In order to acquire the embedded representation at the central timestamp, we design a spatiotemporal fusion net, which consists of a gated temporal fusion module (GTF) and an adaptive spatiotemporal fusion module ($ASTF$). Particularly, GTF aggregates the $Seq[F_{TPN}]$ by $TPNet$ using feature alignment and attention-based gated fusion for producing reliable output, while $ASTF$ is responsible for progressively aligning pyramidal sequence $Seq[F_{PFE}]$ by PFE using adaptive selection, meanwhile forming a skip connection for favoring network training (see Fig. 1).

As shown in Fig. 3, for GTF, we fuse the outputs by $TPNet$: $\{F_0^{TPN}, F_1^{TPN}\}$ via a two-stage process. First, we align the feature F_0^{TPN} to the central timestamp using the deformable neural network (DCN) [10,72]. Then, we concatenate the output of DCN with the central feature F_1^{TPN} to further produce the spatial attention maps $\{SM_0, SM_1\}$ and channel attention maps $\{CM_0, CM_1\}$ for fusing features in both spatial and channel dimensions. For $ASTF$, we use the skip connection to aggregate the outputs of PFE and $TPNet$. Specifically, for each scale (three scales in PFE), the feature sequence generated by PFE is first concatenated and then embedded by a convolutional block to produce three spatial attention maps $\{M_0, M_1, M_2\}$, representing the reliability of input features at different timestamps. We then average the weighted features before adding the aligned feature by GTF for final output as shown in Fig. 1. Mathematically, the

central feature can be derived as

$$F_C = STFNet(Seq[F_{PFE}], Seq[F_{TPN}]). \tag{3}$$

The resultant F_C is then employed to reconstruct the ECM^{SR}_{fine}, which is further redistributed to the final sparse event stream.

Objective Function. We train our RecEvSR in a sequence clip, of which the length is further investigated in Sec. 5. Given the reconstructed $ECM^{\mathrm{SR}}_{fine,t}$ and its ground-truth ECM^{GT}_{t} at timestamp t, we define the loss function $\mathcal{L}$ as

$$\mathcal{L} = \sum_{t=1}^{T} MSE(ECM^{\mathrm{SR}}_{fine,t}, ECM^{\mathrm{GT}}_{t}), \tag{4}$$

where T denotes the number of ECM^{SR}_{fine} in a sequence clip and $MSE(\cdot)$ represents the mean square error function.

4 Experimental Results

To validate the effectiveness of our proposed method, we conduct comprehensive experiments on both synthetic and real-world datasets, and evaluate the performance in both quantitative and qualitative ways.

Datasets and settings. ENFS-real [12] is the first real-world dataset involving multi-scale LR-HR pairs for event SR, captured by a display-camera system. However, the resolution of this dataset is limited by the capturing devices and 8(16)× data pairs are not developed. As a remedy, based on NFS [29], we first generate multi-scale frames and then convert them to events using the event simulator [16] for building a new synthetic dataset, termed as ENFS-syn, which involves 2(4, 8, 16)× LR-HR pairs for training and testing. Our ENFS-syn contains 161 sequences for 65 scenes. The duration of each sequence is no more than 30 seconds. The maximum resolution is 1280 × 720, while the minimum resolution is 80 × 45. Moreover, we also utilize the HR frames from RGB-DAVIS [58] to synthesize another synthetic dataset called RGB-DAVIS-syn. Each aforementioned dataset is randomly splitted for training and testing. Following [12,58], we use RMSE as the evaluation metric. We also apply random horizontal, vertical and polarity flip for data augmentation in training. Please see the supplementary document for details of synthetic datasets and more experimental settings.

Baselines. We make comparisons with EventZoom [12], the first learning-based method for approaching event SR. We use the code provided by the project to conduct all experiments. We retrained the E2I module in EventZoom to construct the whole architecture as suggested. For large factor SR, we run the EventZoom-2× model multiple times, which is also adopted in [12]. We have tried to train a single EventZoom for large factor SR but failed due to expensive training cost of 3D-UNet. Moreover, we construct a variant of EventZoom, termed as EventZoom-cr, by combining EventZoom-1x [12] and coordinate relocation upsampling. As for other event SR methods [33, 34, 58], they either need frames as an auxiliary or only fit simple scenes, posing a challenge to make a fair comparison with them. We also make comparisons with the representative

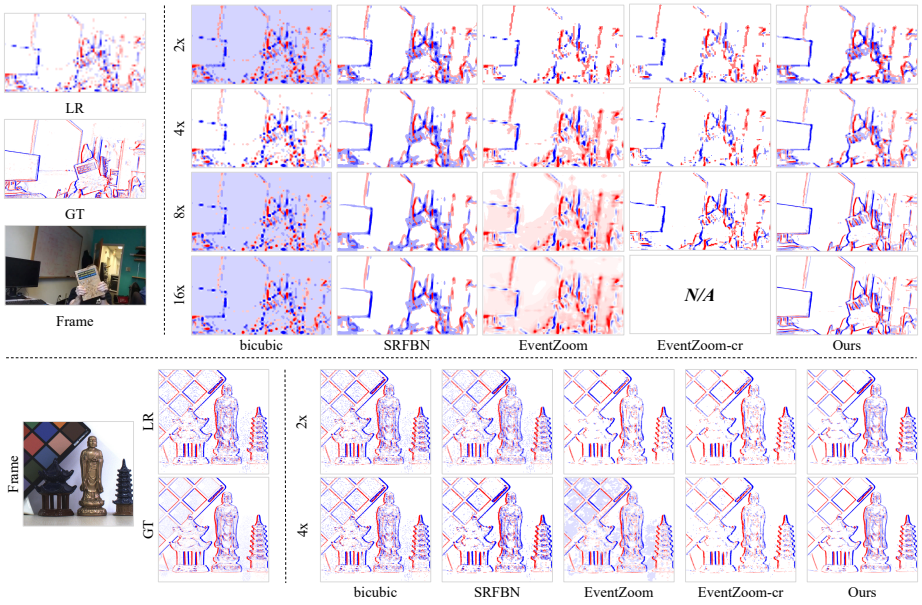


Fig. 4. Visual comparisons on synthetic datasets among bicubic, SRFBN [35], EventZoom [12], EventZoom-cr and ours. The first case (upper) is from ENFS-syn and the second case (below) is from RGB-DAVIS-syn. The 16× SR results of EventZoom-cr are not provided due to high training cost. Blue/red regions denote positive/negative events. Obviously for large factor SR, with severely corrupted LR events as the input, our method still recovers perceptually fine details. Bicubic introduces interpolation noise. SRFBN [35] cannot estimate the visual-satisfying results and EventZoom [12] fires wrong events. Zoom in for best view.

Table 1. Quantitative comparisons among bicubic, SRFBN [35], EventZoom [12], EventZoom-cr and ours on synthetic and real-world datasets in terms of RMSE. Best in bold.

Methods	2×	4×	8×	16×	2×	4×	2×	4×
	ENFS-syn				**RGB-DAVIS-syn**		**ENFS-real**	
bicubic	0.821	0.784	0.791	0.764	0.387	0.378	0.899	0.969
SRFBN	0.694	0.690	0.708	0.678	0.366	0.362	0.669	0.753
EventZoom	0.843	1.036	2.385	5.970	0.583	1.100	0.773	0.910
EventZoom-cr	0.844	0.833	0.823	-	0.604	0.614	0.775	0.828
Ours	**0.686**	**0.653**	**0.617**	**0.582**	**0.352**	**0.329**	**0.663**	**0.663**

methods of frame-based SR, i.e., bicubic and SRFBN [35], which are omitted in [12]. It should be noted that directly applying the framed-based SR methods to super-resolve event streams may fail. For example, we found it hard to train the representative video SR method, RBPN [21], for the event SR task. We attribute it to two reasons: 1) value of ECM represents spatial distribution that is unlimited, while value of frame is typically no more than 255; 2) ECM is sparse

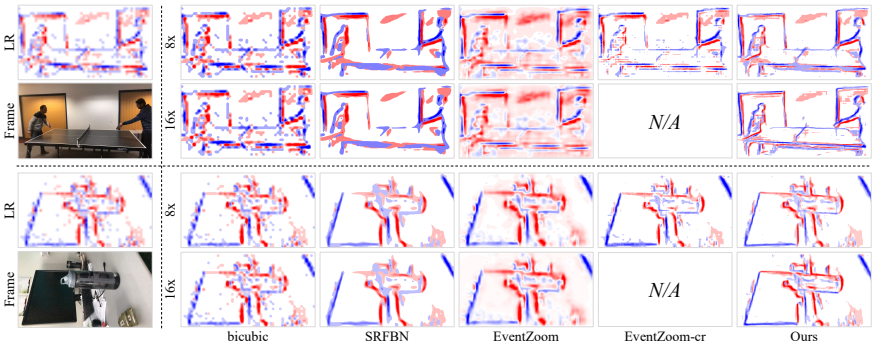


Fig. 5. Visual comparisons for large factor SR (8(16) ×) on real-world dataset (ENFS-real) among bicubic, SRFBN [35], EventZoom [12], EventZoom-cr and ours. The 16× SR results of EventZoom-cr are not provided due to high training cost. Obviously, we achieve superior performance for large factor SR compared with baselines, though our network is trained using synthetic datasets, which demonstrates our outstanding generalization ability against baselines. Zoom in for best view.

and primarily contains edge information, while frame is dense and reflects more conceptual contexts. The essential discrepancy of event data and frame data motivates us to design the specific algorithm for event SR instead of direct adoption of existing frame-based SR methods.

Results on synthetic datasets. We present the quantitative comparison results on the synthetic datasets, i.e., ENFS-syn and RGB-DAVIS-syn, in Tab. 1. Compared with EventZoom, in terms of RMSE, our method achieves near 30% performance boost on average for 2(4)× SR on two datasets. For 8(16)× SR, our method presents clear superiority over EventZoom, yielding over 80% average RMSE gain on ENFS-syn. Compared with frame-based methods, our method still performs favorably against bicubic and SRFBN especially for 8(16)× SR, achieving a relative gain of over 19% in terms of RMSE on ENFS-syn. Furthermore, EventZoom-cr with coordinate relocation upsampling is able to significantly boost EventZoom for large factor SR, achieving 65.49% and 44.18% RMSE gain for 8× SR on ENFS-syn and 4× SR on RGB-DAVIS-syn, respectively. For qualitative comparison, we visualize the super-resolved event streams of all methods in Fig. 4. We can see that our method can restore perceptually better texture details and sharper edges from the severely corrupted LR event stream, accurately presenting the complex scene motion variation, especially for 8(16)× SR. The interpolation noise of bicubic can be obviously observed in Fig. 4, leading to harmful interference for event SR. Notably, although the numerical results in Tab. 1 show that EventZoom underperforms bicubic, the visual results of EventZoom in Fig. 4 are better than those of bicubic. We present more visual results in the supplementary document.

Results on real-world dataset. For real-world evaluation on ENFS-real [12], we give the quantitative results in Tab. 1. In terms of RMSE for 2(4)× SR, our method achieves performance boost on average of 20% over EventZoom

Table 2. Quantitative results of additive noise evaluation in terms of RMSE.

Methods	0	10%	20%	30%
bicubic	0.784	0.797	0.812	0.827
SRFBN	0.689	0.692	0.696	0.701
EventZoom	1.036	1.063	1.082	1.093
EventZoom-cr	0.833	0.830	0.829	0.831
Ours	**0.653**	**0.659**	**0.667**	**0.676**

Table 3. Ablation on upsampling method and recurrent neural network. "bi" means bicubic, "cr" means coordinate relocation.

Methods	2×	4×	8×
	RMSE ↓		
EventZoom	0.843	1.036	2.385
EventZoom-bi	0.845	0.833	0.838
EventZoom-cr	0.844	0.833	0.823
RecEvSR-bi	0.694	0.655	0.619
RecEvSR-cr (Ours)	**0.686**	**0.653**	**0.617**
	upsampling time (ms)		
bicubic	57.6	217.3	892.2
coordinate relocation	**0.2**	**0.2**	**0.2**

Table 4. Ablation on network components. RMSE value is reported.

Variants	LTC	GTC	GTF	ASTF	RMSE
model#A	×	✓	✓	✓	0.704
model#B	✓	×	✓	✓	0.697
model#C	✓	✓	×	✓	0.701
model#D	✓	✓	✓	×	0.698
model#E	✓	✓	✓	✓	**0.695**

Table 5. Ablation on training settings. RMSE value is reported.

Metrics	Sequence length				Augmentation	
	3	6	9	12	w/o aug	w/ aug
RMSE	0.699	0.691	0.689	**0.687**	0.693	**0.688**

(EventZoom-cr) and 17% over frame-based methods. For 8(16)× SR, we cannot provide the quantitative comparisons due to no ground-truth data in the ENFS-real. Therefore, we only exhibit the visual results in Fig. 5, using the pre-trained models on our synthetic ENFS-syn to super-resolve the LR event streams in ENFS-real. It can be clearly observed from Fig. 5 that, although with the severely-corrupted LR event stream as the input, we achieve the visually-satisfying real-world 8(16)× SR results with fine-grained textures against baselines, presenting the strong generalization over baselines. More real-world visual results can be found in the supplementary document.

Results of noise robustness evaluation. We also conduct noise robustness evaluation on ENFS-syn by adding the extra random noise into input event stream. We manually control the noise level, the percentage of input event number, to investigate the effect of noise for different methods. We conduct 4× SR for this experiment. Tab. 2 presents the numerical results. As can be seen, our method achieves best results compared with others, demonstrating the satisfying noise robustness of our method.

5 Ablation Study

In this section, we present more experimental analysis of our method from four aspects: upsampling method, recurrent neural network, recurrent network components and training settings. Before presenting the detailed results, we describe a nomenclature for the variants. The names of variants follow the pattern "A-B", where "B" represents the upsampling method to spatially zooming LR events and "A" represents the backbone network for further refinement. We conduct the ablation experiments on ENFS-syn.

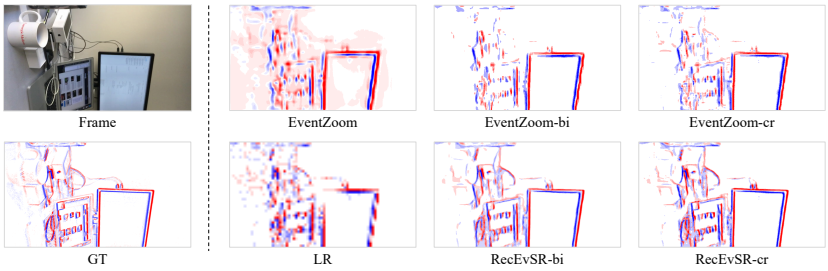


Fig. 6. Visual comparisons of ablation on upsampling method and recurrent network. "bi" means bicubic, "cr" means coordinate relocation. 4× SR results on ENFS-syn are presented. Zoom in for best view.

Ablation on upsampling method. Compared with bicubic upsampling, here we investigate how well our coordinate relocation behaves from two perspectives: **1**) Can coordinate relocation perform better than bicubic? For fair comparisons, we keep the same refinement network except upsampling method. We choose EventZoom and our proposed RecEvSR as refinement network to explore how these two upsampling tools perform. As shown in Tab. 3, EventZoom-cr/RecEvSR-cr perform favorably against EventZoom-bi/RecEvSR-bi for 2(4, 8)× SR in terms of RMSE, reinforcing the effectiveness of coordinate relocation. Particularly, coordinate relocation is more efficient than bicubic in terms of upsampling time, as shown in Tab. 3. The visual results in Fig. 6 show that refinement network with coordinate relocation is able to provide more edges that are suppressed by with bicubic. **2**) Can coordinate relocation be combined with other methods and boost them? In order to validate the generality of coordinate relocation and if it synthesizes well with other methods, we choose "A" as EventZoom. We show the quantitative results in Tab. 3. As can be seen, EventZoom-cr shows favorable performance against EventZoom especially for 4(8)× SR, demonstrating the generality of coordinate relocation to boost other methods. The visual results in Fig. 6 further present that EventZoom-cr provides more perceptually-satisfying details against EventZoom.

Ablation on recurrent neural network. As aforementioned in Sec. 3.3, we build a recurrent neural network to model the internal spatiotemporal correlation of event stream by exploiting the recursive state update mechanism. In order to validate the effectiveness of our RecEvSR, we keep the same upsampling method "A". In such a way, the only different part is the choice of refinement network. It can be clearly observed from Tab. 3 that, RecEvSR-bi/RecEvSR-cr consistently surpasses EventZoom-bi/EventZoom-cr in terms of RMSE, demonstrating the superiority of our RecEvSR. The results in Fig. 6 also provide visual supports.

Ablation on recurrent network components. In order to validate the effectiveness of the designed components of our recurrent neural network, we ablate each sub-network to form different variants to conduct the experiments on ENFS-syn. As shown in Tab. 4, our network with all sub-networks (model#E) achieves the best performance compared with other variants for 2× SR.

Table 6. Quantitative results of downstream applications among bicubic, SRFBN [35], EventZoom [12], EventZoom-cr and ours on NCars [47] for object recognition and on ENFS-syn for video reconstruction. We evaluate object recognition using area under curve (AUC) and recognition accuracy (ACC). For video reconstruction, we show SSIM and LPIPS values. Best in bold, the runner up with underline.

Methods	object recognition					
	2×		4×		8×	
	AUC↑	ACC↑	AUC↑	ACC↑	AUC↑	ACC↑
bicubic	57.25	56.46	56.46	55.66	51.11	50.08
SRFBN	57.39	56.54	56.64	55.82	<u>51.32</u>	<u>50.29</u>
EventZoom	55.98	54.99	50.91	49.92	49.93	48.85
EventZoom-cr	<u>60.24</u>	<u>59.44</u>	<u>57.74</u>	<u>56.94</u>	50.59	49.58
Ours	**63.65**	**62.85**	**62.94**	**62.24**	**53.30**	**52.23**
Ref.	85.29	85.23	93.20	93.38	95.33	95.31
	video reconstruction					
Methods	SSIM↑	LPIPS↓	SSIM↑	LPIPS↓	SSIM↑	LPIPS↓
bicubic	0.562	0.399	0.615	0.516	<u>0.607</u>	0.577
SRFBN	<u>0.596</u>	0.397	0.605	0.493	0.602	<u>0.534</u>
EventZoom	0.555	0.413	0.586	0.480	0.582	0.563
EventZoom-cr	0.583	<u>0.393</u>	**0.657**	<u>0.434</u>	0.593	0.548
Ours	**0.609**	**0.375**	<u>0.643</u>	**0.422**	**0.626**	**0.473**

Ablation on training settings. 1) Sequence length. In order to investigate the influence of length of training sequence clip, we conduct ablation experiments on ENFS-syn and show the 2× SR results in Tab. 5. Obviously, the longer the training sequence is, the better results we can achieve. It implies that a longer training sequence may provide more reliable hidden states, which can be exploited by our recurrent network. However, longer training sequences result in high training cost, thus we choose 9 as the sequence length in all our experiments. **2)** Data augmentation. We also conduct the ablation experiments on the data augmentation as discussed in Sec. 4. As shown in Tab. 5, when disabling the data augmentation, the network shows the performance drop for 2× SR.

6 Downstream Event-driven Applications

Object recognition. We investigate the performance of bicubic, SRFBN [35], EventZoom [12], EventZoom-cr and our method on object recognition application. One popular event-based dataset: NCars [47] are utilized for experiments. We utilize the coordinate relocation reverse operation to down-sample the original event stream for 8×. After that, we perform different SR methods to upsample the LR event stream for 2(4, 8)×. Then we conduct object recognition using the benchmark classifier proposed in [17]. Tab. 6 presents the evaluation results. We report area under curve (AUC) and accuracy (ACC) for evaluation. The row "Ref." means using the event stream directly down-sampled from the original event stream, which is the upper-bound. It can be observed from row "Ref." that AUC (ACC) intensifies as the resolution of the input event stream increases, demonstrating that higher resolution gives rise to better per-

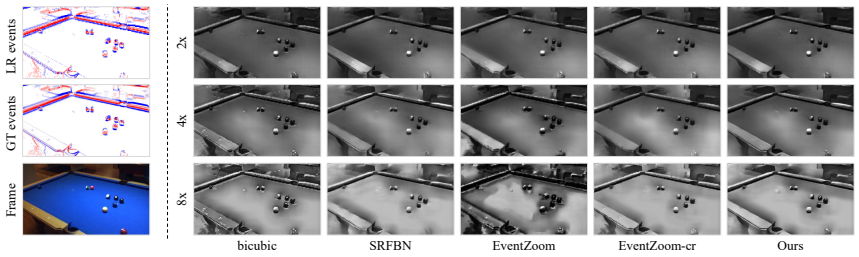


Fig. 7. Visual comparisons of video reconstruction among bicubic, SRFBN [35], EventZoom [12], EventZoom-cr and ours on ENFS-syn. Zoom in for best view.

formance. As for the performance of different SR methods, we can see that our RecEvSR achieves the best performance compared with other methods. Furthermore, EventZoom-cr outperforms EventZoom by a large margin especially for 4(8)× SR, validating the effectiveness of upsampling with coordinate relocation.
Video reconstruction. We also evaluate bicubic, SRFBN [35], EventZoom [12], EventZoom-cr and our method on video reconstruction application. ENFS-syn is employed for this task, because it provides the synchronized ground-truth frames for comparison. The E2VID [45] is chosen as the benchmark algorithm for event-to-video reconstruction with the evaluation metrics of SSIM and LPIPS. Tab. 6 shows the numerical results. Obviously, our method achieves the best result in terms of SSIM and LPIPS for 2(4, 8)× SR except that EventZoom-cr shows best SSIM for 4× SR. Comparing EventZoom with EventZoom-cr, we can see the significant performance boost achieved by EventZoom-cr, further validating the superiority of the combination of EventZoom and coordinate relocation on video reconstruction. The visual results in Fig. 7 clearly show that our method achieves perceptually fine details, in contrast to the artifacts produced by bicubic and EventZoom. We present more visual results in the supplementary document.

7 Conclusion

In this paper, we propose a recurrent neural network for event SR without assistance of frames, which suits for large factor SR. Two elaborate components, i.e., a temporal propagation net and a spatiotemporal fusion net, are built, leading to effective correlation and aggregation of event-aware contexts that enhance event SR. We demonstrate the visually-satisfying event SR results even up to 16× both on synthetic and real-world datasets and validate the superiority of our method against the state-of-the-art methods with extensive experiments, quantitatively and qualitatively. Superior performance is also achieved by our method on two downstream event-driven tasks.
Acknowledgements. We acknowledge funding from National Key R&D Program of China under Grant 2017YFA0700800, National Natural Science Foundation of China under Grants 61901435, 62131003 and 62021001.

References

1. Amir, A., Taba, B., Berg, D., Melano, T., McKinstry, J., Di Nolfo, C., Nayak, T., Andreopoulos, A., Garreau, G., Mendoza, M., et al.: A low power, fully event-based gesture recognition system. In: CVPR (2017) 1
2. Bevilacqua, M., Roumy, A., Guillemot, C., Alberi-Morel, M.L.: Low-complexity single-image super-resolution based on nonnegative neighbor embedding (2012) 3
3. Bi, Y., Chadha, A., Abbas, A., Bourtsoulatze, E., Andreopoulos, Y.: Graph-based object classification for neuromorphic vision sensing. In: ICCV (2019) 1
4. Brandli, C., Berner, R., Yang, M., Liu, S.C., Delbruck, T.: 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
5. Caballero, J., Ledig, C., Aitken, A., Acosta, A., Totz, J., Wang, Z., Shi, W.: Real-time video super-resolution with spatio-temporal networks and motion compensation. In: CVPR (2017) 3
6. Chan, K.C., Wang, X., Yu, K., Dong, C., Loy, C.C.: Basicvsr: The search for essential components in video super-resolution and beyond. In: CVPR (2021) 3
7. Chang, H., Yeung, D.Y., Xiong, Y.: Super-resolution through neighbor embedding. In: CVPR (2004) 3
8. Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using rnn encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078 (2014) 6
9. Choi, J., Yoon, K.J., et al.: Learning to super resolve intensity images from events. In: CVPR (2020) 1
10. Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y.: Deformable convolutional networks. In: ICCV (2017) 7
11. Dong, C., Loy, C.C., He, K., Tang, X.: Learning a deep convolutional network for image super-resolution. In: ECCV (2014) 3
12. Duan, P., Wang, Z.W., Zhou, X., Ma, Y., Shi, B.: Eventzoom: Learning to denoise and super resolve neuromorphic events. In: CVPR (2021) 1, 2, 3, 4, 8, 9, 10, 13, 14
13. Gallego, G., Delbruck, T., Orchard, G.M., Bartolozzi, C., Taba, B., Censi, A., Leutenegger, S., Davison, A., Conradt, J., Daniilidis, K., Scaramuzza, D.: Event-based vision: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence pp. 1–1 (2020). https://doi.org/10.1109/TPAMI.2020.3008413 1
14. Gallego, G., Lund, J.E., Mueggler, E., Rebecq, H., Delbruck, T., Scaramuzza, D.: Event-based, 6-dof camera tracking from photometric depth maps. IEEE transactions on pattern analysis and machine intelligence **40**(10), 2402–2412 (2017) 1
15. Gallego, G., Rebecq, H., Scaramuzza, D.: A unifying contrast maximization framework for event cameras, with applications to motion, depth, and optical flow estimation. In: CVPR (2018) 1
16. Gehrig, D., Gehrig, M., Hidalgo-Carrió, J., Scaramuzza, D.: Video to events: Recycling video datasets for event cameras. In: CVPR (2020) 8
17. Gehrig, D., Loquercio, A., Derpanis, K.G., Scaramuzza, D.: End-to-end learning of representations for asynchronous event-based data. In: ICCV (2019) 13
18. Gerstner, W., Kistler, W.M.: Spiking neuron models: Single neurons, populations, plasticity. Cambridge university press (2002) 2
19. Glasner, D., Bagon, S., Irani, M.: Super-resolution from a single image. In: ICCV (2009) 3

20. Gu, C., Learned-Miller, E., Sheldon, D., Gallego, G., Bideau, P.: The spatio-temporal poisson point process: A simple model for the alignment of event camera data. In: ICCV (2021) 1
21. Haris, M., Shakhnarovich, G., Ukita, N.: Recurrent back-projection network for video super-resolution. In: CVPR (2019) 9
22. He, W., You, K., Qiao, Z., Jia, X., Zhang, Z., Wang, W., Lu, H., Wang, Y., Liao, J.: Timereplayer: Unlocking the potential of event cameras for video interpolation. In: CVPR (2022) 1
23. Heist, S., Zhang, C., Reichwald, K., Kühmstedt, P., Notni, G., Tünnermann, A.: 5d hyperspectral imaging: fast and accurate measurement of surface shape and spectral characteristics using structured light. Optics express **26**(18), 23366–23379 (2018) 3
24. Huang, J.B., Singh, A., Ahuja, N.: Single image super-resolution from transformed self-exemplars. In: CVPR (2015) 3
25. Isobe, T., Jia, X., Gu, S., Li, S., Wang, S., Tian, Q.: Video super-resolution with recurrent structure-detail network. In: ECCV. Springer (2020) 3
26. Isobe, T., Zhu, F., Jia, X., Wang, S.: Revisiting temporal modeling for video super-resolution. BMVC (2020) 3
27. Jiang, Z., Zhang, Y., Zou, D., Ren, J., Lv, J., Liu, Y.: Learning event-based motion deblurring. In: CVPR (2020) 1
28. Jo, Y., Oh, S.W., Kang, J., Kim, S.J.: Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation. In: CVPR (2018) 3
29. Kiani Galoogahi, H., Fagg, A., Huang, C., Ramanan, D., Lucey, S.: Need for speed: A benchmark for higher frame rate object tracking. In: ICCV (2017) 8
30. Kim, H., Leutenegger, S., Davison, A.J.: Real-time 3d reconstruction and 6-dof tracking with an event camera. In: ECCV (2016) 1
31. Kim, M.H., Harvey, T.A., Kittle, D.S., Rushmeier, H., Dorsey, J., Prum, R.O., Brady, D.J.: 3d imaging spectroscopy for measuring hyperspectral patterns on solid objects. ACM Transactions on Graphics (TOG) **31**(4), 1–11 (2012) 3
32. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: CVPR (2017) 3
33. Li, H., Li, G., Shi, L.: Super-resolution of spatiotemporal event-stream image. Neurocomputing **335**, 206–214 (2019) 1, 2, 3, 4, 8
34. Li, S., Feng, Y., Li, Y., Jiang, Y., Zou, C., Gao, Y.: Event stream super-resolution via spatiotemporal constraint learning. In: ICCV (2021) 1, 2, 3, 4, 8
35. Li, Z., Yang, J., Liu, Z., Yang, X., Jeon, G., Wu, W.: Feedback network for image super-resolution. In: CVPR (2019) 9, 10, 13, 14
36. Lim, B., Son, S., Kim, H., Nah, S., Mu Lee, K.: Enhanced deep residual networks for single image super-resolution. In: CVPRW (2017) 3
37. Lin, S., Zhang, J., Pan, J., Jiang, Z., Zou, D., Wang, Y., Chen, J., Ren, J.: Learning event-driven video deblurring and interpolation. In: ECCV (2020) 1
38. Liu, D., Parra, A., Chin, T.J.: Globally optimal contrast maximisation for event-based motion estimation. In: CVPR (2020) 1
39. Messikommer, N., Gehrig, D., Loquercio, A., Scaramuzza, D.: Event-based asynchronous sparse convolutional networks. In: ECCV (2020) 1
40. Orchard, G., Meyer, C., Etienne-Cummings, R., Posch, C., Thakor, N., Benosman, R.: Hfirst: A temporal approach to object recognition. IEEE transactions on pattern analysis and machine intelligence **37**(10), 2028–2040 (2015) 1

41. Pan, L., Scheerlinck, C., Yu, X., Hartley, R., Liu, M., Dai, Y.: Bringing a blurry frame alive at high frame-rate with an event camera. In: CVPR (2019) 1
42. Paredes-Vallés, F., Scheper, K.Y., de Croon, G.C.: Unsupervised learning of a hierarchical spiking neural network for optical flow estimation: From events to global motion perception. IEEE transactions on pattern analysis and machine intelligence **42**(8), 2051–2064 (2019) 1
43. Patrick, L., Posch, C., Delbruck, T.: A 128x 128 120 db 15μ s latency asynchronous temporal contrast vision sensor. IEEE Journal of Solid-state Circuits **43**, 566–576 (2008) 1
44. Rebecq, H., Gallego, G., Scaramuzza, D.: Emvs: Event-based multi-view stereo. In: BMVC (2016) 1
45. Rebecq, H., Ranftl, R., Koltun, V., Scaramuzza, D.: High speed and high dynamic range video with an event camera. IEEE transactions on pattern analysis and machine intelligence (2019) 1, 14
46. Schaefer, S., Gehrig, D., Scaramuzza, D.: Aegnn: Asynchronous event-based graph neural networks. In: CVPR (2022) 1
47. Sironi, A., Brambilla, M., Bourdis, N., Lagorce, X., Benosman, R.: Hats: Histograms of averaged time surfaces for robust event-based object classification. In: CVPR (2018) 1, 13
48. Tian, Y., Zhang, Y., Fu, Y., Xu, C.: Tdan: Temporally-deformable alignment network for video super-resolution. In: CVPR (2020) 3
49. Timofte, R., De Smet, V., Van Gool, L.: Anchored neighborhood regression for fast example-based super-resolution. In: ICCV (2013) 3
50. Timofte, R., De Smet, V., Van Gool, L.: A+: Adjusted anchored neighborhood regression for fast super-resolution. In: ACCV (2014) 3
51. Tong, T., Li, G., Liu, X., Gao, Q.: Image super-resolution using dense skip connections. In: ICCV (2017) 3
52. Tulyakov, S., Bochicchio, A., Gehrig, D., Georgoulis, S., Li, Y., Scaramuzza, D.: Time lens++: Event-based frame interpolation with parametric non-linear flow and multi-scale fusion. In: CVPR (2022) 1
53. Wang, B., He, J., Yu, L., Xia, G.S., Yang, W.: Event enhanced high-quality image recovery. In: ECCV. pp. 155–171 (2020) 1
54. Wang, L., Ho, Y.S., Yoon, K.J., et al.: Event-based high dynamic range image and very high frame rate video generation using conditional generative adversarial networks. In: CVPR (2019) 1
55. Wang, L., Kim, T.K., Yoon, K.J.: Eventsr: From asynchronous events to image reconstruction, restoration, and super-resolution via end-to-end adversarial learning. In: CVPR (2020) 1
56. Wang, T.C., Zhu, J.Y., Kalantari, N.K., Efros, A.A., Ramamoorthi, R.: Light field video capture using a learning-based hybrid imaging system. ACM Transactions on Graphics (TOG) **36**(4), 1–13 (2017) 3
57. Wang, X., Chan, K.C., Yu, K., Dong, C., Change Loy, C.: Edvr: Video restoration with enhanced deformable convolutional networks. In: CVPRW (2019) 3
58. Wang, Z.W., Duan, P., Cossairt, O., Katsaggelos, A., Huang, T., Shi, B.: Joint filtering of intensity images and neuromorphic events for high-resolution noise-robust imaging. In: CVPR (2020) 1, 2, 3, 4, 8
59. Weng, W., Zhang, Y., Xiong, Z.: Event-based video reconstruction using transformer. In: ICCV (2021) 1
60. Xiang, X., Tian, Y., Zhang, Y., Fu, Y., Allebach, J.P., Xu, C.: Zooming slow-mo: Fast and accurate one-stage space-time video super-resolution. In: CVPR (2020) 6

61. Xiao, Z., Fu, X., Huang, J., Cheng, Z., Xiong, Z.: Space-time distillation for video super-resolution. In: CVPR (2021) 3
62. Xiao, Z., Xiong, Z., Fu, X., Liu, D., Zha, Z.J.: Space-time video super-resolution using temporal profiles. In: ACM MM (2020) 3
63. Xu, G., Xu, J., Li, Z., Wang, L., Sun, X., Cheng, M.M.: Temporal modulation network for controllable space-time video super-resolution. In: CVPR (2021) 6
64. Yang, J., Wright, J., Huang, T., Ma, Y.: Image super-resolution as sparse representation of raw image patches. In: CVPR (2008) 1, 4
65. Yang, J., Wright, J., Huang, T.S., Ma, Y.: Image super-resolution via sparse representation. IEEE transactions on image processing **19**(11), 2861–2873 (2010) 3
66. Yao, M., Xiong, Z., Wang, L., Liu, D., Chen, X.: Spectral-depth imaging with deep learning based reconstruction. Optics express **27**(26), 38312–38325 (2019) 3
67. Zeyde, R., Elad, M., Protter, M.: On single image scale-up using sparse-representations. In: International conference on curves and surfaces. pp. 711–730. Springer (2010) 3
68. Zhang, X., Liao, W., Yu, L., Yang, W., Xia, G.S.: Event-based synthetic aperture imaging with a hybrid network. In: CVPR (2021) 1
69. Zhang, X., Yu, L.: Unifying motion deblurring and frame interpolation with events. In: CVPR (2022) 1
70. Zhou, Y., Gallego, G., Rebecq, H., Kneip, L., Li, H., Scaramuzza, D.: Semi-dense 3d reconstruction with a stereo event camera. In: ECCV (2018) 1
71. Zhu, A.Z., Yuan, L., Chaney, K., Daniilidis, K.: Unsupervised event-based learning of optical flow, depth, and egomotion. In: CVPR (2019) 1
72. Zhu, X., Hu, H., Lin, S., Dai, J.: Deformable convnets v2: More deformable, better results. In: CVPR (2019) 7
73. Zihao Zhu, A., Atanasov, N., Daniilidis, K.: Event-based visual inertial odometry. In: CVPR (2017) 1