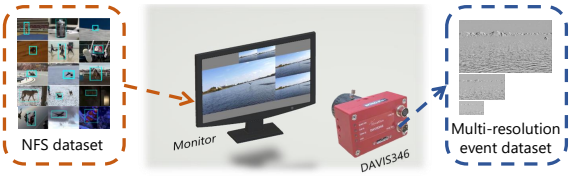
EventZoom: Learning to Denoise and Super Resolve Neuromorphic Events

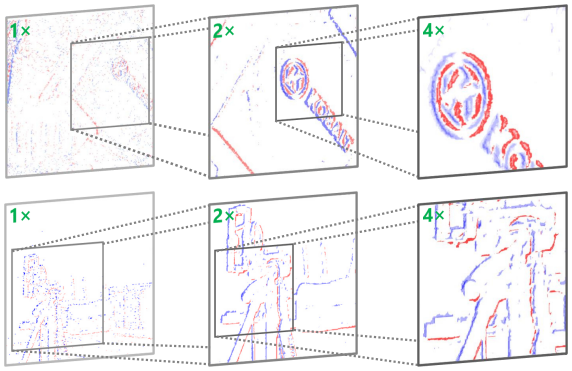
Peiqi Duan[†] Zihao W. Wang[‡] Xinyu Zhou[†] Yi Ma[†] Boxin Shi[†§] ✉
[†]NELVT, Department of Computer Science and Technology, Peking University
[‡]Department of Computer Science and Engineering, Northwestern University
[§]Institute for Artificial Intelligence, Peking University

Abstract

We address the problem of jointly denoising and super resolving neuromorphic events, a novel visual signal that represents thresholded temporal gradients in a space-time window. The challenge for event signal processing is that they are asynchronously generated, and do not carry absolute intensity but only binary signs informing temporal variations. To study event signal formation and degradation, we implement a display-camera system which enables multi-resolution event recording. We further propose EventZoom, a deep neural framework with a backbone architecture of 3D U-Net. EventZoom is trained in a noise-to-noise fashion where the two ends of the network are unfiltered noisy events, enforcing noise-free event restoration. For resolution enhancement, EventZoom incorporates an event-to-image module supervised by high resolution images. Our results showed that EventZoom achieves at least $40\times$ *temporal efficiency compared to state-of-the-art (SOTA) event denoisers. Additionally, we demonstrate that EventZoom enables performance improvements on applications including event-based visual object tracking and image reconstruction. EventZoom achieves SOTA super resolution image reconstruction results while being* $10\times$ *faster.*



(a) Our multi-resolution display-camera system.



(b) EventZoom results. Blue/red: positive/negative events.

Figure 1: We propose EventZoom, a network that performs event denoising and super resolution.

1. Introduction

Neuromorphic events are novel visual signals that address several limitations of mainstream image signals, particularly featuring low power, low latency and high dynamic range (HDR). These are due to the unique sensor design that enables each event pixel to only compare current and last intensity states in log-scale and fire a binary-signed event whenever the log-intensity variation exceeds the preset thresholds [6, 10, 24, 35, 42]. Such sensors are suitable for dynamic visual scenarios thanks to their high sensing speed. Yet event cameras are unable to reveal scene appearance, especially under static conditions. Moreover, events are fired asynchronously, resulting in spatio-temporal point clouds rather than conventional 2D image/video sequences.

As such, the problem of event signal restoration and enhancement has strong deviation from its image-based counterpart, and requires deliberate modifications when applying image-based models. A particular body of the past literature has attended to event-to-image reconstruction [2, 7, 20, 37, 38, 45, 46], and shown that image-based visual algorithms can perform comfortably well on event-reconstructed images [37]. An extended branch explored the benefit events could bring to low-level vision. Such tasks take hybrid inputs of events and images, and perform image/video enhancement in HDR and low-light imaging [15, 50], video synthesis [49], motion deblur [18, 31], and super resolution [44]. Nonetheless, executing visual tasks

✉ Corresponding author: shiboxin@pku.edu.cn
project page: https://sites.google.com/view/EventZoom

by means of event-to-image conversion consumes heavy computational power and time, calling for compact event-to-event restoration and enhancement.

SOTA event restoration and enhancement solutions rely on the intensity signal [1, 48]. Same-resolution images could be employed to label events and a classification network is further leveraged for event denoising [1]. However, the labeling of captured events can only identify and remove wrongly-fired events, while it cannot retrieve unfired events along with previous noisy event removal filters [4, 19, 27]. Guided Event Filtering (GEF) [48] enables retrieving missing events as well as $8\times$ super resolution (SR). However, GEF's performance relies heavily on 1) the quality of high-resolution (HR) images; 2) the accuracy of optical flow estimation, which is computationally expensive. When the HR images are blurry or in lack of spatial features, or when the optical flow fails to register events onto image edges, the filtering output will yield compromised quality.

We propose EventZoom, an end-to-end neural network approach for event denoising and super resolution (EDSR). EventZoom performs event-to-event transformation based on a backbone of 3D U-Net [52]. Although the network does not require image supervision, an event-to-image (E2I) module was designed to study the benefit from image signals. The E2I module is a combination of a same-resolution event-to-image reconstruction network E2VID [36] and an image SR network FSRCNN [9]. The last layer features from the two networks are used as low-resolution (LR) and HR features to be concatenated with the corresponding layers of the 3D U-Net. The input and output are 3D tensors. Two processing steps are involved for converting the raw events to a 3D tensor (event stacking) and reverting the output 3D tensor to events (event re-distributing). Overall, this paper makes the following contributions:

- EventZoom is the first network approach to solve EDSR. EventZoom was built upon 3D U-Net and incorporated an E2I module to leverage HR image information, while preserving computational efficiency.
- We implemented a display-camera system to collect a multi-resolution event dataset (Fig. 1a), and trained the network in a noise-to-noise fashion without ground truth annotation (Fig. 1b).
- EventZoom was applied to event-based visual object tracking and image reconstruction, and achieved significant performance improvement.

2. Related work

This section reviews existing work in event denoising and super resolution (EDSR), event camera systems and datasets, and related neural models for event processing.

EDSR. Existing works were mainly concerned with background activity noise produced by temporal noise and junction leakage currents [1, 4, 19, 24, 27]. Liu *et al*. [27] proposed a denoising filter based on spatiotemporal correlation. Wang *et al*. [47] proposed to filter events by their motion association likelihood. This is based on an assumption that events are triggered by edge motion and therefore shall follow the same spatiotemporal motion projection within a local window if valid [11, 40, 48]. Recently, GEF [48] has further made use of the motion compensation (MC) between the image and event signals, and solved the problem by using guided image filtering techniques. The optimization is performed by maximizing the mutual structures between the LR event and HR image signals. When the image signal has higher spatial resolution than the event signal, GEF enables super resolving the event signal up to the image resolution. Although MC is highly useful for event processing [11,25,34], the computational complexity is beyond practical for downstream visual tasks.

Another pathway for EDSR is first by means of event-to-intensity conversion [29, 46]. The generated high quality images can then be converted back to events via video-to-events simulators [8, 12, 16]. The runtime and the simulation-to-real gap [41] are the main limitations.

Event camera systems and datasets. While the majority of existing datasets have addressed various visual tasks, very few of them focused on EDSR. DVSNOISE20 [1] proposed a noise annotation approach by deriving an event probability mask using APS frames and IMU motion data. In the dataset proposed in [37], HR smartphone videos were provided as reference but were not reversed to raw data form to retrieve intensity information for the need of EDSR. Both MVSEC [51] and RGB-DAVIS [48] have provided HR machine vision images up to $2\times$ and $8\times$ respectively. Particularly, RGB-DAVIS leveraged a beamsplitter to collocate an HR RGB camera and LR DAVIS240 event camera [48]. There has not been a multi-resolution event dataset provided in the literature due to the significant challenges in camera calibration and the lack of HR event camera prototypes. In the benchmark event datasets collected in [17], a display-camera system has been implemented to convert existing video datasets, *e.g*. action recognition, to event datasets. We implement a similar setup with hardware upgrades in both the display and the event camera. To minimize the temporal aliasing induced by large motion, we chose a high frame-rate video dataset Need-for-Speed (NFS) [14].

Event neural models. Events are bio-inspired visual signals resembling the form of asynchronous neural spike trains. Therefore, several bio-inspired learning architectures have been proposed for event-based learning, including SNNs [21, 32], LSTM/RNNs [5, 30, 36], and MLPs [39,43]. CNNs are widely adopted for EDSR-related tasks,

particularly when the output is in image form. Wang *et al*. [49] proposed to use the sigmoid function to approximate the intensity-event relation, and employed a residual net for image enhancement. Before performing convolutions, the input events were first binned or stacked into event frames which induced temporal interruptions [29, 45, 49]. This issue was alleviated by explicitly incorporating inter-stack flow estimation modules [18, 29]. As shown in GEF [48], convolutional SR nets did not perform well on binned event frames as the activation sites are sparse. Messikommer *et al*. [28] adopted sparse convolutions with asynchronous activation mechanism for high-level visual tasks. Gehrig *et al*. [13] proposed volumetric spatio-temporal tensors to form an event feature space that is trained w.r.t. specific tasks.

For EDSR, we employ the 3D U-Net [52] architecture as it has a volumetric encoder-decoder structure and performs third dimension convolutions. Moreover, our work explores the benefit of incorporating the learned event-to-intensity network features.

3. Approach

Here we describe our proposed approach, EventZoom. We first demonstrate the event formation model and its relationship to the image-based counterpart.

3.1. Event formation

In image denoising and super resolution, the basic image formation model assumes that the LR image $\hat{I}^{\text{LR}}$ is the result of a downscaling operation from a degraded HR image I^{HR} added by the noise:

$$\hat{I}^{\text{LR}} = (I^{\text{HR}} * k) \downarrow_{\text{s}} + n_{\text{image}}, \tag{1}$$

where k denotes an unknown image degradation kernel, $\downarrow_{\text{s}}$ denotes a downscaling operation with a scale factor of s, and n_{image} represents the additive image noise. We use $\hat{I}^{\text{LR}}$ to denote I^{LR} has been noise corrupted.

For the case of an event camera, the event sensor output can be described as:

$$E_{\text{t}} = \Gamma\Big\{ \log(\frac{I_{\text{t}} + b}{I_{\text{t-1}} + b}), \epsilon \Big\}, \tag{2}$$

where $\Gamma\{\theta, \epsilon\}$ represents the conversion function from log-intensity to event, and b is an offset value to prevent $\log(0)$. $\Gamma\{\theta, \epsilon\} = 1$ when $\theta >= \epsilon$, indicating a positive event; $\Gamma\{\theta, \epsilon\} = -1$ when $\theta <= -\epsilon$, indicating a negative event; and $\Gamma\{\theta, \epsilon\} = 0$ when $|\theta| < \epsilon$, indicating that no event has been fired. The hot pixels can be interpreted as ϵ being significantly low, and the cold pixels as the opposite.

Equation (2) is the noise-free model of the intensity-to-event conversion. The event formation model considering both the downscaling and noise operation can be represented as:

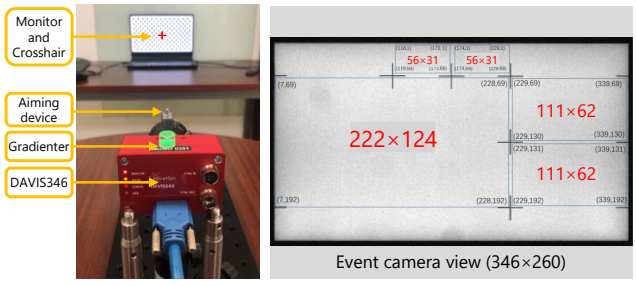
$$\hat{E}^{\text{LR}}_{\text{t}} = \Gamma\Big\{ \log\big(\frac{(I^{\text{HR}}_{\text{t}} * k) \downarrow_{\text{s}} + b}{(I^{\text{HR}}_{\text{t-1}} * k) \downarrow_{\text{s}} + b}\big), \epsilon + n_{\text{event}} \Big\}, \tag{3}$$

where n_{event} represents the perturbation noise pivoted at the firing threshold. According to previous studies [24, 42], n_{event} can be viewed as a Gaussian random process with a mean value of 0. Note that this model does not consider *all* the event sensor noise types but can be used to explain several experimental observations [24] and has been adopted in previous event simulator for generating noise-corrupted events [16]. Our goal is to recover the latent HR event signal $E^{\text{HR}}_{\text{t}} = \Gamma\{\log(\frac{I^{\text{HR}}_{\text{t}} + b}{I^{\text{HR}}_{\text{t-1}} + b}), \epsilon\}$ from the LR noisy signal $\hat{E}^{\text{LR}}_{\text{t}}$.

3.2. A display-camera system for EDSR

The recovery from $\hat{E}^{\text{LR}}_{\text{t}}$ to E^{HR}_{t} is an ill-posed problem as there are many unknown parameters need to be estimated, including the image degradation kernel k, the threshold value ϵ and the event noise n_{event}. Even when all the unknown parameters are correctly estimated, the surjective property of $\Gamma(\cdot)$ mapping from intensity to event makes EDSR elusive and intractable.

To approach EDSR, we developed a display-camera system to observe real-scenario event data at multiple scales. The system setup is presented in Fig. 2a. We used a display (AUO80ed, resolution 1920×1080, 144Hz) and the DAVIS346 monochromatic camera (resolution 346×260) [42]. An F/1.4 lens was mounted on the event camera. The camera was placed at a distance of $\sim$ 180cm away from the display to minimize lens distortion, as shown in Fig. 2a. To calibrate between the camera plane and the display plane, we used a gradienter to limit one rotational degree of freedom. The other two degrees of freedom were limited by the collinearity of the camera view center, aiming device and the crosshair on display center. We chose the Need-for-Speed (NFS) dataset [14] as the source material because it is



(a) Setup (b) Calibrated views

Figure 2: We implemented a display-camera system to study event formation and degradation. The display has been divided into 5 segments with two $1\times$, two $2\times$ and one $4\times$ resolution scales.

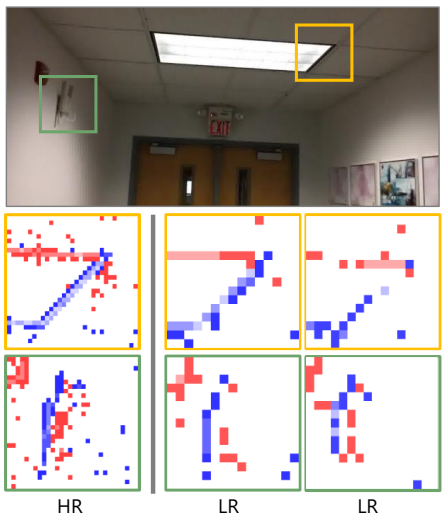


Figure 3: Visualization of an NFS [14] video frame (upper panel), recorded events for the HR video (lower left), and two repeated event recordings of the 2× downsized LR video (lower middle & right).

a high frame-rate visual tracking dataset with all 100 video clips shot using 240FPS video cameras. The original resolution of NFS is 1280 × 720 but three scales (1×, 2× and 4×) were displayed, as shown in Fig. 2b. The original NFS frames were bicubically downsized to avoid spatial aliasing. The new videos were played at 90FPS to avoid frame drop. The display frame-rate imposed a limitation for the temporal resolution of our recorded events. Influence from other light sources was minimized during recording. Eventually we successfully obtained 70 multi-resolution event clips with a total length of about 60 minutes. We refer to the multi-resolution event dataset as "EventNFS".

3.3. Noise-corrupted HR-LR event correspondence

Figure 3 shows an example frame from NFS [14] and its corresponding event patches at two scales. In Fig. 3, the HR patches were recorded at 4× scale while the LR patches were 2×. Although representing the same motion, the two LR patches have different appearance due to noise. Some edge signals were missing due to the increase of the event firing threshold caused by n_{event}, while some noisy events were fired at non-edge positions. Such randomness has made the ground truth data annotation difficult because both the HR and LR event signals have been noise-corrupted.

Now we have obtained a series of noise-corrupted HR-LR event signal pairs, *i.e.* $(\hat{E}^{\text{LR}}_{(i)}, \hat{E}^{\text{HR}}_{(i)})$. Here, the timestamp t is omitted and replaced by the sample index i. According to our noise model in Eq. (3), the event data has an expectation of $\mathbb{E}\big[\hat{E}^{\text{HR}}_{(i)} | \hat{E}^{\text{LR}}_{(i)}\big] = E^{\text{HR}}_{(i)}$ as the noise-corrupted event signal has a zero-mean noise model [22]. This enables us to train a neural regressor Ω that learns to capture a mapping

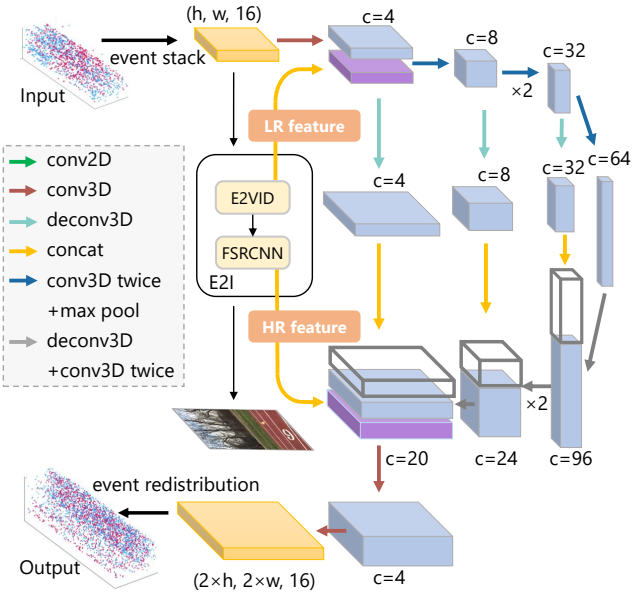


Figure 4: EventZoom-2× network architecture. The input LR events are first binned into a 16-channel event stack. The event stack is then fed into two branches, *i.e.* a 3D U-Net and an E2I module which consists of E2VID [36] and FSRCNN [9]. The LR and HR features from the last layer of E2VID and FSRCNN are concatenated with the beginning and second last layers of the U-Net. Finally, the HR event stack is redistributed to spatiotemporal point clouds.

from noise-corrupted LR event data $\hat{E}^{\text{LR}}$ to noise-free HR event data E^{HR} without ground truth supervision:

$$\underset{\Omega}{\operatorname{argmin}}\, \mathcal{L}\big\{\Omega(\hat{E}^{\text{LR}}_{(i)}), \hat{E}^{\text{HR}}_{(i)}\big\}, \tag{4}$$

where $\mathcal{L}$ denotes a loss function. In our case, we use the mean squared error loss. Note that the hot/cold pixels do not follow this stochastic process and require pre-processing.

3.4. EventZoom neural framework

The network takes as input a spatiotemporal 3D point cloud and outputs its HR enhanced version. The captured events are mostly sparse in space but dense over time. Inspired by previous study [48] where quantitative results showed 2D-CNN-based SR networks are not suitable for EDSR, we employ 3D convolutions for the purpose of learning spatiotemporal features. The neural network is built upon 3D U-Net [52], as shown in Fig. 4 for 2× SR. Compared to other multi-channel 2D-CNN-based approaches, 3D U-Net takes more channels in the time dimension to better exploit temporal coherence. Meanwhile, the limitations of 3D convolution include large network size, more data required for training, and longer inference time.

Both the LR and HR events are binned into a 16-channel event stack to perform supervision. The events are summed

per pixel within each event frame. We have also tried different channel numbers and found 16 achieves the best performance. Since EventNFS has a refresh rate of 90FPS, we chose the time interval of 10ms to enforce each event frame roughly covering one image frame. An event stack then covers a time duration of 160ms.

There are two modules in the network. The main module is the 3D U-Net. The other complementary module performs event-to-image (E2I) conversion, and is composed of two sub-models, *i.e*. E2VID [36] and FSRCNN [9]. E2VID performs event-to-image reconstruction at same resolution. FSRCNN performs $2\times$ image SR. The last layer of features from E2VID is concatenated with the first feature layer of the 3D U-Net. The last layer of features from FSRCNN is concatenated with the second last output layer of the 3D U-Net. Both E2VID and FSRCNN were re-trained with the new NFS data. The 16-channel event stack is binned into a single frame to feed into E2VID. The purpose of E2I is to take advantage of the features learned from event-to-image conversion. E2VID and FSRCNN were chosen on a balance of performance and complexity. There are other alternative network architectures for such purpose.

In $2\times$ SR, the network incorporates additional 3D de-convolution layers for each scale of skip connections, indicated by the light blue arrows in Fig. 4. The output event stack is rounded to integer values and then redistributed by assigning a timestamp for each event. We have experimented with different strategies for timestamp assignment such as random assignment or equal interval, and we found the difference is minimal.

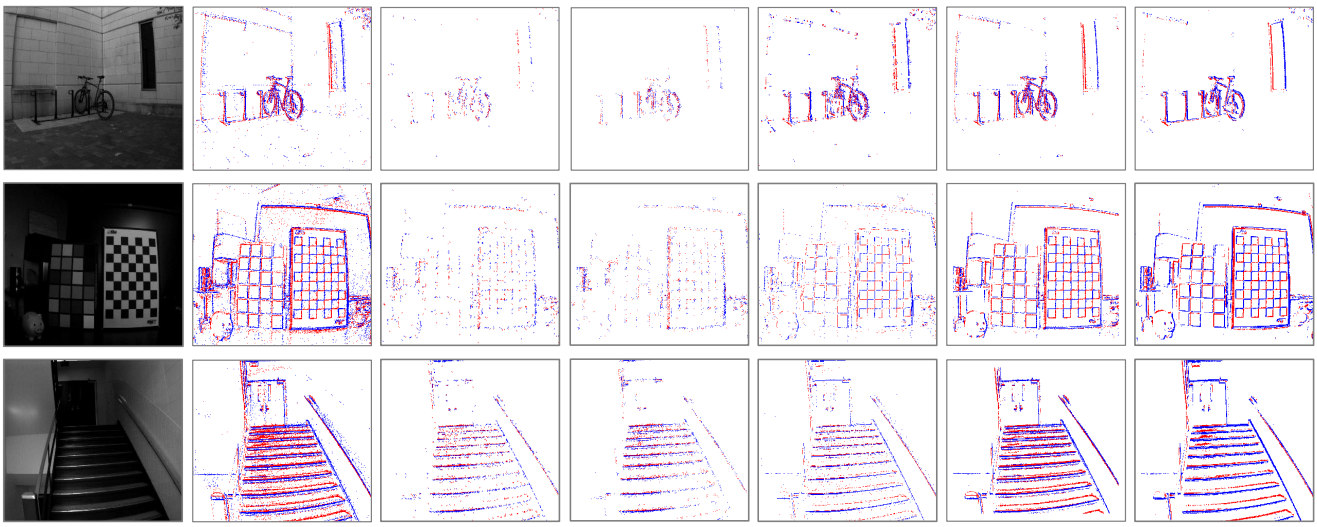
During training, we randomly selected 10 multi-resolution event clips from the EventNFS and generated 2800 LR-HR event pairs as the training set. The validation set has 450 LR-HR pairs extracted from 3 event clips. We used a batch size of 5 and trained EventZoom for 50 epochs. The Adam optimizer is used for minizing the MSE loss with an initial learning rate of 0.01, decayed by a factor of 0.5 every 10 epochs. EventZoom was implemented using PyTorch 1.6 with an NVIDIA 2080 Ti GPU. The training totally took around 2 hours.

4. Results

The experimental results are organized as follows:

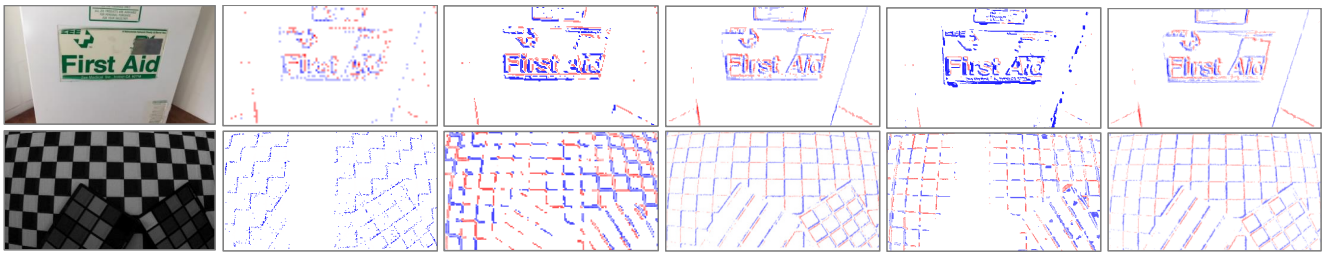
1. EventZoom was compared with SOTA event denoisers on the benchmark dataset DVSNOISE20 [1].
2. For event-to-event SR, EventZoom was compared with GEF [48] up to $4\times$ SR.
3. Ablation studies were conducted to evaluate the effectiveness and tradeoff of the proposed E2I module.

Denoising. EventZoom was compared with four event denoisers, *i.e*., Liu *et al*. [27], EV-gait [47], GEF [48] and EDnCNN [1]. The DVSNOISE20 [1] dataset is used as the benchmark dataset. DVSNOISE20 contains 16 different scenes mostly under static conditions. In this case, the EventZoom-$1\times$ was trained with same-resolution input-output pairs. The 3D de-convolution layers for skip connections shown in Fig. 4 were not used so that the output can keep the original size. We tested 14 out of 16 sequences except Scene-1 and Scene-16 as we found these two sequences



(a) APS frame (b) Raw event data (c) Liu *et al*. [27] (d) EV-gait [47] (e) GEF [48] (f) EDnCNN [1] (g) Ours

Figure 5: Same-resolution denoising results on the DVSNOISE20 [1] dataset. (a) an APS frame captured by a DAVIS346. (b) the event frame accumulating events fired within the exposure time of (a). (c)-(g) the denoising results based on (b).



(a) APS frame (b) Raw event data (c) 2× w/ GEF (d) 2× w/ ours (e) 4× w/ GEF (f) 4× w/ ours

Figure 6: Comparison of 2× and 4× event SR performance between GEF [48] and our method.

Table 1: Denoising runtime comparison on DVSNOISE20 [1]. (unit: second)

	Liu *et al.* [27]	EV-gait [47]	GEF [48]	EDnCNN [1]	Ours
benches	106.86	498.55	5114.73	696.41	**10.62**
bigChecker	1817.88	7790.69	11233.24	907.49	**13.68**
bike	**8.23**	32.91	1007.05	193.54	8.53
bricks	11.06	48.24	1833.77	222.13	**8.47**
checkerFast	968.70	3730.56	7084.50	1077.21	**13.25**
checkerSlow	42.68	142.95	2792.85	526.28	**9.78**
classroom	614.14	2271.18	6337.69	667.74	**10.79**
conference	304.18	1457.25	5427.41	690.29	**11.30**
labFast	228.15	1016.37	5275.22	805.43	**10.78**
labSlow	51.33	209.52	2587.21	586.69	**9.58**
pavers	87.14	281.20	2528.94	355.64	**8.93**
soccer	14.93	52.20	3510.58	352.08	**8.98**
stairs	54.59	225.43	3538.54	631.18	**10.50**
toys	1789.15	6132.71	13049.88	1063.56	**13.92**
average	435.64	1706.41	5094.40	626.83	**10.65**

were severely corrupted by dead pixels. The denoising results are shown in Fig. 5. As can be seen that EventZoom is able to reveal and enhance the scene structures and effectively remove noisy events. Note that EventZoom did not outperform other methods by the metric proposed in EDnCNN [1] as the Event Probablity Mask (EPM) only identifies wrongly-fired events while our EventZoom resulted in additional events which is treated as false positive events by the EPM.

The runtime of all the denoisers was benchmarked in Table 1. GEF [48] was implemented using a coarse-to-fine grid search method for estimating the optical flow for optimal accuracy. All filters were tested on the middle 2% temporal window of the 14 sequences as both GEF and EDnCNN require long run time compared to others. On average, EventZoom takes at least 40× less time than others.

Super resolution. EventZoom was compared with GEF [48] for 2× and 4× event-to-event SR. EventZoom achieved 4× SR by performing EventZoom-2× SR twice. We have also experimented with a single EventZoom-4× network but found the resolution for the 1×-scale data was too low to train. Both DVSNOISE20 [1] and our EventNFS datasets were used for testing. The SR results are shown in Fig. 6. EventZoom filled missing information even within large empty area. This is likely due to the relatively longer time window it processed. On the contrary, GEF was unable to fill large area of missing information as it only extracted mutual information from the image and events.

Table 2: Ablation on the E2I module.

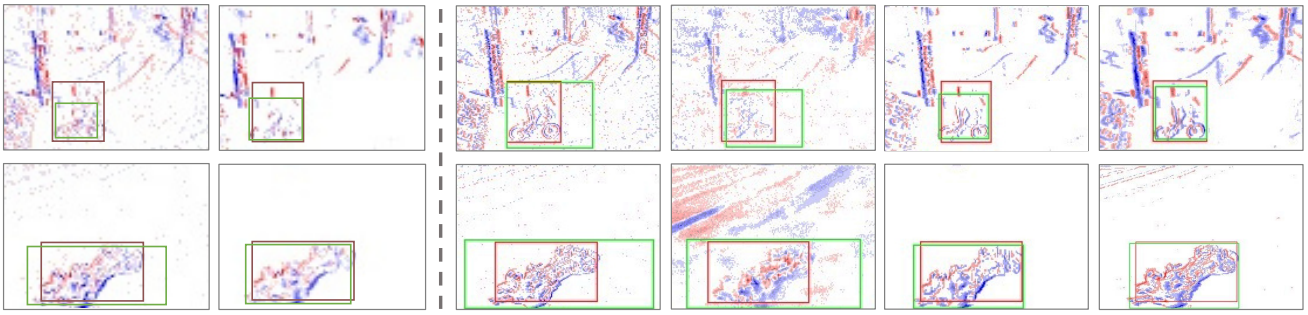
Settings	LR feat.	HR feat.	MSE
Baseline #0	×	×	0.0608
Baseline #1	✓	×	0.0575
Baseline #2	×	✓	0.0580
EventZoom	✓	✓	**0.0568**

Ablation on the E2I module. We evaluated the effectiveness of the proposed E2I module. Since our E2I is a combination of E2VID and FSRCNN, corresponding to LR and HR image features, we considered three baseline models each disabling one/two image feature(s). The minimal loss values during training were used as the evaluation metric, summarized in Table 2. The qualitative results are presented in the supplementary document. In addition, we experimented with other architectural variants, including the partial convolutions [26] and the stacked dilated SPatially-Adaptive DEnormalization mechanism [33]. The results are appended in the supplementary document.

5. Applications

5.1. Event-based visual object tracking

We evaluated the performance improvement that EventZoom can bring to the task of event-based visual object tracking. The original NFS dataset has provided ground truth bounding boxes for each video sequence [14]. The same bounding boxes were copied to the event data. We chose E-MS [3] as the benchmark tracker. E-MS is an efficient event-based tracking method that uses mean-shift clustering and Kalman filters to classify the objects in the scene. In order to calculate the coincidence with bounding box ground truth, we employed the minimum enclosing rectangle of event clusters, and selected the rectangle clos-



(a) 1× raw data (b) 1×, ours (denoised) (c) 2× raw data (d) 1×, E2SRI+V2E (e) 2×, ours (denoised) (f) 1×, ours (SR)

Figure 7: Examples for the tracking results. Red/Green bounding boxes represent the ground truth/prediction. (a) & (b) are 1× resolution scale while (c-f) are 2×. (d) takes 1× data and performs 2× SR by E2SRI [29] and V2E [8]. (e) performs same-resolution event denoising for the 2× data before tracking. (f) directly performs 2× SR before tracking.

Table 3: Visual object tracking performance comparison on 12 samples of EventNFS dataset (greener blocks represent better performance with higher IoU index).

		biker all_1	biker up body	car camaro	car jumping	car_rc rotating	first	horse running	jelly fish_5	motorcross	rubber	running 100_m_2	soccer ball
1×	Raw data	0.271	0.194	0.159	0.221	0.381	0.126	0.176	0.151	0.237	0.094	0.109	0.088
	Ours (denoise)	**0.286**	**0.222**	**0.277**	**0.311**	**0.454**	**0.199**	**0.197**	**0.190**	**0.248**	**0.100**	**0.136**	**0.246**
2×	Raw data	0.250	0.173	0.144	0.131	0.353	0.102	0.171	0.078	0.253	0.065	0.120	0.056
	1×, E2SRI&V2E	0.251	0.205	0.112	0.104	0.367	0.108	0.116	0.070	0.251	0.041	0.086	0.060
	Ours (denoise)	**0.335**	**0.239**	**0.231**	**0.367**	**0.503**	**0.175**	**0.202**	**0.192**	**0.267**	**0.200**	**0.135**	**0.206**
	Ours (1× w/ SR)	**0.353**	**0.241**	**0.220**	**0.270**	**0.529**	**0.146**	**0.178**	**0.200**	**0.254**	**0.228**	**0.159**	**0.248**

est to the target object as the predicted bounding box. Intersection over Union (IoU) was used as the evaluation metric between predicted bounding boxes and the ground truth.

We chose 12 sequences from EventNFS for testing. Each sequence was tested on two resolution scales. The 1× scale represents a resolution of 111 × 62, and the 2× represents 222 × 124. There were 6 cases tested: (a) perform tracking on 1× raw event data; (b) perform same-resolution denoising on 1× data by our EventZoom, then perform tracking; (c) perform tracking on 2× raw data; (d) perform E2I SR by E2SRI [29], convert the 2× video to 2× events by V2E [8], perform tracking; (e) perform same-resolution denoising on 2× data by our EventZoom, then perform tracking; (f) perform 2× EDSR by our EventZoom, then perform tracking.

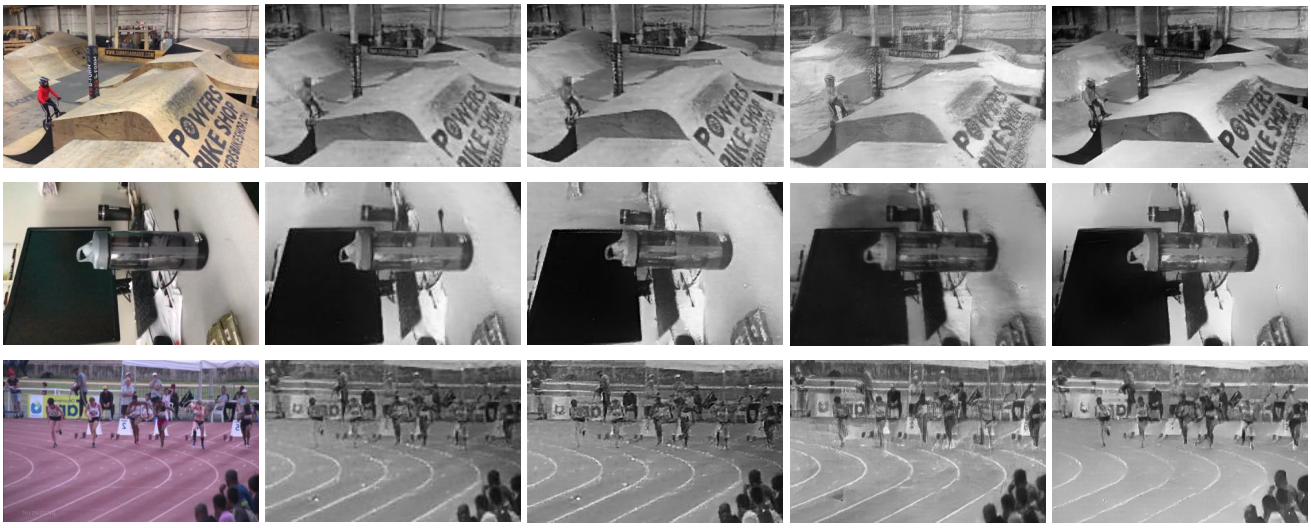
Two examples with the tracking results are shown in Fig. 7. The biker (first row) and the toy car (second row) are better revealed at higher resolution and tracked more accurately. The results for the accuracy of average IoU are reported in Table 3. As shown in the table, both the 1× denoising and 2× SR have achieved improvements compared to those from the raw data. Moreover, we found that the 2× events obtained by {E2SRI+V2E} provided minimal tracking improvements.

Table 4: Image reconstruction performance

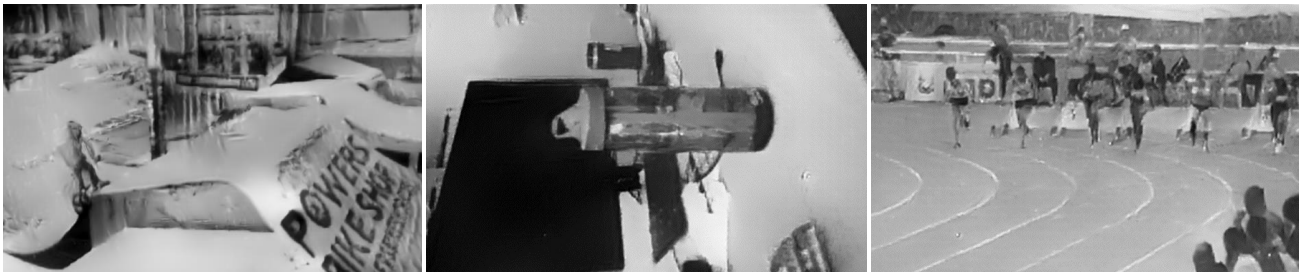
	PSNR	SSIM	MSE	runtime
E2SRI [29]	14.787	0.474	0.041	6.297s
Ours	**15.510**	**0.505**	**0.036**	**0.605s**

5.2. Image reconstruction

We used EventZoom for image reconstruction. The E2VID was chosen as the benchmark 1× event-to-image reconstruction algorithm [36]. For 2× SR image reconstruction, we compared with 1) 1× E2VID + image SR using SRFBN [23] and 2) E2SRI [29], one of the state-of-the-art algorithms that performs super resolved image reconstruction. The results are shown in Fig. 8. As can be seen in the figure, EventZoom achieves the best image reconstruction quality at 2×. We also show 4× result in Fig. 8f which are reconstructed from our 4× SR event data. For quantitative analysis, we benchmarked the reconstruction performance by randomly selecting 1000 images from the EventNFS dataset and calculated several measures including PSNR, SSIM and MSE between reconstructed images and corresponding APS frames. We also recorded the run time and all results are presented in Table 4. Our results showed that



(a) APS frame (b) 1× w/ E2VID (c) 1× w/ E2VID+SRFBN (d) 2× w/ E2SRI (e) Ours (2×)



(f) Ours (4×)

Figure 8: Comparison of event-based image reconstruction performance on our EventNFS dataset. (a) an APS frame. (b) reconstruct 1× image with E2VID [36]. (c) reconstruct 1× image with E2VID [36] and then 2× upsample image with SRFBN [23]. (d) reconstruct 2× image directly with E2SRI [29]. (e) reconstruct 2× event with EventZoom and then reconstruct 2× image with E2VID [36]. (f) reconstruct 4× event with EventZoom and then reconstruct 4× image with E2VID [36].

EventZoom outperformed E2SRI across all metrics with 10× less time on average. Additional video reconstruction results are included in the supplementary material.

6. Conclusion

This paper presented a novel neural framework for event denoising and super resolution, referred as EventZoom. EventZoom used a 3D U-Net as the backbone architecture with an optional event-to-image (E2I) module. The E2I module leveraged SOTA image reconstruction technique. In order to learn the correspondence between the LR and HR event data, we proposed a display-camera system for multi-resolution event data collection. The system was used to convert the high framerate object tracking dataset NFS [14] to an event version (EventNFS) at three scales. By training with the provided noise-corrupted HR-LR pairs, the network was able to effectively perform EDSR up to 4× SR. EventZoom achieves state-of-the-art results with improved time efficiency. The enhanced event streams by EventZoom contribute to improved visual task performance. We have presented two exemplary applications including visual object tracking and SR image reconstruction.

There are several limitations for this work. The dataset quality was compromised by the display, which has relatively low refresh rate and dynamic range. This imposes constraint for applying motion-based algorithms, *e.g.* GEF [48]. Interestingly, we did not find much generalization issue for the trained models after testing on external datasets. A more accurate measure is in need to quantify the effectiveness in event restoration and enhancement.

Acknowledgement

This work was supported by National Natural Science Foundation of China under Grant No. 61872012, 62088102 and Beijing Academy of Artificial Intelligence (BAAI).

References

[1] R Baldwin, Mohammed Almatrafi, Vijayan Asari, and Keigo Hirakawa. Event probability mask (epm) and event denoising convolutional neural network (edncnn) for neuromorphic cameras. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 1701–1710, 2020. 2, 5, 6

[2] Patrick Bardow, Andrew J Davison, and Stefan Leutenegger. Simultaneous optical flow and intensity estimation from an event camera. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 884–892, 2016. 1

[3] F. Barranco, C. Fermuller, and E. Ros. Real-time clustering and multi-target tracking using event-based sensors. In *2018 IEEE/RSJ Intern. Conf. on Intel. Robots and Sys. (IROS)*, pages 5764–5769, 2018. 6

[4] Juan Barrios-Avilés, Alfredo Rosado-Muñoz, Leandro D Medus, Manuel Bataller-Mompeán, and Juan F Guerrero-Martínez. Less data same information for event-based sensors: A bioinspired filtering and data reduction algorithm. *Sensors*, 18(12):4122, 2018. 2

[5] Marco Cannici, Marco Ciccone, Andrea Romanoni, and Matteo Matteucci. A differentiable recurrent surface for asynchronous event-based data. In *Eur. Conf. Comput. Vis.*, 2020. 2

[6] Shoushun Chen and Menghan Guo. Live demonstration: Celex-V: a 1m pixel multi-mode event-based sensor. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, 2019. 1

[7] Jonghyun Choi, Kuk-Jin Yoon, et al. Learning to super resolve intensity images from events. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 2765–2773, 2020. 1

[8] Tobi Delbruck, Hu Yuhuang, and He Zhe. V2E: From video frames to realistic DVS event camera streams. *arxiv*, June 2020. 2, 7

[9] Chao Dong, Chen Change Loy, and Xiaoou Tang. Accelerating the super-resolution convolutional neural network. In *Eur. Conf. Comput. Vis.*, 2016. 2, 4, 5

[10] Guillermo Gallego, Tobi Delbruck, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew Davison, Joerg Conradt, Kostas Daniilidis, et al. Event-based vision: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2020. 1

[11] Guillermo Gallego, Henri Rebecq, and Davide Scaramuzza. A unifying contrast maximization framework for event cameras, with applications to motion, depth, and optical flow estimation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3867–3876, 2018. 2

[12] Daniel Gehrig, Mathias Gehrig, Javier Hidalgo-Carrió, and Davide Scaramuzza. Video to events: Bringing modern computer vision closer to event cameras. 2020. 2

[13] Daniel Gehrig, Antonio Loquercio, Konstantinos G Derpanis, and Davide Scaramuzza. End-to-end learning of representations for asynchronous event-based data. In *Int. Conf. Comput. Vis.*, pages 5633–5643, 2019. 3

[14] Galoogahi Hamed Kiani, Fagg Ashton, Huang Chen, Ramanan Deva, and Lucey Simon. Need for speed: A benchmark for higher frame rate object tracking. In *Int. Conf. Comput. Vis.*, pages 1134–1143, 2017. 2, 3, 4, 6, 8

[15] Jin Han, Chu Zhou, Peiqi Duan, Yehui Tang, Chang Xu, Chao Xu, Tiejun Huang, and Boxin Shi. Neuromorphic camera guided high dynamic range imaging. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2020. 1

[16] Rebecq Henri, Gehrig Daniel, and Scaramuzza Davide. Esim: an open event camera simulator. In *Conf. on Robot Learn. (CoRL)*, page 969–982, 2018. 2, 3

[17] Yuhuang Hu, Hongjie Liu, Michael Pfeiffer, and Tobi Delbruck. Dvs benchmark datasets for object tracking, action recognition, and object recognition. *Front. in neuro.*, 10:405, 2016. 2

[18] Zhe Jiang, Yu Zhang, Dongqing Zou, Jimmy Ren, Jiancheng Lv, and Yebin Liu. Learning event-based motion deblurring. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 3320–3329, 2020. 1, 3

[19] Alireza Khodamoradi and Ryan Kastner. O (n)-space spatiotemporal filter for reducing noise in neuromorphic vision sensors. *IEEE Trans. on Emerg. Topics in Comput.*, 2018. 2

[20] Hanme Kim, Stefan Leutenegger, and Andrew J. Davison. Real-time 3D reconstruction and 6-DoF tracking with an event camera. In *Eur. Conf. Comput. Vis.* Springer, 2016. 1

[21] Chankyu Lee, Adarsh Kosta, Alex Zihao Zhu, Kenneth Chaney, Kostas Daniilidis, and Kaushik Roy. Spike-flownet: Event-based optical flow estimation with energy-efficient hybrid neural networks. In *Eur. Conf. Comput. Vis.*, 2020. 2

[22] Jaakko Lehtinen, Jacob Munkberg, Jon Hasselgren, Samuli Laine, Tero Karras, Miika Aittala, and Timo Aila. Noise2noise: Learning image restoration without clean data. In *Intern. Conf. on Machine Learn. (ICML)*, 2018. 4

[23] Zhen Li, Jinglei Yang, Zheng Liu, Xiaomin Yang, Gwanggil Jeon, and Wei Wu. Feedback network for image super-resolution. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3867–3876, 2019. 7, 8

[24] Patrick Lichtsteiner, Christoph Posch, and Tobi Delbruck. A 128×128 120 db 15 μs latency asynchronous temporal contrast vision sensor. *IEEE J. S.-S. Circuits*, 43(2):566–576, 2008. 1, 2, 3

[25] Daqi Liu, Alvaro Parra, and Tat-Jun Chin. Globally optimal contrast maximisation for event-based motion estimation. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 6349–6358, 2020. 2

[26] Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In *Eur. Conf. Comput. Vis.*, 2018. 6

[27] Hongjie Liu, Christian Brandli, Chenghan Li, Shih-Chii Liu, and Tobi Delbruck. Design of a spatiotemporal correlation filter for event-based sensors. In *2015 IEEE Inter. Sym. on Circ. and Sys. (ISCAS)*, pages 722–725, 2015. 2, 5, 6

[28] Nico Messikommer, Daniel Gehrig, Antonio Loquercio, and Davide Scaramuzza. Event-based asynchronous sparse convolutional networks. In *Eur. Conf. Comput. Vis.*, 2020. 3

[29] S. Mohammad Mostafavi I., Jonghyun Choi, and Kuk-Jin Yoon. Learning to super resolve intensity images from events. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 2768–2786, 2020. 2, 3, 7, 8

[30] Daniel Neil, Michael Pfeiffer, and Shih-Chii Liu. Phased LSTM: Accelerating recurrent network training for long or event-based sequences. In *Adv. Neural Inform. Process. Syst.*, volume 29, pages 3882–3890, 2016. 2

[31] Liyuan Pan, Cedric Scheerlinck, Xin Yu, Richard Hartley, Miaomiao Liu, and Yuchao Dai. Bringing a blurry frame alive at high frame-rate with an event camera. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 1

[32] Federico Paredes-Vallés, Kirk Yannick Willehm Scheper, and Guido Cornelis Henricus Eugene De Croon. Unsupervised learning of a hierarchical spiking neural network for optical flow estimation: From events to global motion perception. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2019. 2

[33] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. *CVPR*, 2019. 6

[34] Xin Peng, Yifu Wang, Ling Gao, and Laurent Kneip. Globally-optimal event camera motion estimation. In *Eur. Conf. Comput. Vis.*, 2020. 2

[35] Christoph Posch, Daniel Matolin, and Rainer Wohlgenannt. A QVGA 143 db dynamic range frame-free PWM image sensor with lossless pixel-level video compression and time-domain CDS. *IEEE J. S.-S. Circuits*, 46(1):259–275, 2010. 1

[36] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. Events-to-video: Bringing modern computer vision to event cameras. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3857–3866, 2019. 2, 4, 5, 7, 8

[37] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. High speed and high dynamic range video with an event camera. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2019. 1, 2

[38] Cedric Scheerlinck, Henri Rebecq, Daniel Gehrig, Nick Barnes, Robert Mahony, and Davide Scaramuzza. Fast image reconstruction with an event camera. In *IEEE/CVF Win. Conf. Appl. Comput. Vis.*, pages 156–163, 2020. 1

[39] Yusuke Sekikawa, Kosuke Hara, and Hideo Saito. Eventnet: Asynchronous recursive event processing. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3887–3896, 2019. 2

[40] Timo Stoffregen, Guillermo Gallego, Tom Drummond, Lindsay Kleeman, and Davide Scaramuzza. Event-based motion segmentation by motion compensation. In *Int. Conf. Comput. Vis.*, 2019. 2

[41] 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 *Eur. Conf. Comput. Vis.*, 2020. 2

[42] 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 Trans. Circuit Syst. II: Express Briefs*, 65(5):677–681, 2018. 1, 3

[43] Stepan Tulyakov, Francois Fleuret, Martin Kiefel, Peter Gehler, and Michael Hirsch. Learning an event sequence embedding for dense event-based deep stereo. In *Int. Conf. Comput. Vis.*, pages 1527–1537, 2019. 2

[44] Bishan Wang, Jingwei He, Lei Yu, Gui-Song Xia, and Wen Yang. Event enhanced high-quality image recovery. In *Eur. Conf. Comput. Vis.*, 2020. 1

[45] Lin Wang, Yo-Sung Ho, Kuk-Jin Yoon, et al. Event-based high dynamic range image and very high frame rate video generation using conditional generative adversarial networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10081–10090, 2019. 1, 3

[46] Lin Wang, Tae-Kyun Kim, and Kuk-Jin Yoon. EventSR: From asynchronous events to image reconstruction, restoration, and super-resolution via end-to-end adversarial learning. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 8315–8325, 2020. 1, 2

[47] Yanxiang Wang, Bowen Du, Yiran Shen, Kai Wu, Guangrong Zhao, Jianguo Sun, and Hongkai Wen. EV-Gait: Event-based robust gait recognition using dynamic vision sensors. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 2, 5, 6

[48] Zihao Winston 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 *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2020. 2, 3, 4, 5, 6, 8

[49] Zihao Winston Wang, Weixin Jiang, Kuan He, Boxin Shi, Aggelos Katsaggelos, and Oliver Cossairt. Event-driven video frame synthesis. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, pages 0–0, 2019. 1, 3

[50] Song Zhang, Yu Zhang, Zhe Jiang, Dongqing Zou, Jimmy Ren, and Bin Zhou. Learning to see in the dark with events. In *Eur. Conf. Comput. Vis.*, 2020. 1

[51] Alex Zihao Zhu, Dinesh Thakur, Tolga Özaslan, Bernd Pfrommer, Vijay Kumar, and Kostas Daniilidis. The multivehicle stereo event camera dataset: An event camera dataset for 3d perception. *IEEE Robot. and Auto. Lett.*, 3(3):2032–2039, 2018. 2

[52] Özgün Çiçek, A. Abdulkadir, S. Lienkamp, T. Brox, and O. Ronneberger. 3d u-net: Learning dense volumetric segmentation from sparse annotation. In *Medic. Image Computi. and Compute. Assi. Inter. Soc. (MICCAI)*, 2016. 2, 3, 4