NEST: Neural Event Stack for Event-Based Image Enhancement

Minggui Teng[1], Chu Zhou[2], Hanyue Lou[1], and Boxin Shi[1,3,4(✉)]

[1] NERCVT, School of Computer Science, Peking University, Beijing, China
shiboxin@pku.edu.cn
[2] KLMP (MOE), School of Artificial Intelligence, Peking University, Beijing, China
[3] Institute for Artificial Intelligence, Peking University, Beijing, China
[4] Beijing Academy of Artificial Intelligence, Beijing, China

Abstract. Event cameras demonstrate unique characteristics such as high temporal resolution, low latency, and high dynamic range to improve performance for various image enhancement tasks. However, event streams cannot be applied to neural networks directly due to their sparse nature. To integrate events into traditional computer vision algorithms, an appropriate event representation is desirable, while existing voxel grid and event stack representations are less effective in encoding motion and temporal information. This paper presents a novel event representation named Neural Event STack (**NEST**), which satisfies physical constraints and encodes comprehensive motion and temporal information sufficient for image enhancement. We apply our representation on multiple tasks, which achieves superior performance on image deblurring and image super-resolution than state-of-the-art methods on both synthetic and real datasets. And we further demonstrate the possibility to generate high frame rate videos with our novel event representation.

Keywords: Event camera · Image enhancement · Event representation

1 Introduction

Event cameras, such as Dynamic Vision Sensor (DVS) [16], can detect brightness changes and trigger events whenever the increase (decrease) of latent irradiance exceeds a preset threshold. They are widely used in image enhancement tasks since they possess clear advantages over traditional cameras in various aspects, such as high temporal resolution, low latency, and high dynamic range (HDR). However, event streams are represented as multiple four coordinates signals (x, y, t, p), and such continuous event signals cannot be processed by traditional computer vision algorithms directly, which brings a natural gap to leverage the advantages of events for image enhancement.

Project page: https://github.com/ChipsAhoyM/NEST.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-20068-7_38.

S. Avidan et al. (Eds.): ECCV 2022, LNCS 13666, pp. 660–676, 2022.
https://doi.org/10.1007/978-3-031-20068-7_38

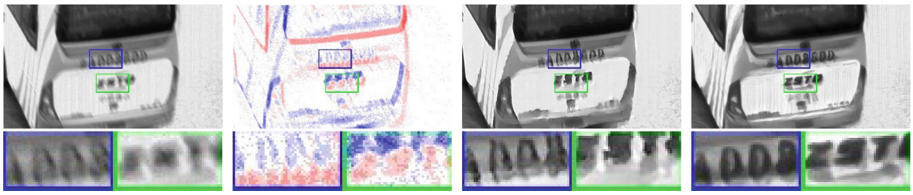


Fig. 1. An example result of NEST-guided image enhancement with 4× super-resolution. (a) Blurry image. (b) Corresponding events (color pair {blue, red} represents the event polarity {positive, negative} throughout this paper). (c) Result of eSL-Net [31]. (d) Our result. (Color figure online)

Finding a favored representation as input is important for event-based image enhancement tasks. Discretizing event signals in the time domain is an intuitive choice. This could be achieved by recording the timestamp of the last event in each pixel location [14], by inserting events into a voxel grid using a linearly weighted accumulation similar to bilinear interpolation [37], or by merging and stacking events within a time interval or a fixed number of events [32]. Despite their simplicity, when the number of channels divided from events increases, noisy events in such hand-crafted representations become hardly distinguishable from useful signals.

Neural representation has become a popular choice in event embedding procedures recently. Useful features could be extracted from event sequences with multi-layer perceptron (MLP) [6,25], spike neural network (SNN) [35], long short-term memory (LSTM) [3,20], and graph neural network (GNN) [1,15]. Despite their effectiveness in object recognition [1,3,6,15,20,35] and segmentation [25], these representations are not supposed to be optimized for image enhancement tasks, since they focus more on preserving semantic information well instead of caring about pixel-wise information, while the latter is crucial for image enhancement. The fact that hand-crafted event representations are prone to noise and neural representations sacrifice contextual information motivates us to propose a tailored representation for event-based image enhancement.

In this paper, we introduce Neural Event STack (**NEST**), which satisfies event physical constraints while faithfully encodes motion and temporal information with less noise involved. We first propose a NEST estimator to transform an event sequence into NESTs by a bidirectional convolutional long short-term memory (ConvLSTM) block [28] in a data-driven manner to fulfill event embedding. Tailored to the NEST, we then propose a NEST-guided Deblurring Net (D-Net) for image deblurring and a NEST-guided Super-resolution Net (S-Net) for image super-resolution, with simple architectures (a NEST-guided image enhancement example is shown in Fig. 1). By parallel processing multiple NESTs with D-Net and S-Net, high frame rate (HFR) videos can be restored with sharper edges and higher resolution.

Overall, this paper contributes in the following aspects:

Table 1. Comparison of LSTM-based event representations. H and W denote the image height and width, C denotes the number of channels, and T denotes the number of temporal bins.

Representation	Dimensions	Characteristics		
		Direction	Resolution	Hidden states
PhasedLSTM [20]	1D vector	Uni-direction	Fixed	Discarded
MatrixLSTM [3]	$C \times H \times W$			
NEST	$T \times C \times H \times W$	Bi-direction	Arbitrary	Preserved

- a neural representation (NEST) comprehensively encoding motion and temporal information from events in a noise-robust manner;
- event-based solutions for image deblurring and super-resolution taking benefit from the new representation;
- a unified framework for HFR video generation guided by NESTs.

We quantitatively and qualitatively evaluate our method on both synthetic and real datasets and demonstrate its superior performance over state-of-the-art methods.

2 Related Work

2.1 Event Representation

Event data possess many attractive advantages such as high speed and high dynamic range. However, it is difficult to apply computer vision algorithms designed for ordinary images to events, since event data are essentially different from image frames. Many algorithms try to find an event representation compatible with frame-based data, and they can be divided into two categories: hand-crafted representation and data-driven representation.

Hand-Crafted Representation. Lagorce *et al.* [14] proposed the time surface representation, obtained by keeping track of the timestamp of the last event that occurred in each location. Based on the time surface representation, Sironi *et al.* [30] proposed using histograms of averaged time surfaces (HATS), preserving more temporal information in histograms. To avoid the "motion overwriting" problem in the time surface representation, Zhu *et al.* [37] proposed the voxel grid representation, which inserts events into a voxel grid using a linearly weighted accumulation similar to bilinear interpolation. Wang *et al.* [32] proposed an event stack representation, which forms events as multiple frame event stacks by merging and stacking them within a time interval or a fixed number of events.

Data-Driven Representation. Recently data-driven models show higher robustness for event representation. Sekikawa *et al.* [25] proposed a recursive architecture and used MLP for computing a recursive formula. Gehrig *et al.* [6]

used MLP to encode time information of event sequences and summed up values from MLP to construct an event spike tensor. Inspired by biological mechanisms, Yao *et al.* [35] encoded events with attention SNN by processing events as asynchronous spikes. To better exploit the topological structure inside events sequences, Bi *et al.* [1] and Li *et al.* [15] used a graph to represent the event cloud with GNN and further conducted graph convolutions to obtain the event representation. Besides, to better exploit temporal information of events sequences, Neil *et al.* [20] proposed PhasedLSTM with a new time gate for processing asynchronous events. Cannici *et al.* [3] proposed the MatrixLSTM representation which integrates event sequences conditionally with LSTM cells. Although these representations show great potential in multiple computer vision tasks (*e.g.*, object recognition, segmentation, and optical flow estimation), hand-crafted representations are still popular for image enhancement tasks, since data-driven representations for these tasks are not readily available. Particularly, LSTM-based methods show great potential in event representation. A comparison of LSTM-based event representations and their design choices are summarized in Table 1. The method in [3] emphasizes preserving sparsity when computing the MatrixLSTM, it is not suitable for image enhancement tasks due to the loss of connection around neighboring pixels. Thus, a proper event representation method tailored to image enhancement tasks is desired.

2.2 Event-Based Image Enhancement

Event-Based Image Deblurring. Pan *et al.* [21] proposed a simple and effective approach, the Event-based Double Integral (EDI) model, to reconstruct an HFR sharp video from a single blurry frame and corresponding event data. Jiang *et al.*[11] proposed a convolutional recurrent neural network and a differentiable directional event filtering module to recover sharp images. Lin *et al.* [17] proposed a deep CNN with a dynamic filtering layer to deblur and generate videos in a frame-aware manner. Wang *et al.* [31] proposed an event-enhanced sparse learning network named eSL-Net to address deblurring, denoising, and super-resolution simultaneously. Shang *et al.* [26] detected the nearest sharp frames with events, and then performed deblurring guided by the nearest sharp frames.

Event-Based Image Super-Resolution. Jing *et al.* [12] proposed an event-based video super-resolution framework, which reconstructs high-frequency low resolution (LR) frames interpolated with events and merges them to form a high resolution (HR) frame. Han *et al.* [7] proposed a two-stage network to fuse event temporal information with images and established event-based single image super-resolution as a multi-frame super-resolution problem.

For these event-based image enhancement methods, event stack is the most widely adopted choice [7,10–12,17,26,33] for representation due to simplicity, despite its poor robustness to noise. In the next section, we will first revisit the formulation of deblurring and super-resolution with events and analyze the demerits of applying the event stack representation method for image enhancement. We then propose the NEST representation to solve these problems.

3 NEST: Representation

In this section, we first derive the formulation of bidirectional event summations, which bridge the gap between low-quality images and high-quality images with events in Sect. 3.1. Based on bidirectional event summations, we briefly analyze the advantages and disadvantages of event stack representation. To avoid noisy events interference, we propose a neural representation to robustly implement bidirectional event summations in Sect. 3.2. Finally, we introduce the model design of our NEST estimator in Sect. 3.3.

3.1 Bidirectional Event Summation

An event e is a quadruple (x, y, t, p) triggered when the log intensity change exceeds a preset threshold c, *i.e.*,

$$|\log(\mathbf{I}^{t}_{x,y}) - \log(\mathbf{I}^{t-\Delta t}_{x,y})| \geq c, \tag{1}$$

in which $\mathbf{I}^{t}_{x,y}$ and $\mathbf{I}^{t-\Delta t}_{x,y}$ represent the instantaneous intensity at time t and $t - \Delta t$ respectively for pixel (x, y), and Δt denotes the time interval since the last event occurred at the same position. Polarity $p \in \{1, -1\}$ indicates the direction (increase or decrease) of intensity change. Eq. (1) applies to each pixel (x, y) independently, and pixel indices are omitted henceforth.

Given two instantaneous intensity frames $\mathbf{I}^{t_r}$ and $\mathbf{I}^{t_i}$, let's assume there are N_e events triggered between time t_r and t_i, denoted as $\{e_k\}_{k=1}^{N_e}$. According to the physical model of the event camera, if $t_r \leq t_i$, the event makes a connection between $\mathbf{I}^{t_r}$ and $\mathbf{I}^{t_i}$ as:

$$\begin{aligned} \mathbf{I}^{t_i} &= \mathbf{I}^{t_r} \cdot \exp(\sum_{k=1}^{N_e} c_r \cdot e_k) \\ &= \mathbf{I}^{t_r} \cdot \tilde{\mathbf{S}}^{c_r}_{r \to i} \quad (t_r \leq t_i), \end{aligned} \tag{2}$$

where $\tilde{\mathbf{S}}^{c_r}_{r \to i}$ denotes event summation from time t_r to t_i in the exponential space with a time-varying threshold c_r. c_r approximately follows a normal distribution over time [22].

Deriving from Eq. (2), we can also obtain $\mathbf{I}^{t_r}$ from $\mathbf{I}^{t_i}$ by reversing the event summation $(t_r > t_i)$. Thus, we formulate the *bidirectional event summation* $\mathbf{S}^{c_r}_{r \to i}$ to consider both cases, *i.e.*,

$$\mathbf{S}^{c_r}_{r \to i} = \begin{cases} \tilde{\mathbf{S}}^{c_r}_{r \to i} & (t_r \leq t_i), \\ 1/\tilde{\mathbf{S}}^{c_i}_{i \to r} & (t_r > t_i). \end{cases} \tag{3}$$

Combining Eq. (3), Eq. (2) can be further expanded to include both forward and reverse event summation:

$$\mathbf{I}^{t_i} = \mathbf{I}^{t_r} \cdot \mathbf{S}^{c_r}_{r \to i}. \tag{4}$$

Image Enhancement with Events. Ill-posedness is a common problem in image enhancement tasks, such as image deblurring and super-resolution. For image deblurring, a blurry image $\mathbf{B}$ can be modeled as the average over a sequence of latent sharp frames $\{\mathbf{I}^{t_i}\}_{i=1}^{N_f}$ [21]:

$$\mathbf{B} \approx \frac{1}{N_f} \sum_{i=1}^{N_f} \mathbf{I}^{t_i}, \tag{5}$$

in which N_f is the number of latent sharp frames. Obviously, there are multiple groups of latent frames satisfying Eq. (5), which brings difficulty to recover sharp frames from a single blurry image.

For image super-resolution, an HR frame can be reconstructed by a sequence of latent sharp frames $\{\mathbf{I}_{LR}^{t_i}\}_{i=1}^{N_f}$, *i.e.*,

$$\mathbf{I}_{SR}^{t_i} = \Uparrow \{\mathbf{I}_{LR}^{t_j}\}_{j=1}^{N_f}, \tag{6}$$

where $\Uparrow$ denotes the multi-frame super-resolution operator, combining information from multiple LR frames to recover details that are missing in individual frames. However, it is hard to record multiple latent sharp frames with traditional cameras, which means we need to generate an HR frame with a single LR frame leading to ill-posedness.

As Eq. (4) has shown the relationship of two latent frames by corresponding events, ill-posedness can be relieved by integrating image and events. By combining Eq. (4) and Eq. (5), we obtain:

$$\mathbf{B} \approx \mathbf{I}^{t_i} \cdot (\frac{1}{N_f} \sum_{j=1}^{N_f} \mathbf{S}_{i \to j}^{c_i}). \tag{7}$$

By substituting Eq. (4), we can rewrite Eq. (6) as follows:

$$\mathbf{I}_{SR}^{t_i} = \Uparrow \{\mathbf{I}_{LR}^{t_i} \cdot \mathbf{S}_{i \to j}^{c_i}\}_{j=1}^{N_f}. \tag{8}$$

Since the bidirectional event summations $\{\mathbf{S}_{i \to j}^{c_i}\}_{j=1}^{N_f}$ are independent of the latent frames, we can restore arbitrary sharp latent frames from a single blurry image or reconstruct arbitrary HR frames from a single LR frame with the corresponding events directly.

3.2 Neural Representation

According to Sect. 3.1, the bidirectional event summation establishes the relationship between low-quality (blurry, LR) images and high-quality (sharp, HR) images. As shown in Eq. (7) and Eq. (8), image deblurring needs the average value of the set, and image super-resolution depends on the magnitude difference of each element in the set for recovering details. Thus, the event signal can be discretized in the time domain to form bidirectional event summations $\{\mathbf{S}_{i \to j}^{c_i}\}_{j=1}^{N_f}$, which can guide image enhancement tasks.

Event stack forms events as multiple frames by merging and stacking them within a time interval or a fixed number of events [32]. Intuitively bidirectional event summations can be seen as a combination of event stacks with linear weights, which can be learned implicitly by a neural network, so that event stack works well in image enhancement tasks. However, event stack will be noise-sensitive when the time resolution increases since they become sparser with more channel numbers, which degrades the restored image quality. Thus, it is necessary to transform event stacks [32] into a more robust representation.

Inspired by data-driven representations in the deep learning field, to fully utilize such information to address these problems, we propose a robust neural representation, named Neural Event STack (**NEST**), to replace $\{\mathbf{S}_{i\to j}^{c_i}\}_{j=1}^{N_f}$ and guide image enhancement. NEST representation explicitly learns the combination parameters of event stack to achieve a robust representation. By substituting bidirectional event summations with NESTs, high-quality frames can be restored according to Eq. (7) and Eq. (8) as below:

$$\mathbf{I}^{t_i} = f_d\left(\mathbf{B}, \mathbf{E}^i\right), \tag{9}$$

$$\mathbf{I}_{SR}^{t_i} = f_s\left(\mathbf{I}_{LR}^{t_i}, \mathbf{E}^i\right), \tag{10}$$

where f_d and f_s are implicit functions derived from Eq. (7) and Eq. (8), and $\mathbf{E}^i$ denotes a NEST.

From Eq. (9) and Eq. (10), we could see that once the NEST $\mathbf{E}^i$ is properly estimated, image enhancement tasks such as deblurring and super-resolution can be solved in a more robust manner. Besides, since the NEST is implemented by deep neural networks in a data-driven manner, it naturally extracts semantic information in the event sequence, which can facilitate the reconstruction of high-quality images. Therefore, our goal turns into estimating NESTs first, and then using NESTs to guide image deblurring and super-resolution procedures. To achieve that goal, we propose three specific sub-networks for estimating NESTs and modeling the implicit functions f_d and f_s respectively, as introduced in the following sections.

3.3 NEST Estimator

To obtain robust event representation, we design a NEST estimator to transform event stacks [32] into NESTs. From Eq. (3), we can divide $\mathbf{E}^i$ into two parts. The preceding part $\{\mathbf{S}_{i\to j}^{c_i}\}_{j=1}^{i-1}$ is represented by $\mathbf{E}_{\mathrm{p}}^i$, and the following part $\{\mathbf{S}_{i\to j}^{c_i}\}_{j=i}^{N_f}$ is represented by $\mathbf{E}_{\mathrm{f}}^i$, which encodes the events before and after time t_i respectively. Therefore, we design the NEST estimator to encode preceding and following events separately as shown in Fig. 2. Such a network can be expressed as:

$$\{\mathbf{E}^i\}_{i=1}^{N_f} = \{(\mathbf{E}_{\mathrm{p}}^i, \mathbf{E}_{\mathrm{f}}^i)\}_{i=1}^{N_f} = f_n\left(\{\mathbf{e}_i^{i+1}\}_{i=1}^{N_f}\right), \tag{11}$$

where f_n denotes our NEST estimator and $\{\mathbf{e}\}_i^{i+1}$ represents the events triggered in t_i to t_{i+1}.

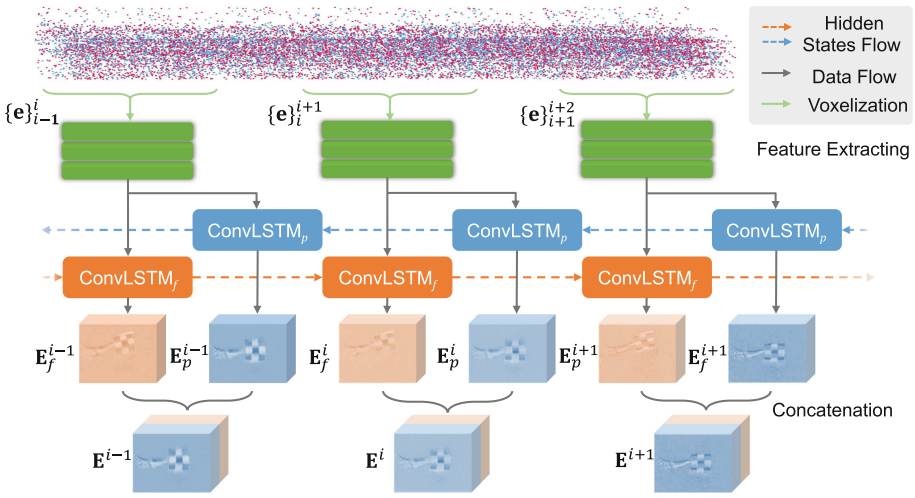


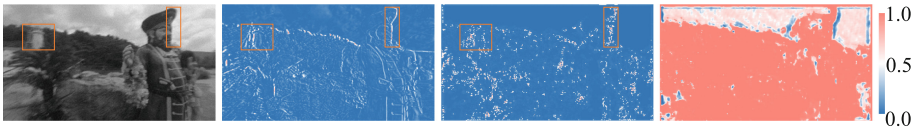
Fig. 2. The architecture of our NEST estimator, which consists of a parameter-shared feature extractor and a bidirectional ConvLSTM block. The input raw events $\{\mathbf{e}\}_i^{i+1}$ (triggered in t_i to t_{i+1}) are first binned into an event stack (voxelization), and then transformed into a NEST $\mathbf{E}^i$. ConvLSTM$_\mathrm{p}$ encodes the preceding part and ConvLSTM$_\mathrm{f}$ for the following part of events.

We first use a feature extractor block, consisting of multiple dense convolution layers [9], to perform local event feature extraction. Recent work has shown that dense convolution can extract high-level features, and filter most noisy events [4]. Then a bidirectional ConvLSTM block [28] is used to construct NESTs, which can not only encode temporal information lying in events but also fuse spatial information and reconstruct gradient information by the convolution operation.

From the event formation model [5], the expectation of event noise is zero. Since NESTs are generated by bi-directional encoding, paired noisy events are combined with temporal-variant thresholds, effectively suppressing noisy events. Besides, thanks to the data-driven encoding operation, NESTs also contain contextual information of the scene, which cannot be encoded by hand-crafted representations like event stacks [32]. As the example shown in Fig. 3, NESTs contain the statistical event information such as event-triggered frequency (Fig. 3 (c)) to indicate the blurry region, and a rough segmentation (Fig. 3 (d)) of the captured frame to distinguish the less blurred background, which both serve as global priors for reconstructing the high-quality image.

4 NEST: Application

In this section, we conduct three experiments: image deblurring (Sect. 4.1), super resolution (Sect. 4.2), and HFR video generation (Sect. 4.3) guided by NESTs to validate the effectiveness of NEST.



(a) Blurry image (b) Error map (c) 27th layer of NEST (d) 94th layer of NEST

Fig. 3. An example of NEST layer visualization. (a) Blurry image. (b) The error map between blurry image and ground truth, indicating the blurry region with higher difference values. (c) Visualization of the 27th layer of NEST, illustrating the blurry region. As highlighted in orange boxes, the blurry region has a higher response value, since more events are generated in this region. (d) Visualization of the 94th layer of NEST, separating less blurry sky apart from the foreground with different response values.

4.1 NEST-Guided Image Deblurring

After embedding events as NESTs, we can use them to conduct image deblurring. Since NESTs contain not only motion information but also global semantic information (an example shown in Fig. 3 (c) and (d)), we propose the NEST-guided D-Net to perform image deblurring by making full use of motion and global semantic information. Guided by NESTs, the image deblurring can be viewed as multi-modality fusion tasks. Thus, we adopt a U-Net-like [23] network architecture to perform image deblurring. We also formulate it as the residual learning with global connection, by fusing motion and intensity information to calculate the residual between the blurry image and the sharp one.[1]

Experiment Result. Our experiment can be divided into 3 parts. The first part (I) compares NEST-guided image deblurring with a state-of-the-art learning-based video deblurring method ESTRNN [36] and three state-of-the-art event-based image deblurring methods: EDI [21], LEDVDI [17], and eSL-Net [31]. To validate the effectiveness of the NEST representation, the second part (II) compares with the event stack representation method and another two data-driven event representations combined with our D-Net (denoted EvST+D/S [32], EST+D/S [6] and MatrixLSTM+D/S [3]). Besides, the third part (III) replaces eSL-Net's event stack representation with NEST representation (named NEST+eSL) to better illustrate the robustness of NEST. For a fair comparison, we retrained ESTRNN [36] on our training dataset.[2]

The quantitative comparison results are shown in Table 2 (a) and qualitative comparisons are shown in Fig. 4. We can see that our method outperforms others on all metrics. Compared to the video deblurring method ESTRNN [36], our method recovers sharper details encoding inside NESTs. As for event-based methods and other event representation methods, our method restored sharp images with fewer artifacts, with NEST's robust event representation. Thanks to the motion and semantic information encoded inside the NESTs, our network can handle blurry images with complicated real scenarios. Besides, as compari-

[1] Detailed D-Net and S-Net configurations are in the supplementary material.
[2] $*$ denotes retraining on our training dataset.

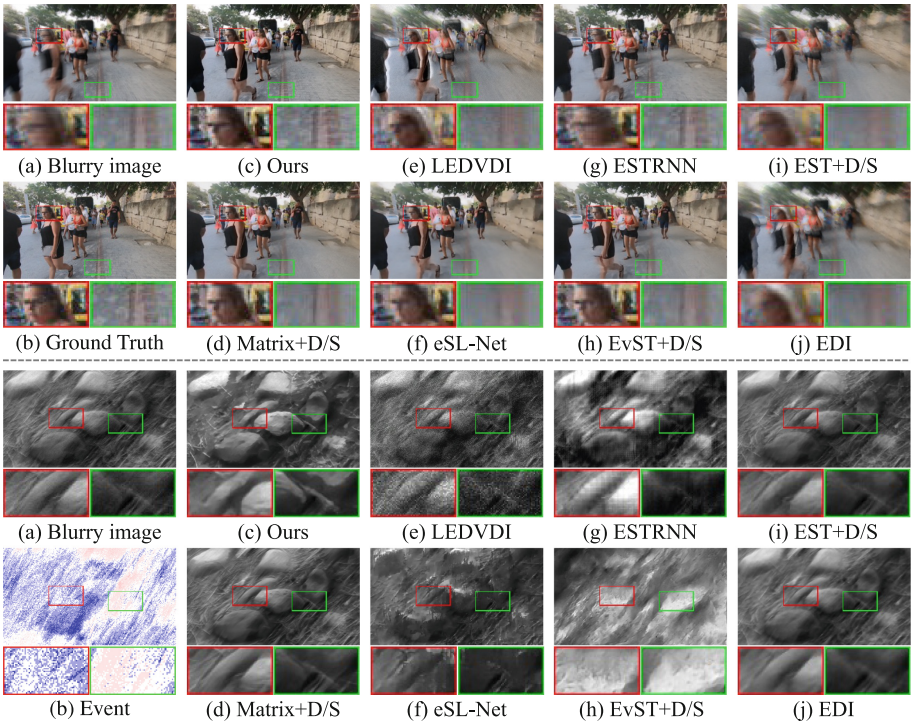


Fig. 4. Qualitative comparisons for deblurring application on synthetic data (upper) and real data (lower). (a) Blurry image. (b) Ground truth (synthetic data) / Event (real data). (c)∼(j) Deblurring results of ours,Matrix+D/S [3], LEDVDI [17], eSL-Net [31], ESTRNN [36], EvST+D/S [32], EST+D/S [6], and EDI [21]. Close-up views are provided below each image.

son between eSL-Net [31] and NEST+eSL has shown Table 2, much lower LPIPS values demonstrate NEST representation can improve the performance.

4.2 NEST-Guided Image Super-Resolution

Event cameras show higher temporal resolution than traditional cameras, which demonstrates the possibility of performing single image super-resolution like multi-frame super-resolution with events to relieve the ill-posed issue. However, frame alignment is an unavoidable difficulty for multi-frame super-resolution. Fortunately, the high temporal resolution property of events only brings slight changes for consecutive latent frames. Besides, our NEST estimator adopts a bidirectional ConvLSTM block, which also aligns temporal information implicitly. To better exploit semantic information hidden in NESTs, we design the NEST-guided S-Net for image super-resolution.

In our S-Net, we use multiple Residual in Residual Dense Blocks (RRDBs) as proposed in ESRGAN [34] to extract different features from NESTs and images

Table 2. Quantitative comparisons for deblurring (a) and super-resolution (b) application on the synthetic testing dataset. ↑ (↓) indicates the higher (lower), the better. The best performances are highlighted in **bold**. Our experiment can be divided into 3 parts: The first part (I) is to compare with state-of-the-art image-based and event-based image enhancement methods; the second part (II) compares "X+D/S", where "X" is other event representation methods; and the third part (III) compares "NEST+X", where "X" is another state-of-the-art event-based image enhancement method;

Methods		Applications					
		(a) Deblurring			(b) Super-resolution		
		PNSR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
I	EDI [21]	20.96	0.5752	0.2537	–	–	–
	LEDVDI [17]	22.08	0.6222	0.1905	–	–	–
	ESTRNN* [36]	30.52	0.8901	0.1105	–	–	–
	SPSR* [18]	–	–	–	27.63	0.7471	0.2763
	RBPN* [8]	–	–	–	27.23	0.7738	0.2956
	EvIntSR [7]	–	–	–	27.52	0.7334	0.2893
II	EvST+D/S [32]	31.09	0.8977	0.0689	28.89	0.7992	0.3150
	EST+D/S [6]	24.10	0.6987	0.2253	13.14	0.6574	0.4765
	Matrix+D/S [3]	31.28	0.9022	0.0596	27.88	0.7966	0.2844
III	eSL-Net [31]	29.73	0.8697	0.1078	28.23	0.7783	0.3950
	NEST+eSL [31]	29.92	0.8935	0.0634	28.87	0.7961	0.3096
Ours		**32.56**	**0.9354**	**0.0422**	**29.43**	**0.8128**	**0.2745**

independently. Besides, we incorporate features extracted from NESTs to the image branch, fusing temporal and global semantic information hidden in the NESTs to guide image super-resolution. Finally, we add a pixel shuffle layer [27] to rearrange features and predict image residual between LR image and HR image. By employing it to the upsampled image with bilinear interpolation, the super-resolved image can be restored.

Experiment Results. Similar to deblurring application, the first part (I) compares NEST-guided image super-resolution with two state-of-the-art learning-based image super-resolution methods SPSR [18] (taking in a single frame) and RBPN [8] (taking in multiple frames from a video), and two state-of-the-art event-based image super-resolution methods: eSL-Net [31] and EvIntSR [7]. The second part (II) compares with event stack representation method and two data-driven event representations combined with our S-Net (denoted EvST+D/S [32], EST+D/S [6] and MatrixLSTM+D/S [3]). The third part (III) replaces eSL-Net's event stack representation with NEST (named NEST+eSL).

The quantitative comparison results are shown in Table 2 (b) and qualitative comparisons are shown in Fig. 5. As experiments on real data show in Fig. 5,

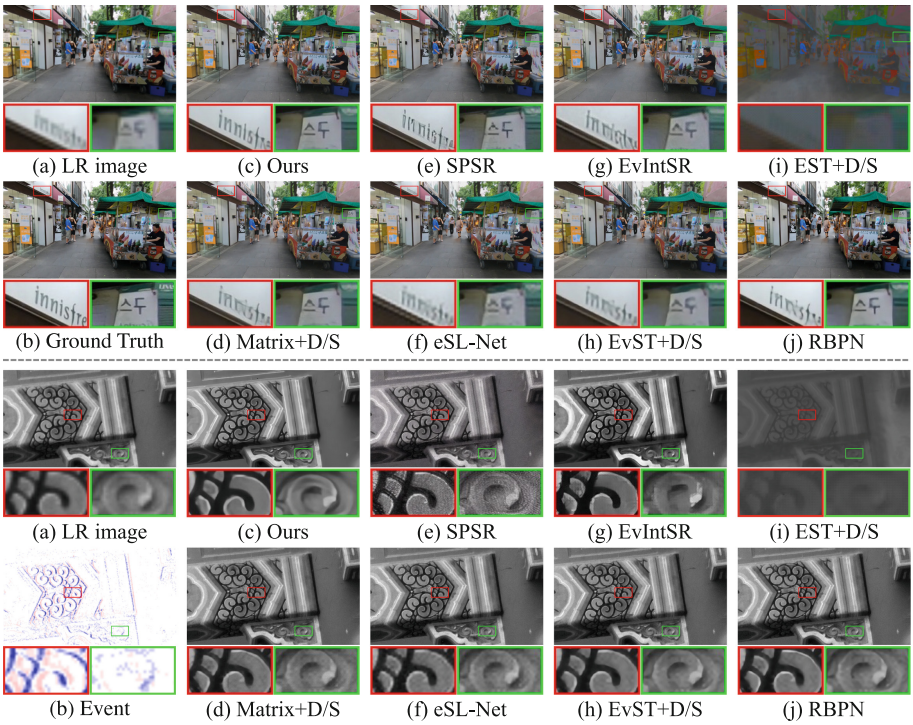


Fig. 5. Qualitative comparisons for super-resolution application on synthetic data (upper) and real data (lower). (a) LR image. (b) Ground truth (synthetic data) / Event (real data). (c)∼(j) Super-resolved 4× results of ours, Matrix+D/S [3], SPSR [18], NEST+eSL [31], EvIntSR [7], EvST+D/S [32], EST+D/S [6], and RBPN [8]. Close-up views are provided below each image.

results obtained by compared methods are distorted by noise, since the quality of intensity frames captured by DAVIS346 cameras is lower than the outputs of traditional cameras. But our method is noise-resistant thanks to NEST's robust representation. Like the deblurring application, eSL-Net [31] can achieve better performance combined with NEST.[3]

4.3 NEST-Guided HFR Video Generation

As Eq. (11) shows, we can obtain multiple NESTs in one pass by ConvLSTM. As shown in Table 1, compared to other LSTM-based event representations such as MatrixLSTM [3] or PhasedLSTM [20], our method preserves the intermediate states of ConvLSTM cells. Therefore, it brings the possibility to extend our D-Net and S-Net to process multiple NESTs in parallel to produce HFR

[3] Qualitative comparison between eSL-Net and NEST+eSL on deblurring and SR applications can be found in the supplementary material.

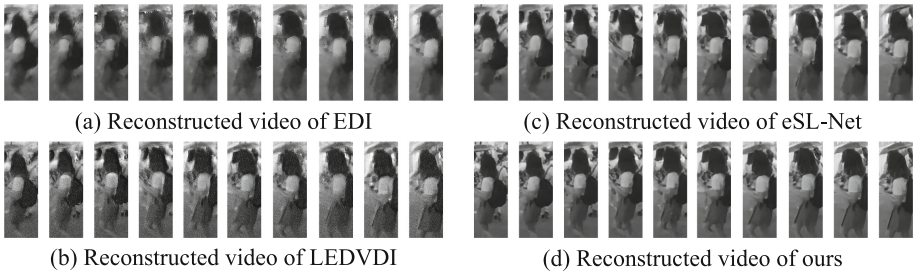


Fig. 6. Qualitative comparisons for HFR video generation application on synthetic data. The crop of reconstructed video frames from (a) EDI [21], (b) LEDVDI [17], (c) eSL-Net [31], and (d) ours are shown.

videos without modifying the original architecture. To implement this, after event sequence was transformed into NESTs. We can then generate multiple sharp images in parallel by D-Net by combining multiple NESTs with a single blurry image. After that, S-Net can generate multiple deblurred HR frames from LR frames to form an HFR video.

Experiment Results. We conduct qualitative comparisons on synthetic data in Fig. 6 for generating HFR videos from a single blurry image, compared with three state-of-the-art event-based HFR video generation methods: EDI [21], LEDVDI [17], and eSL-Net [31]. The results demonstrate that our method can generate frames with sharper edges and better visual quality than other state-of-the-art methods.

4.4 Implementation Details

Loss Function. We use the same loss function for training D-Net and S-Net, which is defined as

$$\mathcal{L} = \alpha \cdot \mathcal{L}_2(\mathbf{I}_o, \mathbf{I}_{gt}) + \beta \cdot \mathcal{L}_{perc}(\mathbf{I}_o, \mathbf{I}_{gt}), \tag{12}$$

where $\mathbf{I}_o$ denotes output image, $\mathbf{I}_{gt}$ for ground truth, and α and β are set to 200 and 0.5 respectively. $\mathcal{L}_2$ denotes the loss on L_2 norm and $\mathcal{L}_{perc}$ for perceptual loss, which is defined as

$$\mathcal{L}_{perc}(\mathbf{I}_o, \mathbf{I}_{gt}) = \mathcal{L}_2(\phi_h(\mathbf{I}_o), \phi_h(\mathbf{I}_{gt})), \tag{13}$$

where ϕ_h denotes the feature map from h-th layer of VGG-19 network [29] pre-trained on ImageNet [24], and we use activations from $VGG_{3,3}$ and $VGG_{5,5}$ convolutional layer here.

Training Details. We implement our method using PyTorch on an NVIDIA 3090Ti GPU. D-Net and S-Net are both trained for 100 epochs and after the first 50 epochs, we linearly decay the learning rate to 0 over the next 50 epochs.

Table 3. Quantitative evaluation results of ablation study on the synthetic testing dataset.

Methods	Applications					
	(a) Deblurring			(b) Super-resolution		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
W/o global	30.74	0.8932	0.0569	28.66	0.7917	0.3075
W/o feature	31.68	0.9064	0.0520	29.20	0.8096	0.2768
Our complete model	**32.56**	**0.9354**	**0.0422**	**29.43**	**0.8128**	**0.2745**

The initial learning rate is set to 1×10^{-3} for D-Net and 1×10^{-4} for S-Net, respectively, and ADAM optimizer [13] is used in the training procedure.

Dataset. Our training and testing datasets are adopted from Wang *et al.* [31]. As their datasets only contain the gray-scale images, we regenerate RGB blurry images and LR images from the original REDS dataset [19] as Wang *et al.* [31] suggested. And our real data are captured by a DAVIS346 camera.

4.5 Ablation Study

We conduct a series of ablation studies. The quantitative comparison results of deblurring application are shown in Table 3 (a) and super-resolution application for Table 3 (b), to verify the validity of each model design choice. We first show the effectiveness of the feature extractor in the NEST estimator by removing it (W/o feature). Next, we show the effectiveness of learning the residual in D-Net and S-Net by removing the global connection (W/o global). As the results show, our complete model achieves the best performance.

5 Conclusion

We propose a novel event representation (NEST) and apply it to event-based image deblurring, super-resolution, and HFR video generation. Thanks to the advantage of NESTs, all these image enhancement methods demonstrate superior performance over state-of-the-art methods.

Discussion. Limited by the low quality of the intensity frame captured by a DAVIS346 camera, although this paper demonstrates convincing evidence of fusing event data to improve the quality of an intensity frame, the final quality still has a gap with sharp frames captured by a modern DLSR camera. In our future work, we hope to build an event-RGB hybrid camera to fuse with high-quality intensity frames. Although event cameras also demonstrate the high dynamic range property (130 dB for DAVIS240 [2]), due to the lack of HDR paired images in our training dataset, we do not optimize the results to handle the HDR issue from a single LDR image with corresponding events. Extending NEST with a well-designed HDR dataset and network is also left as our future work.

Acknowledgement. This work was supported by National Key R&D Program of China (2021ZD0109803) and National Natural Science Foundation of China under Grant No. 62136001, 62088102.

References

1. Bi, Y., Chadha, A., Abbas, A., Bourtsoulatze, E., Andreopoulos, Y.: Graph-based object classification for neuromorphic vision sensing. In: Proc. of International Conference on Computer Vision. pp. 491–501 (2019)
2. 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)
3. Cannici, M., Ciccone, M., Romanoni, A., Matteucci, M.: A differentiable recurrent surface for asynchronous event-based data. In: Proc. of European Conference on Computer Vision. pp. 136–152 (2020)
4. Chen, H., Teng, M., Shi, B., Wang, Y., Huang, T.: Learning to deblur and generate high frame rate video with an event camera. arXiv preprint arXiv:2003.00847 (2020)
5. Duan, P., Wang, Z.W., Zhou, X., Ma, Y., Shi, B.: EventZoom: Learning to denoise and super resolve neuromorphic events. In: Proc. of Computer Vision and Pattern Recognition. pp. 12824–12833 (2021)
6. Gehrig, D., Loquercio, A., Derpanis, K.G., Scaramuzza, D.: End-to-end learning of representations for asynchronous event-based data. In: Proc. of International Conference on Computer Vision. pp. 5633–5643 (2019)
7. Han, J., Yang, Y., Zhou, C., Xu, C., Shi, B.: EvIntSR-Net: Event guided multiple latent frames reconstruction and super-resolution. In: Proc. of International Conference on Computer Vision. pp. 4882–4891 (2021)
8. Haris, M., Shakhnarovich, G., Ukita, N.: Recurrent back-projection network for video super-resolution. In: Proc. of Computer Vision and Pattern Recognition. pp. 3897–3906 (2019)
9. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proc. of Computer Vision and Pattern Recognition. pp. 4700–4708 (2017)
10. I., S.M.M., Choi, J., Yoon, K.: Learning to super resolve intensity images from events. In: Proc. of Computer Vision and Pattern Recognition. pp. 2765–2773 (2020)
11. Jiang, Z., Zhang, Y., Zou, D., Ren, J., Lv, J., Liu, Y.: Learning event-based motion deblurring. In: Proc. of Computer Vision and Pattern Recognition. pp. 3320–3329 (2020)
12. Jing, Y., Yang, Y., Wang, X., Song, M., Tao, D.: Turning frequency to resolution: Video super-resolution via event cameras. In: Proc. of Computer Vision and Pattern Recognition. pp. 7772–7781 (2021)
13. Kingma, D.P., Ba, J.: ADAM: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
14. Lagorce, X., Orchard, G., Galluppi, F., Shi, B.E., Benosman, R.B.: Hots: A hierarchy of event-based time-surfaces for pattern recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **39**(7), 1346–1359 (2016)
15. Li, Y., Zhou, H., Yang, B., Zhang, Y., Cui, Z., Bao, H., Zhang, G.: Graph-based asynchronous event processing for rapid object recognition. In: Proc. of International Conference on Computer Vision. pp. 934–943 (2021)

16. Lichtsteiner, P., Posch, C., Delbruck, T.: A 128 × 128 120 dB 15 μs latency asynchronous temporal contrast vision sensor. IEEE Journal of Solid-State Circuits **43**(2), 566–576 (2008)
17. Lin, S., Zhang, J., Pan, J., Jiang, Z., Zou, D., Wang, Y., Chen, J., Ren, J.: Learning event-driven video deblurring and interpolation. In: Proc. of European Conference on Computer Vision. pp. 695–710 (2020)
18. Ma, C., Rao, Y., Cheng, Y., Chen, C., Lu, J., Zhou, J.: Structure-preserving super resolution with gradient guidance. In: Proc. of Computer Vision and Pattern Recognition. pp. 7766–7775 (2020)
19. Nah, S., Baik, S., Hong, S., Moon, G., Son, S., Timofte, R., Mu Lee, K.: Ntire 2019 challenge on video deblurring and super-resolution: Dataset and study. In: Proc. of Computer Vision and Pattern Recognition Workshops. pp. 1996–2005 (2019)
20. Neil, D., Pfeiffer, M., Liu, S.C.: Phased LSTM: Accelerating recurrent network training for long or event-based sequences. In: Proc. of Neural Information Processing Systems. pp. 3882–3890 (2016)
21. 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: Proc. of Computer Vision and Pattern Recognition. pp. 6820–6829 (2019)
22. Rebecq, H., Gehrig, D., Scaramuzza, D.: ESIM: an open event camera simulator. In: Conference on Robot Learning. pp. 969–982 (2018)
23. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: Proc. of International Conference on Medical Image Computing and Computer Assisted Intervention. pp. 234–241 (2015)
24. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet large scale visual recognition challenge. International Journal of Computer Vision **115**(3), 211–252 (2015)
25. Sekikawa, Y., Hara, K., Saito, H.: Eventnet: Asynchronous recursive event processing. In: Proc. of Computer Vision and Pattern Recognition. pp. 3887–3896 (2019)
26. Shang, W., Ren, D., Zou, D., Ren, J.S., Luo, P., Zuo, W.: Bringing events into video deblurring with non-consecutively blurry frames. In: Proc. of International Conference on Computer Vision. pp. 4531–4540 (2021)
27. Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A.P., Bishop, R., Rueckert, D., Wang, Z.: Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: Proc. of Computer Vision and Pattern Recognition. pp. 1874–1883 (2016)
28. Shi, X., Chen, Z., Wang, H., Yeung, D., Wong, W., Woo, W.: Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In: Proc. of Neural Information Processing Systems. pp. 802–810 (2015)
29. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
30. Sironi, A., Brambilla, M., Bourdis, N., Lagorce, X., Benosman, R.: HATS: Histograms of averaged time surfaces for robust event-based object classification. In: Proc. of Computer Vision and Pattern Recognition. pp. 1731–1740 (2018)
31. Wang, B., He, J., Yu, L., Xia, G.S., Yang, W.: Event enhanced high-quality image recovery. In: Proc. of European Conference on Computer Vision (2020)
32. Wang, L., I., S.M.M., Ho, Y., Yoon, K.: Event-based high dynamic range image and very high frame rate video generation using conditional generative adversarial networks. In: Proc. of Computer Vision and Pattern Recognition. pp. 10081–10090 (2019)

33. 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: Proc. of Computer Vision and Pattern Recognition. pp. 8312–8322 (2020)
34. Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., Change Loy, C.: ESRGAN: Enhanced super-resolution generative adversarial networks. In: Proc. of European Conference on Computer Vision Workshops. pp. 63–79 (2018)
35. Yao, M., Gao, H., Zhao, G., Wang, D., Lin, Y., Yang, Z., Li, G.: Temporal-wise attention spiking neural networks for event streams classification. In: Proc. of International Conference on Computer Vision. pp. 10221–10230 (2021)
36. Zhong, Z., Gao, Y., Zheng, Y., Zheng, B.: Efficient spatio-temporal recurrent neural network for video deblurring. In: Proc. of European Conference on Computer Vision. pp. 191–207 (2020)
37. Zhu, A.Z., Yuan, L., Chaney, K., Daniilidis, K.: Unsupervised event-based learning of optical flow, depth, and egomotion. In: Proc. of Computer Vision and Pattern Recognition. pp. 989–997 (2019)