Learning to Super-Resolve Blurry Images With Events

Lei Yu, *Member, IEEE*, Bishan Wang, Xiang Zhang, Haijian Zhang, *Senior Member, IEEE*, Wen Yang, *Senior Member, IEEE*, Jianzhuang Liu, *Senior Member, IEEE*, and Gui-Song Xia, *Senior Member, IEEE*

***Abstract*—Super-*R*esolution from a single motion *B*lurred image (SRB) is a severely ill-posed problem due to the joint degradation of motion blurs and low spatial resolution. In this article, we employ events to alleviate the burden of SRB and propose an *E*vent-enhanced SRB (E-SRB) algorithm, which can generate a sequence of sharp and clear images with *H*igh *R*esolution (HR) from a single blurry image with *L*ow *R*esolution (LR). To achieve this end, we formulate an event-enhanced degeneration model to consider the low spatial resolution, motion blurs, and event noises simultaneously. We then build an *e*vent-enhanced *S*parse *L*earning *Net*work (eSL-Net++) upon a dual sparse learning scheme where both events and intensity frames are modeled with sparse representations. Furthermore, we propose an event shuffle-and-merge scheme to extend the single-frame SRB to the sequence-frame SRB without any additional training process. Experimental results on synthetic and real-world datasets show that the proposed eSL-Net++ outperforms state-of-the-art methods by a large margin. Datasets, codes, and more results are available at https://github.com/ShinyWang33/eSL-Net-Plusplus.**

***Index Terms*—Deblurring, denoising, event camera, intensity reconstruction, sparse learning, super-resolution.**

I. Introduction

SUPER-RESOLUTION (SR) is a fundamental low-level vision task that aims at recovering a *high-resolution* (HR) image from a *low-resolution* (LR) input [1]. It is known to be an ill-posed problem and often coupled with motion blurs under dynamic scenes with fast-moving objects [2], [3], [4]. The concurrence of multiple image degradations makes the SR problem more challenging. Even though both the problems of image SR and motion deblurring have been investigated separately for decades and promising results have been achieved [1], [4], [5], [6], [7], it is reported that simply superimposing a motion deblurring module on an image SR module may either amplify the unwanted artifacts or lose detailed information [8], [9].

Instead of cascading approaches, it has been demonstrated that *S*uper-*R*esolution from a single motion *B*lurred LR image (SRB) can be better tackled by simultaneously resolving motion ambiguities [9], [10], [11], which is itself severely ill-posed [6], [12]. Recently, promising results for SRB have been obtained by kernel-based approaches if uniform motions are valid [9], [10], [13], [14], [15]. However, most scenes from the real-world scenario are with non-uniform motions, e.g., non-rigid or moving objects, violating the uniform motion assumption. To address this problem, various approaches have been proposed either by estimating motion flows from video sequences [15] or learning end-to-end deep neural networks [16], [17], [18], [19]. These methods are either domain-specific only for face and text images [16], [17] or heavily rely on the performance of the deblurring sub-module [18], [19], and thus are not reliable for general image SR tasks oriented to natural images with complex motions.

In this article, we propose to utilize the event camera to enhance the performance of SRB. Event cameras are bio-inspired sensors that can perceive dynamic scenes and emit asynchronous events, which are triggered to respond to the brightness change and generally represented as compositions of position, time stamp, and polarity [23], [24]. With extremely low latency (in the order of μs), the triggered events inherently encode the intra-frame motions and textures with extremely high temporal resolution [25], [26]. This property motivates us to super-resolve blurry images with events for more challenging SR problems, e.g., complex natural scenes with large and non-uniform motions.

- *Motion Deblurring.* The embedded intra-frame information behind events compensates the erased motions and textures from blurry LR images [27], [28], which significantly relieves the burden of motion deblurring [28].
- *Super-Resolution.* The extremely high temporal resolution of events preserves intra-frame temporal continuities of dynamic scenes when encountering motion blurs [24]. Therefore, similar to video SR [29], the temporal correlation can be leveraged through events to boost the SR performance even with a single motion-blurred image.

Manuscript received 2 April 2022; revised 12 December 2022; accepted 23 January 2023. Date of publication 30 January 2023; date of current version 30 June 2023. This work was supported in part by the National Natural Science Foundation of China under Grants 62271354, 61871297, 61922065, U22B201, 41820104006, and 61871299 and in part by Natural Science Foundation of Hubei Province, China under Grant 2021CFB467. Recommended for acceptance by L. Zhang. *(Corresponding authors: Lei Yu and Gui-Song Xia.)*

Lei Yu, Bishan Wang, Xiang Zhang, Haijian Zhang, and Wen Yang are with the School of Electronic Information, Wuhan University, Wuhan 430072, China (e-mail: ly.wd@whu.edu.cn; wangbs@whu.edu.cn; xiangz@whu.edu.cn; haijian.zhang@whu.edu.cn; yangwen@whu.edu.cn).

Gui-Song Xia is with the School of Computer Science, Wuhan University, Wuhan 430072, China (e-mail: guisong.xia@whu.edu.cn).

Jianzhuang Liu is with the Huawei Noah's Ark Lab, Shenzhen 518000, China (e-mail: liu.jianzhuang@huawei.com).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TPAMI.2023.3240397.

Digital Object Identifier 10.1109/TPAMI.2023.3240397

Besides motion ambiguities, noises and disturbances can lead to visually unpleasant results since these imperfections would be amplified if image SR algorithms are fed with noisy inputs [8], [9]. Due to the special imaging mechanism, the event camera generally contains more noises and disturbances than the conventional frame-based camera [30]. Furthermore, noises of the event camera would be induced from both spatial and temporal domains, which raises the difficulty for event noise suppression [24]. It brings vagueness to the potential of event-based SRB even if the superiority of event-based motion deblurring has been validated [20], [21], [31], [32], [33]. Existing algorithms of event denoising (e.g., [30], [34]) can be applied separately, but it might suffer from over-suppression of events, leading to deteriorated deblurring performance.

Thus it is necessary to cope with motion ambiguities and event noises simultaneously for *E*vent-enhanced SRB (E-SRB). This article is devoted to achieving this end by adopting the *sparse learning* framework. Particularly, the degeneration process from HR sharp images to LR blurry images is revisited by adopting events, where motion blurs and noises are both considered. Based on this *E*vent-enhanced *D*egeneration *M*odel (EDM), the solution of E-SRB is then feasible by imposing sparsity on HR sharp images under specific dictionaries. And finally, an *e*vent-enhanced *S*parse *L*earning *Net*work (eSL-Net++) is proposed by unfolding the iterative algorithm for the ℓ_1-norm penalized optimization problem and training it on a synthetic dataset.

The contributions of this work are three-fold:

- We propose to adopt events to enhance the performance of SRB, where an EDM is presented by taking into account both event noises, motion blurs, and the low spatial resolution.
- We propose an eSL-Net++ to tackle the challenge of E-SRB based on a dual sparse learning scheme, where event noise suppression, motion deblurring, and image SR are simultaneously addressed.
- We propose a rigorous event shuffle-and-merge scheme to extend the eSL-Net++ to high frame-rate HR video sequence recovery from a single blurry LR image without any additional training process. Both synthetic and real-world datasets are built for training and testing.

This article is an extended version of our preliminary work, i.e., eSL-Net [35], with several significant improvements: (i) the event statistics are further considered in the EDM, based on which a *D*ual *S*parse *L*earning scheme (DSL) is proposed to suppress noises from events and blurry images by mutual compensations, and the resultant network is called eSL-Net++; (ii) a rigorous *e*vent *s*huffle-and-*m*erge scheme (ESM) is proposed to extend eSL-Net++ for video sequence recovery which achieves better performance than its previous version, i.e., eSL-Net; (iii) extended evaluations on both synthetic and real-world datasets are implemented. Fig. 1 shows an example by eSL-Net and eSL-Net++.

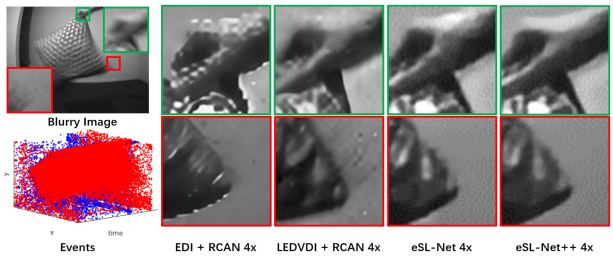


Fig. 1. Our eSL-Net++ reconstructs high-resolution, sharp, and clear intensity images for event cameras by *A*ctive *P*ixel *S*ensor (APS) frames and the corresponding event sequences. The eSL-Net++ performs much better than EDI [20] and LEDVDI [21] that are followed by an SR network RCAN [22].

II. Related Works

A. SR From Blurry LR Images

Image SR is to resolve the spatial ambiguities and recover missing detailed information from the LR images [1], [10], [11], [36]. Early attempts address this problem by optimizing an under-determined problem regularized by various handcrafted priors, which often suffer from undesired artifacts, e.g., over-smoothness [37], [38], [39]. Recent state-of-the-art methods turn to adopt deep priors which often surpass the traditional methods [2], [40], [41], [42]. However, when dealing with blurry LR inputs, the task of image SR becomes more challenging due to motion ambiguities and texture erasures [43], [44], [45]. Thus, it is essential to decouple motion deblurring from image SR.

Even though image SR and motion deblurring have been separately investigated for decades [1], [4], [5], [6], [7], simply cascading existing SR and deblurring methods may amplify unwanted artifacts and suffer from sub-optimal results [8], [9]. It is essential to resolve SR and the deblurring in a joint manner, where decoupling of motion ambiguities plays an important role. The uniform motions are commonly assumed to alleviate the difficulty where blurs can be considered within a unified degeneration model parameterized by a blurring kernel [9], [14], [15], [46]. Jointly optimizing the blurring kernel and image SR can effectively boost the overall performance in either non-blind [9], [46] or blind [10], [14], [15], [47] approaches. Gu et al. [48] propose an iterative kernel correction (IKC) to refine the blur kernel iteratively based on the previous SR results. Flow-based kernel prior (FKP) [10], kernel space representation [49], and mutual affine network (MANet) [50] take into account the spatially variant degradations. To bridge the real-to-synthesis gap, Wang et al. [47] propose a degradation-aware SR network (DASR) by learning degradation representations through contrastive learning. The kernel estimation accuracy plays an important role in blind SR since the degradation error can further be magnified by the SR process [48], which poses challenges to real-world scenarios, especially when encountering large and non-uniform motion blurs.

Image SR from LR images degraded by non-uniform motion blur is more challenging [19], [51]. To tackle this problem, Park et al. [15] propose to predict accurate motion flows from the camera pose and depth, estimated by stereo matching between inter-frame video sequences, which however is only valid for static scenes. On the other hand, several recent works jointly resolve image SR and motion deblurring using end-to-end trained deep networks [16], [17], [18], [19]. Specifically, some domain-specific priors are adopted to alleviate the ill-posedness but they are only valid for blurry face and text images [16], [17].

Deblur-then-SR approaches heavily rely on the performance of the deblurring sub-modules [18], [19], [36], thus not reliable for general image SR tasks oriented to natural images with complex motions.

B. Event-Based Low-Level Image Enhancement

Event-Based Motion Deblurring. Event cameras can "continuously" emit events asynchronously with extremely low latency (in the order of μs) [23], [24], inherently embedding motions and textures [25]. Thus the task of motion deblurring [20], [28] can be essentially alleviated by compensating blurry images with events [31], [35], [52]. In [52], events are approximated as the time differential of intensity frames, a complementary filter is proposed as a fusion engine and nearly continuous-time intensity frames can be generated. Pan et al. [20] propose an event-based deblurring approach by relating blurry *A*ctive *P*ixel *S*ensor (APS) frames and events with an *E*vent-based *D*ouble *I*ntegration (EDI) model. Afterward, a multiple-frame EDI model is proposed for high-rate video reconstruction by further considering inter-frame relations [28]. Recent works turn to convolutional neural networks (CNNs) for event-based motion deblurring by supervised learning from synthesized datasets composed of paired events, blurry inputs, and sharp images [21], [53], [54], [55]. And then the semi-supervised [56] and self-supervised [57] learning frameworks are respectively proposed to bridge the real-to-synthesis gap for event-based motion deblurring. Existing approaches solely focus on event-based motion deblurring, but rarely exploit events to resolve HR clear images from blurred inputs, i.e., E-SRB.

Event SR. Even though event cameras have an extremely high temporal resolution, their spatial (pixel) resolution is relatively low and yet not easy to be resolved physically [24]. Directly resolving the resolution purely from events is very challenging due to the heavily interfered events [34]. To achieve this end, Duan and Wang et al. [34], [58] propose spatial guidance for events by leveraging gradients and motions provided by images with a high spatial resolution from traditional cameras. If further providing camera poses, intensity images with a high resolution can be directly reconstructed purely from events by leveraging the Poisson equation [59]. On the other hand, recent works turn to exploit deep neural networks for event-based intensity recovery [21], [26], [53] and correspondingly, E2SRI [60] and EventSR [32] are proposed for event-based image recovery and super-resolution where classical techniques of generative adversary network (GAN), recurrent neural network (RNN) and U-Net are respectively exploited. Both E2SRI and EventSR are devoted to recovering images purely from events. Benefiting from the high temporal resolution, one can adopt events to enhance video frame interpolation (VFI) [61] and video super-resolution (VSR) [62]. However, they only accept sharp and clear inputs and require two consecutive image frames. Instead, only a single blurry LR image is available in this article, making the problem more challenging.

Event Denoising. The collected events from event cameras often suffer from noises and disturbances due to the thermal effects and the environmental brightness fluctuations [30]. And it becomes the main obstacle to follow-up applications, especially for low-level imaging tasks [24], [63], [64]. Temporal and spatial consistencies are commonly employed to remove noisy events [65], [66], but real events may violate such consistencies, especially with complex textures. To address this problem, the event time surface [67] provides a smooth manifold [68] while intensity images generate an event occurrence probability [30], [34], with which one can largely improve the robustness to event noises [30], [34], [68]. On the other hand, various methods have been proposed to address the problem of event noises jointly within the framework of event-based motion deblurring. Barua et al. [69] propose a learning-based approach to smooth the image gradients by imposing sparsity regularization and then exploit Poisson integration to recover the intensity image from gradients. Instead of sparsity, Munda et al. [68] introduce the manifold regularization over the event time surface [67] and propose a real-time intensity reconstruction algorithm. Even though the problem of SRB can be essentially alleviated by events, motion deblurring and event denoising are two challenges. Straightforwardly, concatenating modules of event denoising [30] and event-based motion deblurring [56] can tackle the task of event-based SRB. However, errors or artifacts might be aggregated, leading to sub-optimal solutions. It motivates us to resolve the event-based SRB and jointly tackle the problems of motion deblurring and super-resolution in the presence of tremendous event noises.

III. Problem Statement

Let $\boldsymbol{Y}$ be the image frame captured during the exposure time interval $\mathcal{T} \triangleq [0, T)$. Due to the imperfection of image sensors, the observed image $\boldsymbol{Y}$ may suffer from non-negligible quality degeneration due to noises, motion blurs, and low spatial resolution. And the degraded observation $\boldsymbol{Y}$ can be related to the high-quality (sharp, clear, and high-resolution) latent images $\boldsymbol{X}(t)$ as follows,

$$\begin{aligned} \boldsymbol{Y} &= \frac{1}{T}\int_{\mathcal{T}} \boldsymbol{I}(t)dt + \boldsymbol{\varepsilon}_Y, \\ \boldsymbol{I}(t) &= \boldsymbol{P}\boldsymbol{X}(t), \end{aligned} \tag{1}$$

where $\boldsymbol{I}(t)$ denotes the down-sampled version of $\boldsymbol{X}(t)$ via the operator $\boldsymbol{P}$ at time t, the integration over $\mathcal{T}$ represents the motion blur process [70] and $\boldsymbol{\varepsilon}_Y \sim \mathcal{N}(0, \sigma_Y^2)$ is the Gaussian noise with standard deviation σ_Y. Thus the super-resolution from an LR blurred image (SRB) can be represented as,

$$\boldsymbol{X}(t) = \mathrm{SRB}(\boldsymbol{Y}). \tag{2}$$

Obviously, finding a single HR image $\boldsymbol{X}(t), t \in \mathcal{T}$ or a sequence of HR images $\{\boldsymbol{X}(t)\}_{t\in\mathcal{T}}$ from a single blurry image $\boldsymbol{Y}$ is severely ill-posed.

To tackle this problem, we propose to alleviate the SRB problem (2) via the event camera, which outputs events whenever the logarithm of the intensity changes over a pre-setting threshold $c > 0$,

$$\log(\boldsymbol{I}(t)) - \log(\boldsymbol{I}(\tau)) = p \cdot c, \tag{3}$$

where $\boldsymbol{I}(t)$ and $\boldsymbol{I}(\tau)$ denote the latent instantaneous intensities at time t and τ, respectively, and $p \in \{+1, -1\}$ is the polarity representing the direction (increase or decrease) of the intensity change. Correspondingly, we have the following relationship,

$$\log\left(\boldsymbol{I}(t)\right) = \log\left(\boldsymbol{I}(f)\right) + c\int_f^t e(s)ds, \tag{4}$$

where $\boldsymbol{I}(f)$ is the latent image of time f, and $e(t) \triangleq \sum_{\tau \in \mathcal{T}_{f\to t}} p(\tau)\cdot\delta(t-\tau)$ denotes the event stream (represented as the spike train) triggered at position x during $\mathcal{T}_{f\to t} \triangleq [f, t)$. Finally, one can derive the following relation from (4),

$$\boldsymbol{I}(t) = \boldsymbol{I}(f) \circ \exp\left(c\int_f^t e(s)ds\right), \tag{5}$$

where $\circ$ represents the Hadamard product. Then according to (1) and (5), it has

$$\begin{aligned}\boldsymbol{Y} &= \boldsymbol{E}(f)\circ\boldsymbol{I}(f) + \boldsymbol{\varepsilon}_Y,\\ \boldsymbol{I}(f) &= \boldsymbol{P}\boldsymbol{X}(f),\end{aligned} \tag{6}$$

with

$$\boldsymbol{E}(f) \triangleq \frac{1}{T}\int_{\mathcal{T}} \exp\left(c\int_f^t e(s)ds\right)dt, \tag{7}$$

which is called the EDI in [20] representing the average of accumulated events. Benefiting from the high temporal resolution, (7) can provide missing intra-frame information caused by the blurring process and thus largely alleviate the difficulty of SRB. Moreover, (6) implies that one can even super-resolve $\boldsymbol{X}$ at any specific time $f \in \mathcal{T}$, which provides a convenient approach to recover a sequence of HR images $\boldsymbol{X}(f)$ from one single blurry LR image $\boldsymbol{Y}$.

However, the thermal effects or current leakage largely disturb the triggered events [30], resulting in violation of (6). Define the event accumulation $\boldsymbol{\Lambda}(f,t) \triangleq \int_f^t e(s)ds$. Then it can be assumed as a random variable drawn from the Poisson distribution [71],

$$\boldsymbol{\Lambda}(f,t) \sim \text{Poisson}(\lambda|t-f|), \tag{8}$$

with λ representing the event firing rate, i.e., the number of events triggered in a unit time interval. Applying the Taylor expansion, one can obtain $\exp(c\boldsymbol{\Lambda}(f,t)) \approx 1 + c\boldsymbol{\Lambda}(f,t)$ and then (7) becomes,

$$\boldsymbol{E}(f) \approx 1 + \frac{c}{T}\int_{\mathcal{T}} \boldsymbol{\Lambda}(f,t)dt. \tag{9}$$

Since $\boldsymbol{\Lambda}(f,t)$ is a random variable, the integration on the right side of (9) is still a random variable and obeys the Poisson distribution,

$$\int_{\mathcal{T}} \boldsymbol{\Lambda}(f,t)dt \sim \text{Poisson}\left(\lambda\rho\right),$$

with $\rho = f^2 - fT + \frac{T^2}{2}$. And when λ is large enough, Poisson distribution can be approximated by a Gaussian distribution with mean and variance equal to $\lambda\rho$. According to (9), we can finally derive that $\boldsymbol{E}(f)$ is approximately drawn from a Gaussian distribution

$$\boldsymbol{E}(f) \sim \mathcal{N}\left(\mu, \sigma^2\right), \tag{10}$$

where $\mu = 1 + c\lambda\rho/T$ and $\sigma = \frac{c\sqrt{\lambda\rho}}{T}$.

Thus we can assume an additive Gaussian noise $\boldsymbol{\varepsilon}_E \sim \mathcal{N}(0, \sigma^2)$ for $\boldsymbol{E}(f)$ and finally, obtain the event-enhanced degeneration model,

$$\begin{aligned}\boldsymbol{Y} &= \bar{\boldsymbol{E}}(f)\circ\boldsymbol{I}(f) + \boldsymbol{\varepsilon}_Y,\\ \boldsymbol{E}(f) &= \bar{\boldsymbol{E}}(f) + \boldsymbol{\varepsilon}_E,\\ \boldsymbol{I}(f) &= \boldsymbol{P}\boldsymbol{X}(f),\end{aligned} \tag{11}$$

where $\bar{\boldsymbol{E}}(f) \triangleq \mu$ represents the estimator of $\boldsymbol{E}(f)$. And our goal is to solve (11) for $\boldsymbol{X}(f)$. Specifically, given the observed image $\boldsymbol{Y}$ and the corresponding triggered events $\mathcal{E}_{\mathcal{T}} \triangleq \{e(t), t \in \mathcal{T}\}$, our goal is to reconstruct a high-quality intensity image $\boldsymbol{X}$ of the specified time $f \in \mathcal{T}$, i.e.,

$$\boldsymbol{X}(f) = \text{E-SRB}\left(\boldsymbol{Y}, \mathcal{E}_{\mathcal{T}}, f\right), \tag{12}$$

where the events $\mathcal{E}_{\mathcal{T}}$ and the blurry image $\boldsymbol{Y}$ are calibrated in the spatial and temporal domains [34], [61].

IV. Event-Based Sparse Learning for SRB

In this section, we first formulate the E-SRB problem using a DSL scheme by imposing sparsity on both noiseless events and latent clear intensities. Then, we build the eSL-Net++ network accordingly by unfolding the iteration stages of the sparse recovery algorithm and training eSL-Net++ with synthetic datasets. Finally, we present a rigorous approach to extend eSL-Net++ from the single to sequence E-SRB by event shuffles without further training.

A. Dual Sparse Learning With Events

Exploiting sparsity techniques, we assume that the LR sharp clear image $\boldsymbol{I}$, the HR sharp clear image $\boldsymbol{X}$ and the accumulated events $\bar{\boldsymbol{E}}$ can be sparsely represented by convolutional kernels or dictionaries $\boldsymbol{d}_I, \boldsymbol{d}_X, \boldsymbol{d}_E$, i.e.,

$$\begin{aligned}\boldsymbol{I} &= \boldsymbol{d}_I * \boldsymbol{\alpha}_I,\\ \boldsymbol{X} &= \boldsymbol{d}_X * \boldsymbol{\alpha}_X,\\ \bar{\boldsymbol{E}} &= \boldsymbol{d}_E * \boldsymbol{\beta},\end{aligned} \tag{13}$$

where $*$ denotes the convolutional operator and $\boldsymbol{\alpha}_I$, $\boldsymbol{\alpha}_X$, and $\boldsymbol{\beta}$ are correspondingly the sparse representations. Assume that the LR sharp clear image $\boldsymbol{I}$ and the HR sharp clear image $\boldsymbol{X}$ share the same sparse representation, i.e., $\boldsymbol{\alpha} = \boldsymbol{\alpha}_I = \boldsymbol{\alpha}_X$, if $\boldsymbol{d}_I$ and $\boldsymbol{d}_X$ are defined properly. Therefore, given an observed blurry image $\boldsymbol{Y}$ and the corresponding events $\mathcal{E}_{\mathcal{T}}$, we can find the sparse coefficients $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ over the convolutional dictionaries $\boldsymbol{d}_I$ and $\boldsymbol{d}_E$ by solving the following problem:

$$\min_{\boldsymbol{\alpha},\boldsymbol{\beta}} \frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{I}\circ\bar{\boldsymbol{E}}\|_2^2 + \frac{\lambda_1}{2}\|\boldsymbol{E} - \bar{\boldsymbol{E}}\|_2^2 + \lambda_2\|\boldsymbol{\alpha}\|_1 + \lambda_3\|\boldsymbol{\beta}\|_1, \tag{14}$$

where $\circ$ denotes the Hadamard product. The first two terms represent the data fidelity and the last two terms are the sparse

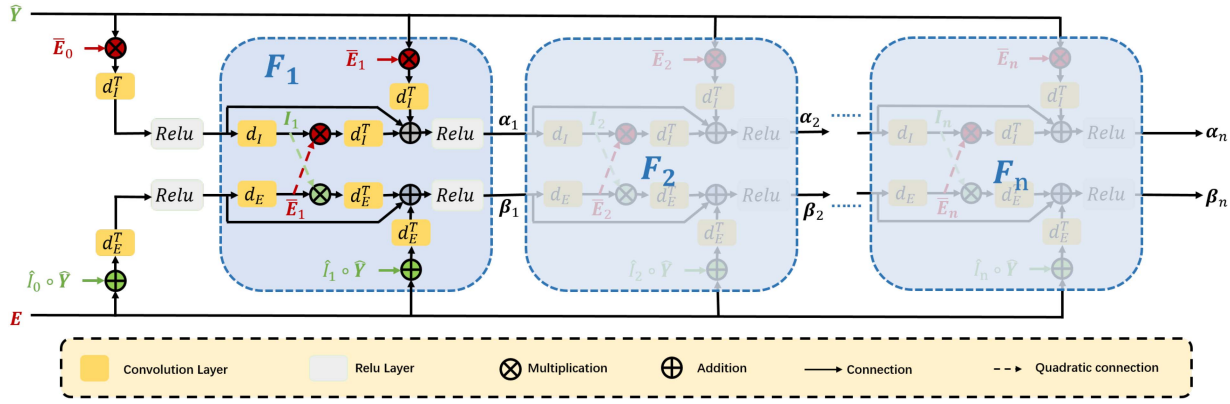


Fig. 2. The *D*ual *S*parse *L*earning (DSL) scheme, where (14) is optimized through the deep neural network by unfolding (15). Note that quadratic connection means that the input is squared and then fed into the block to which the arrow connects.

regularization with $\|\cdot\|_p$ denoting the ℓ_p-norm. The coefficients $\lambda_1, \lambda_2, \lambda_3$ balance the data fidelity and the sparse regularization. Actually, we can consider (14) as two sub-problems with respect to $\boldsymbol{I}$ and $\boldsymbol{E}$, and each is a typical LASSO problem [72], [73]. Thus, the classical solver for (14) is to use the iterative shrinkage thresholding algorithm (ISTA) [74] and the solutions can be found via the following iterative updates:

$$\begin{aligned}\boldsymbol{\alpha}^{+} &= \Gamma_{\eta\lambda_2}\left(\boldsymbol{\alpha} - \eta \boldsymbol{d}_I^T * \bar{\boldsymbol{E}} \circ \left(\bar{\boldsymbol{E}} \circ (\boldsymbol{d}_I * \boldsymbol{\alpha}) - \boldsymbol{Y}\right)\right), \\ \boldsymbol{\beta}^{+} &= \Gamma_{\eta\lambda_3}\left(\boldsymbol{\beta} - \eta \boldsymbol{d}_E^T * \boldsymbol{I} \circ (\boldsymbol{I} \circ (\boldsymbol{d}_E * \boldsymbol{\beta}) - \boldsymbol{Y}) \right. \\ &\quad \left. - \eta\lambda_1 \boldsymbol{d}_E^T * (\boldsymbol{d}_E * \boldsymbol{\beta} - \boldsymbol{E})\right), \end{aligned} \tag{15}$$

with $\bar{\boldsymbol{E}} = \boldsymbol{d}_E * \boldsymbol{\beta}$, $\boldsymbol{I} = \boldsymbol{d}_I * \boldsymbol{\alpha}$, η denoting the step size, and $\Gamma_\theta(\iota) = \text{sign}(\iota)\max(|\iota| - \theta, 0)$ denoting the element-wise soft-thresholding function. After obtaining the optimum solution of the sparse coefficients $\boldsymbol{\alpha}$, we can finally recover the LR sharp clear image $\boldsymbol{I}$ and the HR sharp clear image $\boldsymbol{X}$ according to (13).

The iterative updates (15) provide a DSL scheme for the task of E-SRB, which implicitly unifies noise suppression and motion deblurring. In the DSL scheme, the sparse coefficients $\boldsymbol{\alpha}$ of the intensity images and $\boldsymbol{\beta}$ of the events can mutually compensate each other by leveraging the smoothness of frames and high temporal resolution of events. On the other hand, when eliminating the updates of $\boldsymbol{\beta}$, (15) degenerates to the preliminary version, i.e., eSL-Net [35], where $\boldsymbol{E}$ is directly fed without the noise suppression and mutual enhancement.

B. Network Modules

Instead of directly optimizing (14), we unfold the iterations (15) and build a CNN architecture for sparse learning, which consists of a fixed number of phases corresponding to the iterations of (15).

Dual Sparse Learning Module. The DSL scheme is shown in Fig. 2, where $\{\bar{\boldsymbol{E}}_i\}_{i=0,\ldots,n}$ and $\{\boldsymbol{I}_i\}_{i=0,\ldots,n}$ are respectively the reconstructions of the accumulated events $\bar{\boldsymbol{E}}$ and the restored sharp image $\boldsymbol{I}$ at the ith iteration via updated sparse coefficients $\boldsymbol{\alpha}_i$ and $\boldsymbol{\beta}_i$. Obviously, DSL jointly optimizes both $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ in a manner of mutual compensation that efficiently leverages information from image frames and events. Specifically, the DSL of (15) degenerates to its preliminary version [35] where noisy events $\boldsymbol{E}$ are directly used to update $\boldsymbol{\alpha}$ and thus may produce noisy reconstructions.

Instead of directly inputting image frame $\boldsymbol{Y}$ to the DSL module, we linearly transform it into the feature domain, i.e., $\hat{\boldsymbol{Y}}$. Then both $\hat{\boldsymbol{Y}}$ and $\boldsymbol{E}$ are fed into the DSL block to compute the sparse coefficients. The DSL block is implemented in a recursive manner that is composed of n recursive blocks $\boldsymbol{F}_1, \ldots, \boldsymbol{F}_n$ and the unfolded details are depicted in Fig. 2 where the ReLU layer is employed to implement Γ_θ. In summary, the DSL module accepts $\boldsymbol{E}$ and $\hat{\boldsymbol{Y}}$ and outputs sparse coefficients $\boldsymbol{\alpha}, \boldsymbol{\beta}$, i.e.,

$$\boldsymbol{\alpha}, \boldsymbol{\beta} = \text{DSL}(\boldsymbol{E}, \hat{\boldsymbol{Y}}).$$

Learnable Double Integral Module. According to (7), calculating $\boldsymbol{E}(f)$ from events $\boldsymbol{\mathcal{E}}_{\mathcal{T}}$ requires two integral operations. To approximate such integration, a *l*earnable *d*ouble *i*ntegral (LDI) network is proposed where two convolution layers and two sigmoid layers are adopted as shown in Fig. 3. And the input and output relation of LDI can be written as

$$\boldsymbol{E} = \text{LDI}(\boldsymbol{\mathcal{E}}_{\mathcal{T}}).$$

Reconstruction Module. Then, the outputs of the last recursion $\boldsymbol{F_n}$ are optimized sparse codes $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. Then we use convolutional layers followed by the shuffle layer [75] as the HR conventional kernel $\boldsymbol{d}_X$ to reconstruct the final sharp clear HR image frame $\boldsymbol{X}$. Besides, the denoised events $\bar{\boldsymbol{E}}$ can also be easily obtained by $\boldsymbol{d}_E * \boldsymbol{\beta}$. Finally, the input and output relation of the *REC*onstruction (REC) module can be written as

$$\boldsymbol{X}, \bar{\boldsymbol{E}} = \text{REC}(\boldsymbol{\alpha}, \boldsymbol{\beta}).$$

C. Overall eSL-Net++ and Its Training Strategy

The overall architecture of the proposed network for E-SRB is shown in Fig. 3. The eSL-Net++ is fed with a pair of inputs, i.e., the LR blurry image frame $\boldsymbol{Y}$ and the corresponding events during the exposure time interval $\mathcal{T}$, and outputs the HR sharp clear intensity image $\boldsymbol{X}$, i.e.,

$$\boldsymbol{X}, \bar{\boldsymbol{E}} = \text{eSL-Net++}\,(\boldsymbol{Y}, \boldsymbol{\mathcal{E}}_{\mathcal{T}})$$

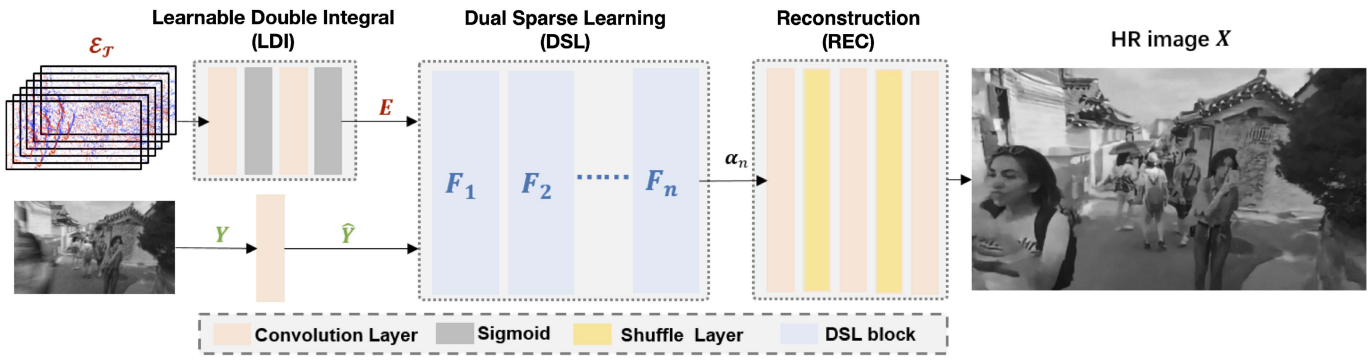


Fig. 3. Event-enhanced Sparse Learning Network for super-resolving blurry images with events, i.e., eSL-Net++. Coefficients β and event reconstruction $\bar{\boldsymbol{E}}$ are omitted here.

$$\triangleq \text{REC}\left(\text{DSL}\left(\text{LDI}\left(\boldsymbol{\mathcal{E}}_{\mathcal{T}}\right), \hat{\boldsymbol{Y}}\right)\right). \tag{16}$$

The proposed eSL-Net++ is trained over the synthetic dataset, where noise free events $\boldsymbol{\mathcal{E}}_{\mathcal{T}}^{gt}$ and the ground truth HR sharp image $\boldsymbol{X}^{gt}$ are both available. Correspondingly, the total training loss is composed of two aspects, i.e., event denoising error and image SR error:

$$\mathcal{L} = \zeta_1 \mathcal{L}_E + \zeta_2 \mathcal{L}_X, \tag{17}$$

where the event denoising error is defined as $\mathcal{L}_E = \|\boldsymbol{E}^{gt} - \bar{\boldsymbol{E}}\|_1$ with $\boldsymbol{E}^{gt} = \text{LDI}(\boldsymbol{\mathcal{E}}_{\mathcal{T}}^{gt})$, the image SR error is defined as $\mathcal{L}_X = \|\boldsymbol{X}^{gt} - \boldsymbol{X}\|_1$, and ζ_1, ζ_2 are balancing parameters. In our experiments, we set $[\zeta_1, \zeta_2] = [1, 1]$.

D. From Single to Sequence E-SRB by Event Shuffles

Training eSL-Net++ with the ground truth HR sharp image $\boldsymbol{X}^{gt}(f)$ at the specific time $f \in \mathcal{T}$, one can obtain a solver for the task of the single-frame E-SRB (2) at time f. Thus, we name the resultant network to resolve the HR image at time f as *eSL-Net++*$_f$ and the corresponding LDI as *LDI*$_f$, while the modules of DSL and REC are independent of f.

Even though one can train eSL-Net++ for different time f to get the sequence solver of E-SRB, it is time consuming for the training stage and cannot solve E-SRB for arbitrary time stamps. Thus we propose an easy method to extend the single-frame solver to the sequence-frame solver without any additional training procedure.

Suppose that eSL-Net++$_0$ has been trained, LDI_0 can be considered as the approximation of $\boldsymbol{E}(0) \approx \text{LDI}_0(\boldsymbol{\mathcal{E}}_{\mathcal{T}})$, i.e.,

$$\text{LDI}_0\left(\boldsymbol{\mathcal{E}}_{\mathcal{T}}\right) \approx \frac{1}{T}\int_0^T \exp\left(c\int_0^t e(s)ds\right)dt. \tag{18}$$

From (7), we have

$$\begin{aligned}\boldsymbol{E}(f) = \frac{1}{T}\Bigg(&\int_0^f \exp\left(c\int_f^t e(s)ds\right)dt \\ &+ \int_f^T \exp\left(c\int_f^t e(s)ds\right)dt\Bigg)\end{aligned}$$

$$\begin{aligned}= \frac{1}{T}\int_0^f \exp\left(c\int_0^{t'} -e(-s+f)ds\right)dt' \\ + \frac{1}{T}\int_0^{T-f} \exp\left(c\int_0^{t'} e(s+f)ds\right)dt,'\end{aligned} \tag{19}$$

where the first term is related to the events triggered in $[0, f)$ and the second term is related to the events in $[f, T)$. Correspondingly, we can divide events $\boldsymbol{\mathcal{E}}_{\mathcal{T}}$ into two subsets, i.e., $\boldsymbol{\mathcal{E}}_{[0,f)}$ and $\boldsymbol{\mathcal{E}}_{[f,T)}$. Then one can approximate $\boldsymbol{E}(f)$ via LDI_0 according to (18) and (19), which corresponds to a *rigorous event shuffle-and-merge* (RESM) scheme, i.e.,

$$\boldsymbol{E}(f) \triangleq \text{RESM}\left(\boldsymbol{\mathcal{E}}_{\mathcal{T}}, f\right) \tag{20}$$

$$\begin{aligned}\approx \frac{f}{T}\cdot \text{LDI}_0\left(\mathcal{R}\left(\boldsymbol{\mathcal{E}}_{[0,f)}\right)\right) \\ + \left(1-\frac{f}{T}\right)\cdot \text{LDI}_0\left(\mathcal{S}\left(\boldsymbol{\mathcal{E}}_{[f,T)}\right)\right),\end{aligned} \tag{21}$$

where $\mathcal{R}$ and $\mathcal{S}$ are event shuffle operators, i.e., $\mathcal{R}(\boldsymbol{\mathcal{E}}_{[0,f)}) \triangleq \{-e(-t+f), t \in [0, f)\}$ represents the event operator consisting of time shift, flip, and polarity reversal, and $\mathcal{S}(\boldsymbol{\mathcal{E}}_{[f,T)}) \triangleq \{e(t+f), t \in [0, T-f)\}$ represents the event operator of time shift, as shown in Fig. 4(b).

It implies by (20) that the double integral $\boldsymbol{E}(f)$ can be approximately calculated with LDI_0 by simply shuffling collected events $\boldsymbol{\mathcal{E}}_{\mathcal{T}}$. Thus for arbitrary time f, we can get the corresponding HR sharp reconstruction $\boldsymbol{X}(f)$,

$$\boldsymbol{X}(f), \bar{\boldsymbol{E}}(f) = \text{REC}\left(\text{DSL}\left(\boldsymbol{E}(f), \hat{\boldsymbol{Y}}\right)\right),$$

with $\bar{\boldsymbol{E}}(f)$ the denoised version of $\boldsymbol{E}(f)$. Theoretically, we can generate a video with frame-rate as high as the DVS's (*D*ynamic *V*ision *S*ensor) eps (*e*vents *p*er *s*econd).

V. Experiments

In this section, we evaluate the proposed E-SRB approaches, i.e., eSL-Nets (eSL-Net [35] and its extension eSL-Net++) and compare them with existing state-of-the-art SRB methods. The datasets, codes, and more results are available at https://github.com/ShinyWang33/eSL-Net-Plusplus. The performances of our

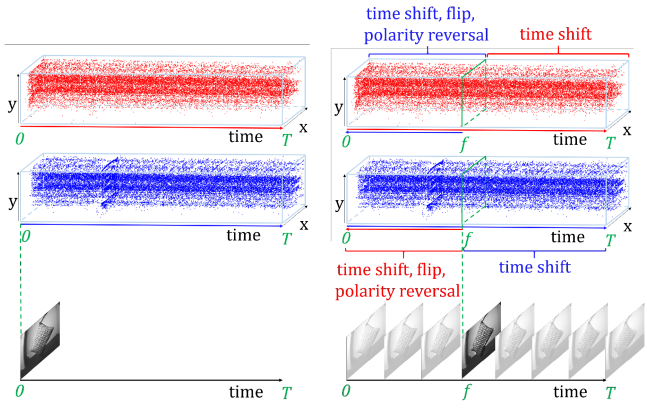


Fig. 4. The *R*igorous *E*vent *S*huffle-and-*M*erge (RESM) scheme: (a) the input event tensors when reconstructing the image of time $f = 0$ (from top to bottom are positive event stream, negative event stream, and recovered latent image of time $f = 0$); (b) the shuffle of events to input tensors when reconstructing the image of time $f > 0$.

methods are quantitatively evaluated by PSNR and SSIM [76]. In addition, the qualitative evaluation is also given by visualizing reconstructed HR images.

A. Datasets and Implementation Details

In this article, both synthetic and real-world datasets are built. To train the proposed eSL-Net, we first synthesize the GoPro dataset consisting of LR blurry noisy images with labeled HR ground-truth images as well as the event streams, which are simulated from high frame-rate video sequences via ESIM [77]. Meanwhile, the real-world dataset is also provided to validate the effectiveness of our proposed network in real-world scenarios.

Dataset With Synthetic Events. We build the synthetic GoPro dataset containing *HR clear images*, *LR blurry images*, and *Event streams*.

HR clear images. We choose the continuous sharp clear images with resolution of 1280×720 from the GoPro dataset [4] as our ground truth. It consists of 25,650 HR sharp clear frames with various natural and manmade objects captured at different locations.

LR blurry images. Similar to [70], we first increase the frame-rate of LR sharp clear images to 960 fps using the method proposed in [78], and then generate motion blurred images by averaging 17 consecutive frames. *Event streams.* For each blurry image frame, the ESIM [77] is employed to synthesize concurrent events from the interpolated high frame-rate LR clear images. We also add noise to the synthesized events to imitate real scenarios. Specifically, we first calculate the number of the original events (denoted by N_o) in the exposure time of the corresponding blurry frames and then generate N_n noise events with $N_n \triangleq \omega N_o$ with the event noise ratio ω ($\omega = 0.3$ in our GoPro dataset). Afterward, we assign the noise events with pixel coordinates and polarities randomly sampled from the uniform distribution. Finally, we apply the rounding operation to the sampled coordinates and polarities to keep them in the integer format and add the noise events to the original event streams.

According to the partitions of the GoPro dataset [4], images and event streams in the synthetic dataset from 240 video sequences are used for training and the rest from 30 video sequences are for evaluation.

Dataset With Real-World Events. To validate the effectiveness of the proposed eSL-Net in real-world scenarios, the performance is further evaluated on two datasets with real-world events, i.e., HQF from [79] and RWS partially built by ourselves.

HQF. The HQF dataset [79] contains real-world events and sharp LR clear ground-truth frames that are well-exposed to avoid motion blurs, while the motion blurs can be synthesized by averaging over 49 consecutive sharp clear image frames. Finally, the HQF dataset contains real-world events and synthetic blurry images as well as the ground-truth frames, enabling quantitative evaluation with real-world events. Note that HQF only provides LR clear images and thus we first up-sample the LR clear images by RCAN [22] and utilize the generated HR images as the ground truth.

RWS. The real-world scenes (RWS) dataset is built mainly to validate the effectiveness of our proposed method on different event cameras (a DAVIS346 camera and a DAVIS240 camera [20]) over different scenes. Since RWS only contains real-world events and real-world blurry APS frames while the HR clear images can not be provided as the ground truth, we only evaluate SRB methods qualitatively on the RWS dataset.

Implementation Details. We implement the proposed eSL-Nets using PyTorch on NVIDIA Titan-RTX 3090 GPUs. The step size η in (15) is set to 0.01 for stabilization of eSL-Nets. The parameter λ_1 in (14) is a learnable parameter in our model, and λ_2 and λ_3 serve as the offsets in (15), which are implicitly embedded in the convolutional layers. Thus, the parameters λ_1, λ_2, and λ_3 can be automatically learned during network training. Both eSL-Net [35] and eSL-Net++ are trained over the synthetic GoPro dataset. Adam optimizer is used and the maximum epoch of training iterations is set to 50. The learning rate starts at 8×10^{-4} and then decays by 50% every 10 epochs. To determine the number of recursions n, we train 5 different models respectively with 5, 10, 15, 20, and 25 recursion blocks and their SRB performance in terms of PSNR is respectively 25.13 dB, 25.63 dB, 25.73 dB, 25.74 dB, and 25.78 dB. Apparently, increasing the recursion number n can improve reconstruction performance, but at the sacrifice of the computation load. Thus, we choose $n = 15$ to balance the computational burden and reconstruction performance.

B. Results of Single Frame SRB

SRB is a joint task where both image SR and motion deblurring should be tackled. Thus we compare our eSL-Net and eSL-Net++ to state-of-the-art methods that can achieve SRB including the end-to-end methods, i.e., GFN [18], DASR [47], FKP [10], and MANet [50], and the two-stage Deblur-then-SR methods by cascading the motion deblurring with the image SR. Different motion deblurring methods are compared including the frame-based methods, i.e., SRN [80] and CDVD [81], and the event-based methods, i.e., EDI [20], RED-Net [56], and LED-VDI [21]. The task of the image SR is fulfilled by RCAN [22].

TABLE I
QUANTITATIVE COMPARISON OF OUR PROPOSED ESL-NET AND ESL-NET++ WITH THE STATE-OF-THE-ART SR METHODS WITH INPUTS OF PURE IMAGES, PURE EVENTS, AND FUSION OF IMAGES AND EVENTS

Method	Inputs		Single Frame Reconstruction				Sequence Reconstruction				Params
			Synthetic events		Real events		Synthetic events		Real events		
	Events	Image	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	
GFN	✗	✓	22.43	0.5295	21.10	0.7507	/	/	/	/	12.2M
DASR	✗	✓	20.87	0.5124	17.65	0.7128	/	/	/	/	5.97M
FKP	✗	✓	19.40	0.4717	17.25	0.6646	/	/	/	/	**0.59M**
MANet	✗	✓	20.01	0.4993	17.48	0.6786	/	/	/	/	9.89M
SRN+RCAN$_{4\times}$	✗	✓	22.86	0.5601	20.91	0.7562	/	/	/	/	34.9M
CDVD+RCAN$_{4\times}$	✗	✓	22.20	0.5967	20.95	0.7577	/	/	/	/	47.3M
E2VID+RCAN$_{4\times}$	✓	✗	12.15	0.3385	9.88	0.5392	12.14	0.3369	9.86	0.5382	41.8M
E2SRI+RCAN$_{2\times}$	✓	✗	11.90	0.4183	11.28	0.5832	11.85	0.4162	11.34	0.5853	40.9M
EDI+RCAN$_{4\times}$	✓	✓	22.48	0.6196	21.34	0.7689	21.77	0.6021	19.79	0.7473	31.1M
LEDVDI+RCAN$_{4\times}$	✓	✓	23.75	0.5676	21.81	0.7942	23.30	0.5670	21.34	**0.7914**	36.1M
RED-Net+RCAN$_{4\times}$	✓	✓	23.54	0.5608	22.78	0.7644	22.82	0.5443	21.64	0.7475	25.3M
EFNet+DASR$_{4\times}$	✓	✓	23.85	0.5556	22.73	0.7996	/	/	/	/	14.4M
LEDVDI+RealBasicVSR	✓	✓	22.76	0.6286	18.11	0.6995	22.23	0.6150	16.97	0.6841	26.9M
RED-Net+RealBasicVSR	✓	✓	23.41	0.5999	19.80	0.7273	20.49	0.5262	21.19	0.7450	16.0M
eSL-Net (Ours)	✓	✓	25.32	0.6705	23.91	0.8075	23.80	0.6455	22.16	0.7790	1.32M
eSL-Net++ (Ours)	✓	✓	**25.73**	**0.6824**	**23.99**	**0.8087**	**24.69**	**0.6602**	**22.99**	0.7913	1.41M

Evaluations are conducted for single frame and sequence reconstructions, respectively.

We further make comparisons to the methods of intensity restoration from pure events, i.e., E2VID [26] and E2SRI [60]. The performance of single-frame SRB is evaluated quantitatively and qualitatively in the following of this subsection.

Quantitative Results. The quantitative performances of our proposed eSL-Net and eSL-Net++ are evaluated over the GoPro dataset and the HQF dataset, where the ground-truth clear HR images are available. The PSNRs and SSIMs over different datasets are given in Table I. Although our networks are trained on the synthetic GoPro dataset with synthesized motion blurs and events, they still perform well on datasets with real events as shown in Table I, which validates the ability of our models to generalize from synthetic to real-world scenes.

Both eSL-Net and eSL-Net++ outperform the state-of-the-art methods by a large margin. Specifically, the problem of motion blur can be well tackled with events and thus event-based SRB methods (RED-Net+RCAN, LEDVDI+RCAN, eSL-Net, and eSL-Net++) perform better than SRB methods without events, including joint SR-and-deblurring method (GFN), blind SR methods (DASR, FKP, and MANet), and deblurring-then-SR methods (SRN+RCAN and CDVD+RCAN). Even though EDI is a traditional method, cascading EDI with RCAN still achieves comparable results to deep networks including SRN and CDVD. Among event-based SRB methods, our proposed eSL-Net and eSL-Net++ exhibit significant improvements over the two-stage event-based SRB methods cascaded with RCAN, i.e., EDI+RCAN, RED-Net+RCAN, and LEDVDI+RCAN, which validates the effectiveness of our unified framework with motion deblurring and super-resolution.

Furthermore, we also compare our eSL-Net and eSL-Net++ with the blind SR method, i.e., DASR [47], with a motion-deblurring pre-processing, where the most recent state-of-the-art method for event-based motion deblurring, i.e., EFNet [54], is applied to reduce the difficulty of SRB. The quantitative results are presented in Table I. It is shown that the deblurring pre-processing by EFNet improves the performance of blind SR approaches, but our eSL-Net and eSL-Net++ still perform better than EFNet+DASR.

On the other hand, eSL-Net++ outperforms its previous version eSL-Net on both datasets with synthetic and real events, which shows the effectiveness of the dual sparse learning scheme for event denoising.

Qualitative Results. We further evaluate the performances qualitatively on the GoPro, HQF, and RWS datasets. The results are visualized in Fig. 5 for the GoPro and HQF datasets and Fig. 6 for the RWS dataset.

Fig. 5 illustrates the qualitative results on the GoPro (first two rows) and HQF (second two rows) datasets with synthesized motion blurs. Both eSL-Net and eSL-Net++ can give precise HR reconstructions which exhibit the most similar appearances to the ground-truth HR images. Furthermore, eSL-Net++ still outperforms eSL-Net, producing less noises and artifacts, e.g., halo effects.

Fig. 6 visualizes the SR results from blurry images of the RWS dataset with real-world motion blurs and events, where the ground-truth HR images are not available. We only present the SR results of GFN, LEDVDI+RCAN, and our proposed eSL-Nets that have comparable quantitative results in Table I. Compared to traditional SRB methods, i.e., GFN, E-SRB methods, i.e., LEDVDI+RCAN and eSL-Nets, can largely remove motion blurs benefiting from introducing events, which is consistent to the quantitative results. Furthermore, eSL-Nets produce better visualization results than cascading approaches, e.g., LEDVDI+RCAN, while eSL-Net++ achieves the best visual quality.

C. Results of Sequence Frame SRB

The task of sequence frame SRB is more challenging than single frame SRB. To the best of our knowledge, existing approaches for SRB can only tackle single frame SRB,

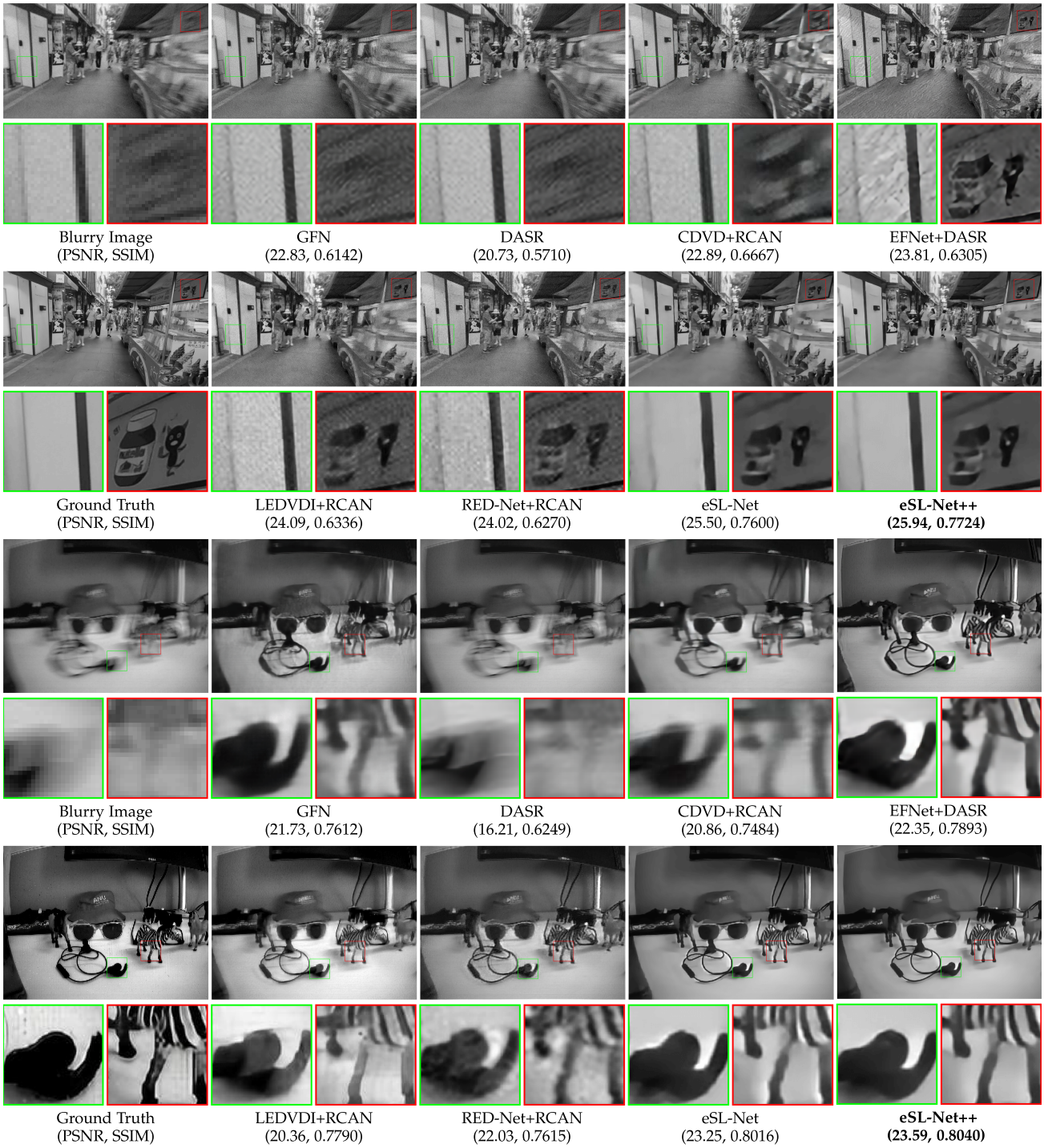


Fig. 5. Quantitative and qualitative results on the GoPro dataset (top two rows) and the HQF dataset (bottom two rows), where our proposed eSL-Net and eSL-Net++ are compared to GFN, DASR, CDVD+RCAN, RED-Net+RCAN, LEDVDI+RCAN, and EFNet+DASR.

e.g., GFN, DASR, FKP, and MANet. Thus, we only compare our proposed eSL-Nets with the event-based SRB methods, i.e., EDI+RCAN, LEDVDI+RCAN, and RED-Net+RCAN. Quantitative and qualitative results are respectively given in Table I and Fig. 7.

Regarding the quantitative results, we evaluate the performance of sequence frame SRB methods with respect to three reconstructed latent sharp HR images from a single blurry image, respectively corresponding to the start, middle, and end of the exposure time. Accordingly, EDI and LEDVDI are utilized to reconstruct latent HR images of the same timestamps. For the datasets with synthetic or real events, our eSL-Nets are with significantly better performances compared to the other E-SRB methods in terms of both PSNR and SSIM.

For qualitative comparisons on sequence frame SRB, we show the sequence frame reconstructions from a single blurry

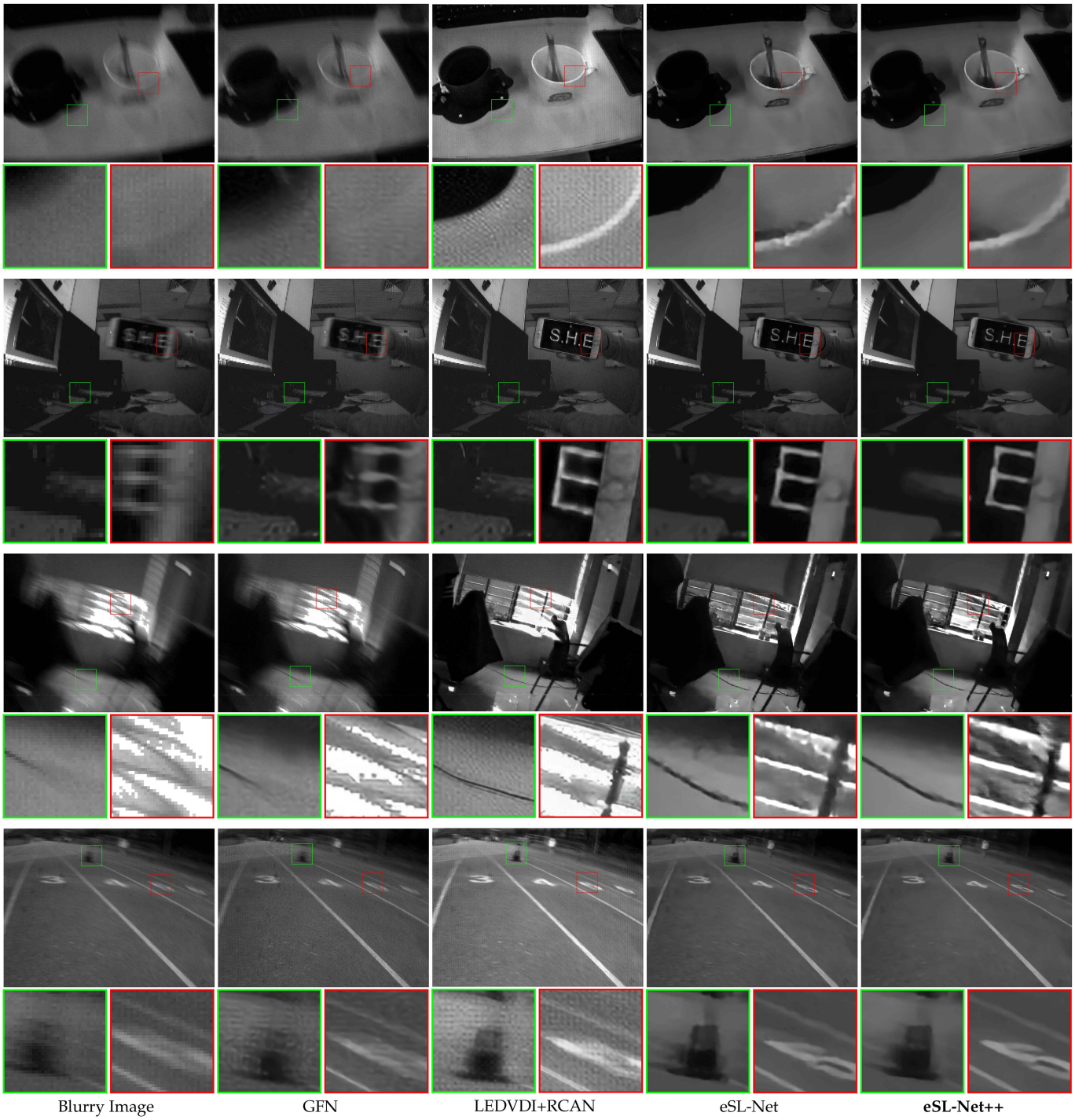


Fig. 6. Qualitative results on the RWS dataset, where our proposed eSL-Net and eSL-Net++ are compared with GFN and LEDVDI+RCAN.

image of the RWS dataset by four methods, i.e., EDI+RCAN, LEDVDI+RCAN, eSL-Net and eSL-Net++, as shown in Fig. 7. As EDI and eSL-Nets can output arbitrary number of frames, while LEDVDI only outputs 6 frames, we reconstruct 13 frames for EDI and eSL-Nets to facilitate the frame alignment. Obviously, the reconstructions of EDI+RCAN still suffer from blurry effects and noises as illustrated in the 2nd row of Fig. 7. LEDVDI+RCAN also produces noisy outputs as illustrated in the third row of Fig. 7 and reconstruction of higher frame-rate HR videos requires additional training phase. Our eSL-Net is likely to generate halo artifacts around black edges and the reconstructed images are still noisy as illustrated in the 4th row of Fig. 7. Benefiting from the dual sparse learning module and the rigorous event shuffle-and-merge module, eSL-Net++ can alleviate the effects of event noises and halo artifacts, leading to the best visualization performance on sequence reconstruction.

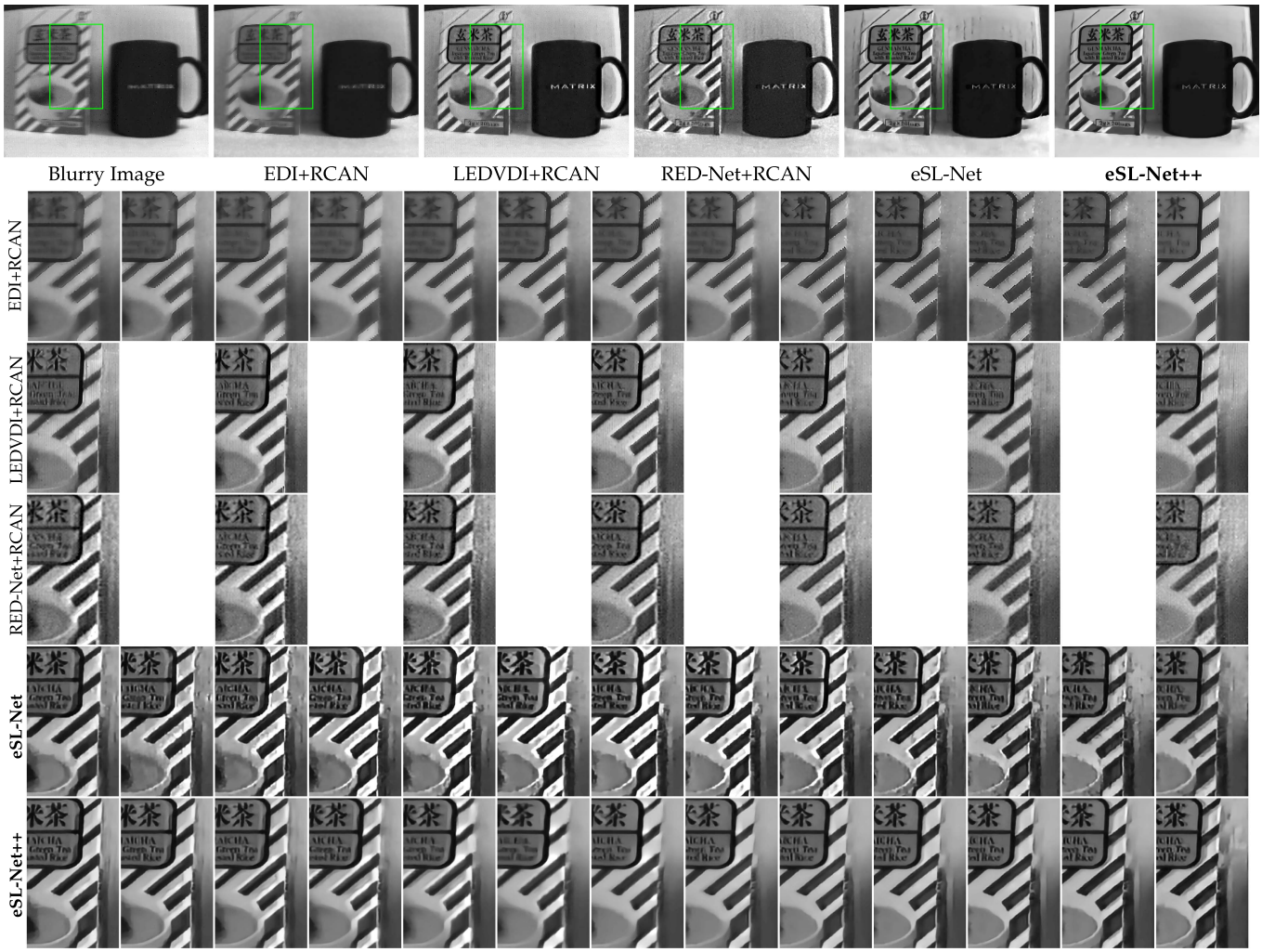


Fig. 7. Qualitative results of single frame reconstructions (1st row) and their corresponding sequence reconstructions (2nd-6th rows) on the RWS dataset, where our proposed eSL-Net (5th row) and eSL-Net++ (6th row) are compared to EDI+RCAN (2nd row), LEDVDI+RCAN (3 rd row) and RED-Net+RCAN (4th row).

D. Comparisons to Video SR

The video SR (VSR) can boost SR performance by leveraging the temporal inter-frame correlations [29]. Thus we also make comparisons of our eSL-Net and eSL-Net++ to Deblur-then-VSR approaches, where we first exploit two event-based motion deblurring methods, i.e., LEDVDI and RED-Net, to reconstruct sharp LR video sequences from a single blurry image, and then apply the VSR approach, i.e., RealBasicVSR [29] as the consecutive SR procedure. The quantitative results are given in Table I and our eSL-Net and eSL-Net++ perform much better than the two Deblur-then-VSR approaches. According to the qualitative results in Fig. 8, we can observe that eSL-Net and eSL-Net++ give clearer upscaled faces in the top row and ox feet in the bottom row than the Deblur-then-VSR methods. The performance of two-stage Deblur-then-VSR methods is confined and the deblurring errors may even be magnified by the SR procedure, e.g., halo effects on the right shoulder of the girl in the top row and the vertical stripes on the wall in the bottom row.

E. Ablation Study

Different from the preliminary version [35], i.e., eSL-Net, eSL-Net++ further contains the *D*ual *S*parse *L*earning (DSL) scheme to take into account event noises and a *R*igorous *E*vent *S*huffle-and-*M*erge (RESM) scheme for sequence reconstruction. In this section, ablation studies on these two schemes, i.e., DSL and RESM, are analyzed with the sequence reconstruction on the GoPro dataset with synthetic events and the HQF dataset with real events. Four different experiments are implemented to analyze the effectiveness of DSL and RESM, as shown in Table II.

Importance of DSL. Since events $\mathcal{E}_{\mathcal{T}}$ and image frames $\boldsymbol{I}$ are coupled together in a multiplicative manner as (11), noises in events $\mathcal{E}_{\mathcal{T}}$ would inevitably affect the final SRB results, as shown in Fig. 9. Without DSL, eSL-Net directly utilizes the output of the LDI module which suffers from serious noises, as shown in Fig. 9(b). And accordingly, these noises would be aggregated and degrade the final SRB performance, as shown in Fig. 9(f). On the other hand, the DSL scheme utilized in eSL-Net++ alleviates the

Fig. 8. Qualitative comparisons with Deblur-then-VSR methods on the GoPro (top row) and HQF (bottom row) datasets, where the VSR algorithm employed in this experiment is RealBasicVSR [29].

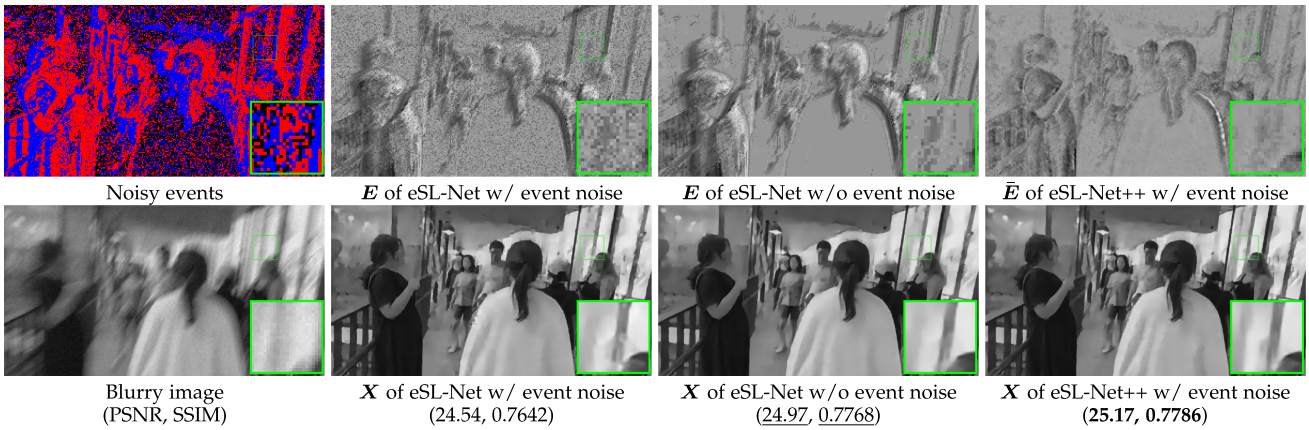


Fig. 9. Qualitative ablation study on the DSL module. (b) and (f) are respectively the output of LDI, i.e., $\boldsymbol{E}$ and the final SR result of eSL-Net [35] when inputting noisy synthesized events (a) and the corresponding blurry image (e). Accordingly, (c) and (g) are the results of eSL-Net when inputting noiseless events, while (d) and (h) are the results of eSL-Net++ when inputting noisy events.

TABLE II
ABLATION STUDIES ON *DSL* AND *RESM* OVER THE GOPRO DATASET WITH SYNTHETIC EVENTS AND THE HQF DATASET WITH REAL EVENTS

DSL	RESM	Synthetic events		Real events	
		PSNR	SSIM	PSNR	SSIM
✗	✗	23.80	0.6455	22.16	0.7790
✗	✓	24.42	0.6530	22.86	0.7890
✓	✗	24.12	0.6579	22.43	0.7898
✓	✓	**24.69**	**0.6602**	**22.99**	**0.7913**

burden of event noises by leveraging the smoothness of images (Fig. 9(h)), finally leading to better SRB reconstructions than eSL-Net (Fig. 9(f)). We further compare eSL-Net++ to eSL-Net without event noise (i.e., Fig. 9(c) and (g)) which outputs less noisy $\boldsymbol{E}$ and $\boldsymbol{X}$ than eSL-Net with event noise (i.e., Fig. 9(b) and (f)). Nevertheless, eSL-Net++ with noisy events still achieves the best performance, which validates the effectiveness of the DSL module.

Quantitative ablations of the DSL scheme are shown in Table II. Specifically, we respectively implement two types of experiments, i.e., with or without the RESM scheme. For methods without the RESM scheme (1st and 3 rd rows in Table II), the DSL scheme brings remarkable improvements in terms of PSNR and SSIM on both GoPro and HQF datasets, i.e., $0.32/0.0122$ on the GoPro dataset and $0.27/0.0108$ on the HQF dataset. When with the RESM scheme (2nd and 4th rows in Table II), the DSL scheme can further boost the SRB accuracy with improvements of $0.27/0.0072$ on the GoPro dataset and $0.13/0.0023$ on the HQF dataset.

Importance of Rigorous ESM (RESM). The eSL-Nets are originally trained for the task of single-frame SRB. To extend eSL-Nets to sequence SRB without additional training process, an event shuffle scheme has been proposed according to the chronological order of events and image frames in [35]. Instead, we further analyze the theoretical relationship between single-frame E-SRB and sequence-frame E-SRB and accordingly propose a rigorous ESM method as (20), extending eSL-Net++ to the task of sequence SRB. To validate its effectiveness, two groups of experiments are done with different settings, i.e., with or without DSL, and the quantitative results are given in Table II. It shows that the RESM scheme brings notable improvements on both experimental settings, i.e., up to 0.7 dB in PSNR and 0.01 in SSIM without the DSL scheme, and 0.57 dB in PSNR and 0.0023 in SSIM with the DSL scheme.

To further illustrate the effectiveness of DSL and RESM, we conduct qualitative comparisons of the sequence restorations respectively by eSL-Net (Baseline), eSL-Net+RESM (RESM),

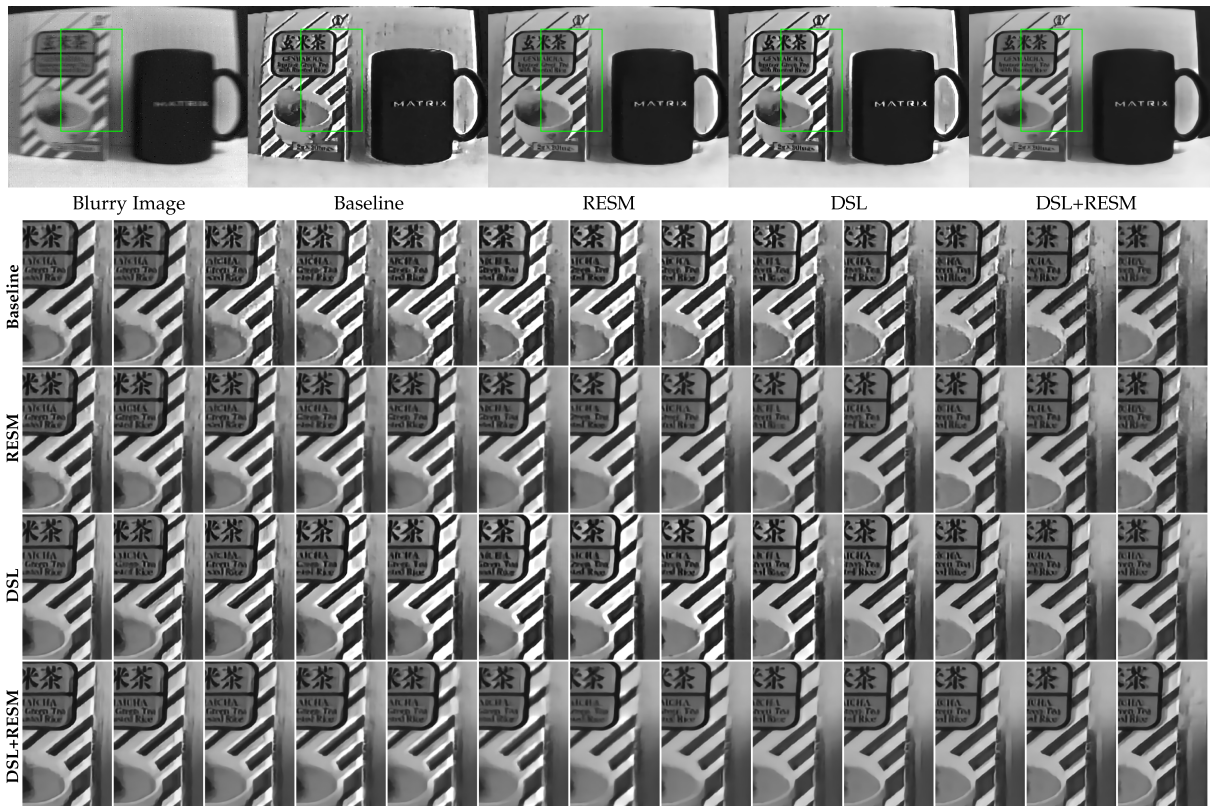
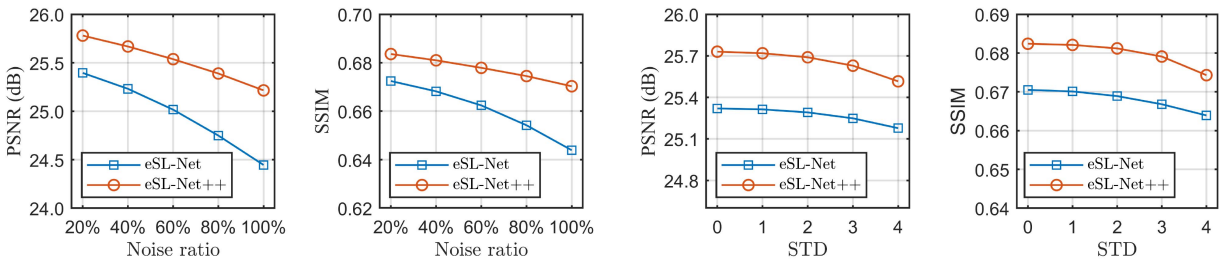


Fig. 10. Qualitative results of single frame reconstructions (1st row) and their corresponding sequence reconstructions (2nd-5th rows) on the RWS dataset, where Baseline indicates eSL-Net without DSL and RESM.



(a) Results under different event noise (b) Results under different image noise

Fig. 11. Quantitative comparisons of single frame reconstruction under different levels of event and image noise on the GoPro dataset. (a) Results under different event noise, where the noise ratio indicates the ratio of noise events to the original events. (b) Results under different image noise, where STD represents the standard deviation of the white Gaussian noise added to images.

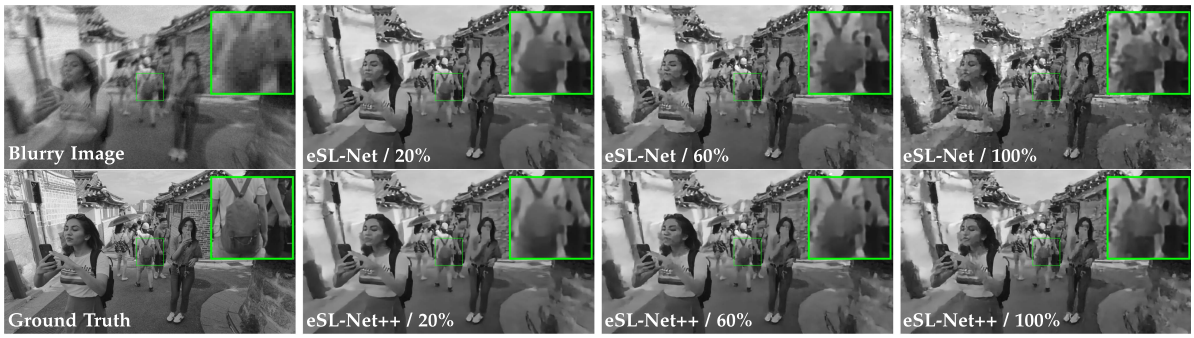
eSL-Net+DSL (DSL), and eSL-Net++ (DSL+RESM). The results on the RWS dataset are shown in Fig. 10 .

- The improvement from DSL can be found by comparing DSL to Baseline. Clearly, the restorations of DSL are less noisy than those of Baseline. Similarly, DSL+RESM gives less noisy restorations than RESM due to the usage of the DSL module.
- The inaccurate ESM scheme used in eSL-Net often leads to halo effects on high-contrast edges for sequence restoration, as shown in the intermediate frames restored by Baseline and DSL, while these halo effects can be effectively suppressed by the RESM scheme used in eSL-Net++, as shown in the intermediate frames restored by RESM and DSL+RESM.

Either the DSL scheme or the RESM scheme can improve the SRB performance. Specifically, DSL can effectively reduce performance degradation caused by noises, while RESM can avoid halo effects brought by the inaccurate ESM in [35] and thus improves the accuracy of sequence restoration without additional network training process. Thus, eSL-Net++ achieves the best performance benefiting from these two schemes.

F. Robustness to Event and Image Noise

To validate the robustness of our proposed algorithms, we conduct quantitative and qualitative comparisons of single frame reconstruction under different levels of event and image noise on the GoPro dataset. Regarding the event noise, we add event



(a) Results under different event noise

(b) Results under different image noise

Fig. 12. Qualitative comparisons under different levels of event noise and image noise. (a) Results under different event noise, where 20%, 60%, and 100% indicate that the noise events are 20%, 60%, and 100% of the original events, respectively. (b) Results under different image noise, where 0, 2, and 4 indicate the standard deviations of the white Gaussian noise added to images are 0, 2, and 4, respectively.

noise with varying ratios of noise ranging from 20% to 100% and evaluate the performance of eSL-Net and eSL-Net++. The quantitative results are plotted in Fig. 11(a). Regarding the image noise, we add Gaussian noise to the blurry image with different standard deviations. The corresponding performance of eSL-Net and eSL-Net++ can be referred from Fig. 11(b). The quantitative results validate that eSL-Net++ outperforms eSL-Net at different levels of noise no matter on events or images, which validates that eSL-Net++ is more robust than eSL-Net.

Correspondingly, we also provide qualitative comparisons under different levels of event noise and image noise as shown in Fig. 12, where we can draw conclusions consistent with the quantitative comparisons.

VI. Conclusion

In this article, we have proposed a novel network named *eSL-Net++* for E-SRB. We start from formulating an *E*vent-enhanced *D*egeneration *M*odel (EDM) to simultaneously take into account the image degradation caused by noises, motion blurs, and down-sampling. We present a *D*ual *S*parse *L*earning scheme (DSL) by assuming the sparsity over latent images and events, and then built a deep neural network, i.e., eSL-Net++, by unfolding DSL iterations. Compared to its previous version, i.e., eSL-Net, eSL-Net++ further takes into account event noises and extends to sequence frame SRB by a rigorous event shuffle-and-merge scheme, leading to superior performance to eSL-Net. Extensive experiments on synthetic and real-world data have demonstrated the effectiveness and superiority of our eSL-Net++.

References

[1] Z. Wang, J. Chen, and S. C. Hoi, "Deep learning for image super-resolution: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3365–3387, Oct. 2020.
[2] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2016.
[3] B. Bascle, A. Blake, and A. Zisserman, "Motion deblurring and super-resolution from an image sequence," in *Proc. Eur. Conf. Comput. Vis.*, 1996, pp. 573–582.
[4] S. Nah et al., "NTIRE 2019 challenge on video deblurring and super-resolution: Dataset and study," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2019, pp. 1996–2005.
[5] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.
[6] S. Nah, S. Son, S. Lee, R. Timofte, and K. M. Lee, "NTIRE 2021 challenge on image deblurring," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2021, pp. 149–165.
[7] C. Tian, L. Fei, W. Zheng, Y. Xu, W. Zuo, and C.-W. Lin, "Deep learning on image denoising: An overview," *Neural Netw.*, vol. 131, pp. 251–275, 2020.
[8] A. Singh, F. Porikli, and N. Ahuja, "Super-resolving noisy images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 2846–2853.
[9] K. Zhang, W. Zuo, and L. Zhang, "Learning a single convolutional super-resolution network for multiple degradations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3262–3271.
[10] J. Liang, K. Zhang, S. Gu, L. V. Gool, and R. Timofte, "Flow-based kernel prior with application to blind super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, Art. no. 10.
[11] W. Niu, K. Zhang, W. Luo, Y. Zhong, X. Yu, and H. Li, "Blind motion deblurring super-resolution: When dynamic spatio-temporal learning meets static image understanding," *IEEE Trans. Image Process.*, vol. 30, no. 5, pp. 7101–7111, Aug. 2021.
[12] S. Gu and R. Timofte, "A brief review of image denoising algorithms and beyond," in *Proc. Inpainting Denoising Challenges*, 2019, pp. 1–21.

[13] J. Pan, H. Bai, J. Dong, J. Zhang, and J. Tang, "Deep blind video super-resolution," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 4811–4820.
[14] T. Yamaguchi, H. Fukuda, R. Furukawa, H. Kawasaki, and P. Sturm, "Video deblurring and super-resolution technique for multiple moving objects," in *Proc. Asian Conf. Comput. Vis.*, 2010, pp. 127–140.
[15] H. Park and K. M.U. Lee, "Joint estimation of camera pose, depth, deblurring, and super-resolution from a blurred image sequence," in *Proc. IEEE Int. Conf. Comput. Vis*., 2017, pp. 4613–4621.
[16] X. Yu, B. Fernando, R. Hartley, and F. Porikli, "Super-resolving very low-resolution face images with supplementary attributes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 908–917.
[17] X. Xu, D. Sun, J. Pan, Y. Zhang, H. Pfister, and M.-H. Yang, "Learning to super-resolve blurry face and text images," in *Proc. IEEE Int. Conf. Comput. Vis*., 2017, pp. 251–260.
[18] X. Zhang, H. Dong, Z. Hu, W.-S. Lai, F. Wang, and M.-H. Yang, "Gated fusion network for joint image deblurring and super-resolution," in *Proc. Brit. Mach. Vis. Conf.*, 2018, pp. 1–13.
[19] X. Zhang, F. Wang, H. Dong, and Y. Guo, "A deep encoder-decoder networks for joint deblurring and super-resolution," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process*., 2018, pp. 1448–1452.
[20] L. Pan, C. Scheerlinck, X. Yu, R. Hartley, M. Liu, and Y. Dai, "Bringing a blurry frame alive at high frame-rate with an event camera," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 6820–6829.
[21] S. Lin et al., "Learning event-driven video deblurring and interpolation," in *Proc. Eur. Conf. Comput. Vis*., 2020, pp. 695–710.
[22] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proc. Eur. Conf. Comput. Vis*., 2018, pp. 286–301.
[23] P. Lichtsteiner, C. Posch, and T. Delbruck, "A 128 120 db 15 μs latency asynchronous temporal contrast vision sensor," *IEEE J. Solid-State Circuits*, vol. 43, no. 2, pp. 566–576, Feb. 2008.
[24] G. Gallego et al., "Event-based vision: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 154–180, Jan. 2020.
[25] R. Benosman, C. Clercq, X. Lagorce, S.-H. Ieng, and C. Bartolozzi, "Event-based visual flow," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 2, pp. 407–417, Feb. 2013.
[26] H. Rebecq, R. Ranftl, V. Koltun, and D. Scaramuzza, "Events-to-video: Bringing modern computer vision to event cameras," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3857–3866.
[27] A. Z. Zhu, L. Yuan, K. Chaney, and K. Daniilidis, "EV-FlowNet: Self-supervised optical flow estimation for event-based cameras," in *Proc. Robot.: Sci. Syst.*, 2018, pp. 1–9.
[28] L. Pan, R. Hartley, C. Scheerlinck, M. Liu, X. Yu, and Y. Dai, "High frame rate video reconstruction based on an event camera," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 5, pp. 2519–2533, May 2022.
[29] K. C. Chan, S. Zhou, X. Xu, and C. C. Loy, "Investigating tradeoffs in real-world video super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 5962–5971.
[30] R. Baldwin, M. Almatrafi, V. Asari, and K. Hirakawa, "Event probability mask (EPM) and event denoising convolutional neural network (EDNCNN) for neuromorphic cameras," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 1701–1710.
[31] Z. Jiang, Y. Zhang, D. Zou, J. Ren, J. Lv, and Y. Liu, "Learning event-based motion deblurring," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3320–3329.
[32] L. Wang, T.-K. Kim, and K.-J. Yoon, "EventSR: From asynchronous events to image reconstruction, restoration, and super-resolution via end-to-end adversarial learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 8315–8325.
[33] L. Wang, S. M. M. I., Y.-S. Ho, and K.-J. Yoon, "Event-based high dynamic range image and very high frame rate video generation using conditional generative adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 10081–1090.
[34] Z. W. Wang, P. Duan, O. Cossairt, A. Katsaggelos, T. Huang, and B. Shi, "Joint filtering of intensity images and neuromorphic events for high-resolution noise-robust imaging," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 1609–1619.
[35] B. Wang, J. He, L. Yu, G.-S. Xia, and W. Yang, "Event enhanced high-quality image recovery," in *Proc. Eur. Conf. Comput. Vis*., 2020, pp. 155–171.
[36] X. Zhang, H. Dong, Z. Hu, W.-S. Lai, F. Wang, and M.-H. Yang, "Gated fusion network for degraded image super resolution," *Int. J. Comput. Vis.*, vol. 128, no. 6, pp. 1699–1721, Jun. 2020.
[37] J. Yang, J. Wright, T. Huang, and Y. Ma, "Image super-resolution as sparse representation of raw image patches," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2008, pp. 1–8.
[38] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Trans. Image Process.*, vol. 19, no. 11, pp. 2861–2873, Nov. 2010.
[39] F. Shi, J. Cheng, L. Wang, P.-T. Yap, and D. Shen, "LRTV: MR image super-resolution with low-rank and total variation regularizations," *IEEE Trans. Med. Imag.*, vol. 34, no. 12, pp. 2459–2466, Dec. 2015.
[40] X. Mao, C. Shen, and Y.-B. Yang, "Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 2802–2810.
[41] H. Kim, S. Leutenegger, and A. J. Davison, "Real-time 3D reconstruction and 6-DOF tracking with an event camera," in *Proc. Eur. Conf. Comput. Vis*., 2016, pp. 349–364.
[42] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2472–2481.
[43] M. Jin, G. Meishvili, and P. Favaro, "Learning to extract a video sequence from a single motion-blurred image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6334–6342.
[44] M. Jin, Z. Hu, and P. Favaro, "Learning to extract flawless slow motion from blurry videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 8112–8121.
[45] K. Purohit, A. Shah, and A. Rajagopalan, "Bringing alive blurred moments," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 6830–6839.
[46] K. Zhang, W. Zuo, and L. Zhang, "Deep plug-and-play super-resolution for arbitrary blur kernels," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, 2019, pp. 1671–1681.
[47] L. Wang et al., "Unsupervised degradation representation learning for blind super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 10581–10590.
[48] J. Gu, H. Lu, W. Zuo, and C. Dong, "Blind super-resolution with iterative kernel correction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 1604–1613.
[49] V. Cornillère, A. Djelouah, W. Yifan, O. Sorkine-Hornung, and C. Schroers, "Blind image super-resolution with spatially variant degradations," *ACM Trans. Graph.*, vol. 38, no. 6, pp. 1–13, 2019.
[50] J. Liang, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "Mutual affine network for spatially variant kernel estimation in blind image super-resolution," in *Proc. IEEE Int. Conf. Comput. Vis*., 2021, pp. 4096–4105.
[51] K. Zhang, Y. Li, W. Zuo, L. Zhang, L. Van Gool, and R. Timofte, "Plug-and-play image restoration with deep denoiser prior," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 6360–6376, Oct. 2022.
[52] C. Scheerlinck, N. Barnes, and R. Mahony, "Continuous-time intensity estimation using event cameras," in *Proc. Asian Conf. Comput. Vis.*, 2018, pp. 308–324.
[53] Z. W. Wang, W. Jiang, K. He, B. Shi, A. Katsaggelos, and O. Cossairt, "Event-driven video frame synthesis," in *Proc. Int. Conf. Comput. Vis. Workshops*, 2019, pp. 4320–4329.
[54] L. Sun et al., "Event-based fusion for motion deblurring with cross-modal attention," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 412–428.
[55] C. Song, Q. Huang, and C. Bajaj, "E-CIR: Event-enhanced continuous intensity recovery," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 7803–7812.
[56] F. Xu et al., "Motion deblurring with real events," in *Proc. IEEE Int. Conf. Comput. Vis*., 2021, pp. 2583–2592.
[57] X. Zhang and L. Yu, "Unifying motion deblurring and frame interpolation with events," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 17765–17774.
[58] P. Duan, Z. W. Wang, X. Zhou, Y. Ma, and B. Shi, "EventZoom: Learning to denoise and super resolve neuromorphic events," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 12824–12833.
[59] H. Kim, A. Handa, R. Benosman, S. Ieng, and A. Davison, "Simultaneous mosaicing and tracking with an event camera," in *Proc. Brit. Mach. Vis. Conf.*, 2014, pp. 1–12.
[60] S. M. MostafaviI, J. Choi, and K.-J. Yoon, "Learning to super resolve intensity images from events," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 2768–2786.
[61] S. Tulyakov et al., "Time lens: Event-based video frame interpolation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 16155–16164.
[62] Y. Jing, Y. Yang, X. Wang, M. Song, and D. Tao, "Turning frequency to resolution: Video super-resolution via event cameras," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 7772–7781.

[63] C. Brandli, L. Muller, and T. Delbruck, "Real-time, high-speed video decompression using a frame- and event-based DAVIS sensor," in *Proc. Proc. IEEE Int. Symp. Circuits Syst.*, Melbourne VIC, Australia, 2014, pp. 686–689.
[64] Y. Hu, S.-C. Liu, and T. Delbruck, "V2E: From video frames to realistic DVS events," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2021, pp. 1312–1321.
[65] D. Czech and G. Orchard, "Evaluating noise filtering for event-based asynchronous change detection image sensors," in *Proc. IEEE 6th Int. Conf. Biomed. Robot. Biomechatron.*, 2016, pp. 19–24.
[66] J. Wu, C. Ma, L. Li, W. Dong, and G. Shi, "Probabilistic undirected graph based denoising method for dynamic vision sensor," *IEEE Trans. Multimedia*, vol. 23, no. 11, pp. 1148–1159, May 2020.
[67] X. Lagorce, G. Orchard, F. Galluppi, B. E. Shi, and R. B. Benosman, "HOTS: A hierarchy of event-based time-surfaces for pattern recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 7, pp. 1346–1359, Jul. 2016.
[68] G. Munda, C. Reinbacher, and T. Pock, "Real-time intensity-image reconstruction for event cameras using manifold regularisation," *Int. J. Comput. Vis.*, vol. 126, no. 12, pp. 1381–1393, 2018.
[69] S. Barua, Y. Miyatani, and A. Veeraraghavan, "Direct face detection and video reconstruction from event cameras," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2016, pp. 1–9.
[70] S. Nah, T. Hyun Kim, and K. Mu Lee, "Deep multi-scale convolutional neural network for dynamic scene deblurring," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3883–3891.
[71] A. Khodamoradi and R. Kastner, "O(N)-space spatiotemporal filter for reducing noise in neuromorphic vision sensors," *IEEE Trans. Emerg. Topics Comput.*, vol. 9, no. 1, pp. 15–23, First Quarter 2021.
[72] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Statist. Soc. Ser. B-Methodological*, vol. 58, no. 1, pp. 267–288, 1996.
[73] E. J. Candes, M. B. Wakin, and S. P. Boyd, "Enhancing sparsity by reweighted ℓ_1 minimization," *J. Fourier Anal. Appl.*, vol. 14, no. 5/6, pp. 877–905, 2008.
[74] I. Daubechies, M. Defrise, and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Commun. Pure Appl. Math.*, vol. 57, no. 11, pp. 1413–1457, 2004.
[75] N. Ahn, B. Kang, and K.-A. Sohn, "Fast, accurate, and lightweight super-resolution with cascading residual network," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 256–272.
[76] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Proc. IEEE 37th Asilomar Conf. Signals Syst. Comput.*, 2003, pp. 1398–1402.
[77] H. Rebecq, D. Gehrig, and D. Scaramuzza, "ESIM: An open event camera simulator," in *Proc. Conf. Robot Learn.*, 2018, pp. 969–982.
[78] S. Niklaus, L. Mai, and F. Liu, "Video frame interpolation via adaptive separable convolution," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 261–270.
[79] T. Stoffregen et al., "Reducing the sim-to-real gap for event cameras," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 534–549.
[80] X. Tao, H. Gao, X. Shen, J. Wang, and J. Jia, "Scale-recurrent network for deep image deblurring," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8174–8182.
[81] J. Pan, H. Bai, and J. Tang, "Cascaded deep video deblurring using temporal sharpness prior," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3043–3051.

Lei Yu (Member, IEEE) received the BS and PhD degrees in signal processing from Wuhan University, Wuhan, China, in 2006 and 2012, respectively. From 2013 to 2014, he has been a Postdoc Researcher with the VisAGeS Group at the Institut National de Recherche en Informatique et en Automatique (INRIA) for one and half years. He is currently working as an associate professor with the School of Electronics and Information, Wuhan University. From 2016 to 2017, he has also been a visiting professor with Duke University for one year. He has been working as a guest professor in the École Nationale Supérieure de l'Électronique et de ses Applications (ENSEA), Cergy, France, for one month in 2018. His research interests include neuromorphic vision and computation.

Bishan Wang received the BE degree in communication engineering from Wuhan University, Wuhan, China, in 2019. She is currently working toward the MS degree in information and communication engineering with the electronic information school, Wuhan University. Her research interests include image processing and computer vision.

Xiang Zhang received the BE degree in communication engineering from Wuhan University, Wuhan, China, in 2020. He is currently working toward the MS degree in information and communication engineering with the electronic information school, Wuhan University. His research interests include computer vision and neuromorphic computation.

Haijian Zhang (Senior Member, IEEE) received the BEng degree in electronic information engineering from Wuhan University, Wuhan, China, in 2006, and the joint the PhD degree from the Conservatoire National des Arts et Metiers, Paris, France, and Wuhan University, in 2011. From 2011 to 2014, he was a Research Fellow with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. He is currently an associate professor with the School of Electronic Information, Wuhan University. His main research interests include time–frequency analysis, array signal processing, and multimedia forensics and security.

Wen Yang (Senior Member, IEEE) received the BS degree in electronic apparatus and surveying technology, and the MS degree in computer application technology, and the PhD degree in communication and information system from Wuhan University, Wuhan, China, in 1998, 2001, and 2004, respectively. From 2008 to 2009, he worked as a visiting scholar with the Apprentissage et Interfaces (AI) Team, Laboratoire Jean Kuntzmann, Grenoble, France. From 2010 to 2013, he worked as a post-doctoral researcher with the State Key Laboratory of Information Engineering, Surveying, Mapping and Remote Sensing, Wuhan University. Since then, he has been a Full Professor with the School of Electronic Information, Wuhan University. He is also a guest professor of the Future Lab AI4EO in Technical University of Munich. His research interests include object detection and recognition, multisensor information fusion, and remote sensing image processing.

Jianzhuang Liu (Senior Member, IEEE) received the PhD degree in computer vision from the Chinese University of Hong Kong, Hong Kong, in 1997. From 1998 to 2000, he was a Research Fellow with Nanyang Technological University, Singapore. From 2000 to 2012, he was a postdoctoral fellow, an assistant professor, and an adjunct associate professor with the Chinese University of Hong Kong. In 2011, he joined Shenzhen Institute of Advanced Technology, University of Chinese Academy of Sciences, Shenzhen, China, as a professor. He is currently a principal researcher with Huawei Technolo- gies Company Ltd., Shenzhen. He has authored more than 150 articles. His research interests include computer vision, image processing, deep learning, and graphics.

Gui-Song Xia (Senior Member, IEEE) received the PhD degree in image processing and computer vision from CNRS LTCI, Télécom ParisTech, Paris, France, in 2011. From 2011 to 2012, he has been a postdoctoral researcher with the Centre de Recherche en Mathématiques de la Decision, CNRS, Paris-Dauphine University, Paris, for one and a half years. He is currently working as a full professor in computer vision and photogrammetry with Wuhan University. He has also been working as visiting scholar with DMA, École Normale Supérieure (ENS-Paris) for two months, in 2018. He is also a guest professor of the Future Lab AI4EO in Technical University of Munich (TUM). His current research interests include mathematical modeling of images and videos, structure from motion, perceptual grouping, and remote sensing image understanding. He serves on the Editorial Boards of several journals, including *ISPRS Journal of Photogrammetry and Remote Sensing, Pattern Recognition, Signal Processing: Image Communications, EURASIP Journal on Image & Video Processing, Journal of Remote Sensing, and Frontiers in Computer Science: Computer Vision*.