

EventSR: From Asynchronous Events to Image Reconstruction, Restoration, and Super-Resolution via End-to-End Adversarial Learning

Lin Wang¹, Tae-Kyun Kim², and Kuk-Jin Yoon¹

¹Visual Intelligence Lab., KAIST, Korea

²ICVL Lab., Imperial College London, UK

wanglin@kaist.ac.kr, tk.kim@imperial.ac.uk, kjyoon@kaist.ac.kr

Abstract

Event cameras sense intensity changes and have many advantages over conventional cameras. To take advantage of event cameras, some methods have been proposed to reconstruct intensity images from event streams. However, the outputs are still in low resolution (LR), noisy, and unrealistic. The low-quality outputs stem broader applications of event cameras, where high spatial resolution (HR) is needed as well as high temporal resolution, dynamic range, and no motion blur. We consider the problem of reconstructing and super-resolving intensity images from LR events, when no ground truth (GT) HR images and down-sampling kernels are available. To tackle the challenges, we propose a novel end-to-end pipeline that reconstructs LR images from event streams, enhances the image qualities, and upsamples the enhanced images, called EventSR. For the absence of real GT images, our method is primarily unsupervised, deploying adversarial learning. To train EventSR, we create an open dataset including both real-world and simulated scenes. The use of both datasets boosts up the network performance, and the network architectures and various loss functions in each phase help improve the image qualities. The whole pipeline is trained in three phases. While each phase is mainly for one of the three tasks, the networks in earlier phases are fine-tuned by respective loss functions in an end-to-end manner. Experimental results show that EventSR reconstructs high-quality SR images from events for both simulated and real-world data. A video of the experiments is available at https://youtu.be/OShS_MwHeCs.

1. Introduction

Event cameras are bio-inspired sensors that sense the changes of intensity at the time they occur and produce asynchronous event streams [24, 44, 18], while conventional cameras capture intensity changes at a fixed frame rate. This distinctive feature has sparked a series of meth-

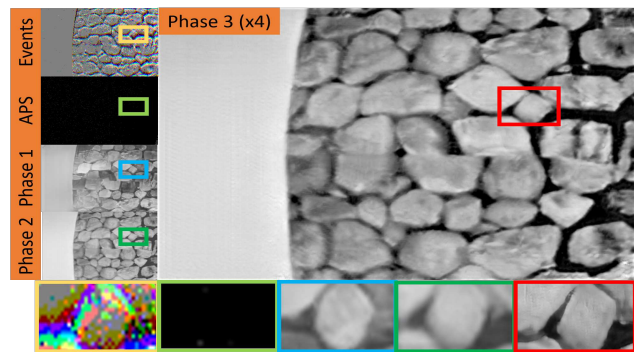


Figure 1: Reconstructing realistic HDR SR intensity image from pure events. EventSR reconstructs LR HDR intensity image, restores realistic LR image and finally generates SR image (with scale factor of $\times 4$) from events in phase 1, 2 and 3, respectively.

ods developed specific for event cameras [37], and only recently, generic learning algorithms were successfully applied to event-based problems [44, 53, 46, 32, 7].

Event cameras (e.g., DAVIS 240) convey clear advantages such as very high dynamic range (HDR) (140dB) [24], no motion blur and high temporal resolution ($1\mu s$), and it has been shown that an event camera alone is sufficient to perform high-level tasks such as object detection [27], tracking [14], and SLAM [19]. In addition, as its potential, event streams might contain complete visual information for reconstructing high quality intensity images and videos with HDR and no motion blur. However, state-of-the-arts (SOTA) [44, 32, 29, 3] for intensity image reconstruction suffer due to accumulated noise and blur (out of focus) in stacked events and low resolution (LR) of event cameras. The active pixel sensor (APS) images are with low dynamic range, LR and blur. The reconstructed images thus typically are in LR and with artifacts. Although [19, 35] focused HR for event cameras, namely spherical HR image mosaicing and HR panorama of events, respectively, they did not consider image-plane HR intensity image reconstruction and its perceptual realism.

In this work, we strive to answer the question, ‘*is it pos-*

sible to directly super-resolve LR event streams to reconstruct image-plane high quality high resolution (HR) intensity images?' The challenges aforementioned render the reconstruction of HR intensity images ill-posed. The problem of reconstructing, restoring (e.g. denoising/deblurring), and super-resolving intensity images from pure event streams has not been convincingly excavated and substantiated. We delve into the problem of reconstructing high-quality SR intensity images with HDR and no motion blur.

For conventional camera images, deep learning (DL) based methods have achieved significant performance gains on single image super-resolution (SISR) using LR and HR image pairs [36, 23, 45]. Most of the works assume that the downsampling methods are available and LR images are pristine. When it comes to event cameras, either stacked events or APS images are noisy and blurred, and GT HR images are unavailable, let alone the degradation models. It is less clear if such DL methods work for event cameras.

Inspired by the development of DL on image translation [54, 43], denoising/deblurring [49, 22], and SISR [47, 52], and some recent successes in DL on event camera data [53, 44], we probe unsupervised adversarial learning to the problem of reconstructing HR intensity images from LR event streams. The results obtained demonstrate the efficacy of our method. To the best of our knowledge, this is the *first* work for reconstructing HR intensity images by super-resolving LR event streams. The proposed pipeline consists of three major tasks. First, 1) we reconstruct LR images from LR event streams. However, these reconstructed images are usually noisy, blurred and unrealistic. 2) So, we then restore (deblur/denoise) realistic LR intensity images from events. 3) Finally, we super-resolve the restored LR images to SR images from events as shown in Fig. 1. Our framework is an end-to-end learning approach and, for more efficient training, we propose phase-to-phase network training strategy. The losses of later phases are back-propagated to the networks of earlier phases. The various loss functions and detailed network architectures are also important to best qualities. We build an open dataset containing 110K images for event to SR image reconstruction, using an event camera simulator [31], event camera dataset [28], and also RGB SR dataset [48, 41]. The conjunctive and alternative use of both real-world and simulated data for EventSR effectively boosts up the network performance. Experimental results using both the simulated dataset [44] and real-world dataset [28] show that EventSR achieves significantly better results than the SOTAs [44, 3, 29]. In summary, our contributions are: 1) the *first* pipeline of reconstructing image-plane HR intensity images from LR events considering image restoration, 2) an *open* dataset to train EventSR for event-based super-resolution and the *skills* of using it for high performance training, 3) the proposed detail architecture, loss functions and end-to-end learning strategy, and 4)

better results than the SOTA works for image reconstruction. Our dataset is open at https://github.com/wl082013/ESIM_dataset.

2. Related Works

Events to intensity image reconstruction The first attempt of reconstructing intensity images from events was done by [8] using rotating visual interpretations. Later on [18] tried to reconstruct 2D panoramic gradient images from a rotating event camera and [19] further delved into reconstructing HR *mosaic* images based on *spherical* 3D scenes. Besides, Bardow *et al.* [3] proposed to estimate optical flow and intensity changes simultaneously via a variational energy function. Similarly, Munda *et al.* [29] regarded image reconstruction as an energy minimization problem defined on manifolds induced by event timestamps. Compared to [29], Scheerlinck *et al.* [37] proposed to filter events with a high-pass filter prior to integration. Recently, DL-based approaches brought great progress on intensity image and video reconstruction. Wang *et al.* [44] proposed to use GAN [15, 4, 43] to reconstruct intensity images and achieved the SOTA performance. In contrast, Rebecq *et al.* [32] exploited recurrent networks to reconstruct video from events. They also used an event sensor with VGA (640 × 480 pixels) resolution to reconstruct higher resolution video, however, the problem is essentially different from our work.

Deep learning on event-based vision Reinbacher *et al.* [35] considered sub-pixel resolution to create a panorama for tracking with much higher spatial resolution of events, however, *not of intensity image reconstruction*. Alonso *et al.* [1] further used an encoder-decoder structure for event segmentation. In contrast, Zhu *et al.* [53] utilized an encoder-decoder network for optical flow, depth and ego-motion estimation via unsupervised learning. Besides, Cannici *et al.* [7] refined YOLO [33] for event-based object detection. Moreover, [46] and [6] both utilized CNNs for human pose and action recognition. Meanwhile, to analyze event alignment, Gallego *et al.* [12, 11] proposed some loss and optimization functions, which are further applied to motion compensation [39], flow estimation [53], etc.

Deep learning on image restoration/enhancement Image restoration addresses the problem of unsatisfactory scene representation, and the goal is to manipulate an image in such a way that it will in some sense more closely depict the scene that it represents [34] by deblurring and denoising from a degraded version. While the objective of image enhancement is to process the image (e.g. contrast improvement, image sharpening, super-resolution) so that it is better suited for further processing or analysis [2]. Recently, CNN has been broadly applied to image restoration and enhancement. The pioneering works include a multilayer perception for image denoising [5] and a three-layer CNN for image SR [9]. Deconvolution was adopted to save computa-

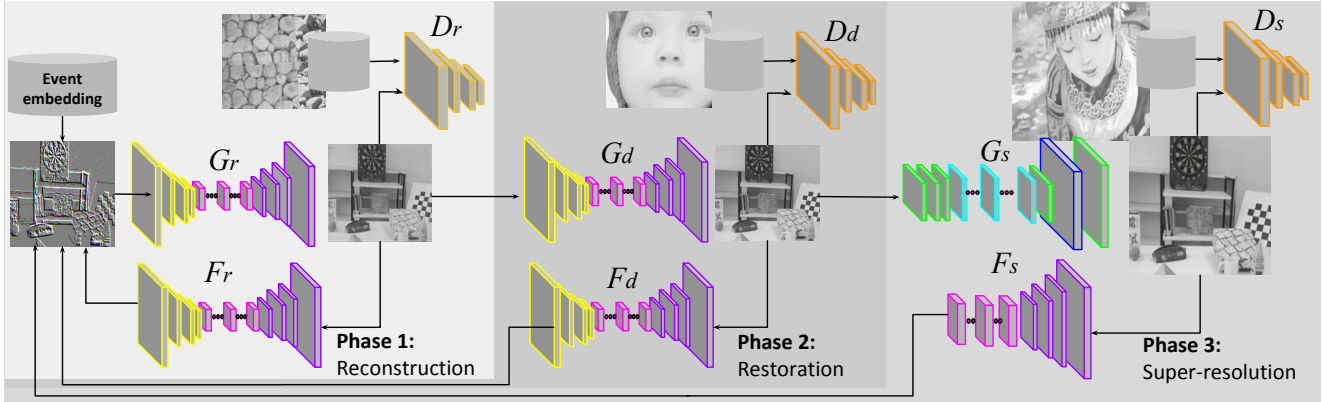


Figure 2: An illustration of the proposed EventSR consisting of three phases: event to image reconstruction (Phase 1), event to image restoration (Phase 2), and event to image super-resolution (Phase 3) via unsupervised adversarial learning. With well designed training and test dataset, EventSR not only works well for simulated but also for real-world data with HDR effects and motion blur.

tion cost and accelerate inference speed [10, 38]. Very deep networks were designed to boost SR accuracy in [20, 25]. Dense connections among various residual blocks were included in [51]. Similarly, CNN- and GAN-based methods were developed for image denoising in [26, 22, 47, 52].

3. Proposed Methods

Our goal is to reconstruct SR images \mathcal{I}^{SR} from a stream of events \mathcal{E} . To feed events to the network, we consider merging events based on the number of incoming events to embed them into images as done in [44, 53]. We then propose a novel unsupervised framework that incorporates namely, event to image reconstruction (Phase 1), event to image restoration (Phase 2), and event to image super-resolution (Phase 3) as shown in Fig. 2. We train the whole system in a sequential phase-to-phase manner, than learning all from scratch. This gradually increases the task difficulty to finally reconstruct SR images. In each phase, the networks of earlier phases are updated thus in an end-to-end manner. More details are given in Sec. 3.2.

3.1. Event embedding and datasets

Event embedding To process event streams using CNNs, we need to stack events into an image or fixed tensor representation as in [44, 53]. An event camera interprets the intensity changes as asynchronous event streams. An event e is represented as a tuple (u, t, p) , where $u = (x, y)$ are pixel coordinates, t is the timestamp of the event, and $p = \pm 1$ is the polarity indicating the sign of brightness change. A natural choice is to encode the events in a spatial-temporal 3D volume or voxel grid [53, 44]. Here, we consider representing 3D event volume by merging events based on the number of events as shown in Fig 3. We reduce event blur (out of focus) by adjusting event sharpness and also variance (contrast) as in [11]. The first N_e events are merged into frame one, and next N_e are merged into frame 2, which

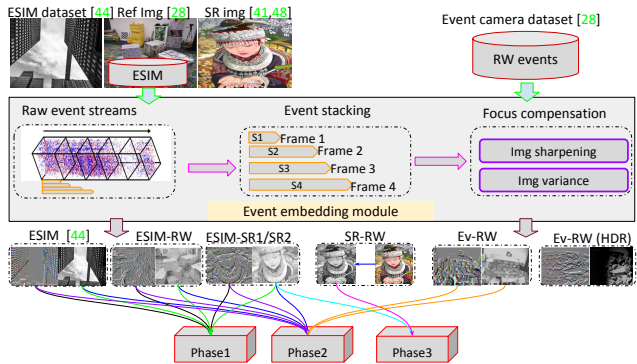


Figure 3: An illustration of event embedding and dataset creation for training EventSR. More details are in the main context.

is repeated up to frame n to create one stack with n frames. Thus, the stack that contains nN_e events will be fed as input to EventSR. In Fig 3, S_1, S_2, S_3 and S_4 are the frames containing different number of events $N_e, 2N_e, 3N_e, 4N_e$, respectively. The event embedding method guarantees rich event data as inputs for EventSR and allows us to adaptively adjust N_e in each frame and n in one stack.

EventSR dataset One crucial contribution of this work is to build a dataset including both simulated and real-world scenes for training EventSR. As mentioned in Sec. 1, real events are noisy and out of focus. Besides, real APS images are degraded with blur, noise and low dynamic range. Therefore, training with only real-world data is not optimal, also shown in [44], and not enough to reconstruct SR images. We propose a novel EventSR dataset including both real and simulated events. We utilize both data conjunctively and alternatively in each phase of training as shown in Fig. 3 and Table. 1, and demonstrate that it works well. For the simulated data, there are three categories for different purposes. First, we use the dataset proposed by [44] for comparisons in intensity image reconstruction. Second, in order to better handle the ill-posed problem caused by real-world data [44], we utilize the reference color images

Table 1: Data source used for training EventSR. (R/S for real/synthetic, P1/P2/P3 for phase 1/2/3, Eval for numerical evaluation, Gen. for generalization to real data, \checkmark / \times for yes/no, and \checkmark indicates very crucial for training EventSR.)

Data name	Resolution	R/S	P1	P2	P3	Eval	Gen.
ESIM data [44]	256x256	S	\checkmark	\checkmark	\checkmark	\checkmark	\times
ESIM-RW	256x256	S	\checkmark	\checkmark	\checkmark	\times	\checkmark
ESIM-SR1	256x256	S	\checkmark	\checkmark	\checkmark	\times	\checkmark
ESIM-SR2	1024x1024	S	\checkmark	\checkmark	\checkmark	\checkmark	\times
Ev-RW(HDR)	256x256	R	\times	\checkmark	\times	\checkmark	\checkmark
SR-RW	1024x1024	R	\times	\checkmark	\checkmark	\times	\checkmark

from the event camera dataset [28]. This brings a simulated dataset called ESIM-RW (around 60K) using the event simulator(ESIM) [31]. The networks trained using the dataset generalises well to real event data. We also take the standard RGB SR dataset [48, 41] to make ESIM-SR dataset (around 50K). However, note that ESIM generates multiple synthetic events and APS images (cropped and degraded) given one HR color image, which renders the SR problem without real GT, thus making it difficult to evaluate the quality of reconstructing SR images from events. We use ESIM to create ESIM-SR1 dataset with image size (256 \times 256) for training phase 1 and phase 2. To numerically evaluate the SR quality, we create ESIM-SR2 dataset, where we set ESIM to output ‘HR’ APS images with larger size (e.g. 1024x1024) as shown in Table. 1, ¹ which are then down-sampled (e.g. bicubic) to smaller size (e.g. 256x256) as LR images. However, reconstructing LR images up to the quality level of these ‘HR’ APS images does not achieve our goal since we want to generate realistic SR images. Thus, we exploit a real-world dataset for phase 3. For the real-world dataset, we directly make Ev-RW dataset using the event camera dataset [28] including general, motion blur and HDR effects. It has been shown that using real event and APS pairs for reconstructing SR images is difficult [44]. Instead, in phase 1, we use ESIM-RW dataset, which is crucial for training EventSR. In phase 2, we first refine the real APS images through phase 1 to get clean APS images (the reason is given in Eq. (4)), then use them for event to image restoration. Lastly, for phase 3, we convert the RGB SR images to grayscale as SR-RW dataset, and it turns out they are crucial for training EventSR. The trained EventSR generalizes well for both simulated and real data, and also the data with HDR effects as shown in Fig. 3 and Table. 1.

3.2. Loss functions and training strategy of EventSR

As shown in Fig. 2, EventSR consists of three phases: event to image reconstruction, event to image restoration and event to image super-resolution. EventSR includes three network functionals G , F , and D in each phase.

Event to image reconstruction (Phase 1) In order to ob-

¹Different from generic SR problems, these ‘HR’ APS images are in low quality (unclear edges and corners) due to the inherent properties of event camera. They, however, can be used to evaluate the quality of SR.

tain SR images, we first reconstruct images from the event streams. Our goal is to learn a mapping $\mathcal{I}^{LR} = G_r(\mathcal{E})$, aided by an event feedback mapping $\mathcal{E} = F_r(\mathcal{I}^{LR})$, and the discriminator $D_r(\mathcal{I}^{LR})$. The inputs are unpaired training events \mathcal{E} and the LR intensity images \mathcal{I}^{LR} .

Event to image restoration (Phase 2) Since the reconstructed images are noisy, blurry, and unrealistic, we then aim to restore (denoise/ deblur) images using both events \mathcal{E} and clean LR images \mathcal{I}^{cLR} . The goal of phase 2 is to learn a mapping $\mathcal{I}^{cLR} = G_d(G_r(\mathcal{E}))$, an event feedback mapping $\mathcal{E} = F_d(\mathcal{I}^{cLR})$, and the discriminator $D_d(\mathcal{I}^{cLR})$. The inputs are unpaired events \mathcal{E} and the clean images \mathcal{I}^{cLR} .

Event to image super-resolution (Phase 3) We then reconstruct SR images from events, utilizing the stacked events \mathcal{E} and real-world HR images \mathcal{I}^{HR} . The problem is to learn a mapping $\mathcal{I}^{SR} = G_s(G_d(G_r(\mathcal{E})))$, an event feedback mapping $\mathcal{E} = F_d(\mathcal{I}^{SR})$, and the discriminator $D_d(\mathcal{I}^{SR})$.

3.2.1 Loss functions for EventSR training

The loss functional for each phase is defined as a linear combination of four losses as:

$$\mathcal{L} = \mathcal{L}_{Adv}(\bar{G}, D) + \lambda_1 \mathcal{L}_{Sim}(\bar{G}, F) + \lambda_2 \mathcal{L}_{Id}(\bar{G}) + \lambda_3 \mathcal{L}_{Var}(\bar{G}) \quad (1)$$

where \mathcal{L}_{Adv} , \mathcal{L}_{Sim} , \mathcal{L}_{Id} , \mathcal{L}_{Var} are the discriminator, event similarity, identity, and total variation losses, respectively. Note D and F are the relevant networks of each stage, and \bar{G} is an accumulated one i.e. G_r , $G_d(G_r)$, $G_s(G_d(G_r))$, in phase 1, 2, and 3. The loss for phase 1,2 and 3 is denoted as \mathcal{L}^r , \mathcal{L}^d and \mathcal{L}^s , respectively.

Adversarial loss \mathcal{L}_{Adv} Given stacked events \mathcal{E} , the generator \bar{G} learns to generate what are similar to given the dataset i.e. the reconstructed, the restored, and the super-resolved, respectively. The discriminator D in this case learns to distinguish the generated images from the given target images via discriminator loss \mathcal{L}_D . The adversarial loss is:

$$\mathcal{L}_{Adv}(\bar{G}, D) = -\mathbb{E}[\log(1 - D(\bar{G}(\mathcal{E})))]. \quad (2)$$

We observe standard GAN training is difficult in phase 3. To stabilize the training and make optimization easier, we use the adversarial loss based on the Relativistic GAN [17].

Event similarity loss \mathcal{L}_{Sim} Since events are usually sparse, we found using pixel-level loss too restrictive and less effective. Here, we propose a new event similarity loss that is based on the interpolation of the pixel-level loss and the perceptual loss based VGG19 inspired by [16]. Namely, we measure the similarity loss of the reconstructed events $F(\bar{G}(\mathcal{E}))$ and the input events \mathcal{E} . We linearly interpolate the pixel-wise loss like L_2 and the perceptual loss as:

$$\mathcal{L}_{Sim}(\bar{G}, F) = \mathbb{E} \left[\alpha \|F(\bar{G}(\mathcal{E})) - \mathcal{E}\|_2 + (1 - \alpha) \frac{1}{C_i W_i H_i} \|\Phi_i(F(\bar{G}(\mathcal{E}))) - \Phi_i(\mathcal{E})\|_2 \right] \quad (3)$$

where Φ_i is the feature map from i -th VGG19 layer, and C_i , H_i , and W_i are the number of channel, height, and width of the feature maps, respectively.

Identity loss \mathcal{L}_{Id} For better learning from events, and also to avoid brightness and contrast variation among different iterations, we utilize the identity loss \mathcal{L}_{Id} . Besides, since *Ev-RW APS images are noisy*, we use \mathcal{L}_{Id} to optimize G_r as a denoiser using clean synthetic APS images. When G_r is trained, the *Ev-RW APS images are fed to the denoiser to get clean real-world images \mathcal{I}^{cLR} to train G_d in phase 2.*

$$\mathcal{L}_{Id}(\bar{G}) = \mathbb{E}[\|\bar{G}(\mathcal{I}) - \mathcal{I}\|_2] \quad (4)$$

where \mathcal{I} and \bar{G} are the target image and the generator in each phase. Since there is the upsampling operation in G_s of phase 3, we propose to use the downsampled HR images as input to G_s . The identity loss helps preserve the shading and texture composition between the $\bar{G}(\mathcal{I})$ and \mathcal{I} .

Total variation loss \mathcal{L}_{Var} Since stack events are sparse, the generated images are spatially not smooth. To impose the spatial smoothness of the generated images, we add a total variation loss:

$$\mathcal{L}_{Var}(\bar{G}) = \mathbb{E}[\|\nabla_h \bar{G}(\mathcal{E}) + \nabla_w \bar{G}(\mathcal{E})\|_2], \quad (5)$$

where ∇_h and ∇_w are the gradients of \bar{G} .

3.2.2 Learning strategy and network structure

End-to-end learning We have described the pipeline of reconstructing, restoring, and attaining SR images from events. We then explore how to unify three phases and train EventSR in an end-to-end manner. Under the unified learning, the second phase becomes auxiliary to the first and the third stage auxiliary to the second and the first. The total loss is:

$$\mathcal{L}^{total} = \mathcal{L}^r + \mathcal{L}^d + \mathcal{L}^s \quad (6)$$

Phase-to-phase learning Rather than learning all network parameters from scratch all together, to facilitate the training, we propose a learning strategy called phase-to-phase learning where we start with an easy task and then gradually increase the task difficulty. Specifically, we first start with G_r with D_r, F_r . We then strengthen the task difficulty by fusing G_r and G_d . We train G_d and D_d, F_d from scratch, meanwhile, fine-tuning G_r . Note each loss term has \bar{G} which is the cascaded reconstruction function i.e. $G_d(G_r)$ in the phase 2. The loss gradients back-propagated to G_r , and D_r, F_r are also updated respectively. We lastly fuse G_s with both G_r and G_d from events. We train the G_s, D_s, F_s from scratch, while fine-tuning both G_r, G_d simultaneously. The generation function $\bar{G} = G_s(G_d(G_r))$.

Network Architecture As shown in Fig. 2, EventSR includes three generators, G_r, G_d and G_s , and three discriminators, D_r and D_d and D_s . For convenience and efficiency,

we design G_r, G_d to share the same network structure. For G_s , we adopt the SOTA SR networks [45, 23]. We also set D_r, D_d , and D_s to share the same network architecture. To better utilize the rich information in events, we also design an event feedback module including F_r, F_d , and F_s , sharing the same network structures based on ResNet blocks. However, for F_s , it has down-sampling operation, so we set the stride with 2 instead. Through the event feedback module, the generators learn to fully utilize the rich information from events to reconstruct, restore, and super-resolve images from events.

4. Experiments and Evaluation

Implementation and training details To facilitate the efficient training of our network, we utilize the proposed phase-to-phase training strategy to achieve the goal of end-to-end learning. In phase 1, we train G_r and D_r with feedback network F_r . We set $\alpha = 0.6$ in Eq. 3 and $\lambda_1 = 10$, $\lambda_2 = 5$ and $\lambda_3 = 0.5$ in Eq. 1. We then train $G_d(G_r(\mathcal{E}))$ and D_d from scratch with F_d in phase 2. We set the $\lambda_1 = 10$, $\lambda_2 = 5$ and $\lambda_3 = 2$ in Eq. 1. In phase 3, we train and $G_r(G_d(G_s(\mathcal{E})))$ from scratch with F_s . The parameters in this phase are set with $\lambda_1 = 10$, $\lambda_2 = 5$ and $\lambda_3 = 3$ in Eq. 1. We initialize the network with dynamic learning rate. we set the batch size of 1 for single GPU, and augment the training data by random rotation and horizontal flipping. We use Adam solver [21] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ to optimize our framework. 3 stacks ($N_e = 10K$ events per stack) are used to get an event image. We assess the quality of each phase outputs using the SSIM, FSIM, and PSNR. To compare with SOTA works [32, 29, 3], we also use LPIPS [50] for measuring the image quality. For all datasets, in order to measure the similarity, each APS image is matched with corresponding reconstructed images with the closest timestamp. We mainly focus on scaling factor of $\times 4$ since it is more challenging and meaningful as studied in SOTA SR works [45, 23].

4.1. Evaluation on simulated datasets

We first compared with [44, 32], which is supervised-learning-based, using dataset proposed in [44]. Figure 4 shows qualitative results on event to image reconstruction (phase 1), restoration (phase 2), and SR with scaling factor of 4 (phase 3). It is shown that EventSR is able to recover the lattices and alleviates the blurry artifacts, which can be visually verified in the cropped patches (second row). Besides, the generated LR images in phase 1 and 2 are close to APS image. Table. 2 shows the quantitative evaluation of phase 1 and 2 results and comparison with [44, 32]. It turns out that our phase 1 (Ours-Rec ($n = 3$)) is comparable to [44, 32] regarding image reconstruction. Since the stacked event images are noisy and out of focus, the reconstructed images are also noisy, blurred and unrealistic. However, our phase 2 successfully handles these problems and achieves

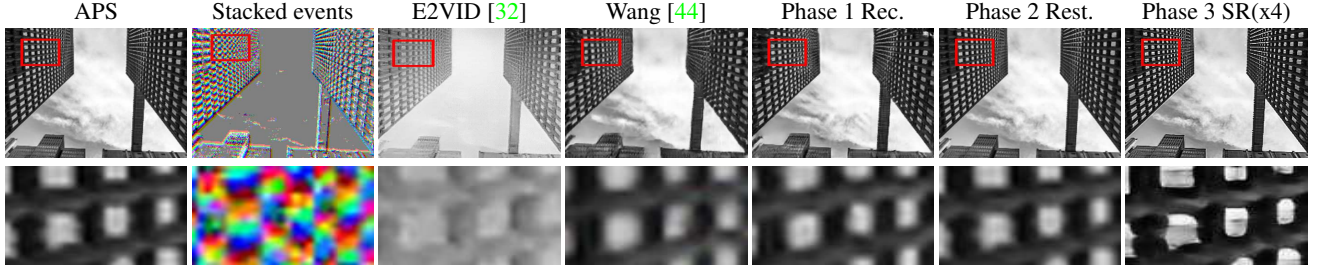


Figure 4: Visual comparison on ESIM dataset [44]. The first row shows our results and the second row shows the cropped patches. EventSR achieves similar performance regarding phase 1 and better results in phase 2.

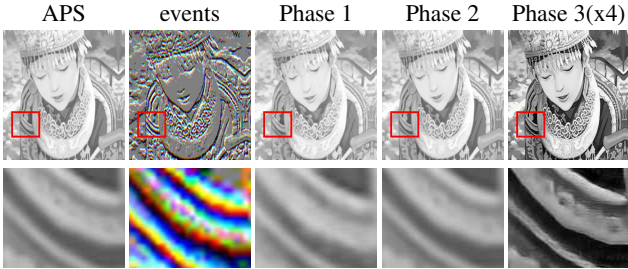


Figure 5: Visual results on our open ESIM-SR dataset. First row shows our results and the second row shows the cropped patches.

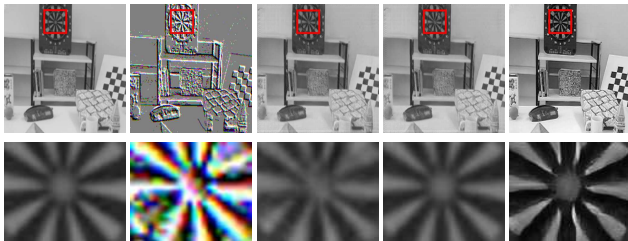


Figure 6: Results on our ESIM-RW dataset. EventSR recovers significant visual structure from events. With ESIM-RW dataset for training EventSR, it also works well on real-world events.

Table 2: Quantitative comparison of phase 1 and 2 with [44, 32] (supervised) based on dataset [44]. Our phase 1 achieves comparable results with [44, 32] and phase 2 achieves much better results.

	PSNR (\uparrow)	FSIM (\uparrow)	SSIM (\uparrow)
E2VID [32]	22.74 \pm 1.96	0.84 \pm 0.06	0.75 \pm 0.10
Wang [44]($n = 1$)	20.51 \pm 2.86	0.81 \pm 0.09	0.67 \pm 0.20
Wang [44]($n = 3$)	24.87 \pm 3.15	0.87 \pm 0.06	0.79 \pm 0.12
Ours-Rec ($n = 3$)	23.26 \pm 3.60	0.85 \pm 0.09	0.78 \pm 0.24
Ours-Rest ($n = 3$)	26.75\pm2.85	0.89\pm0.05	0.81\pm0.23

much better results than [44, 32] and phase 1.

Evaluation on ESIM-SR dataset We also validate EventSR on our ESIM-SR dataset. Figure 5 shows the qualitative results on ESIM-SR1 dataset. Our method can recover very complex objects such as human face. We can see EventSR could utilize the high-frequency information (*e.g.* edge/corner) in the events to reconstruct SR images better than APS images (second row). As mentioned in Sec. 3.1, there are no GT images for ESIM-SR1, thus making quantitative evaluation of SR images difficult. However, we use the ESIM to output ‘HR’ APS images (1024×1024) and

Table 3: Quantitative evaluation of phase 3 on our ESIM-RW dataset with BI degradation model.

	PSNR (\uparrow)	SSIM (\uparrow)
Bicubic	44.27 \pm 2.56	0.98 \pm 0.19
Ours-Phase.3 SR x4 ($n = 3$)	47.68\pm2.17	0.99\pm0.12

Table 4: Quantitative comparison of phase 1 (Rec.) of EventSR with state-of-the-art works based on Ev-RW dataset [28].

	LPIPS (\downarrow)	FSIM (\uparrow)	SSIM (\uparrow)
HF [37]	0.53	–	0.42
MR [29]	0.55	–	0.46
E2VID [32]	0.42	–	0.56
Wang [44]($n = 3$)	–	0.85 \pm 0.05	0.73 \pm 0.16
Ours-Rec ($n = 3$)	0.35	0.86 \pm 0.07	0.75 \pm 0.20
Ours-Rest ($n = 3$)	0.32	0.88\pm0.09	0.78\pm0.18

then downsample (*e.g.* bicubic) them to LR images with scale factor of 4 (ESIM-SR2 dataset). So, we could quantitatively evaluate the quality of EventSR on SR as shown Table. 3. Although the ‘LR’ images do not really differ from ‘HR’ images (high PSNR and SSIM values), our method outperforms the BI method, showing better performance.

Results on ESIM-RW dataset We also evaluate the performance of EventSR on our ESIM-RW dataset as mentioned in Sec. 3.1. This novel dataset is made using the reference color images from event camera dataset [28], aiming to enhance the performance of EventSR on real-world data. We train G_r of phase 1 using this dataset, and surprisingly G_r performs well not only on ESIM-RW events but also on real world events. Figure 6 shows the experimental results on our ESIM-RW dataset. EventSR can recover the correct lines and textures from events, which can be visually verified in the cropped patches in the second row.

4.2. Evaluation on real-world dataset

Evaluation on Ev-RW dataset We demonstrate EventSR shows more impressive results on real-world data. As mentioned in Sec. 3.1, using real-world data alone is not able to handle the challenges in three phases. We show that training G_r in phase 1 with ESIM-RW dataset and G_d with clean real-world APS images in phase 2 are more advantageous. Note that since we use \mathcal{L}_{Id} in phase 1, we obtain clean RW APS images through G_r to get clean RW APS

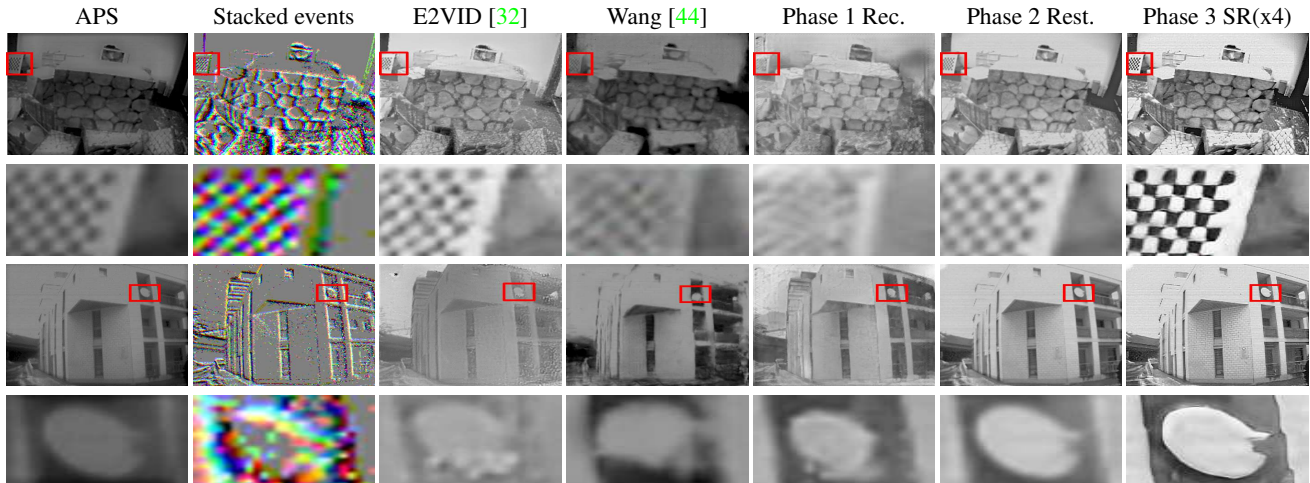


Figure 7: Visual comparison on Ev-RW dataset [28]. With phase 1 trained using ESIM-RW, our method is capable of reconstruct the visual features like edge and corner, etc, and achieves better performance.

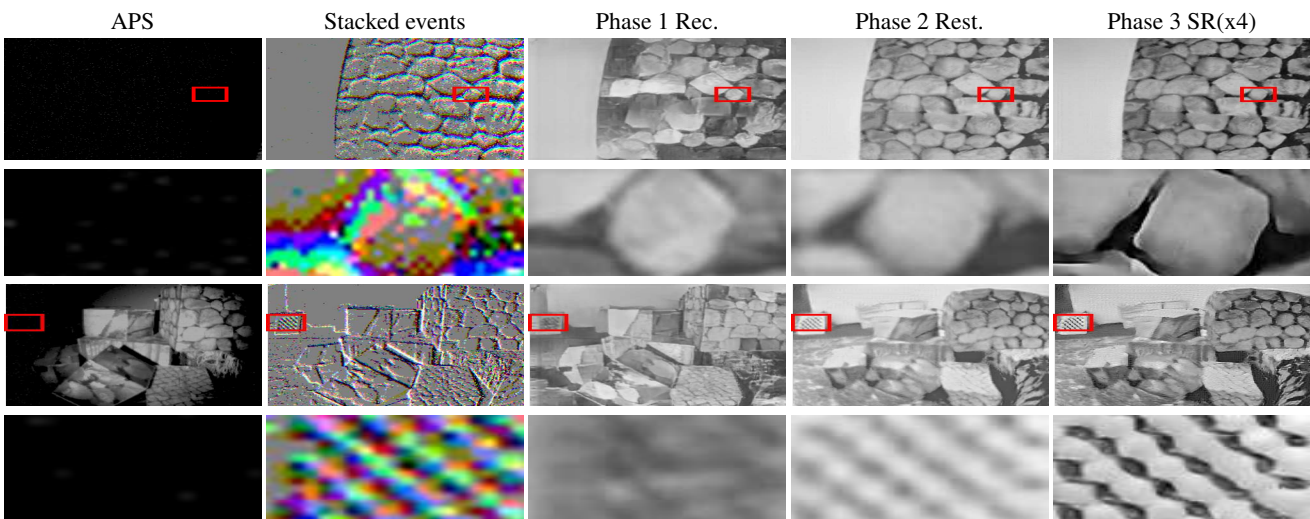


Figure 8: Experimental results on HDR effects with event camera dataset [28]. EventSR also works well on reconstructing HDR images.

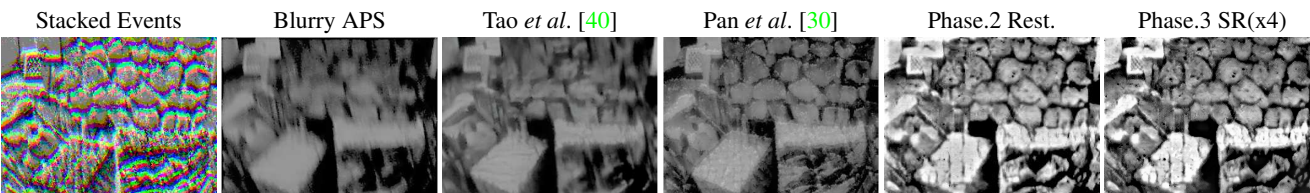


Figure 9: Qualitative results for motion blur on RW dataset [28]. EventSR achieves better quality than Tao *et al.* [40] and Pan *et al.* [30].

images before training phase 2. Figure 7 shows the experimental results on Ev-RW dataset. In phase 1, our method can successfully reconstruct shapes, building, etc, however, the reconstructed images are quite noisy, blurry and unrealistic. In phase 2, the EventSR could restore realistic LR images from events. *This indicates that restoring realistic images from events in phase 2 is a very crucial step for SR image reconstruction.* Although real events are noisy, in phase 3, EventSR can recover the high-frequency structures (e.g. lines and textures) and non-blurry SR images from

events. The cropped patches in the second and fourth rows clearly depict the effectiveness of each stage, in contrast to the APS images. Table 4 quantitatively shows EventSR achieves better results than the prior-arts [44, 32, 29, 3] regarding phase 1. Our unsupervised method shows lower LPIPS and higher SSIM/FSIM scores than the supervised methods, indicating better reconstruction in phase 1.

High dynamic range image In this work, it is apparently shown that events have rich information for HDR image reconstruction, restoration, and super-resolution. Although

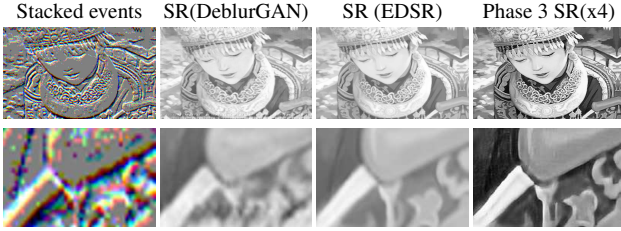


Figure 10: Visual comparison using existing deblur and SR nets.

some parts of the scenes are invisible in APS images due to the low dynamic range, many events do exist in those regions, and EventSR can fully utilize the rich information contained in events to reconstruct HDR SR images. We evaluate the HDR effects [28] with G_r (phase 1) trained with ESIM-RW dataset and with G_d (phase 2) trained using Ev-RW dataset. Figures 8 and 1 show that EventSR successfully reconstructs HDR SR images from pure events. Although in phase 1 the reconstructed images are noisy, blurry and unrealistic, phase 2 recovers the correct shapes and textures of wall poster and boxes. In phase 3, the right structure and informative details are recovered, which can be verified in the cropped patches in second and fourth rows.

Motion deblur We also demonstrate that EventSR can restore deblurred images from events. As shown in Fig. 9, we visually compare our approach with Tao *et al.* [18] and Pan *et al.* [30] (energy minimization) based on Ev-RW blur effects [28]. Although, the APS image is blurry, our method can restore sharp and realistic images (clear edge, corner, texture, etc) from events in phase 2, which is further enhanced to the non-blurry SR image from events in phase 3.

4.3. Ablation study

To validate the effectiveness of the proposed loss functions in EventSR, we compare different network structures by removing loss functions selectively.

Remove F_s and D_s We remove G_s and D_s , namely removing \mathcal{L}_{Adv}^s and \mathcal{L}_{Sim}^s . We map the embedded events to clean LR images and reconstruct SR images using $G_s(G_d(G_r))$. However, without F_s and D_s , some noise in the events are mapped to the SR images, affecting visual quality.

Remove F_d and D_d We also remove F_d and D_s from EventSR, namely, removing \mathcal{L}_{Adv}^d and \mathcal{L}_{Sim}^d . We use EventSR for event to SR image reconstruction, where the forward network is $G_s(G_d(G_r))$ and F_s is the feedback network. We load the pre-trained G_r and G_d and add them to G_s . However, without \mathcal{L}_{Adv}^d and \mathcal{L}_{Sim}^d , the G_d is unable to get the clean images from events.

Remove F_r and D_r We lastly remove F_r and D_r , namely \mathcal{L}_{Adv}^r and \mathcal{L}_{Sim}^r . However, it shows that \mathcal{I}^r is always with undesired artifacts and training is unstable. It is hard to reconstruct SR images from events without these losses.

5. Discussions

Computational cost, observation window, and latency

Our model has around 17M parameters, which is less than some SOTA SR DNNs, such as EDSR [25], RDN [51], thus the training time is comparable with others. The inference time is around 300 ~ 500ms on average when using NVIDIA 1080 Ti GPU. In our experiments, 10K events are gathered in 5 ms time duration on average. Events can be stacked with the fixed observation window as SBT in [44].

Using the existing deblur and SR nets One might think that the SR results can be achieved by directly using the existing deblur and SR networks after phase 1. However, we need to clarify that our method is not just combining networks in a naive way and the phase 2 is not just to deblur but to restore. To verify this, we replace phase 2 and 3 with pretrained SOTA deblur and SR networks (*e.g.* DeblurGAN [22] and EDSR [25]). As in Fig. 10, one can clearly see that the proposed method (4th column) is superior to such a naive combination. Without continuously utilized event information, applying existing deblur and SR nets magnifies the noise level and fails to enhance the image quality.

SR video from events In this work, we focus on super-resolving HR images from LR events and we do not fully consider the temporal consistency for video. However, we will investigate enforcing temporal consistency for super-resolving video from events in our future work.

6. Conclusion and Future Work

In this paper, we presented the first and novel framework for event to SR image reconstruction. Facing up with the challenges of no GT images for real-world data in all three phases, we proposed EventSR to learn a mapping from events to SR images in an unsupervised manner. To train EventSR, we made an open dataset including both simulated and real-world scenes. The conjunctive and alternative use of them boosted up the performance. Experimental results showed that EventSR achieved impressive results even on phase 1 and phase 2, and desirable results in phase 3. However, in this work, we have not deeply considered how the forms of event stacks affect the overall performance of EventSR. We will investigate better ways to embed events as input to EventSR as in [13, 42] and its potential applications to other tasks in the following work. Besides, we are also aiming to reconstruct SR video from event streams.

Acknowledgement

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (NRF-2018R1A2B3008640) and the Next-Generation Information Computing Development Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Science, ICT (NRF-2017M3C4A7069369).

References

- [1] Iñigo Alonso and Ana C Murillo. Ev-segnet: Semantic segmentation for event-based cameras. In *IEEE International Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019. 2
- [2] Saeed Anwar, Salman Khan, and Nick Barnes. A deep journey into super-resolution: A survey. *arXiv preprint arXiv:1904.07523*, 2019. 2
- [3] Patrick Bardow, Andrew J Davison, and Stefan Leutenegger. Simultaneous optical flow and intensity estimation from an event camera. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 884–892, 2016. 1, 2, 5, 7
- [4] Patrick Alexander Bardow. *Estimating general motion and intensity from event cameras*. PhD thesis, Imperial College London, 2018. 2
- [5] Harold C Burger, Christian J Schuler, and Stefan Harmeling. Image denoising: Can plain neural networks compete with bm3d? In *2012 IEEE conference on computer vision and pattern recognition*, pages 2392–2399. IEEE, 2012. 2
- [6] Enrico Calabrese, Gemma Taverni, Christopher Awai Easthope, Sophie Skriabine, Federico Corradi, Luca Longinotti, Kynan Eng, and Tobi Delbruck. Dhp19: Dynamic vision sensor 3d human pose dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 2
- [7] Marco Cannici, Marco Ciccone, Andrea Romanoni, and Matteo Matteucci. Asynchronous convolutional networks for object detection in neuromorphic cameras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 1, 2
- [8] Matthew Cook, Luca Gugelmann, Florian Jug, Christoph Krautz, and Angelika Steger. Interacting maps for fast visual interpretation. In *The 2011 International Joint Conference on Neural Networks*, pages 770–776. IEEE, 2011. 2
- [9] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *European conference on computer vision*, pages 184–199. Springer, 2014. 2
- [10] Chao Dong, Chen Change Loy, and Xiaoou Tang. Accelerating the super-resolution convolutional neural network. In *European conference on computer vision*, pages 391–407. Springer, 2016. 3
- [11] Guillermo Gallego, Mathias Gehrig, and Davide Scaramuzza. Focus is all you need: Loss functions for event-based vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12280–12289, 2019. 2, 3
- [12] Guillermo Gallego, Henri Rebecq, and Davide Scaramuzza. A unifying contrast maximization framework for event cameras, with applications to motion, depth, and optical flow estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2
- [13] Daniel Gehrig, Antonio Loquercio, Konstantinos G Derpanis, and Davide Scaramuzza. End-to-end learning of representations for asynchronous event-based data. *arXiv preprint arXiv:1904.08245*, 2019. 8
- [14] Daniel Gehrig, Henri Rebecq, Guillermo Gallego, and Davide Scaramuzza. Asynchronous, photometric feature tracking using events and frames. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 750–765, 2018. 1
- [15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 2
- [16] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 4
- [17] Alexia Jolicoeur-Martineau. The relativistic discriminator: a key element missing from standard gan. *arXiv preprint arXiv:1807.00734*, 2018. 4
- [18] Hanme Kim, Ankur Handa, Ryad Benosman, Sio-Hoi Ieng, and Andrew J Davison. Simultaneous mosaicing and tracking with an event camera. *British Machine Vision Conference (BMVC)*, 43:566–576, 2014. 1, 2, 8
- [19] Hanme Kim, Stefan Leutenegger, and Andrew J Davison. Real-time 3d reconstruction and 6-dof tracking with an event camera. In *European Conference on Computer Vision*, pages 349–364. Springer, 2016. 1, 2
- [20] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1646–1654, 2016. 3
- [21] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5
- [22] Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin, and Jiří Matas. Deblurgan: Blind motion deblurring using conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8183–8192, 2018. 2, 3, 8
- [23] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017. 2, 5
- [24] Patrick Lichtsteiner, Christoph Posch, and Tobi Delbruck. A 128×128 120 db 15μs latency asynchronous temporal contrast vision sensor. *IEEE journal of solid-state circuits*, 43(2):566–576, 2008. 1
- [25] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 136–144, 2017. 3, 8
- [26] Ding Liu, Bihan Wen, Xianming Liu, Zhangyang Wang, and Thomas S. Huang. When image denoising meets high-level vision tasks: A deep learning approach. In *IJCAI*, 2018. 3
- [27] Anton Mitrokhin, Cornelia Fermüller, Chethan Parameshwara, and Yiannis Aloimonos. Event-based moving object

- detection and tracking. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1–9. IEEE, 2018. [1](#)
- [28] Elias Mueggler, Henri Rebecq, Guillermo Gallego, Tobi Delbruck, and Davide Scaramuzza. The event-camera dataset and simulator: Event-based data for pose estimation, visual odometry, and slam. *The International Journal of Robotics Research*, 36(2):142–149, 2017. [2](#), [4](#), [6](#), [7](#), [8](#)
- [29] Gottfried Munda, Christian Reinbacher, and Thomas Pock. Real-time intensity-image reconstruction for event cameras using manifold regularisation. *International Journal of Computer Vision*, 126(12):1381–1393, 2018. [1](#), [2](#), [5](#), [6](#), [7](#)
- [30] Liyuan Pan, Cedric Scheerlinck, Xin Yu, Richard Hartley, Miaomiao Liu, and Yuchao Dai. Bringing a blurry frame alive at high frame-rate with an event camera. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6820–6829, 2019. [7](#), [8](#)
- [31] Henri Rebecq, Daniel Gehrig, and Davide Scaramuzza. Esim: an open event camera simulator. In *Conference on Robot Learning*, pages 969–982, 2018. [2](#), [4](#)
- [32] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. Events-to-video: Bringing modern computer vision to event cameras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3857–3866, 2019. [1](#), [2](#), [5](#), [6](#), [7](#)
- [33] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. [2](#)
- [34] Stanley J Reeves. Image restoration: fundamentals of image restoration. In *Academic press library in signal processing*, volume 4, pages 165–192. Elsevier, 2014. [2](#)
- [35] Christian Reinbacher, Gottfried Munda, and Thomas Pock. Real-time panoramic tracking for event cameras. In *2017 IEEE International Conference on Computational Photography (ICCP)*, pages 1–9. IEEE, 2017. [1](#), [2](#)
- [36] Mehdi SM Sajjadi, Bernhard Scholkopf, and Michael Hirsch. Enhancenet: Single image super-resolution through automated texture synthesis. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4491–4500, 2017. [2](#)
- [37] Cedric Scheerlinck, Nick Barnes, and Robert Mahony. Continuous-time intensity estimation using event cameras. In *Asian Conference on Computer Vision*, pages 308–324. Springer, 2018. [1](#), [2](#), [6](#)
- [38] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1874–1883, 2016. [3](#)
- [39] Timo Stoffregen and Lindsay Kleeman. Event cameras, contrast maximization and reward functions: an analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12300–12308, 2019. [2](#)
- [40] Xin Tao, Hongyun Gao, Xiaoyong Shen, Jue Wang, and Ji-aya Jia. Scale-recurrent network for deep image deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8174–8182, 2018. [7](#)
- [41] Radu Timofte, Shuhang Gu, Jiqing Wu, and Luc Van Gool. Ntire 2018 challenge on single image super-resolution: Methods and results. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 852–863, 2018. [2](#), [4](#)
- [42] Stepan Tulyakov, Francois Fleuret, Martin Kiefel, Peter Gehler, and Michael Hirsch. Learning an event sequence embedding for dense event-based deep stereo. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1527–1537, 2019. [8](#)
- [43] Lin Wang, Wonjune Cho, and Kuk-Jin Yoon. Deceiving image-to-image translation networks for autonomous driving with adversarial perturbations. *arXiv preprint arXiv:2001.01506*, 2020. [2](#)
- [44] Lin Wang, S. Mohammad Mostafavi I., Kuk-Jin Yoon, et al. Event-based high dynamic range image and very high frame rate video generation using conditional generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10081–10090, 2019. [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#)
- [45] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018. [2](#), [5](#)
- [46] Yanxiang Wang, Bowen Du, Yiran Shen, Kai Wu, Guanrong Zhao, Jianguo Sun, and Hongkai Wen. Ev-gait: Event-based robust gait recognition using dynamic vision sensors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6358–6367, 2019. [1](#), [2](#)
- [47] Yuan Yuan, Siyuan Liu, Jiawei Zhang, Yongbing Zhang, Chao Dong, and Liang Lin. Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 701–710, 2018. [2](#), [3](#)
- [48] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *International conference on curves and surfaces*, pages 711–730. Springer, 2010. [2](#), [4](#)
- [49] Li Zhang, Alok Deshpande, and Xin Chen. Denoising vs. deblurring: Hdr imaging techniques using moving cameras. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 522–529. IEEE, 2010. [2](#)
- [50] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. [5](#)
- [51] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2472–2481, 2018. [3](#), [8](#)
- [52] Tianyu Zhao, Changqing Zhang, Wenqi Ren, Dongwei Ren, and Qinghua Hu. Unsupervised degradation learn-

ing for single image super-resolution. *arXiv preprint arXiv:1812.04240*, 2018. [2](#), [3](#)

- [53] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Unsupervised event-based learning of optical flow, depth, and egomotion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 989–997, 2019. [1](#), [2](#), [3](#)
- [54] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. [2](#)