EvIntSR-Net: Event Guided Multiple Latent Frames Reconstruction and Super-resolution

Jin Han[1] Yixin Yang[2] Chu Zhou[1] Chao Xu[1] Boxin Shi[2,3,4*]
[1]Key Lab of Machine Perception (MOE), Dept. of Machine Intelligence, Peking University
[2]NELVT, Dept. of Computer Science and Technology, Peking University
[3]Institute for Artificial Intelligence, Peking University
[4]Beijing Academy of Artificial Intelligence

Abstract

An event camera detects the scene radiance changes and sends a sequence of asynchronous event streams with high dynamic range, high temporal resolution, and low latency. However, the spatial resolution of event cameras is limited as a trade-off for these outstanding properties. To reconstruct high-resolution intensity images from event data, we propose EvIntSR-Net that converts Event data to multiple latent Intensity frames to achieve Super-Resolution on intensity images in this paper. EvIntSR-Net bridges the domain gap between event streams and intensity frames and learns to merge a sequence of latent intensity frames in a recurrent updating manner. Experimental results show that EvIntSR-Net can reconstruct SR intensity images with higher dynamic range and fewer blurry artifacts by fusing events with intensity frames for both simulated and real-world data. Furthermore, the proposed EvIntSR-Net is able to generate high-frame-rate videos with super-resolved frames.

1. Introduction

Event cameras with bio-inspired silicon retina sensors work radically different from conventional frame-based cameras. The unconventional sensor design enables them to measure scene radiance changes in an asynchronous manner [12, 31], instead of capturing images at a fixed frame rate. Event cameras detect brightness changes in a scene in log scale, and send a stream of event data that are binary-signed recordings of brightness changes ("+1" for brightness increase and "-1" for brightness decrease). The particular properties of event sensors include: very high dynamic range (HDR, up to 140 dB), high temporal resolution (in the order of μs), low latency, and low power consumption. The latest Dynamic and Active Pixel Vision Sensor (DAVIS [4]) combines a conventional Active Pixel Sensor (APS) with the event sensor, which can capture intensity frames simultaneously with event data. However, most of the available event cameras bear low spatial resolution (*e.g.*, 240×180 for DAVIS240, 346×260 for DAVIS346) partially due to the consideration of data transmission efficiency.

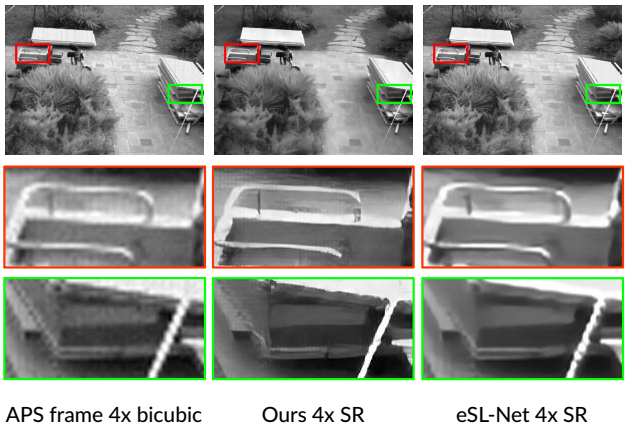

Figure 1: An example of $4\times$ super-resolution results from eSL-Net [40] and our reconstruction on real-world data. Both of them take APS and event data as inputs.

Event data contains visual information that can be utilized for reconstructing high-quality intensity images. Previous reconstruction approaches [32, 35, 41] can only achieve low-resolution (LR) intensity reconstruction that is restricted by the spatial resolution of event cameras. However, high-resolution (HR) intensity images with higher quality (more structural details, higher dynamic range, less blurry artifacts) significantly contribute to many other event-based vision tasks (*e.g.*, object recognition [6], detection[3], tracking[2], *etc.*). It is therefore of practical interest to conduct super-resolution (SR) for event-guided intensity image reconstruction.

Super-resolving intensity images for event cameras can be achieved in several ways. One kind of approach is first converting event data $\mathbf{E}$ to intensity images $\mathbf{I}$ [32, 35,

*Corresponding author: shiboxin@pku.edu.cn

41] at the same spatial scale, then using existing SR approaches [19, 39, 48] to get the SR results $\mathbf{S}$. Such an approach can be expressed as:

$$\mathbf{S} = \uparrow (\Gamma(\mathbf{E})), \tag{1}$$

where $\uparrow$ $()$ and $\Gamma()$ represent SR operation and conversion from events to intensity images, respectively. Another method is directly super-resolving LR event streams to reconstruct HR intensity images without the assistance of intensity frames [7, 42], which is expressed using the following expression:

$$\mathbf{S} = \uparrow (\mathbf{E}). \tag{2}$$

Moreover, hybrid signals (*e.g.*, APS images $\hat{\mathbf{I}}$ and event data $\mathbf{E}$) can be taken as input to realize spatial resolution enhancement of intensity images [40]:

$$\mathbf{S} = \Gamma_{\uparrow}(\hat{\mathbf{I}}, \mathbf{E}), \tag{3}$$

where $\Gamma_{\uparrow}()$ denotes jointly reconstructing and super resolving operation.

However, the three types of event-based intensity image SR described as Eq. (1)$\sim$Eq. (3) have some disadvantages. Firstly, Eq. (1) relies heavily on the performance of $\Gamma()$ due to the domain gap between upsampling events and intensity data independently. Secondly, Eq. (2) does not take intensity information into account. Ignoring the intensity information from APS frames that faithfully record scene radiance with less motion results in fewer details and unstable intensity for video reconstruction. Thirdly, method [40] like Eq. (3) jointly achieving image deblurring, denoising, and SR may not produce high-quality images because different degradation problems are caused by various reasons, as shown in Fig. 1.

In this paper, we propose to fuse intensity frames with event data to achieve high-quality super-resolution of intensity images by utilizing the information provided by hybrid types of input data. The APS frames record spatial irradiance with rich semantic information of a scene at each pixel, while event data encode the rapid temporal irradiance changes along the edges of objects. The static intensity values and dynamic events are complementary to each other. We turn the SR problem into a better-posed multiple image version, as described in Eq. (4):

$$\mathbf{S} = \uparrow_{\Sigma} \left(\hat{\Gamma}(\hat{\mathbf{I}}, \mathbf{E}^{i}) \right), \tag{4}$$

where $\uparrow_{\Sigma}()$ denotes the multi-image super-resolution (MISR) operator. $\hat{\Gamma}()$ differs from $\Gamma()$ in that the conversion from event data to multiple latent intensity frames is provided with the assistance of APS, and i is the index of different batches of events.

We therefore propose *EvIntSR-Net*, a neural network that learns to convert Event data to multiple latent Intensity frames to achieve SR for reconstructing high-resolution intensity images. As described Eq. (4), such a merging process can be described as two steps: 1) The events represent *residuals* from low-quality APS frame to high-quality latent frames. Given APS frame and its preceding and/or following event streams, we can reconstruct multiple latent frames with higher dynamic range and sharper details. 2) The reconstructed latent frames could then be treated as a sequence of video frames, which benefit from MISR to enhance the resolution of a target APS frame. EvIntSR-Net is composed of two sub-networks: latent frame reconstruction network (*LFR-Net*), which estimates the residuals between intensity frames and reconstructs multiple latent frames; and multi-image fusion network (*MIF-Net*), which solves the misalignment issue among latent frames first, then learns to effectively merge them in a recurrent updating manner.

Extensive experiments on synthetic data, as well as real-world data (*e.g.*, DAVIS346) demonstrate EvIntSR-Net can successfully reconstruct $2\times$ and $4\times$ super-resolved intensity images with higher fidelity comparing to state-of-the-art approaches. An example is shown in Fig. 1, compared to eSL-Net [40], which also takes $\hat{\mathbf{I}}$ and $\mathbf{E}$ as input data, EvIntSR-Net (ours) recovers much sharper edges and richer structural details in $4\times$ SR on images captured by a real event camera. In addition to image SR, EvIntSR-Net can generate high-frame-rate (HFR) videos with SR frames.

2. Related Work

2.1. Intensity Image Super-resolution

Intensity image SR algorithms can be divided into two categories: single image super-resolution (SISR) and multiple image super-resolution (MISR) methods. We focus on MISR approaches in this section since our method belongs to this category. Please refer to comprehensive surveys in [30, 44] for a summary of SISR methods.

MISR methods aim to extract temporal information along with contextual features in a series of frames, which are generally more successful in recovering missing details. Previous approaches [5, 13, 20, 25, 26, 46] usually compute optical flow and perform image warping based motion compensation to get well-aligned frames, then use convolutional neural networks (CNNs) to fuse them for HR results. However, these optical flow-based approaches still have limited performance because they rely on accurate motion estimation and the motion compensation will introduce undesired artifacts around image structures. Besides, several methods [23, 38] take advantage of 3D convolutions to extract features from concatenated frames without any explicit alignment. However, the noise introduced by misalignment especially on the edges impacts the reconstruction results. And 3D convolutions require huge computational re-

sources. Apart from that, some MISR approaches [39, 43] conduct implicit motion compensation by using deformable convolutional filters [9], which reshape the configuration of filters with computed offsets to handle the geometric transformations. The EDVR [43] extracts features of two frames at multiple levels and computes the offsets between feature maps for deformable convolutional filters, then warps the adjacent frame to the target frame with deformable convolutions.

2.2. Event-based Super-resolution

Since event cameras are restricted by their low spatial resolution, some works focus on the super-resolution of event streams in both spatial and temporal domains. Li *et al*. [22] used Poisson point process to model the event sequence, and sample the events according to a nonhomogeneous Poisson process. EventZoom [11] collected a multi-resolution event dataset and used a noise-to-noise fashion to learn the denoising and upsampling mapping of event signals. Wang *et al*. [45] proposed guided event filtering (GEF), which built a hybrid camera and took advantage of HR RGB signals to guide the upsampling of events. Intensity image reconstruction from event data has been exploited in many ways [1, 8, 18, 32, 41]. All of them reconstruct intensity images in the same spatial resolution as event data. To achieve higher resolution intensity images, Mohammad *et al*. [7] used a recurrent neural network to iteratively add details to intensity frames for SR. Wang *et al*. proposed a three-phase reconstruction network named EventSR [42], which used unsupervised adversarial learning to upsample the enhanced intensity image. To recover high-quality HR images, Wang *et al*. proposed eSL-Net [40], an event-enhanced sparse learning network, to address deblurring, denoising, and super-resolution simultaneously.

3. Proposed Method

In Sec. 3.1, we first formulate the physical model of event data generation and the relationship between intensity frames and event data. Then we describe the event-guided image super-resolution model that can be viewed as a two-stage process: 1) latent frame reconstruction, and 2) multi-image fusion. In the first stage, we reconstruct multiple latent frames from the current APS frame with its neighboring event data. Then we merge latent images to achieve high-resolution intensity image reconstruction. To realize high-quality event-guided image super-resolution, we propose the EvIntSR-Net in Sec. 3.2, which is designed according to the two-stage process. Sec. 3.3 and Sec. 3.4 describes the details of dataset generation and training strategy, respectively.

3.1. Formulation

3.1.1 Physical model of event data and intensity frame

An event data $e_i(x_i, y_i, t_i, p_i)$ is triggered when the log intensity change exceeds the dispatched threshold θ, where i represents the index of the event in a stream, x and y are the spatial coordinates of the event, t denotes the triggering time stamp, $p \in \{-1, +1\}$ is the polarity that indicates the increase or decrease of intensity changes. The output of event sensor $\mathbf{E}^t$ can be represented using

$$\mathbf{E}^t_{(x,y)} = \Phi\left\{\log\left(\frac{\mathbf{I}^t_{(x,y)} + \epsilon^t}{\mathbf{I}^{t-1}_{(x,y)} + \epsilon^{t-1}}\right), \theta\right\}, \tag{5}$$

where $\mathbf{I}^t$ is the captured intensity values, and ϵ is an offset value that prevents division by zero. $\Phi\{\alpha, \theta\}$ represents the condition of events generation. When $\alpha \geq \theta$, positive events are generated, while $\alpha \leq -\theta$ will trigger negative events.

Given two consecutive captured intensity values $\mathbf{I}^{t_i}$ and $\mathbf{I}^{t_j}$, the events triggered by intensity changes between them during a time slot $t_i \rightarrow t_j$ are represented as $\mathbf{E}^{t_i \rightarrow t_j}$. According to the physical model of event generation, the relationship between $\mathbf{I}^{t_i}$ and $\mathbf{I}^{t_j}$ can be formulated as

$$\begin{aligned}\mathbf{I}^{t_j} &= \mathbf{I}^{t_i} \cdot \exp\left(\theta \int \mathbf{E}^{t_i \rightarrow t_j} dt\right) \\ &= \mathbf{I}^{t_i} \cdot Res^{t_i \rightarrow t_j},\end{aligned} \tag{6}$$

where $Res^{t_i \rightarrow t_j}$ is the residual between two intensity frames, which is computed from the integral events in the time slot.

Therefore, given the current intensity frame and its neighboring event streams, the latent frames can be reconstructed using Eq. (6).

3.1.2 Image super-resolution guided by event data

We aim to super-resolve the intensity image $\mathbf{I}^t$ with the information provided by neighboring event data. The event streams represent the log intensity changes, which are sparse and in quite different type of data format. Therefore, it is difficult to bridge the domain gap by directly fusing event data with intensity images. As Eq. (4) stated, we divide the process into two steps. We firstly improve the quality of the APS frame, which is blurry and in low dynamic range (LDR). Then we convert the neighboring event streams into latent frames that are in the intensity domain by estimating the residuals between latent frames and current $\mathbf{I}^t$. The number of latent frames depends on the number of events and the time slot we choose. With the reconstructed latent frames, we can treat this problem as the MISR process. To avoid ghosting artifacts, the latent frames are warped to the improved $\mathbf{I}^t$ at first. Then we fuse

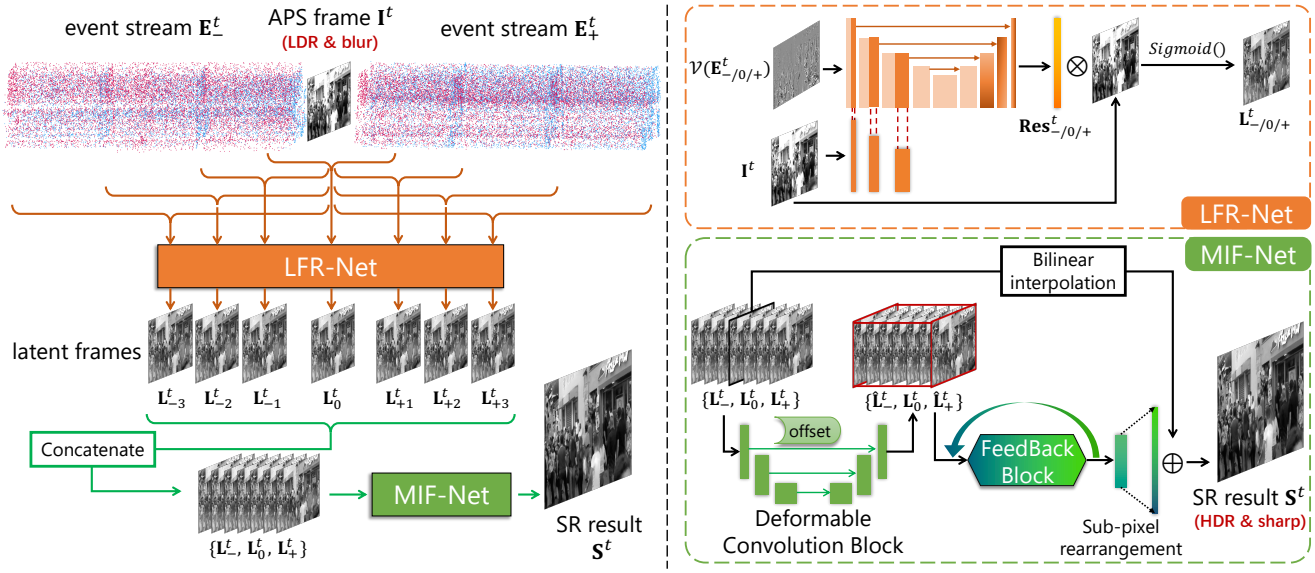

Figure 2: The pipeline of EvIntSR-Net. Left: the MISR process of our proposed method, which is composed of latent frames reconstruction and multi-image fusion. The inputs are the central APS frame $\mathbf{I}^t$, which is in low-quality (LDR and blur), and its neighboring event streams $\mathbf{E}^t_{-/+}$. The output is the SR result $\mathbf{S}^t$ of the enhanced APS frame, which is improved in spatial resolution, dynamic range, and sharpness. Right: the detailed architectures of LFR-Net and MIF-Net.

the stack of frames in an iterative manner to get the super-resolved reconstruction results.

3.2. Network Architecture

The network architecture of EvIntSR-Net is shown in Fig. 2. EvIntSR-Net takes an LR intensity frame $\mathbf{I}^t$ and its neighboring event stream $\mathbf{E}^t_{-/+}$ as the input, where "$-$" and "$+$" represent the foregoing and following event streams of the intensity frame.

3.2.1 Latent frame reconstruction

The latent frames are reconstructed using intensity residuals estimated from neighboring event data. However, directly stacking the event data and multiplying with $\mathbf{I}^t$ leads to ringing artifacts in edges.

Hence, we propose to estimate the intensity residuals and reconstruct latent frames by the latent frame reconstruction network (LFR-Net). Because of the sparsity in spatial domain of event streams, we need to bridge the relationship between intensity frames and event data. So we convert the stream-like events into a frame-like representation, which can be easily processed by the convolutional filters. However, simply stacking a series of events to form a frame-like tensor is not a rational representation. Whether the event streams are time-based (SBT) or number-based (SBN) stacking proposed in [41], they ignored the important time-stamp information and the polarities may be counteracted in a pixel. To maintain the temporal information and encode it into event tensors, we choose to use the spatial-temporal voxel grid [49] as the representation format with temporal bins B of 10.

The LFR-Net takes a sequence of voxel grids $\mathcal{V}(\mathbf{E}^t_{-/0/+})$ and the APS frame $\mathbf{I}^t$ as the input, where $\mathcal{V}(\mathbf{E}^t_0)$ represents the combination of both preceding and following events around APS frame. The basic architecture of LFR-Net is a U-Net [33] with different encoders for $\mathbf{I}^t$ and $\mathcal{V}(\mathbf{E}^t_{-/0/+})$, respectively. The feature maps extracted from $\mathbf{I}^t$ are concatenated at different scales with those from $\mathcal{V}(\mathbf{E}^t_{-/0/+})$, and activated by SE blocks [14] to estimate more accurate intensity residuals.

The skip-connections help the decoder of LFR-Net to fuse the feature maps from encoders and output a 1-channel intensity residual $\mathbf{Res}^t_{-/0/+}$ for each of the voxel grids $\mathcal{V}(\mathbf{E}^t_{-/0/+})$. Then the latent frames $\mathbf{L}^t_{-/0/+}$ could be reconstructed by element-wise multiplication $\otimes$ of $\mathbf{I}^t$ and $\mathbf{Res}^t_{-/0/+}$. Finally, the product is activated by a *Sigmoid()* function that restrains the intensity values to the range of $[0, 1]$, as expressed in Eq. (7):

$$\begin{aligned}\mathbf{L}^t_{-/0/+} &= LFR\left(\mathbf{I}^t, \mathcal{V}(\mathbf{E}^t_{-/0/+})\right) \\ &= Sigmoid\left(\mathbf{I}^t \otimes \mathbf{Res}^t_{-/0/+}\right). \end{aligned} \tag{7}$$

Note that $\mathbf{L}^t_0$ is the enhanced APS frame $\mathbf{I}^t$, which is deblurred and includes HDR information due to the HDR property and high temporal resolution of event data. The intensity residuals $\mathbf{Res}^t$ here hold different mathematical meaning from that in Eq. (6), because of the *Sigmoid()* activation function added in the final layer for network training.

3.2.2 Multi-image fusion

The structural information encoded in APS frames and event data is converted to intensity values of latent frames. Given a sequence of reconstructed latent intensity frames $\{\mathbf{L}_{-}^{t}, \mathbf{L}_{0}^{t}, \mathbf{L}_{+}^{t}\}$, the super-resolution problem can be treated as the MISR process. Therefore, we propose the multi-image fusion network (MIF-Net) to merge latent frames.

Firstly, the $\mathbf{L}_{-/0/+}^{t}$ are concatenated in channel axis. We can regard the concatenated tensor as a sequence of high-frame-rate video frames. Since there are multiple frames taken into consideration, the temporal alignment plays a crucial role to avoid blur artifacts for the next step of fusion. We use the deformable convolutional layers[9] to accomplish alignment between frames. We follow the EDVR [43] network that extracts feature maps of different frames in multiple pyramid levels, and compute the offsets between them for conducting alignment. The deformable convolutional layers can be easily embedded into our networks and trained end-to-end without any additional supervision. The aligned latent frames are expressed as:

$$\hat{\mathbf{L}}_{-/+}^{t} = DC\left(\mathbf{L}_{-/+}^{t}, \mathbf{L}_{\mathrm{o}}^{t}\right), \tag{8}$$

where $DC(x, y)$ denotes the deformable convolution block, which warps image x to the target image y with the computed offsets.

The aligned multiple frames $\{\hat{\mathbf{L}}_{-}^{t}, \mathbf{L}_{0}^{t}, \hat{\mathbf{L}}_{+}^{t}\}$ are fed into the fusion layers in the following process, which merges them in a channel-wise manner and reconstructs the high-resolution images. To reconstruct finer details of upsampled results, we use recurrent convolutional networks in this stage. Recurrent structure maintains a hidden state that is modulated by the input feature maps in each iteration to reconstruct finer detailed images. We apply feedback block [24] with dense connections, which retains the reconstructed details of the previous iteration in a hidden state and fuses with the feature maps in the next iteration. The feedback block outputs the residuals between the SR result $\mathbf{S}^{t}$ and the LR enhanced APS frame $\mathbf{L}_{0}^{t}$. So the final super-resolved intensity image $\mathbf{S}^{t}$ is the summation of residuals and the interpolated enhanced frame $\mathbf{L}_{0\uparrow}^{t}$, which can be represented as:

$$\begin{aligned}\mathbf{S}^{t} &= MIF\left(\mathbf{L}_{-}^{t}, \mathbf{L}_{0}^{t}, \mathbf{L}_{+}^{t}\right) \\ &= FB\left(\hat{\mathbf{L}}_{-}^{t}, \mathbf{L}_{\mathrm{o}}^{t}, \hat{\mathbf{L}}_{+}^{t}\right) \oplus \mathbf{L}_{\mathrm{o}\uparrow}^{t},\end{aligned} \tag{9}$$

where $FB()$ and $\oplus$ denote the feedback block and element-wise addition, respectively. Here we use bilinear interpolation to get $\mathbf{L}_{0\uparrow}^{t}$ from $\mathbf{L}_{0}^{t}$. The residual feature maps from $FB()$ are rearranged by a periodic shuffling operator [36] to match the size of $\mathbf{L}_{0\uparrow}^{t}$.

Since EvIntSR-Net reconstructs SR intensity images in a frame-by-frame manner, it naturally extends to video reconstruction. The frame rate of SR videos is determined by the number of merged events. So we can generate high-frame-rate videos in high resolution by interpolating more latent frames.

3.3. Dataset Preparation

Considering the end-to-end training of the proposed network, we need a sufficient dataset with the inputs of LR intensity images $\mathbf{I}^{T} = \{\mathbf{I}^{0}, ..., \mathbf{I}^{t}\}$ and the neighbouring event sequences $\mathbf{E}_{-/+}^{T} = \{\mathbf{E}_{-}^{0}, \mathbf{E}_{+}^{0} ..., \mathbf{E}_{-}^{t}, \mathbf{E}_{+}^{t}\}$. The ground truth are HR intensity images $\mathbf{H}^{T} = \{\mathbf{H}^{0}, ..., \mathbf{H}^{t}\}$.

However, there is no public large-scale dataset that consists of LR event data with the corresponding HR intensity images. What's more, the spatial resolution of APS and event data in a DAVIS sensor are both in low-resolution. So we cannot acquire the HR intensity images as ground truth if we collect the dataset with a real event sensor. Therefore, we use synthetic dataset to train our network as done in previous works [7, 32]. We choose event simulator V2E [10] to generate event data in arbitrary spatial resolution. V2E [10] synthesizes realistic event data from any real or synthetic conventional frame-based video using an accurate DVS model, which optionally uses Super-SloMo [15] to upsample the temporal resolution of a standard camera video. Since the synthetic frames interpolated from Super-SloMo [15] rely heavily on the pretrained model, we use high frame rate (240 FPS) and high-resolution (1280×720) videos (*e.g.*, Need for Speed [17] and GoPro [29] datasets) as input sources without frame interpolation.

Consequently, the high-resolution intensity images $\mathbf{H}^{T}$ are readily available. To simulate the real APS frames $\mathbf{I}^{T}$, we downsample the frame size of videos to 128×128 to generated LR event streams $\mathbf{E}_{-/+}^{T}$ using V2E [10]. Then we introduce LDR and blurry artifacts to the sharp frames by multiplying random exposures and averaging several consecutive frames. The corresponding sharp HR intensity images $\mathbf{H}^{T}$ are simply downsampled to 256×256 or 512×512 based on the training upscale factors ($2\times$ or $4\times$). A 240 FPS source video is viewed as a video that has 30 APS frames per-second. The other frames are regarded as latent frames' ground truth, as shown in Fig. 3.

The synthetic dataset has 3828 $\{\mathbf{I}^{T}, \mathbf{E}_{-/+}^{T}, \mathbf{H}^{T}\}$ data tuples generated from 132 video sequences. To improve the generalization ability of EvIntSR-Net to real event data, we randomly set the positive and negative contrast thresholds θ when generating events by sampling according to a normal distribution with mean 0.15 and standard deviation 0.03.

3.4. Training Strategy

3.4.1 Loss functions

There are three basic loss functions during the training process. ℓ_1 loss $\mathcal{L}_{\ell_1}$, ℓ_2 loss $\mathcal{L}_{\ell_2}$, and perceptual loss [16] $\mathcal{L}_{perc}$. $\mathcal{L}_{\ell_1}$ and $\mathcal{L}_{\ell_2}$ compute the mean absolute error (MAE)

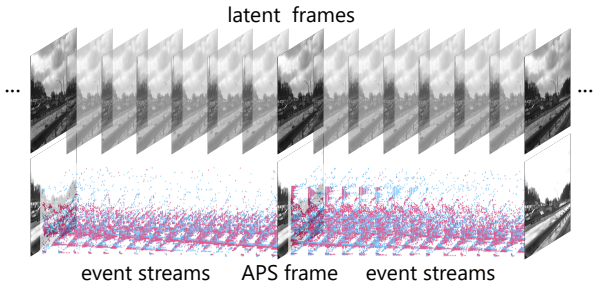

Figure 3: Top row: latent frames from a high-frame-rate video [17]. Bottom row: event streams generated from V2E [10] and the simulated APS frames by degrading (LDR and blur) some of these latent frames.

and mean square error (MSE) between the output of networks and the ground truth, respectively. The $\mathcal{L}_{perc}$ is defined based on the feature maps of images extracted by the VGG-16 network [37] pre-trained on ImageNet [34]:

$$\mathcal{L}_{perc} = \sum_{h} \Big(\big\| \phi_h(\mathbf{S}) - \phi_h(\mathbf{H}) \big\|_2^2 + \big\| G_h^{\phi}(\mathbf{S}) - G_h^{\phi}(\mathbf{H}) \big\|_2^2 \Big), \tag{10}$$

where ϕ_h denotes the feature map convoluted from h-th layer of VGG-16, G_h^{ϕ} is the Gram matrix of feature maps ϕ_h of two input images. Both of the two parts are computed by ℓ_2 norm. The layers we use to compute $\mathcal{L}_{perc}$ are 'relu4_3' and 'relu5_3' of VGG-16 network in our experiments.

For LFR-Net, we aim to output intermediate latent frames with more additional details for multi-frame fusion. The loss function is a combination of $\mathcal{L}_{\ell_1}$ and $\mathcal{L}_{perc}$:

$$\mathcal{L}_{LFR} = \alpha_1 \mathcal{L}_{\ell_1} + \alpha_2 \mathcal{L}_{perc}, \tag{11}$$

where α_1 and α_2 are weights for different losses, which are set to 100.0 and 5.0, respectively. As for MIF-Net, using ℓ_2 norm as the regularizer makes the SR results smoother. So we choose $\mathcal{L}_{\ell_2}$ and $\mathcal{L}_{perc}$ as the loss function for MIF-Net:

$$\mathcal{L}_{MIF} = \beta_1 \mathcal{L}_{\ell_2} + \beta_2 \mathcal{L}_{perc}, \tag{12}$$

where β_1 and β_2 are 200.0 and 3.0, respectively.

3.4.2 Implementation details

The proposed network is implemented by PyTorch with an NVIDIA 2080 Ti GPU. Our framework is an end-to-end learning approach. Instead of training the whole network simultaneously, we perform phase-to-phase training for better learning efficiency. The LFR-Net was pre-trained firstly with the supervision of latent frame ground truths. In the second phase, we fix the parameters of LFR-Net and use the output of LFR-Net to train the MIF-Net. Each of the sub-networks is trained for 200 epochs. We use the ADAM optimizer [21] with an initial learning rate of 10^{-5}. After the first 160 epochs, the learning rate linearly decays to 0 over the last 40 epochs.

4. Experimental Results

We compare the proposed method to several state-of-the-art image super-resolution approaches. Since there are just a few works aiming to reconstruct SR images for event cameras, we also compare with some image-based SR methods [13, 24]. The comparing approaches are listed below:

1) eSL-Net [40], which reconstructs HR intensity images from APS and event data.
2) E2SRI [7], which directly uses the pure event data as input to reconstruct HR intensity images.
3) EV [32]+SISR [24], which reconstructs the intensity images from event streams by E2VID [32] at first, then super-resolves using a trained SISR network [24].
4) APS+MISR [13], which is a multiple images super-resolution algorithm, the input are the target APS with its adjacent 7 intensity frames.

We can regard method 1) as the implementation of Eq. (3). Method 2) directly reconstructs SR intensity images from event streams, which is the same as the category of Eq. (2). Method 3) divides the SR process into 2 steps as described in Eq. (1). The non-learning-based reconstruction methods from events to images [1, 28] are omitted since they have been compared in [32] and shown lower-quality reconstruction.

4.1. Evaluation on Synthetic Dataset

Since there are ground truth HR images in the synthetic dataset, we conduct quantitative evaluation using peak-signal-to-noise (PSNR), structural-similarity (SSIM), and learned perceptual image patch similarity (LPIPS) [47] on the synthetic test dataset. The total test dataset consists of 841 intensity images and the event streams between two consecutive frames from 19 high-frame-rate video sequences [17, 29]. Table 1 shows the comparison of different methods on the evaluation metrics. Our model outperforms other comparing methods in all the metrics. Compared with the approaches listed above, quantitative evaluation shows that the proposed EvIntSR-Net is able to reconstruct $2\times$ and $4\times$ HR intensity images with higher quality on experiments of synthetic data.

Fig. 4 shows $2\times$ and $4\times$ SR results of EvIntSR-Net and other comparing approaches. Visual quality comparisons reveal that the fusion of intensity images and event data can achieve higher-quality image super-resolution with much more structural details. The APS frames lose HDR information and sharp details, which are encoded in event frames, as shown in the second column of Fig. 4. The event data capture much more details of a scene and contributes to the

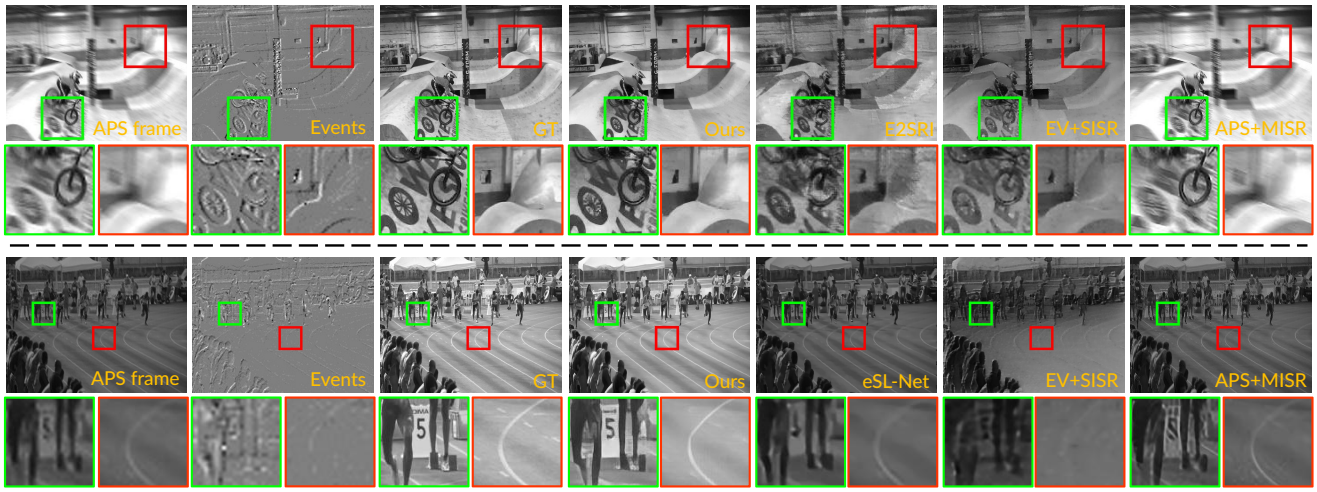

Figure 4: Visual quality comparison of $2\times$ SR (top part) and $4\times$ SR (bottom part) on synthetic dataset between EvIntSR-Net and other state-of-the-art super-resolution methods, including both event-based approaches: eSL-Net [40], E2SRI [7] ($2\times$ SR model weights of eSL-Net and $4\times$ SR model weights of E2SRI are unavailable) and EV [32]+SISR [24], as well as the image-based MISR method: APS+MISR [13]. The APS frames (first column) and event stacks (second column) are upsampled with bicubic interpolation to the corresponding scale for reference.

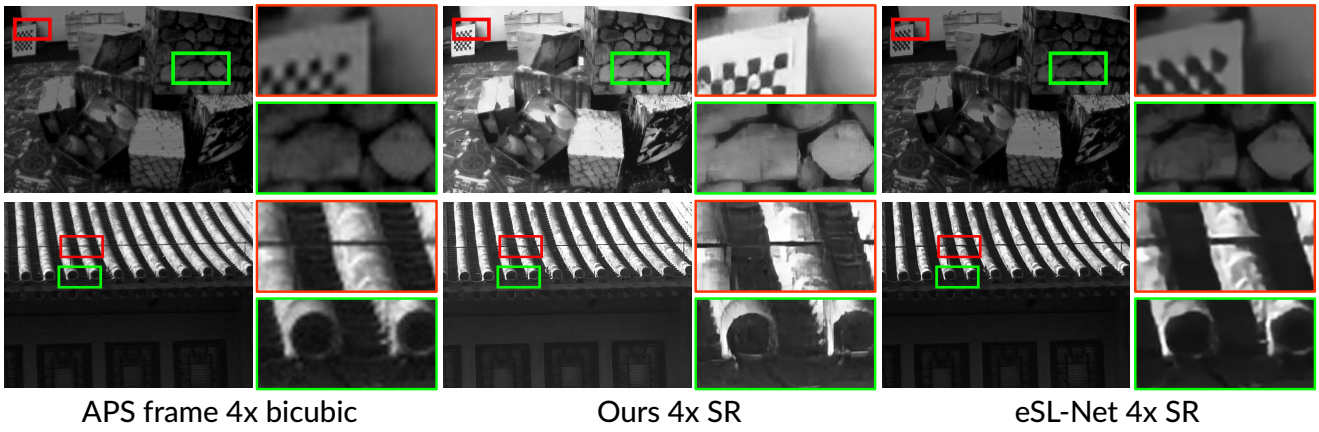

Figure 5: $4\times$ SR comparisons between EvIntSR-Net (Ours) and eSL-Net [40] on real-captured data by event cameras. The APS frames are upsampled with bicubic interpolation for reference.

final SR results. For example, the letter "W" in $2\times$ SR and the number "5" in $4\times$ SR comparing results of Fig. 4 can be restored finer by the EvIntSR-Net. However, eSL-Net [40] did not take advantage of using multiple latent frames. E2SRI [7] and produce blurry edges. EV [32]+SISR [24] take pure event data as input without the assistance of intensity images, the super-resolved results are estimated from high-frequency events densely distributed at the edges of the objects. The reconstructed images are unstable and highly influenced by the number of events collected in the stacks. Because few events cannot provide sufficient information for reconstruction, while too many events stacked on edges can easily induce blur artifacts. The MISR [13] approach merges the adjacent 7 APS frames to reconstruct the central target APS frame, while our EvIntSR-Net just takes advantage of the event data generated between the adjacent 2 APS frames. We use the central APS frame as the base image and do not need any additional intensity information from other APS frames. Compared with other methods, EvIntSR-Net can reconstruct much more structural details and clearer edges of objects. Our method outperforms both SISR and MISR approaches in qualitative comparisons and quantitative evaluations.

4.2. Evaluation on Real Dataset

Reconstruction results on real-world data are shown in Fig. 5. The testing samples are captured using a real event camera DAVIS346 on various scenarios. We compare our

Table 1: Quantitative evaluation of $2\times$ and $4\times$ SR (note the $2\times$ SR results of eSL-Net are unavailable, which are downsampled from $4\times$ results using bicubic interpolation. The $4\times$ SR of E2SRI model are also unavailable, so its $4\times$ results not provided.) results on synthetic dataset among ours and the comparing methods. ↑(↓) means the higher (lower) the better results throughout this paper.

Scale	Method	PSNR↑	SSIM↑	LPIPS↓
$2\times$	eSL-Net [40]	14.99	0.587	0.354
	E2SRI [7]	15.35	0.547	0.320
	EV [32]+SISR [24]	14.73	0.555	0.422
	APS+MISR [13]	15.69	0.673	0.303
	Ours	**23.12**	**0.776**	**0.130**
$4\times$	eSL-Net [40]	14.94	0.583	0.465
	E2SRI [7]	-	-	-
	EV [32]+SISR [24]	14.73	0.582	0.516
	APS+MISR [13]	15.18	0.609	0.427
	Ours	**23.25**	**0.745**	**0.231**

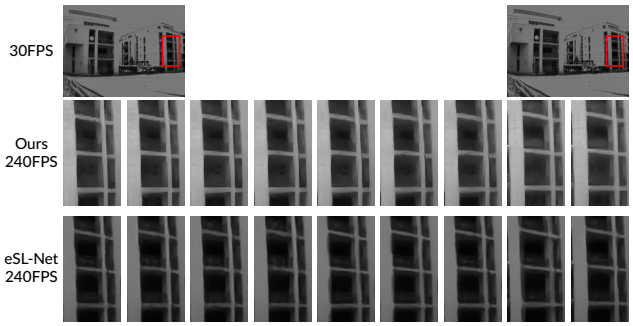

Figure 6: HFR video (240 FPS) generation with $4\times$ SR frames. The first row is a 30 FPS LR video from [27]. The bottom two rows are interpolated SR frames from EvIntSR-Net (Ours) and eSL-Net [40].

reconstruction results with eSL-Net [40]. The SR results demonstrate the effectiveness of EvIntSR-Net's ability to reconstruct sharper edges with more details and handle challenging SR reconstruction of real-world scenarios, such as HDR scenes (our reconstruction on roof tiling recovers better texture details than eSL-Net in under-exposed area).

Our method supports HFR video generation by reconstructing a sequence of SR latent frames,[1] as shown in Fig. 6. The number of latent frames depends on the number of event bins. We can reconstruct a 240 FPS video from a 30 FPS video by dividing the events between 2 consecutive frames into 8 bins. We put the target latent frame to be super-resolved in the central position of a sequence of latent frames. Then the other latent frames are merged to the target frame by deformable convolutional filters to achieve SR.

[1] More HFR videos and SR images are provided in the supplement.

Table 2: Quantitative evaluation of ablation study.

	PSNR↑	SSIM↑	LPIPS↓
W/o LFR-Net	22.97	0.769	0.134
W/o DC align	23.08	0.774	0.131
W/o FB block	23.08	0.766	0.151
Complete model	**23.12**	**0.776**	**0.130**

The 240 FPS SR videos look smoother and restore more realistic details compared to eSL-Net [40] and the original 30 FPS LR videos.

4.3. Ablation Study

To demonstrate the effectiveness of the proposed model design, we ablate different modules from the complete model and evaluate it quantitatively in Table 2. Firstly, the latent frames can be computed directly from the APS frame and event stacks using Eq. (6). To prove the necessity of LFR-Net, we remove LFR-Net and use the latent frames reconstructed from Eq. (6) to achieve MISR (W/o LFR-Net). In addition, we remove the deformable convolution block (W/o DC align) and feedback block (W/o FB block) to verify the significance of latent frames alignment and recurrent manner in MIF-Net, respectively.

5. Conclusion

This paper presents how to fuse event data with multiple latent intensity frames of event cameras to achieve reconstruction of SR intensity images. The SR process is separated into two steps: latent frame reconstruction and multi-frame fusion, which bridges the domain gap between event streams and intensity images, and achieves SR reconstruction from the MISR fusion manner, which is achieved by the proposed EvIntSR-Net. Extensive experiments on synthetic data and real-world data demonstrate that the proposed method can deal with HDR scenes and blurry artifacts, and outperforms various state-of-the-art comparing methods.

Limitations and future work. We focus on reconstructing SR intensity images using event data. However, when there is too fast camera motion, the APS frames are prone to have severe blurry artifacts. The proposed EvIntSR-Net has limited ability to handle huge blur, which is left as our future fork.

6. Acknowledgement

This work is supported by National Key R&D Program of China (2020AAA0105200), and National Natural Science Foundation of China under Grant No. 61872012, 61876007, 62088102.

References

[1] Patrick Bardow, Andrew J Davison, and Stefan Leutenegger. Simultaneous optical flow and intensity estimation from an event camera. In *Proc. of Computer Vision and Pattern Recognition*, 2016. 3, 6

[2] Francisco Barranco, Cornelia Fermuller, and Eduardo Ros. Real-time clustering and multi-target tracking using event-based sensors. In *International Conference on Intelligent Robots and Systems*, 2018. 1

[3] F. Barranco, C. L. Teo, C. Fermuller, and Y. Aloimonos. Contour detection and characterization for asynchronous event sensors. In *Proc. of Internatoinal Conference on Computer Vision*, 2015. 1

[4] Christian Brandli, Raphael Berner, Minhao Yang, Shih-Chii Liu, and Tobi Delbruck. A 240 $\times$ 180 130 db 3 μs latency global shutter spatiotemporal vision sensor. *Journal of Solid-State Circuits*, 2014. 1

[5] Jose Caballero, Christian Ledig, Andrew Aitken, Alejandro Acosta, Johannes Totz, Zehan Wang, and Wenzhe Shi. Real-time video super-resolution with spatio-temporal networks and motion compensation. In *Proc. of Computer Vision and Pattern Recognition*, 2017. 2

[6] Marco Cannici, Marco Ciccone, Andrea Romanoni, and Matteo Matteucci. Attention mechanisms for object recognition with event-based cameras. In *Winter Conference on Applications of Computer Vision*, 2019. 1

[7] Jonghyun Choi, Kuk-Jin Yoon, et al. Learning to super resolve intensity images from events. In *Proc. of Computer Vision and Pattern Recognition*, 2020. 2, 3, 5, 6, 7, 8

[8] Matthew Cook, Luca Gugelmann, Florian Jug, Christoph Krautz, and Angelika Steger. Interacting maps for fast visual interpretation. In *The International Joint Conference on Neural Networks*, 2011. 3

[9] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen We. Deformable convolutional networks. In *Proc. of Internatoinal Conference on Computer Vision*, 2017. 3, 5

[10] Tobi Delbruck, Yuhuang Hu, and Zhe He. V2E: From video frames to realistic DVS event camera streams. *arXiv preprint arXiv:2006.07722*, 2020. 5, 6

[11] Peiqi Duan, Zihao W Wang, Xinyu Zhou, Yi Ma, and Boxin Shi. EventZoom: Learning to denoise and super resolve neuromorphic events. In *Proc. of Computer Vision and Pattern Recognition*, 2021. 3

[12] Guillermo Gallego, Tobi Delbruck, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew Davison, Jörg Conradt, Kostas Daniilidis, et al. Event-based vision: A survey. *arXiv preprint arXiv:1904.08405*, 2019. 1

[13] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Recurrent back-projection network for video super-resolution. In *Proc. of Computer Vision and Pattern Recognition*, 2019. 2, 6, 7, 8

[14] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proc. of Computer Vision and Pattern Recognition*, 2018. 4

[15] Huaizu Jiang, Deqing Sun, Varun Jampani, Ming-Hsuan Yang, Erik Learned-Miller, and Jan Kautz. Super slomo: High quality estimation of multiple intermediate frames for video interpolation. In *Proc. of Computer Vision and Pattern Recognition*, 2018. 5

[16] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Proc. of European Conference on Computer Vision*, 2016. 5

[17] Hamed Kiani Galoogahi, Ashton Fagg, Chen Huang, Deva Ramanan, and Simon Lucey. Need for speed: A benchmark for higher frame rate object tracking. In *Proc. of Internatoinal Conference on Computer Vision*, 2017. 5, 6

[18] Hanme Kim, Ankur Handa, Ryad Benosman, Sio-Hoi Ieng, and Andrew Davison. Simultaneous mosaicing and tracking with an event camera. In *Proceedings of the British Machine Vision Conference*, 2014. 3

[19] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Deeply-recursive convolutional network for image super-resolution. In *Proc. of Computer Vision and Pattern Recognition*, 2016. 2

[20] Tae Hyun Kim, Mehdi SM Sajjadi, Michael Hirsch, and Bernhard Schölkopf. Spatio-temporal transformer network for video restoration. In *Proc. of European Conference on Computer Vision*, 2018. 2

[21] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[22] Hongmin Li, Guoqi Li, and Luping Shi. Super-resolution of spatiotemporal event-stream image. *Neurocomputing*, 2019. 3

[23] Sheng Li, Fengxiang He, Bo Du, Lefei Zhang, Yonghao Xu, and Dacheng Tao. Fast spatio-temporal residual network for video super-resolution. In *Proc. of Computer Vision and Pattern Recognition*, 2019. 2

[24] Zhen Li, Jinglei Yang, Zheng Liu, Xiaomin Yang, Gwanggil Jeon, and Wei Wu. Feedback network for image super-resolution. In *Proc. of Computer Vision and Pattern Recognition*, 2019. 5, 6, 7, 8

[25] Renjie Liao, Xin Tao, Ruiyu Li, Ziyang Ma, and Jiaya Jia. Video super-resolution via deep draft-ensemble learning. In *Proc. of Internatoinal Conference on Computer Vision*, 2015. 2

[26] Ziyang Ma, Renjie Liao, Xin Tao, Li Xu, Jiaya Jia, and Enhua Wu. Handling motion blur in multi-frame super-resolution. In *Proc. of Computer Vision and Pattern Recognition*, 2015. 2

[27] 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*, 2017. 8

[28] Gottfried Munda, Christian Reinbacher, and Thomas Pock. Real-time intensity-image reconstruction for event cameras using manifold regularisation. *International Journal of Computer Vision*, 2018. 6

[29] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *Proc. of Computer Vision and Pattern Recognition*, 2017. 5, 6

[30] Kamal Nasrollahi and Thomas B Moeslund. Super-resolution: A comprehensive survey. *Machine vision and applications*, 2014. 2

[31] Lichtsteiner Patrick, Christoph Posch, and Tobi Delbruck. A 128 $\times$ 128 120 db 15 μs latency asynchronous temporal contrast vision sensor. *IEEE journal of solid-state circuits*, 2008. 1

[32] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. Events-to-video: Bringing modern computer vision to event cameras. In *Proc. of Computer Vision and Pattern Recognition*, 2019. 1, 3, 5, 6, 7, 8

[33] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Proc. of International Conference on Medical Image Computing and Computer-assisted Intervention*, 2015. 4

[34] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 2015. 6

[35] Cedric Scheerlinck, Nick Barnes, and Robert Mahony. Continuous-time intensity estimation using event cameras. In *Proc. of Asian Conference on Computer Vision*, 2018. 1

[36] 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 *Proc. of Computer Vision and Pattern Recognition*, 2016. 5

[37] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 6

[38] Kim Soo Ye, Lim Jeongyeon, Na Taeyoung, and Kim Munchurl. 3DSRnet: Video super-resolution using 3D convolutional neural networks. *arXiv preprint arXiv:1812.09079*, 2018. 2

[39] Yapeng Tian, Yulun Zhang, Yun Fu, and Chenliang Xu. TDAN: Temporally-deformable alignment network for video super-resolution. In *Proc. of Computer Vision and Pattern Recognition*, 2020. 2, 3

[40] Bishan Wang, Jingwei He, Lei Yu, Gui-Song Xia, and Wen Yang. Event enhanced high-quality image recovery. In *Proc. of European Conference on Computer Vision*, 2020. 1, 2, 3, 6, 7, 8

[41] Lin Wang, Yo-Sung Ho, Kuk-Jin Yoon, et al. Event-based high dynamic range image and very high frame rate video generation using conditional generative adversarial networks. In *Proc. of Computer Vision and Pattern Recognition*, 2019. 1, 2, 3, 4

[42] Lin Wang, Tae-Kyun Kim, and Kuk-Jin Yoon. EventSR: From asynchronous events to image reconstruction, restoration, and super-resolution via end-to-end adversarial learning. In *Proc. of Computer Vision and Pattern Recognition*, 2020. 2, 3

[43] Xintao Wang, Kelvin C.K. Chan, Ke Yu, Chao Dong, and Chen Change Loy. EDVR: Video restoration with enhanced deformable convolutional networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019. 3, 5

[44] Zhihao Wang, Jian Chen, and Steven CH Hoi. Deep learning for image super-resolution: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 2

[45] Zihao Wang, Peiqi Duan, Oliver Cossairt, Aggelos Katsaggelos, Tiejun Huang, and Boxin Shi. Joint filtering of intensity images and neuromorphic events for high-resolution noise-robust imaging. In *Proc. of Computer Vision and Pattern Recognition*, 2020. 3

[46] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. Video enhancement with task-oriented flow. *International Journal of Computer Vision*, 2019. 2

[47] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proc. of Computer Vision and Pattern Recognition*, 2018. 6

[48] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proc. of European Conference on Computer Vision*, 2018. 2

[49] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Unsupervised event-based optical flow using motion compensation. In *Proc. of European Conference on Computer Vision*, 2018. 4

Supplementary Material: EvIntSR-Net: Event Guided Multiple Latent Frames Reconstruction and Super-resolution

Jin Han[1] Yixin Yang[2] Chu Zhou[1] Chao Xu[1] Boxin Shi[2,3,4]
[1]Key Lab of Machine Perception (MOE), Dept. of Machine Intelligence, Peking University
[2]NELVT, Dept. of Computer Science and Technology, Peking University
[3]Institute for Artificial Intelligence, Peking University
[4]Beijing Academy of Artificial Intelligence

6. Additional Results of EvIntSR-Net

6.1. Results on Image Super-resolution

In addition to Fig. 4 and Fig. 5 of the main paper, we provide more comparisons on synthetic and real data between the proposed EvIntSR-Net and other state-of-the-art methods, including E2SRI [1], eSL-Net [6], EV [5]+SISR [3], and APS+MISR [2]. Fig. 7 $\sim$ Fig. 10 show $2\times$ and $4\times$ SR results on synthetic data. Fig. 11 compares our SR results with those from eSL-Net [6] on real data.

6.2. High-frame-rate Video Generation

High-frame-rate (HFR) videos with super-resolved frames are shown in the supplementary video[1]. We first reconstruct multiple latent frames, then put each latent frame on the central position, which is viewed as the target frame to super-resolve. We interpolate continuous frames with frame-rate 8 times higher than the original APS frames (*e.g.*, 240 FPS videos from 30 FPS videos). In supplementary video, we compare our HFR videos with those generated from eSL-Net [6] on both simulated data and real-captured data. Results show that our reconstruction videos look smoother and reserve more details than eSL-Net [6].

7. Ablation Study on Loss Functions

We ablate different loss functions from the complete model and evaluate them quantitatively in Table 3. The comparing results show that using the combination of L2 loss and perceptual loss helps the network to perform better in reconstructing SR images.

Table 3: Ablation study on loss functions. "Perc. loss" means perceptual loss in this table.

	PSNR↑	SSIM↑	LPIPS↓
L2 loss only	23.03	0.767	0.170
Perc. loss only	22.65	0.740	0.132
L1 loss + perc. loss	22.35	0.764	0.140
L2 loss + perc. loss (Ours)	**23.12**	**0.776**	**0.130**

References

[1] Jonghyun Choi, Kuk-Jin Yoon, et al. Learning to super resolve intensity images from events. In *Proc. of Computer Vision and Pattern Recognition*, 2020. 1

[2] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Recurrent back-projection network for video super-resolution. In *Proc. of Computer Vision and Pattern Recognition*, 2019. 1

[3] Zhen Li, Jinglei Yang, Zheng Liu, Xiaomin Yang, Gwanggil Jeon, and Wei Wu. Feedback network for image super-resolution. In *Proc. of Computer Vision and Pattern Recognition*, 2019. 1

[4] 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*, 2017. 6

[5] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. Events-to-video: Bringing modern computer vision to event cameras. In *Proc. of Computer Vision and Pattern Recognition*, 2019. 1

[6] Bishan Wang, Jingwei He, Lei Yu, Gui-Song Xia, and Wen Yang. Event enhanced high-quality image recovery. In *Proc. of European Conference on Computer Vision*, 2020. 1, 6

[1] https://youtu.be/3Uc1MMiYiO4

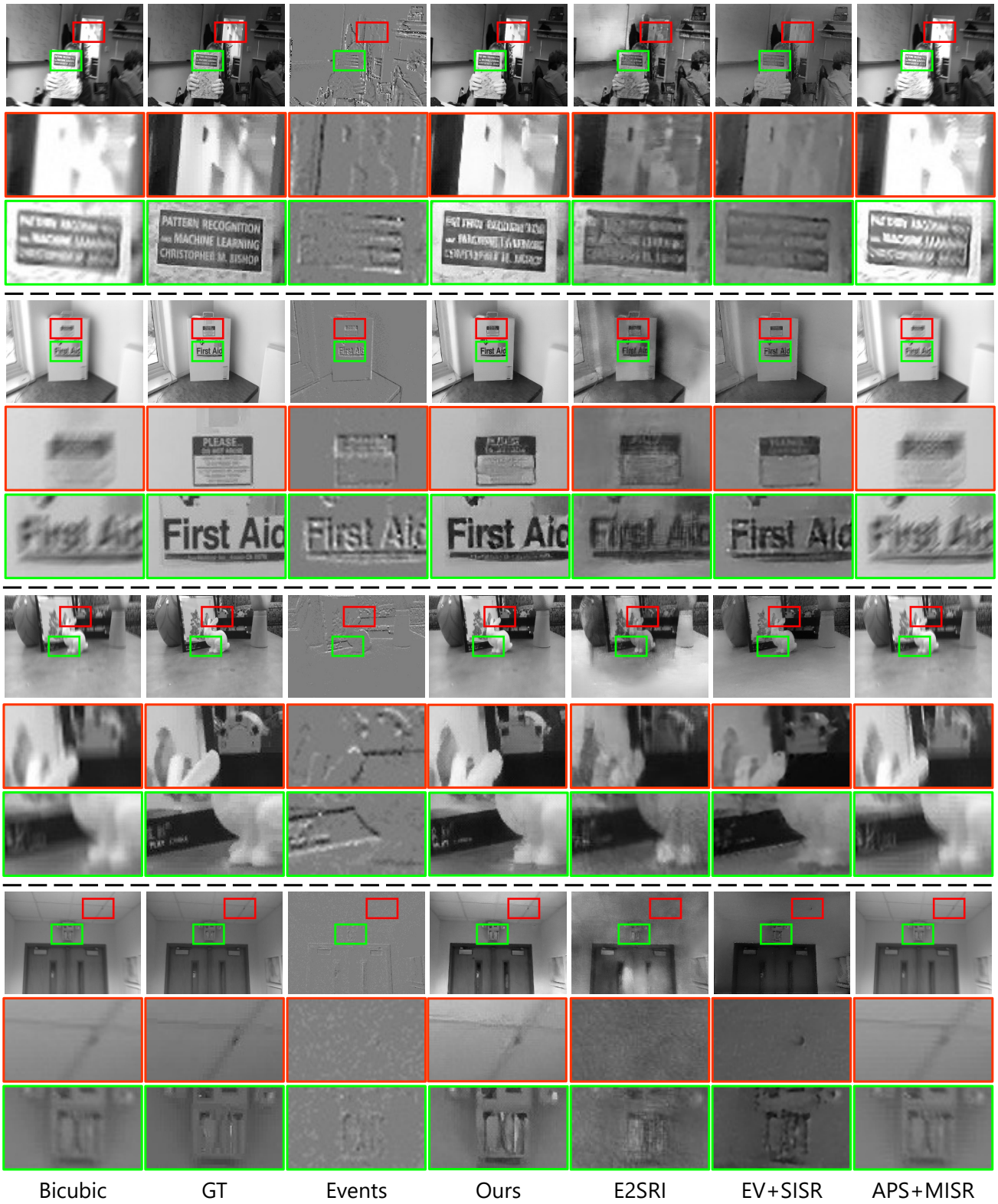

Figure 7: Visual quality comparison of 2× SR on synthetic dataset between EvIntSR-Net and other state-of-the-art super-resolution methods, including both event-based approaches and image-based methods. The APS frames (first column) and event stacks (third column) are upsampled with bicubic interpolation to the corresponding scale for reference.

Figure 8: Visual quality comparison of 2× SR on synthetic dataset between EvIntSR-Net and other state-of-the-art super-resolution methods.

Figure 9: Visual quality comparison of $4\times$ SR on synthetic dataset between EvIntSR-Net and other state-of-the-art super-resolution methods.

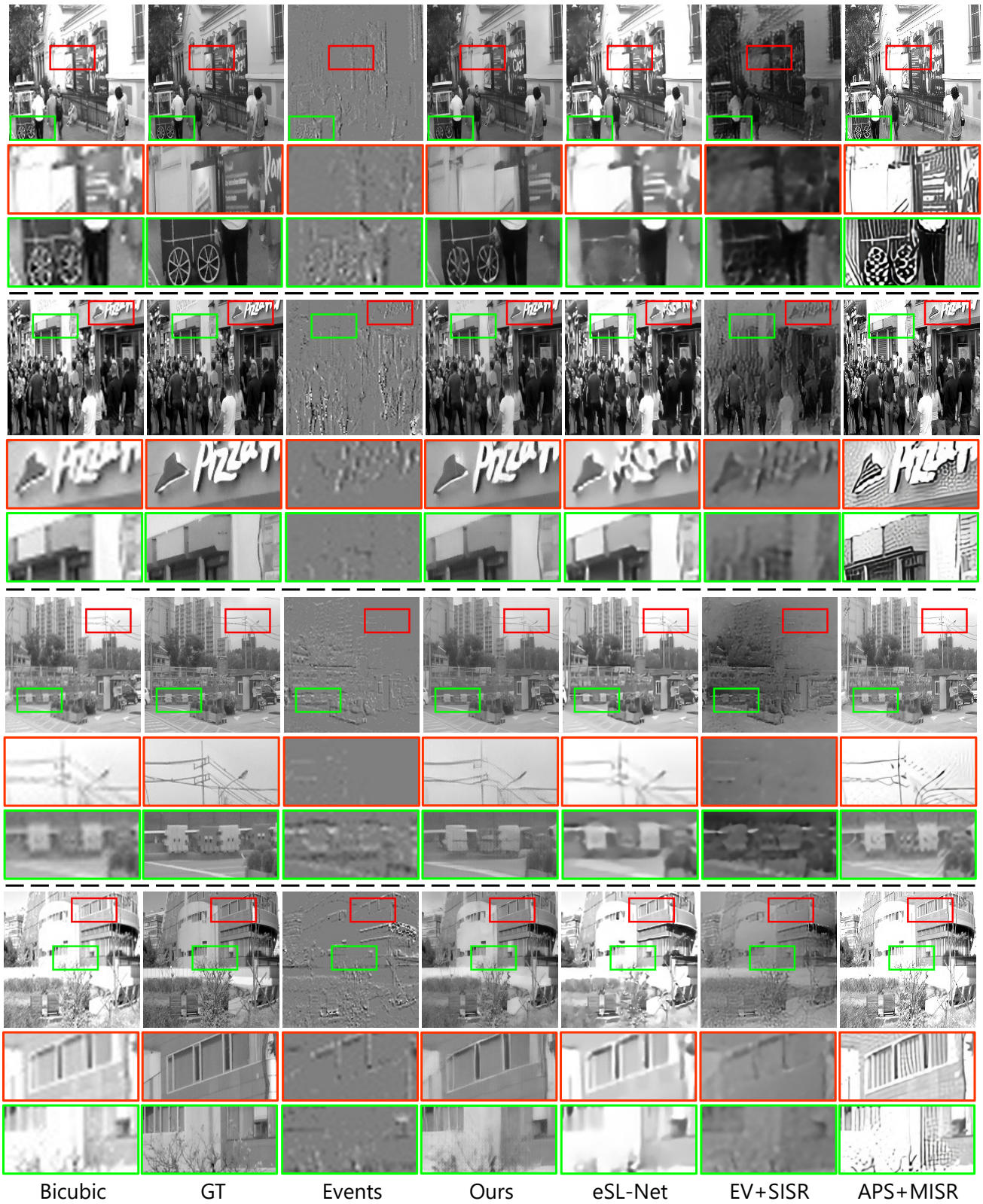

Figure 10: Visual quality comparison of $4\times$ SR on synthetic dataset between EvIntSR-Net and other state-of-the-art super-resolution methods.

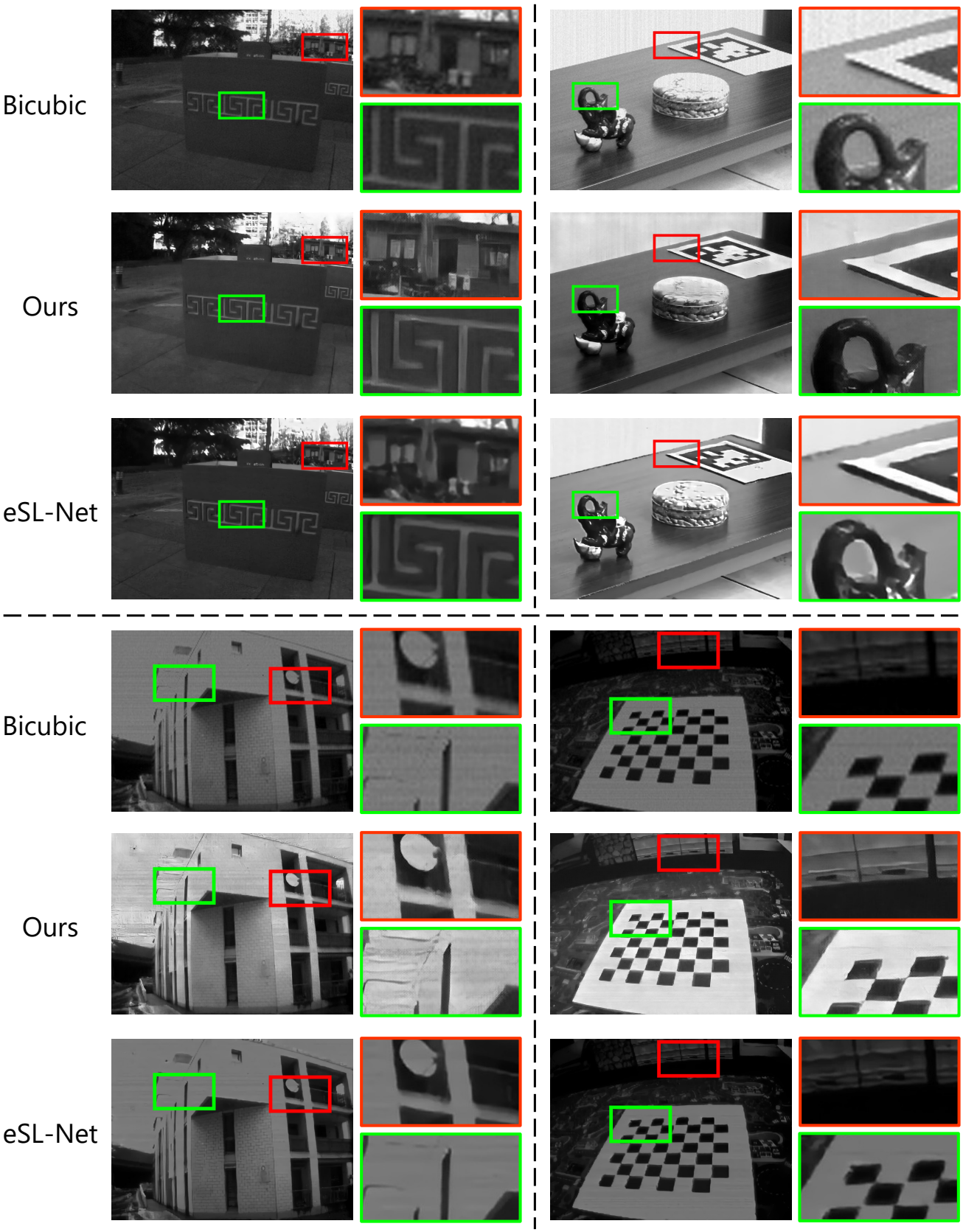

Figure 11: 4× SR visual quality comparison between EvIntSR-Net and eSL-Net [6] on real samples from DAVIS346 captured by us (top 2 cases) and public dataset [4] (bottom 2 cases).