Learning to Super-resolve Dynamic Scenes for Neuromorphic Spike Camera

Jing Zhao[1,2], Ruiqin Xiong[1,2], Jian Zhang[2,3], Rui Zhao[1,2], Hangfan Liu[4], Tiejun Huang[1, 2, 5]

[1] Institute of Digital Media, School of Computer Science, Peking University
[2] National Engineering Research Center of Visual Technology (NERCVT), Peking University
[3] School of Electronic and Computer Engineering, Peking University Shenzhen Graduate School
[4] Center for Biomedical Image Computing and Analytics, University of Pennsylvania
[5] Beijing Academy of Artificial Intelligence
{jzhaopku, rqxiong, zhangjian.sz, tjhuang}@pku.edu.cn, ruizhao@stu.pku.edu.cn, hfliu@upenn.edu

Abstract

Spike camera is a kind of neuromorphic sensor that uses a novel "integrate-and-fire" mechanism to generate a continuous spike stream to record the dynamic light intensity at extremely high temporal resolution. However, as a trade-off for high temporal resolution, its spatial resolution is limited, resulting in inferior reconstruction details. To address this issue, this paper develops a network (SpikeSR-Net) to super-resolve a high-resolution image sequence from the low-resolution binary spike streams. SpikeSR-Net is designed based on the observation model of spike camera and exploits both the merits of model-based and learning-based methods. To deal with the limited representation capacity of binary data, a pixel-adaptive spike encoder is proposed to convert spikes to latent representation to infer clues on intensity and motion. Then, a motion-aligned super resolver is employed to exploit long-term correlation, so that the dense sampling in temporal domain can be exploited to enhance the spatial resolution without introducing motion blur. Experimental results show that SpikeSR-Net is promising in super-resolving higher-quality images for spike camera.

Introduction

The newly emerged real-time computer vision applications, such as autonomous driving and unmanned aerial vehicle, require the machines to record high-speed motion clearly and response quickly. This makes the inherent limitations of conventional digital cameras evident. Conventional cameras are mostly based on one-time exposure imaging, which can produce clear images for still and slow-motion scenes. However, for dynamic scenes with high-speed motion, a single point on a moving object may be projected to different pixels on the image sensor, resulting in blurry images.

To capture high-motion dynamic scenes clearly, a novel neuromorphic camera called spike camera was recently invented (Dong, Huang, and Tian 2017; Dong et al. 2019; Huang et al. 2022). Unlike conventional cameras that compact all the visual information within an exposure interval into a single snapshot, spike camera abandons the concept of exposure window. Instead, it monitors the incoming light persistently and continuously, and outputs binary spike streams to record the dynamic scenes at extremely high temporal resolution (40000Hz). Additionally, different from the neuromorphic event cameras (Lichtsteiner, Posch, and Delbruck 2008b; Brandli et al. 2014b) that use *differential* sampling model to record the *brightness changes*, spike cameras fire spikes to record the arrival of a certain amount of photons. Such an *integral* sampling model enables the spike cameras to report *per-pixel luminance intensity*, providing more explicit information for recovering texture details.

For spike camera, image reconstruction is an important topic to bridge the domain gap between binary spikes and light intensity. This has attracted great attention in the past few years (Zhu et al. 2019, 2020; Zheng et al. 2021; Zhao et al. 2021b; Zhu et al. 2021; Zhao et al. 2021a). However, the previous works usually suffer from noise, blur or unsatisfactory details, due to the following issues:

(1) **Ultra high-speed motion.** Due to the existence of high-speed motion, the spikes fired by a single sensor pixel no longer describe the same point on the moving objects. Thus, special attention needs be paid to avoiding the mixing of lights from different object points.

(2) **Undesired noise.** Due to the Poisson effect of photon arrivals and the existence of thermal noises, the firing interval of neighboring spikes exhibits a strong fluctuation. For this reason, the reconstruction tends to be noisy, unstable, and spatio-temporally incoherent.

(3) **Limited spatial resolution.** As a trade-off for high temporal resolution, the spatial resolution of spike camera is limited, which results in inferior reconstruction details.

To address the above issues, this paper develops an end-to-end spike camera super-resolution network (SpikeSR-Net), aiming to super-resolve spike streams to high-quality high-resolution intensity images. In particular, SpikeSR-Net is designed based on the observation model of spike camera super-resolution imaging, which takes both the merits of model-based and learning-based methods. To be specific, we first propose a pixel-adaptive spike encoder (PASE) to convert the raw binary spikes to latent intensity representation. Considering the diversity of scene content, PASE adaptively exploits local spatio-temporal correlation to cope with different motion and light conditions. Then, we propose a motion-aligned super resolver (MASR), which conducts motion-aligned gradient descent and proximal mapping iteratively. MASR exploits long-term temporal correlation to convert the dense sampling in temporal domain to spatial

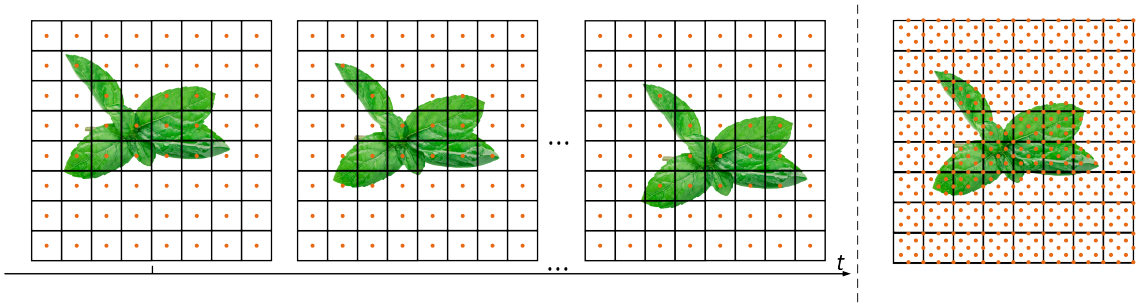


Figure 1: The principle of spike camera super-resolution imaging. Left: Continuous sampling. Due to the object motion, the spike camera sensor samples at different locations at different moments. Right: Motion-aligned resampling. Benefiting from the extremely high temporal resolution, the displacements are continuous and sub-pixel information are recorded, which enables more texture details to be reconstructed.

domain, which super resolves high-resolution images with fine details, high signal-to-noise ratio (SNR) and less blur. Experiments on both real-world and synthesized spike data demonstrate the promising performance of the SpikeSR-Net.

The main contributions of this paper are:

(1) We develop an optimization-inspired network, combining the merits of model-based and learning-based methods, to super-resolve binary spike streams to intensity images with texture details beyond the sensor resolution.

(2) We propose a pixel-adaptive spike encoder to convert binary spikes to latent representation. The encoder adaptively exploits spatio-temporal correlation to cope with different motion and light conditions. Particularly, it can be flexibly applied to other spike-camera-based tasks.

(3) We design an unfolding motion-aligned super resolver, employing gradient descent and proximal mapping iteratively, so that the network exploits the long-term temporal correlation to improve the SNR and texture details.

Related Works

Event camera reconstruction. Event cameras are neuromorphic sensors that send asynchronous events to record light intensity changes at a high temporal resolution (Litzenberger et al. 2006; Lichtsteiner, Posch, and Delbruck 2008a; Posch, Matolin, and Wohlgenannt 2008; Brandli et al. 2014a; Gallego et al. 2019). Recovering intensity images from events is an active topic. Kim *et al.* (Kim et al. 2008) proposed a Extended Kalman Filter to reconstruct gradient images from events. Bardow *et al.* (Bardow, Davison, and Leutenegger 2016) employed primal-dual algorithm to simultaneously estimate optical flow and light intensity. Some other works (Reinbacher, Graber, and Pock 2016; Scheerlinck, Barnes, and Mahony 2018) reconstructed images with direct event integration. Recently, many works (Scheerlinck et al. 2020; Rebecq et al. 2019a,b; Stoffregen et al. 2020; Wang et al. 2019a; Ahmed et al. 2021; Choi, Yoon et al. 2020; Jiang et al. 2020) explored to use deep convolutional networks for event camera reconstruction. However, as event cameras only provide relative beightness changes, these methods can hardly reconstruct texture details. Different from the event cameras, spike camera fires spikes to represent the arrival of a certain amount of photons, which provides a more explicit input for reconstructing textures.

Spike camera reconstruction. Recently, many spike camera reconstruction methods have been proposed. With an analysis of spike data, the work (Zhu et al. 2019) inferred the light intensity by estimating the spiking frequency of each pixel. However, there is a trade-off between improving SNR and removing motion blur. To address this issue, Zhao *et al.* (Zhao, Xiong, and Huang 2020) proposed a motion-aligned temporal filtering to handle the conflict brought by high-speed motion and noise. In addition, with the development of neural networks, some SNN-based (Zhu et al. 2020; Zheng et al. 2021) and CNN-based (Zhao et al. 2021b; Zhu et al. 2021) methods have been proposed to achieve high-quality imaging. However, these methods mainly focused on removing blur and noise artifacts, ignoring the issue of low resolution. To improve spatial resolution, Zhao *et al.* (Zhao et al. 2021a) develop an MGSR framework to super-resolve low-resolution (LR) spike streams to high-resolution (HR) images for the first time. However, MGSR is dependent on complex optimization and rigorous assumptions (e.g. motion consistency), which limits its applicability.

Image and video super-resolution. Over the past decades, many image super-resolution (SR) methods have been proposed to improve the spatial resolution. The early works employed interpolation techniques with low computation complexity. Some other works proposed more complex mapping functions between LR and HR images based on neighbor embedding (e.g. (Xie, Feris, and Sun 2015)) or sparse coding (e.g. (Dong et al. 2013)) to recover realistic textures. Recently, many CNN-based methods (Dong et al. 2014; Haris, Shakhnarovich, and Ukita 2018) have been proposed and made impressive progress. Additionally, based on the success of image SR methods, many video SR methods (Caballero et al. 2017; Jo et al. 2018; Kappeler et al. 2016; Liao et al. 2015; Tao et al. 2017) were proposed to jointly exploit the spatio-temporal correlation. An intuitive approach to generate HR images for spike camera is to combine the spike camera reconstruction methods with the image/video SR algorithms. However, such pipelined schemes do not take fully advantage of the spike characteristics.

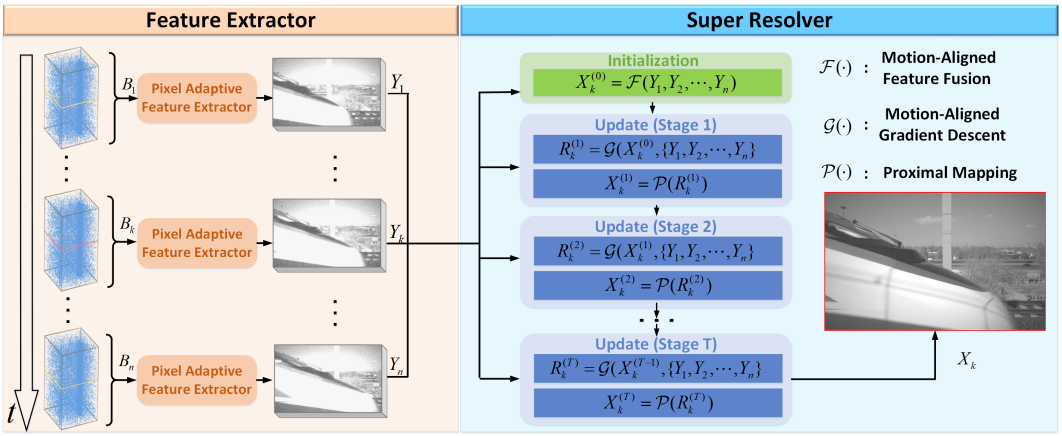


Figure 2: Overall pipeline of the proposed SpikeSR-Net. Firstly, the pixel-adaptive spike encoders convert the binary spikes to latent representations. Then, the motion-aligned super resolver iteratively exploits the long-term temporal correlation among different representations, aiming to reconstruct the high-quality HR image with fine details.

Problem Formulation

Basics of Spike Camera

Spike camera uses a "integrate-and-fire" mechanism to accumulate the incoming light $I(t)$ persistently. Whenever the dispatch threshold θ is reached, a spike is fired and the integrator is reset. Suppose $\{t_1, t2, \cdots, t_n\}$ are the firing time of the generated spikes, the "integrate-and-fire" process of the $k-$th spike can be formulated as:

$$\int_{t_{k-1}}^{t_k} \eta I(x)dx = \theta. \tag{1}$$

Here η denotes the photoelectric conversion rate. A pixel on spike image sensor may fire spikes at arbitrary time, but the camera can only read out the spikes as a discrete-time binary signal $S(n)$. To be specific, the k-th spike fired at the time t_k will be read out as $S(n_k) = 1$ with $n_k = \lceil t_k/\delta \rceil$. δ is the readout interval of spike camera (i.e., 25μs).

Spike Camera Super-resolution Imaging

As shown in Fig. 1, due to the relative motion between the camera and objects, the spike sensor may sample at different locations at different moments. Furthermore, the camera records the dynamic scene with an extremely high temporal resolution (40000Hz), which produces sub-pixel displacements and provides sub-pixel coverage for SR.

Suppose X_k is the HR image at time k. Based on the assumption of brightness constancy, the observation model of spike camera super-resolution (SCSR) imaging can be described as:

$$\begin{aligned} Y_i &= \mathcal{D}_s \mathcal{H}_i \mathcal{T}_{k\to i} X_k + N_i, \quad i = 1, 2, \cdots, n \\ \{S_1, S_2, \cdots, S_n\} &= \Gamma\left(\{Y_1, Y_2, \cdots, Y_n\}, \theta\right) \end{aligned} \tag{2}$$

Here $\mathcal{H}_i$ and $\mathcal{T}_{k\to i}$ represent the blur degradation and motion transform matrix, respectively. $\mathcal{D}_s$ denotes downsampling with scale factor s. N_i is additive noise. $\{Y_1, Y_2, \cdots, Y_n\}$ are the degraded LR intensity information at different moments. $\Gamma(\cdot)$ denotes the "integrate-and-fire" operation of the spike camera, generating a stream of binary spikes. The goal of SCSR is to convert the dense sampling in temporal domain to spatial domain, and super-resolve LR spike data $\{S_1, S_2, \cdots, S_n\}$ to high-quality HR intensity image $\hat{X}_k$.

Method

Overall Pipeline

Inspired by the SCSR observation model, this paper develops an end-to-end trainable SCSR network, dubbed SpikeSR-Net, to super-resolve the HR images from the LR spike streams. Fig. 2 shows the overview of SpikeSR-Net. It takes a set of consecutive spike frames $\{S_{k+i}\}, i = \pm 1, \pm 2, \cdots$ around k as input, so that the network can exploit the temporal correlation to generate the high-quality HR intensity image $\hat{X}_k$. The network mainly consists of two components, namely pixel-adaptive spike encoder (PASE) and motion-aligned super resolver (MASR). Firstly, the input spike stream is split into several overlapping short-term spike blocks $\{B_i\}, i \in \{1, 2, \cdots, k, \cdots, n-1, n\}$. The PASE $\mathcal{E}(\cdot)$ adaptively exploits the local spatio-temporal correlation among each binary spike block, producing latent intensity representation $\{Y_i\}$:

$$Y_i = \mathcal{E}(B_i). \tag{3}$$

Then, the target of the super resolver can be described as:

$$\arg\min_{X_k} \sum_{i=1}^{n} \omega_i \cdot \|\mathcal{D}_s \mathcal{H}_i \mathcal{T}_{k\to i} X_k - Y_i\|_2^2 + \lambda \Psi(X_k), \tag{4}$$

where ω_i denotes the weight related to the noise N_i. $\Psi(X_k)$ represents the HR image prior and λ denotes regularization parameter. The above inverse problem can be solved by iterating between the following update steps:

$$\begin{cases} R_k^{(t)} = X_k^{(t-1)} - \beta \sum_{i=1}^{n} \omega_i \mathcal{T}_{k\to i}^{\mathrm{T}} \mathcal{H}_i^{\mathrm{T}} \mathcal{D}_s^{\mathrm{T}} \left(\mathcal{D}_s \mathcal{H}_i \mathcal{T}_{k\to i} X_k^{(t-1)} - Y_i\right), \\ X_k^{(t)} = \arg\min_{X_k} \|X_k - R_k^{(t)}\|_2^2 + \lambda \Psi(X_k). \end{cases} \tag{5}$$

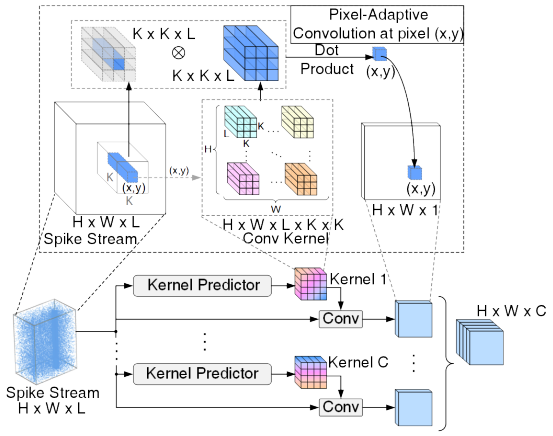


Figure 3: Illustration of pixel-adaptive spike encoder (PASE). The convolution kernels adapt to different motion and light conditions, aiming to convert the binary spikes to efficient latent representation.

where β is the update step size.

Inspired by the optimization process, we design an iterative CNN-based super-resolver to further exploit the long-term temporal correlation and reconstruct the HR image with fine details. The procedure can be expressed as:

$$\begin{cases} R_k^{(t)} = \mathcal{G}\left(X_k^{(t-1)}, \{Y_0, Y_1, \cdots, Y_n\}\right), \\ X_k^{(t)} = \mathcal{P}\left(R_k^{(t)}\right). \end{cases} \tag{6}$$

$\mathcal{G}(\cdot)$ and $\mathcal{P}(\cdot)$ represent gradient decent module (GDM) and proximal mapping module (PMM), respectively. The details of $\mathcal{E}(\cdot)$, $\mathcal{G}(\cdot)$ and $\mathcal{P}(\cdot)$ are discussed in the following.

Pixel-adaptive Spike Encoder

In spike camera, the firing of each spike represents the arrival of a certain amount of photons, where the accumulation process may occur in multiple readout intervals. That is, instead of representing the instantaneous intensity, each spike implicitly records the visual information, which has limited representation capacity. To address this issue, we design a pixel-adaptive spike encoder (PASE) to convert the binary spike streams into latent intensity representation.

Fig. 3 illustrates the proposed PASE. In order to enhance the representation capacity, convolutions are employed to exploit the local spatio-temporal correlation among spike streams. However, the diversity of scene contents may result in the diversity of correlation structure, and fixed convolution kernels can not well handle the various contents. For example, for the regions with high-speed motion or high-frequency edges, special attention should be paid to avoiding blur artifacts. Thus, rather than directly extract features using a fixed convolutional network, PASE uses the convolutional network to analyze the correlation structure among spike block B_i and generates a group of learnable kernel features $\mathcal{K}_i^{H\times W\times LK^2}$. H, W and L are the height, width and length of the input spike block. K is filter size. Then, the kernel features are reshaped into $H \times W$ liner filters $\mathcal{K}_{i,j}^{'K\times K\times L}$ and each filter is applied to corresponding input spike patches to extract the instantaneous intensity information:

$$Y_i(x,y) = \mathcal{K}_{i,j}^{'}(x,y) \otimes P_i(x,y). \tag{7}$$

Here $\mathcal{K}_{i,j}^{'}(x,y)$ denotes the $j-$th kernel for $P_i(x,y)$ and $P_i(x,y)$ denotes the $K \times K \times L$ binary spike patch centered at (x,y). Additionally, to enhance the capacity of PASE, we use several parallel branches to generate multiple groups of kernels. In particular, the filter sizes in different groups can be different, so that the PASE can be with various respective fields. The output of PASE is a stack of the features derived from different branches, which can be formulated as $Y_i = [Y_i^1, Y_i^2, \cdots, Y_i^C]$. Here C is the number of branches. In this paper, the PASE is applied to all the short-term spike blocks, generating a sequence of latent intensity representation $\{Y_1, Y_2, \cdots, Y_n\}$.

Motion-aligned Super Resolver

The target of super resolver is to reconstruct the HR intensity image with fine details. To this end, we propose to further exploit the long-term temporal correlation among different moments, so as to convert the dense sampling in temporal domain to spatial domain. In particular, considering the high-speed motion in the outer scene, special attention should be paid to avoiding motion blur. Inspired by the optimization model, we develop a motion-aligned GDM $\mathcal{G}(\cdot)$ and a PMM $\mathcal{P}(\cdot)$ to take advantage of long-term correlation and refine details iteratively.

Motion-aligned gradient descent module. Fig. 4 illustrates the architecture of the motion-aligned GDM $\mathcal{G}(\cdot)$. According to Equation (5), we first adopt gradient calculation units to calculate the gradients based on the latent representation at different moments. The gradient calculation unit in the i-th branch is defined as:

$$\text{alignment:} \quad E_i^{(t)} = \mathcal{T}_{k\to i}\left(\left[X_k^{(t-1)}, \mathcal{D}_s^{-1}(Y_i)\right]\right) \tag{8}$$

$$\text{scale down:} \quad L_i^{(t)} = \mathcal{D}_s\left(\mathcal{H}_i\left(E_i^{(t)}\right)\right) \tag{9}$$

$$\text{scale residual up:} \quad F_i^{(t)} = \mathcal{H}_s^{-1}\left(\mathcal{D}_i^{-1}\left(L_i^{(t)} - Y_i\right)\right) \tag{10}$$

$$\text{alignment:} \quad G_i^{(t)} = \mathcal{T}_{k\to i}^{-1}\left(\left[F_i^{(t)}, X_k^{(t-1)}\right]\right) \tag{11}$$

where $[,]$ represents concatenation. $\mathcal{D}_s(\cdot)$ and $\mathcal{D}_s^{-1}(\cdot)$ are the downsampling and upscaling operations, which are based on convolution and de-convolution with the stride of s. $\mathcal{H}_i(\cdot)$ and $\mathcal{H}_i^{-1}(\cdot)$ are the degradation and refinement operation based on residual blocks. $\mathcal{T}_{k\to i}$ represents the deformable convolution layers (Dai et al. 2017), that align the features of time k to the features of time i, and $\mathcal{T}_{k\to i}^{-1}(\cdot) := \mathcal{T}_{i\to k}(\cdot)$. In deformable convolution, additional offsets are learned to augment the spatial sampling locations, which has shown great transformation modeling capability in low-level computer vision (Wang et al. 2019b; Tian et al. 2020; Xiang et al. 2020; Shim, Park, and Kweon 2020). Here we exploit deformable convolution to align the features of different moments to avoid motion blur.

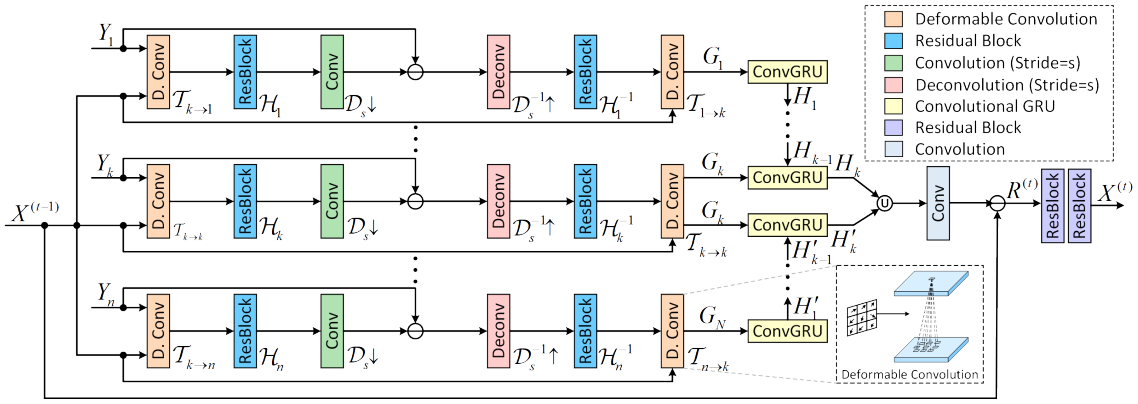


Figure 4: Illustration of (t)-th stage in motion-aligned super resolver. The architecture is designed based on Equation (5), where $R_k^{(t)}$ is the output of GDM and the input of PMM.

Then, to further exploit the long-term temporal correlation, we introduce a bi-directional ConvGRU-based feature fusion to aggregate the gradients $\{G_1^{(t)}, G_2^{(t)}, \cdots, G_n^{(t)}\}$ adaptively. Each ConvGRU (Tokmakov, Alahari, and Schmid 2017; Shen et al. 2019) consists of a learnable update gate and a learnable reset gate. The forward aggregation can be formulated as:

$$\begin{aligned} Q_i^{(t)} &= \tanh\left(\text{Conv}\left([C_i^{(t)} \odot H_{i-1}, G_i^{(t)}]\right)\right), \\ H_i^{(t)} &= (1 - Z_i^{(t)}) \odot H_{i-1}^{(t)} + Z_i^{(t)} \odot Q_i^{(t)}. \end{aligned} \tag{12}$$

The update gate $Z_i^{(t)}$ controls that how much the current information is incorporated into the new state. The reset gate $C_i^{(t)}$ controls that how much the previous state is preserved. $Z_i^{(t)}$ and $C_i^{(t)}$ are calculated based on $H_{i-1}^{(t)}$ and $G_i^{(t)}$ via convolution and Sigmoid activation operations. The backward aggregation is similar to the forward one. The output of GDM at stage t can be formulated as:

$$R_k^{(t)} = X_k^{(t-1)} - \text{Conv}\left(\left[H_k^{(t)}, H_k^{'(t)}\right]\right), \tag{13}$$

where $H_k^{(t)}$ and $H_k^{'(t)}$ represent the forward and the backward aggregated gradients, respectively.

Proximal mapping module. Then, a PMM $\mathcal{P}(\cdot)$ is proposed to solve the proximal mapping problem in Equation (5). The target of PMM is to restore $R_k^{(t)}$ closer to the desired signal. To this end, we adopt concatenated residual blocks to refine details, generating the output $X_k^{(t)} = \mathcal{P}\left(R_k^{(t)}\right)$.

Thus, the final reconstruction can be expressed as $\hat{X}_k = X_k^{(T)}$, with T representing the stage number.

Initialization. Before applying the iterations, it requires an initialization for $X_k^{(0)}$. In this paper, we present a motion-aligned feature fusion $\mathcal{F}(\cdot)$ to initialize $X_k^{(0)}$. To be specific, we first exploit deformable convolution to align each neighboring representation Y_i to Y_k. Then, we use the bi-directional ConvGRU-based feature fusion (depicted above) to aggregate the aligned feature. Finally, we upscale the aggregated features and derive $X_k^{(0)}$.

Implementation

Data Preparation

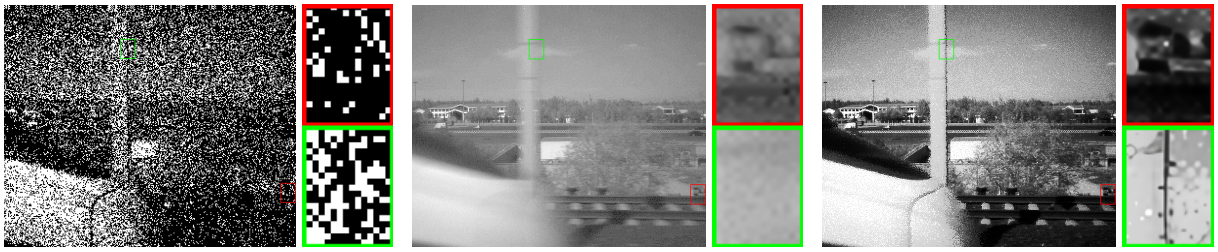
To train and evaluate the SpikeSR-Net, spike datasets, including LR spike streams and corresponding HR images, are required. However, it is challenging to capture the high-quality HR images, especially in high-motion dynamic scenes. To solve this problem, we modify the spike camera simulator in (Zhao et al. 2021b) to simultaneously generate LR spike streams and HR ground-truth. In this paper, we consider two modes, i.e., image-based synthesis and video-based synthesis. For image-based synthesis, we regard each selected image as the scene to record and suppose that there is a global motion between the scene and the sensor. We use the images from DIV2K(Agustsson and Timofte 2017) and the videos from REDS (Nah et al. 2020) and $\times$4K1000FPS (Sim, Oh, and Kim 2021) as the virtual scenes. The training dataset consists of 600 spike streams, which are generated based on all the three datasets to enhance diversity. For the testing datasets, we build a DIV2K-based spike dataset and a REDS-based spike dataset. Each testing dataset consists of 40 spike sequences. Note that there is no overlap between the testing datasets and the training dataset.

Training Details

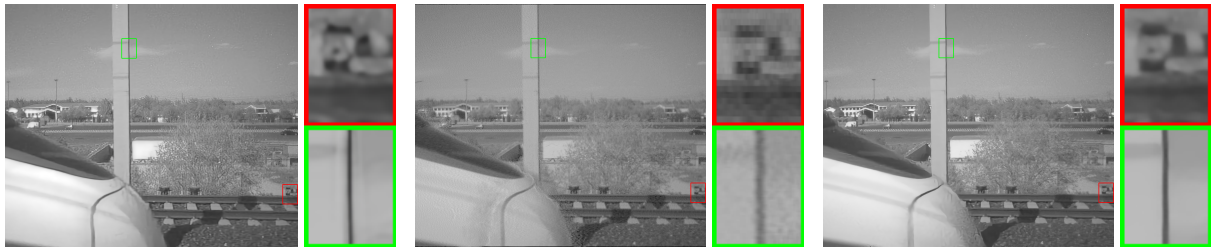
In our implementation, four residual blocks are used in kernel predictor. The stage number of super-resolver is set to 4. The loss function is defined as:

$$\mathcal{L} = \lambda_t \sum_{t=1}^{T} \|X_k^{(t)} - I_k\|. \tag{14}$$

I_k is the HR ground-truth at time k. When $t < T$, λ_t is set to 0.1. Otherwise, λ_t is set to 1. We crop the spike frames into 40×40 patches and set the batch size to 6. During training, data augmentation is performed by randomly rotating $90°$, $180°$, $270°$ and horizontally flipping. We use Adam optimizer and implement our experiments using PyTorch with two GTX 1080Ti GPUs. The model is trained for 30 epochs.

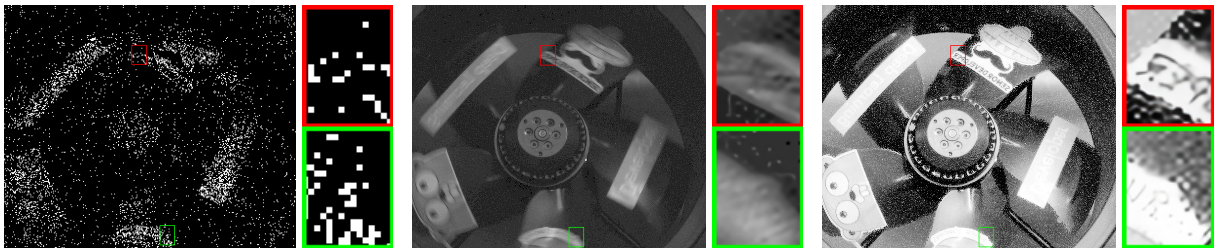


(a) Spike (b) TFP+MISR (c) STP+MISR

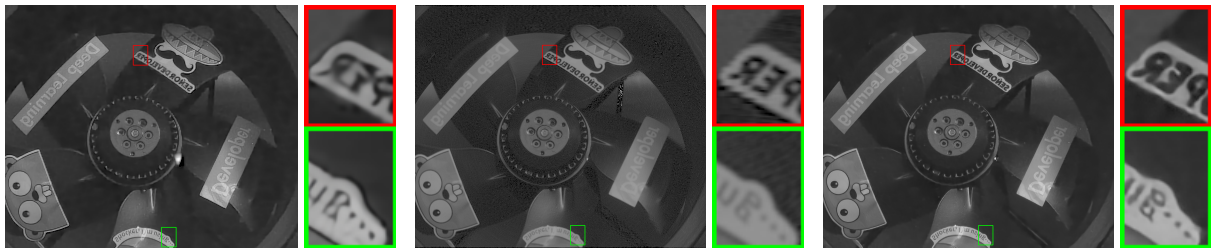


(d) Spk2ImgNet+MISR (e) MGSR (f) SpikeSR-Net (Proposed)

Figure 5: $\times 4$ SR results of the high-speed moving train (350km/h).



(a) Spike (b) TFP+MISR (c) STP+MISR



(d) Spk2ImgNet+MISR (e) MGSR (f) SpikeSR-Net (Proposed)

Figure 6: $\times 4$ SR results of the high-speed rotating fan (2600rpm).

Experiments

Comparative Results

To evaluate our SpikeSR-Net, we compare it with the existing spike SR method, i.e., MGSR (Zhao et al. 2021a). Additionally, we also combine existing spike camera reconstruction methods, i.e., TFP (Zhu et al. 2019), STP (Zheng et al. 2021) and Spk2ImgNet (Zhao et al. 2021b), with the competitive single image super resolution (SISR)(Luo et al. 2020) and multiple image super resolution (MISR)(Haris, Shakhnarovich, and Ukita 2019) for better comparison.

Synthesized data. Table 1 shows the PSNR and SSIM results on synthesized data. The results of STP+# methods are inferior, which is because that the inherent nonlinear operation would change the luminance and influence the measurement of PSNR and SSIM. Spk2ImgNet is the SOTA spike camera reconstruction method and Spk2ImgNet+# methods also achieve promising performance on spike camera SR. In particular, the additional MISR, which can exploit the temporal correlation, achieves better performance than SISR. MGSR is the first attempt to directly super-resolve LR spikes to HR intensity images, but the rigorous model assumption limits its effectiveness. Our SpikeSR-Net achieves

(a) Hybrid system (b) CMOS camera (c) Spike camera

Figure 7: Comparison with conventional camera for a high-speed dynamic scene. Our proposed method can reconstruct the high-speed moving object with fine details.

Method	DIV2K-based Spike Data		REDS-based Spike Data	
	×2	×4	×2	×4
TFP + SISR	20.18/0.3538	20.01/0.4132	23.09/0.5055	23.14/0.5230
TFP + MISR	20.53/0.4168	19.99/0.4087	24.15/0.5810	23.15/0.5180
STP + SISR	14.61/0.3585	14.37/0.3214	14.82/0.4237	14.84/0.4111
STP + MISR	15.03/0.4027	14.46/0.3296	15.44/0.4876	14.97/0.4260
Spk2ImgNet + SISR	24.48/0.7806	24.59/0.7064	25.39/0.8165	25.97/0.7637
Spk2ImgNet + MISR	27.14/0.8197	25.85/0.7253	28.76/0.8687	27.24/0.7881
MGSR	23.96/0.6603	22.93/0.5999	23.63/0.6341	22.73/0.5921
Ours	**28.70/0.8413**	**26.88/0.7451**	**30.79/0.8883**	**28.46/0.8012**

Table 1: PSNR(dB) and SSIM comparison with the state-of-the-arts.

Setting	PNSR/SSIM
W./O. pixel adaptive spike encoder	29.48/0.8585
W./O. gradient decent module	29.71/0.8643
W./O. motion alignment	30.00/0.8744
W./O. proximal mapping module	29.83/0.8663
Our final model	**30.79/0.8883**

Table 2: Quantitative evaluation results of ablation study on REDS-based dataset (×2 SR).

the best performance on the both datasets.

Real-world captured data. Fig. 5 and Fig. 6 show the ×4 SR results on real-world captured high-motion scenes. The TFP-based method and STP-based method suffer blur and noise, respectively. Moreover, compared to Spk2ImgNet-based method and MGSR, our SpikeSR-Net reconstructs clearer textures, achieving the best visual quality.

Ablation Study

To verify the validity of each module, we conduct a series of ablation studies and show comparisons in Table 2. We first show the effectiveness of PASE by comparing with a model without learning the latent representation. Then, we investigate the effectiveness of GDM and the necessity of the motion-alignment by removing the whole GDM and removing the alignment, respectively. Finally, we validate the necessity of PMM. We could tell that without the PMM, the performance degenerates.

Comparison with Conventional Cameras

We build a hybrid system to compare spike camera with a conventional CMOS camera as shown in Fig. 7. The conventional camera adopts auto-exposure mode and sets frame rate to 120 fps. As the conventional CMOS camera ignores the object motion within the exposure window, the image of high-speed object is blurry. In contrast, the spike camera produces a continuous spike stream to record the high-speed dynamic scene. By properly modeling the spatio-temporal correlation and motion, our proposed method can reconstruct the spike stream to a clear image with fine textures.

Conclusion

Spike camera is a novel neuromorphic sensor that outputs spike streams to record dynamic scenes at extremely high temporal resolution. In this paper, we present an end-to-end SpikeSR-Net to super-resolve the dynamic scenes from the continuous spike streams. By analyzing the observation model of SCSR, the SpikeSR-Net is designed from an optimization perspective, which combines the merits of both model-based and learning-based methods. Experiments on both real-world captured and synthesized spike data demonstrate its promising performance on reconstructing high-motion dynamic scenes.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants 22127807, 62072009, and in part by the National Key R&D Program of China under Grant 2021YFF0900501.

References

Agustsson, E.; and Timofte, R. 2017. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 126–135.

Ahmed, S. H.; Jang, H. W.; Uddin, S. N.; and Jung, Y. J. 2021. Deep Event Stereo Leveraged by Event-to-Image Translation. In *AAAI Conference on Artificial Intelligence (AAAI)*, 882–890.

Bardow, P.; Davison, A. J.; and Leutenegger, S. 2016. Simultaneous optical flow and intensity estimation from an event camera. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 884–892.

Brandli, C.; Berner, R.; Yang, M.; Liu, S.-C.; and Delbruck, T. 2014a. A 240×180 130 dB 3μs Latency Global Shutter Spatiotemporal Vision Sensor. *IEEE Journal of Solid-State Circuits (JSSC)*, 49(10): 2333–2341.

Brandli, C.; Berner, R.; Yang, M.; Liu, S.-C.; and Delbruck, T. 2014b. A 240x180 130db 3us latency global shutter spatiotemporal vision sensor. *IEEE Journal of Solid-State Circuits (JSSC)*, 49(10): 2333–2341.

Caballero, J.; Ledig, C.; Aitken, A.; Acosta, A.; Totz, J.; Wang, Z.; and Shi, W. 2017. Real-time video super-resolution with spatio-temporal networks and motion compensation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4778–4787.

Choi, J.; Yoon, K.-J.; et al. 2020. Learning to super resolve intensity images from events. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2768–2776.

Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; and Wei, Y. 2017. Deformable convolutional networks. In *IEEE International Conference on Computer Vision (ICCV)*, 764–773.

Dong, C.; Loy, C. C.; He, K.; and Tang, X. 2014. Learning a deep convolutional network for image super-resolution. In *European Conference on Computer Vision (ECCV)*, 184–199.

Dong, S.; Huang, T.; and Tian, Y. 2017. Spike camera and its coding methods. In *Data Compression Conference (DCC)*, 437–437.

Dong, S.; Zhu, L.; Xu, D.; Tian, Y.; and Huang, T. 2019. An efficient coding method for spike camera using inter-spike intervals. *arXiv preprint arXiv:1912.09669*.

Dong, W.; Zhang, L.; Lukac, R.; and Shi, G. 2013. Sparse representation based image interpolation with nonlocal autoregressive modeling. *IEEE Transactions on Image Processing (TIP)*, 22(4): 1382–1394.

Gallego, G.; Delbruck, T.; Orchard, G.; Bartolozzi, C.; Taba, B.; Censi, A.; Leutenegger, S.; Davison, A.; Conradt, J.; Daniilidis, K.; et al. 2019. Event-based vision: A survey. *arXiv preprint arXiv:1904.08405*.

Haris, M.; Shakhnarovich, G.; and Ukita, N. 2018. Deep back-projection networks for super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1664–1673.

Haris, M.; Shakhnarovich, G.; and Ukita, N. 2019. Recurrent Back-Projection Network for Video Super-Resolution. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Huang, T.; Zheng, Y.; Yu, Z.; Chen, R.; Li, Y.; Xiong, R.; Ma, L.; Zhao, J.; Dong, S.; Zhu, L.; et al. 2022. 1000x Faster Camera and Machine Vision with Ordinary Devices. *arXiv preprint arXiv:2201.09302*.

Jiang, Z.; Zhang, Y.; Zou, D.; Ren, J.; Lv, J.; and Liu, Y. 2020. Learning event-based motion deblurring. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3320–3329.

Jo, Y.; Oh, S. W.; Kang, J.; and Kim, S. J. 2018. Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3224–3232.

Kappeler, A.; Yoo, S.; Dai, Q.; and Katsaggelos, A. K. 2016. Video super-resolution with convolutional neural networks. *IEEE Transactions on Computational Imaging (TCI)*, 2(2): 109–122.

Kim, H.; Handa, A.; Benosman, R.; Ieng, S.-H.; and Davison, A. J. 2008. Simultaneous mosaicing and tracking with an event camera. *IEEE Journal of Solid-State Circuits (JSSC)*, 43: 566–576.

Liao, R.; Tao, X.; Li, R.; Ma, Z.; and Jia, J. 2015. Video super-resolution via deep draft-ensemble learning. In *IEEE International Conference on Computer Vision (ICCV)*, 531–539.

Lichtsteiner, P.; Posch, C.; and Delbruck, T. 2008a. A 128×128 120 dB 15μs Latency Asynchronous Temporal Contrast Vision Sensor. *IEEE Journal of Solid-State Circuits (JSSC)*, 43(2): 566–576.

Lichtsteiner, P.; Posch, C.; and Delbruck, T. 2008b. A 128×128 120dB 15us latency asynchronous temporal contrast vision sensor. *IEEE Journal of Solid-State Circuits (JSSC)*, 43(2): 566–576.

Litzenberger, M.; Posch, C.; Bauer, D.; Belbachir, A. N.; Schon, P.; Kohn, B.; and Garn, H. 2006. Embedded Vision System for Real-Time Object Tracking using an Asynchronous Transient Vision Sensor. In *Digital Signal Processing Workshop-signal Processing Education Workshop*, 173–178.

Luo, Z.; Huang, Y.; Li, S.; Wang, L.; and Tan, T. 2020. Unfolding the Alternating Optimization for Blind Super Resolution. *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 33.

Nah, S.; Baik, S.; Hong, S.; Moon, G.; Son, S.; Timofte, R.; and Lee, K. M. 2020. NTIRE 2019 Challenge on Video Deblurring and Super-Resolution: Dataset and Study. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*.

Posch, C.; Matolin, D.; and Wohlgenannt, R. 2008. An asynchronous time-based image sensor. *IEEE International Symposium on Circuits and Systems (ISCAS)*, 2130–2133.

Rebecq, H.; Ranftl, R.; Koltun, V.; and Scaramuzza, D. 2019a. Events-to-video: Bringing modern computer vision to event cameras. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3857–3866.

Rebecq, H.; Ranftl, R.; Koltun, V.; and Scaramuzza, D. 2019b. High speed and high dynamic range video with an event camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*.

Reinbacher, C.; Graber, G.; and Pock, T. 2016. Real-time intensity-image reconstruction for event cameras using manifold regularisation. *arXiv preprint arXiv:1607.06283*.

Scheerlinck, C.; Barnes, N.; and Mahony, R. 2018. Continuous-time intensity estimation using event cameras. In *Asian Conference on Computer Vision (ACCV)*, 308–324.

Scheerlinck, C.; Rebecq, H.; Gehrig, D.; Barnes, N.; Mahony, R.; and Scaramuzza, D. 2020. Fast image reconstruction with an event camera. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 156–163.

Shen, T.; Li, X.; Zhong, Z.; Wu, J.; and Lin, Z. 2019. R-Net: Recurrent and Recursive Network for Sparse-View CT Artifacts Removal. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 319–327.

Shim, G.; Park, J.; and Kweon, I. S. 2020. Robust Reference-Based Super-Resolution With Similarity-Aware Deformable Convolution. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 8425–8434.

Sim, H.; Oh, J.; and Kim, M. 2021. XVFI: Extreme video frame interpolation. In *IEEE International Conference on Computer Vision (ICCV)*, 14489–14498.

Stoffregen, T.; Scheerlinck, C.; Scaramuzza, D.; Drummond, T.; Barnes, N.; Kleeman, L.; and Mahony, R. 2020. Reducing the sim-to-real gap for event cameras. In *European Conference on Computer Vision (ECCV)*, 534–549.

Tao, X.; Gao, H.; Liao, R.; Wang, J.; and Jia, J. 2017. Detail-revealing deep video super-resolution. In *IEEE International Conference on Computer Vision (ICCV)*, 4472–4480.

Tian, Y.; Zhang, Y.; Fu, Y.; and Xu, C. 2020. TDAN: Temporally-Deformable Alignment Network for Video Super-Resolution. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3360–3369.

Tokmakov, P.; Alahari, K.; and Schmid, C. 2017. Learning video object segmentation with visual memory. In *IEEE International Conference on Computer Vision (ICCV)*, 4481–4490.

Wang, L.; Ho, Y.-S.; Yoon, K.-J.; et al. 2019a. Event-based high dynamic range image and very high frame rate video generation using conditional generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 10081–10090.

Wang, X.; Chan, K. C.; Yu, K.; Dong, C.; and Change Loy, C. 2019b. EDVR: Video restoration with enhanced deformable convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 0–0.

Xiang, X.; Tian, Y.; Zhang, Y.; Fu, Y.; Allebach, J. P.; and Xu, C. 2020. Zooming Slow-Mo: Fast and Accurate One-Stage Space-Time Video Super-Resolution. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3370–3379.

Xie, J.; Feris, R. S.; and Sun, M.-T. 2015. Edge-guided single depth image super resolution. *IEEE Transactions on Image Processing (TIP)*, 25(1): 428–438.

Zhao, J.; Xie, J.; Xiong, R.; Zhang, J.; Yu, Z.; and Huang, T. 2021a. Super Resolve Dynamic Scene From Continuous Spike Streams. In *IEEE International Conference on Computer Vision (ICCV)*, 2533–2542.

Zhao, J.; Xiong, R.; and Huang, T. 2020. High-Speed Motion Scene Reconstruction for Spike Camera via Motion Aligned Filtering. In *IEEE International Symposium on Circuits and Systems (ISCAS)*, 1–5.

Zhao, J.; Xiong, R.; Liu, H.; Zhang, J.; and Huang, T. 2021b. Spk2ImgNet: Learning To Reconstruct Dynamic Scene From Continuous Spike Stream. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 11996–12005.

Zheng, Y.; Zheng, L.; Yu, Z.; Shi, B.; Tian, Y.; and Huang, T. 2021. High-Speed Image Reconstruction Through Short-Term Plasticity for Spiking Cameras. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6358–6367.

Zhu, L.; Dong, S.; Huang, T.; and Tian, Y. 2019. A retina-inspired sampling method for visual texture reconstruction. In *IEEE International Conference on Multimedia and Expo (ICME)*, 1432–1437.

Zhu, L.; Dong, S.; Li, J.; Huang, T.; and Tian, Y. 2020. Retina-like visual image reconstruction via spiking neural model. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1438–1446.

Zhu, L.; Li, J.; Wang, X.; Huang, T.; and Tian, Y. 2021. NeuSpike-Net: High Speed Video Reconstruction via Bio-Inspired Neuromorphic Cameras. In *IEEE International Conference on Computer Vision (ICCV)*, 2400–2409.