Spike Stream Memory Transfer for Dynamic Scene Reconstruction

Yanchen Dong[1], Ruiqin Xiong[1]*, Rui Zhao[1,2], Xinfeng Zhang[3], Tiejun Huang[1]

[1]State Key Laboratory for Multimedia Information Processing, School of Computer Science, Peking University
[2]College of Computing and Data Science, Nanyang Technological University
[3]School of Computer Science and Technology, University of Chinese Academy of Sciences
yanchendong@stu.pku.edu.cn, {rqxiong, tjhuang}@pku.edu.cn, zhao.rui@ntu.edu.sg, xfzhang@ucas.ac.cn

Abstract

As a retina-inspired sensor with ultra-high temporal resolution, spike camera can continuously capture dynamic scenes with high-speed motion. It is a key task to restore clear images from spike streams. The quantization effects in spike readout bring degradation to the visual quality of restored images. To tackle the degradation without introducing motion blur, existing methods often employ a short-term temporal window to infer the light intensity at a certain time point. However, these methods only focus on the spike signals within the current window, which limits their performance. Motivated by the human-like memory mechanism for visual signals from the retina, we explore Spike Stream Memory Transfer (SSMT) to restore the dynamic scenes, considering spike signals beyond the window. Specifically, we design a framework that leverages temporal memory by transferring previously inferred light intensity and motion to enhance current reconstruction. The framework enables a long-term temporal perception of spike streams to handle the spike quantization effects. Besides, we utilize the estimated motion to suppress the potential blur from inter-stream clips, considering the underlying motion of spike streams. We also develop a spike interval-guided alignment module to tackle the blur from intra-stream clips. Experimental results on both synthetic and real-captured data demonstrate that our method can restore high-quality images from spike streams.

Code — https://github.com/csycdong/SSMT

Introduction

With the prevalence of high-speed vision applications, such as robotics, autonomous driving and industrial inspection, the demand for cameras that can capture high-speed motion and respond quickly is increasing. Due to exposure time limitations, conventional digital cameras often produce motion blur when capturing dynamic scenes, making them unsuitable for many high-speed vision applications.

Spike camera (Zhu et al. 2019; Zhao et al. 2021c; Zhu et al. 2025b,a) shows great potential in high-speed imaging, which is a neuromorphic sensor with ultra-high frequency (*e.g*, 40,000Hz). Mimicking the human retina, spike camera records dynamic scenes by continuously accumulating incoming photons and generating a binary spike signal stream. In particular, a spike will be fired whenever the accumulated photons reach a predetermined threshold. ***Different*** from another neuromorphic sensor called event camera, spike camera captures absolute light intensity rather than intensity changes. Recently, many studies of spike cameras have been performed, such as image reconstruction (Zhao et al. 2021b; Dong et al. 2022b; Hu et al. 2024a; Zhao et al. 2024b), and motion estimation (Zhao et al. 2022; Hu et al. 2022; Zhao et al. 2024a; Xia et al. 2024).

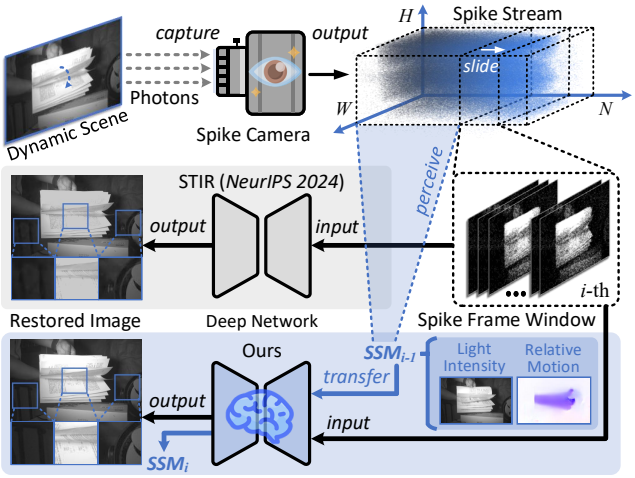

Figure 1: Dynamic scene reconstruction for spike cameras. Most existing methods focus on a short window of spike frames to restore images. We propose to transfer the spike stream memory (SSM) for long-term temporal perception.

Image reconstruction is a key task for spike cameras, which aims to restore clear images from the output spike streams. The restored high-frame-rate images of dynamic scenes can be used for many downstream tasks. In real spike camera sensors, the readout of spike signals is not at any time but is controlled by a discrete clock signal. Thus, there are quantization effects in spike streams. The quantization effects bring randomness to the spike signals, resulting in visual quality degradation of restored images.

To handle the degradation by quantization effects, exist-

*Corresponding author.

ing reconstruction methods (Zhao, Xiong, and Huang 2020; Zhao et al. 2021c,b; Zheng et al. 2021; Chen et al. 2022; Zhang et al. 2023) often employ a temporal window of neighboring spike signals. Considering the underlying motion of spike streams, the window tends to be short to avoid introducing potential motion blur. However, most of these methods only focus on the spike signals within the short-term temporal window, as shown in Fig. 1. The lack of exploitation of long-term temporal perception, which means utilizing spike signals from non-local time points, may limit the visual performance of their restored texture details, especially for areas of sparse spikes.

The human retina continuously encodes incoming light into neural signals, which are transmitted to the brain to perceive light intensity and motion of the dynamic scene. The memory mechanism of the brain allows it to interpret the current scene with the aid of previous observations. Motivated by the mechanism, we explore a Spike Stream Memory Transfer (SSMT) strategy for spike camera image reconstruction, which enables long-term temporal perception.

In our paper, we design a framework that transfers the spike stream memory of previously inferred light intensity and motion for dynamic scene reconstruction of spike cameras. As neighboring stream clips share similar texture details, the memory of intensity can enhance the current reconstruction. Considering the motion continuity, the motion memory can also aid motion estimation. By utilizing the estimated optical flows, we suppress potential blur from inter-stream clips during the intensity transfer. In most cases, the restored previous intensity is imperfect, resulting in an inaccurate warped result. Therefore, we employ a learnable flow head (LFH) to make the flow adaptive for reconstruction. Besides, spike intervals can better reflect the intensity structure in the spatiotemporal domain for feature matching. Thus, we input spike intervals for motion estimation. We also design a spike interval-guided alignment (SIA) module to handle potential blur from intra-stream clips.

Experiments demonstrate that our SSMT-based method achieves state-of-the-art (SOTA) performance on both synthetic and real-captured spike streams. Our main contributions can be summarized as follows:

- We explore the spike stream memory transfer method and propose a framework that utilizes previously inferred light intensity and motion to enhance reconstruction, achieving long-term temporal perception.
- We present an SSMT-based motion estimation strategy for spike cameras. By motion estimation and our spike interval-guided alignment module, we tackle potential blurs from inter- and intra-stream clips when aggregating long-term spike signals, respectively.
- Experiments on both synthetic and real-captured spike streams demonstrate that our method achieves the SOTA reconstruction performance of spike cameras.

Related Work

Dynamic Scene Reconstruction of Spike Camera

Non-Learning-based Methods. As the first attempt, TFI and TFP (Zhu et al. 2019) employ spike intervals and spike counts within a temporal window to infer the light intensity from spike streams, respectively. Zhao, Xiong, and Huang (2020) develop a motion-aligned filter (MAF) for spike streams to suppress the motion blur. Inspired by the short-term plasticity (STP) mechanism, TFSTP (Zheng et al. 2021) can restore high-quality images with low computational complexity. Besides, Zhao et al. (2021a) leverage relative motion and derive the relationship between intensity and each spike to restore high-resolution images. Designed for color spike cameras (Dong et al. 2024c,b,a, 2025b,c,a), 3DRI (Dong et al. 2022a) is proposed to restore color images from the Bayer-pattern spike stream. These methods usually explicitly exploit temporal properties of spike streams.

Learning-based Methods. With a progressive structure, Spk2ImgNet (Zhao et al. 2021b) is the first reconstruction network for spike cameras. Zhang et al. (2023) utilize temporal-robust features in the time-frequency domain with wavelet transforms, resulting in another network WGSE. Based on the spike neural network, SSIR (Zhao et al. 2023b) achieves low energy consumption. Fan et al. (2024) design the spatiotemporal interactive network STIR to jointly perform feature alignment and filtering, achieving the SOTA performance. In addition, Xiang et al. (2021) and Zhao et al. (2023a) explore end-to-end networks to reconstruct high-resolution images from low-resolution spike streams. Besides, Chen et al. (2022) propose SSML to enable training reconstruction models without ground truth images.

Dynamic Scene Reconstruction of Other Sensors

Event Camera. As another type of neuromorphic sensor, event cameras (Kai, Zhang, and Sun 2023; Kai et al. 2024, 2025; Xiao et al. 2024; Ding et al. 2022) capture intensity changes of dynamic scenes. To restore images, Kim et al. (2008) introduce an Extended Kalman Filter, leveraging photometric constancy. Rebecq et al. (2019) propose the classic E2VID model and achieve promoting visual quality. Hyper-E2VID (Ercan et al. 2024) employs hypernetworks for adaptive filters on a per-pixel basis. Besides, Weng, Zhang, and Xiong (2021) present a hybrid CNN-Transformer network for event-based image reconstruction.

Quanta Image Sensor (QIS). Unlike spike cameras and event cameras, QIS (Dutton et al. 2015, 2014) is designed for low-light applications. In particular, the high-frequency sensor can detect individual photons through spatial and temporal oversampling. Choi, Elgendy, and Chan (2018) propose the first reconstruction network to restore images with a QIS. To address the challenges of dynamic scenes under low-light conditions, Chi et al. (2020) introduce a training strategy where a student network learns by distilling knowledge from motion and denoising teachers.

Preliminary of Spike Camera

The spike camera contains an array of pixels that work independently. As shown in Fig. 2, the spike camera mimics the human retina structure and mechanism. Specifically, each pixel includes a receptor, an integrator and a comparator. To capture dynamic scenes, the pixels continuously accumulate incoming photons and fire spikes. Specifically, the pixel receives photons through the receptor and converts them into

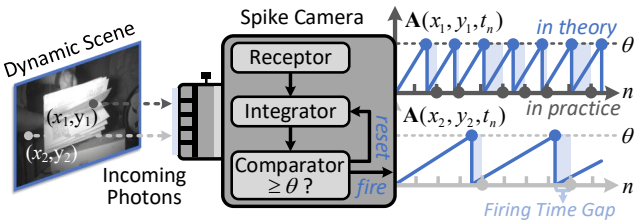

Figure 2: Spike camera mechanism, where (x_1, y_1) and (x_2, y_2) denote pixels with dense and sparse spikes, respectively. The firing time gaps lead to quantization effects.

electric charges for the integrator. When the accumulated charges reach a threshold θ according to the comparator, the pixel triggers a spike-firing flag. Subsequently, the integrator is reset, initiating a new "integrate-and-fire" cycle. The accumulation can be written as (see (Zhao et al. 2021c))

$$\mathbf{A}(x, y, t) = \int_0^t \eta \cdot \mathcal{P}(\mathbf{I}(x, y, \tau))d\tau \mod \theta, \quad (1)$$

where (x, y) denotes the spatial coordinate of the pixel, t denotes its time point, η denotes the photoelectric conversion rate, $\mathcal{P}(\cdot)$ denotes Poisson process of photon arrival, and $\mathbf{I}(x, y, \tau)$ denotes the light intensity.

In theory, the firing flags should be checked immediately. However, the flag status is read out as a discrete-time signal under the control of a clock in real implementation, as shown in Fig. 2. The time gaps result in the quantization effects of spike signals. At the n-th checking time point, we have the spike signal $\mathbf{S}_n(x, y) = 1$ if the firing flag is presented; otherwise, the signal is read out as 0. As a result, the spike camera generates a spike stream $\{\mathbf{S}_i\}_{i=1}^N$ with a shape of $N \times H \times W$, where N denotes the number of spike frames in the stream, and $H \times W$ denotes the spatial resolution.

Method

Motivation of Spike Stream Memory Transfer

The human visual system continuously receives incoming photons through the retina. Then the visual signals are encoded into neural signals and transmitted to the brain for imaging. Beyond current visual perception, the brain can retain the memory of recent observations, involving previous light intensity and motion of the dynamic scene. The memory can aid in understanding the current scene.

The spike camera mimics the structure and mechanism of the retina. The incoming photons are encoded into continuous spike signals by the spike camera. Existing methods have shown the effectiveness of neural networks for imaging from spike signals. Inspired by the memory mechanism of the brain, we explore introducing the spike stream memory (SSM) to networks to assist image reconstruction. As spike streams also contain the underlying intensity and motion, we can transfer SSM based on previous stream clips.

As shown in Fig. 3, most existing methods only consider spike signals within the current temporal window, limiting their temporal perception ranges. Due to the quantization effects and sensor noise, we need to perceive sufficient spike signals in the temporal domain for high-quality imaging. In contrast, the SSMT-based strategy can achieve long-term temporal perception. For example, the memory of the first clip can be transferred to the last inference stage.

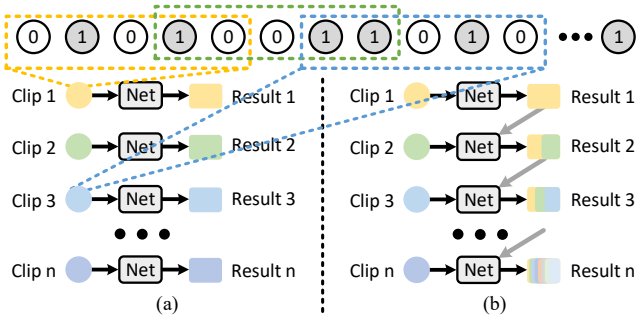

Figure 3: (a) Inference from current spike stream clip. (b) Inference with SSMT from previous clips. The colors of a certain result indicate the perceived clips for inference.

Overall Architecture

Inspired by the analysis above, we propose a Spike Stream Memory Transfer (SSMT)-based method to restore images from the spike stream. The overall architecture is presented in Fig. 4, which involves an initial reconstruction network and our SSMT-based reconstruction and motion estimation networks. In our method, we perform pre-training for the initial reconstruction and SSMT-based motion estimation networks. With the pre-trained models with frozen parameters, we train the SSMT-based reconstruction model.

As shown in Fig. 4, we employ a sliding spike frame window with the aid of previously inferred light intensity and motion to restore the image of each time point. For the first clip, we restore the image by the pre-trained model of an existing reconstruction method (*e.g.*, STIR (Fan et al. 2024)). For the following clips, we propose the SSMT-based reconstruction network to restore images. As neighboring stream clips share similar texture details, we transfer the previous light intensity to the current stage for better visual quality. Since spike intervals can better reflect the intensity structure for feature matching, we develop a spike interval-guided alignment (SIA) module in our reconstruction network to tackle the potential motion blur from intra-clips. To suppress the potential blur from inter-clips during the transfer, we estimate the optical flow between the two clips to warp the previous intensity before encoding, which also employs spike intervals. Considering the motion continuity between neighboring time points, we transfer previously inferred optical flows to aid current motion estimation.

SSMT-based Reconstruction

We propose to exploit the spike stream memory of light intensity to handle the quantization effects of spike signals and restore high-quality images with fine texture details.

Due to the underlying motion of the spike stream, there is potential blur when directly restoring an image from the whole input stream clip. Therefore, we first split the clip

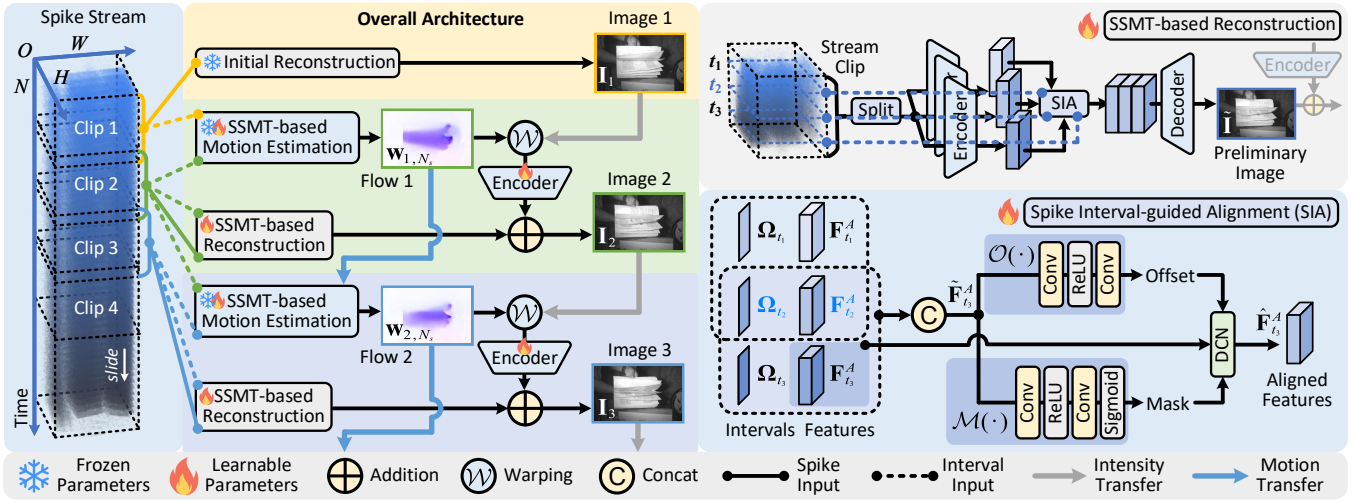

Figure 4: The overall architecture of our method, the structure of the SSMT-based reconstruction network, and its SIA module.

into three subclips and pass them to an encoder to extract features, respectively. The features are denoted as $\mathbf{F}^A_{t_j}, j \in \{1,2,3\}$, where t_j is the middle time point of a certain subclip. By our spike interval-guided alignment (SIA) module, we align the other features to the features of the middle time point to handle the feature offsets by intra-clip motion.

The spike interval (Zhu et al. 2019) is a basic spike stream representation, which can be formulated as

$$\mathbf{\Omega}_n(x,y) = \mathbf{R}_n(x,y) - \mathbf{L}_n(x,y), \tag{2}$$

$$\mathbf{R}_n(x,y) = \arg\min_i \{\mathbf{S}_i(x,y) = 1, i \geq n\}, \tag{3}$$

$$\mathbf{L}_n(x,y) = \arg\max_i \{\mathbf{S}_i(x,y) = 1, i < n\}, \tag{4}$$

where $\mathbf{\Omega}_n(x,y)$ denotes the spike interval for a certain pixel (x,y) of the n-th read-out time point, and $\mathbf{R}_n(x,y)$ and $\mathbf{L}_n(x,y)$ denote the time points of the left and right boundaries of the spike interval, respectively. Compared to binary spike signals, spike intervals have more gradation levels and can better reflect the intensity structure in the spatiotemporal domain. As a result, spike intervals are more efficient for feature matching. Thus, we guide the motion alignment by spike intervals in the SIA module, which is based on deformable convolution (Dai et al. 2017). As shown in Fig. 4, we concentrate the features and interval frames of t_2 and another time point t_a, which can be written as follows:

$$\tilde{\mathbf{F}}^A_{t_a} = \text{Concat}\left(\mathbf{F}^A_{t_a}, \mathbf{F}^A_{t_2}, \mathbf{\Omega}_{t_a}, \mathbf{\Omega}_{t_2}\right), \quad a \in \{1,3\}, \tag{5}$$

where $\mathbf{\Omega}_{t_a}$ denotes the interval frame of the time point t_a, and Concat$(\cdot)$ denotes the channel-wise concatenation. Guided by the intervals, we estimate the mask and offset for deformable convolution and align the features as

$$\hat{\mathbf{F}}^A_{t_a} = \text{DCN}\left(\mathbf{F}^A_{t_a}, \mathcal{O}\left(\tilde{\mathbf{F}}^A_{t_a}\right), \mathcal{M}\left(\tilde{\mathbf{F}}^A_{t_a}\right)\right), \quad a \in \{1,3\}, \tag{6}$$

where DCN$(\cdot)$ denotes the deformable convolution layer, and $\mathcal{O}(\cdot)$ and $\mathcal{M}(\cdot)$ denote the layers (see Fig. 4) for offset and mask estimation, respectively. Then we concentrate the middle time point features $\mathbf{F}^A_{t_2}$ and the aligned features $\{\hat{\mathbf{F}}^A_{t_1}, \hat{\mathbf{F}}^A_{t_3}\}$, and pass them to the decoder. After that, we restore the preliminary image $\tilde{\mathbf{I}}_i$ from the i-th stream clip. The process can be formulated as

$$\tilde{\mathbf{I}}_i = \mathcal{F}^R\left(\{\mathbf{S}_j, \mathbf{\Omega}_j\}_{j=(i-1)N_s+1}^{iN_s}\right), \tag{7}$$

where $\mathcal{F}^R(\cdot)$ denotes the main part of the reconstruction network, and N_s denotes the spike frame count of each input clip. The preliminary image is restored just from the current clip. To restore finer texture details, we perceive more spike signals by transferring the previous intensity $\mathbf{I}_{i-1}$ to the current stage. Due to the inter-clip motion, it is likely to introduce blur during the intensity transfer. Therefore, we warp $\mathbf{I}_{i-1}$ by our estimated optical flow before employing an encoder to extract the temporal information. Finally, we obtain the restored image of the current stage by

$$\mathbf{I}_i = \tilde{\mathbf{I}}_i + \mathcal{E}_w\left(\mathcal{W}\left(\mathbf{I}_{i-1}, \mathbf{w}_{i-1,N_r}\right)\right), \tag{8}$$

where $\mathcal{E}_w(\cdot)$ denotes the encoder for the warped image, $\mathcal{W}(\cdot)$ denotes the warping operation, and $\mathbf{w}_{i-1,N_r}$ denotes the optical flow from the i-th clip to the $(i-1)$-th clip.

As we focus more on an extensive SSMT-based framework, the encoders/decoders are just implemented by several (4/8) residual blocks (He et al. 2016). They can be replaced by some more effective modules for better performance.

SSMT-based Motion Estimation

We also explore SSMT for more accurate motion estimation, which is applied to the above intensity transfer. In particular, we propose the SSMT-based motion estimation network for spike cameras, built on the classic image motion estimation network, RAFT (Teed and Deng 2020).

The key problem of motion estimation is pixel-level feature matching between two time points. As discussed above, spike intervals can better reflect the intensity structure and present advantages for feature matching. As a result, we estimate optical flow from the middle time point of the i-th clip

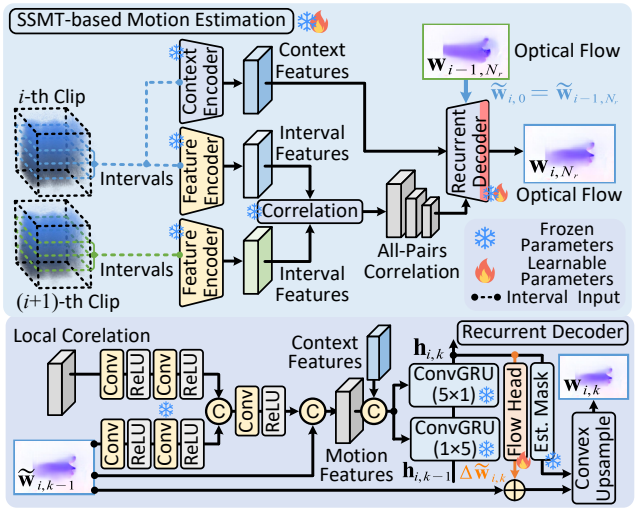

Figure 5: The structure of the SSMT-based motion estimation network. The symbol $N_r = 8$ denotes the number of iterations of the recurrent decoder.

to that of the $(i+1)$-th clip also by spike intervals, which can be formulated as follows:

$$\mathbf{w}_{i,N_r} = \mathcal{F}^M\left(\{\mathbf{\Omega}_j\}_{j=(i-1)N_s+1}^{2iN_s}, \mathbf{w}_{i-1,N_r}\right), \quad (9)$$

where $\mathcal{F}^M(\cdot)$ denotes our SSMT-based motion estimation network, $\mathbf{w}_{i-1,N_r}$ denotes the optical flow of the previous stage, and N_r denotes the iteration count of the recurrent decoder. As shown in Fig. 5, we keep most modules and the main structure of RAFT. We adapt it to spike camera motion estimation with SSMT. We first extract features from intervals of the two clips by two encoders and calculate the all-pairs correlation. Guided by context features from the intervals of the i-th clip, we estimate the optical flow $\mathbf{w}_{i,N_r}$ from the all-pairs correlation by the decoder.

Considering the motion continuity between neighboring clips, we can estimate more accurate optical flow with the aid of the previous optical flow. Thus, we propose to employ the downsampled previous optical flow $\widetilde{\mathbf{w}}_{i-1,N_r}$ as the initial low-resolution optical flow of the decoder, *i.e.*, $\widetilde{\mathbf{w}}_{i,0} = \widetilde{\mathbf{w}}_{i-1,N_r}$. Then the local correlation is looked up from the all-pairs correlation. We calculate the motion features $\mathbf{F}^M_{i,k-1}$ from the local correlation and the current flow. With the context features $\mathbf{F}^C_i$, we update the hidden states of the k-th iteration by ConvGRUs (Cho et al. 2014) as

$$\mathbf{h}_{i,k} = \text{ConvGRU}\left(\mathbf{h}_{i,k-1}, \text{Concat}\left(\mathbf{F}^M_{i,k-1}, \mathbf{F}^C_i\right)\right). \quad (10)$$

After that, we can obtain the residual between $\widetilde{\mathbf{w}}_{i,k}$ and $\widetilde{\mathbf{w}}_{i,k-1}$ by the flow head. With the residual, we can calculate the optical flow of the k-th iteration as follows:

$$\widetilde{\mathbf{w}}_{i,k} = \widetilde{\mathbf{w}}_{i,k-1} + \text{FlowHead}\left(\mathbf{h}_{i,k}\right). \quad (11)$$

By upsampling the optical flow of the last iteration, *i.e.*, $\widetilde{\mathbf{w}}_{i,N_s}$, we can obtain the estimated optical flow $\mathbf{w}_{i,N_s}$ to warp the previously inferred image for light intensity transfer. However, the image is not entirely accurate in most cases, leading to an inaccurate warped result by the estimated optical flow. To make the flow adaptive for reconstruction, we employ a flow head with learnable parameters during the end-to-end training of the reconstruction model.

Experiments

Experimental Settings

Training Details. For model training, we employ ℓ_1 loss function, Adam optimizer, and learning rate 10^{-4}, which decays to 0.9 times every 25 epochs. The batch size and patch size are 8 and 96×96, respectively. According to previous work (Zhao et al. 2021b; Chen et al. 2022; Zhang et al. 2023) and discussion in (Zhao et al. 2023b), the spike frame number N_s of each input clip is 41. The stride s for each stream clip in our approach is 20. The clip number of each training sample N_c is 10. The pre-training of the initial reconstruction models employs the same settings. The pre-training of our motion estimation models follows Spike2Flow (Zhao et al. 2022). Besides, we use the PyTorch framework and train the models via an NVIDIA RTX 3090 GPU.

Datasets. Since it is hard to collect lots of spike streams with high-quality ground truth, we develop a spike camera simulator to generate the datasets. Previous learning-based methods (Zhao et al. 2021b; Chen et al. 2022; Zhang et al. 2023; Hu et al. 2024b; Fan et al. 2024) have verified the effectiveness of simulators. To be specific, we employ the training set (240 scenes) and evaluation set (30 scenes) of REDS-120FPS (Nah et al. 2019) to obtain our training and evaluation sets. For each scene in the *training set* of REDS-120FPS, we synthesize 15 spike stream samples from different spatial areas (256×256). Each sample consists of $N_s + s(N_c - 1) = 221$ spike frames. To verify the model's generalization ability on other datasets, we generate another two evaluation sets with different motion and brightness distributions based on DAVIS (Pont-Tuset et al. 2017) and GoPro (Nah, Hyun Kim, and Mu Lee 2017). The three synthetic evaluation sets are denoted as REDS-SPK, DAVIS-SPK and GoPro-SPK. To further verify the generalization, we experimented with a set of widely used real-captured samples (book, car, train, fan, and so on), denoted as Real-SPK.

Compared Methods. The compared methods consist of three types. (A) non-learning-based methods: TFI (Zhu et al. 2019), TFP (Zhu et al. 2019), TFSTP (Zheng et al. 2021) and MAF (Zhao, Xiong, and Huang 2020). (B) learning-based methods (event cameras): ETNet (Weng, Zhang, and Xiong 2021) and Hyper-E2VID (Ercan et al. 2024). (C) learning-based methods: SSML (Chen et al. 2022), WGSE (Zhang et al. 2023), Spk2ImgNet (Zhao et al. 2021b) and STIR (Fan et al. 2024). In particular, WGSE, Spk2ImgNet and STIR serve as the SOTA image reconstruction methods for spike cameras. As shown in Table 1, we employ the three SOTA networks as our initial reconstruction methods for the first spike stream clip, respectively. The input of the event-based methods is also binary frames. We split the N_s binary frames into 5 overlapped parts to adapt to their recurrent input.

Method	REDS-SPK			DAVIS-SPK			GoPro-SPK			Real-SPK		Running Time
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	NIQE ↓	BRISQUE ↓	
TFI	23.93dB	0.5994	0.4013	22.65dB	0.4685	0.6077	23.86dB	0.5545	0.5093	10.680	39.273	0.0131s
TFP	28.35dB	0.7046	0.3012	24.83dB	0.6448	0.4490	30.76dB	0.7649	0.2954	12.700	41.801	**0.0003s**
TFSTP	22.87dB	0.5272	0.4072	21.86dB	0.4340	0.6038	22.97dB	0.5089	0.5052	14.955	42.633	16.026s
MAF	34.31dB	0.9104	0.0835	28.47dB	0.8378	0.2167	36.79dB	0.9281	0.0691	5.6037	34.381	1.0193s
ETNet	37.67dB	0.9562	0.0379	34.52dB	0.9140	0.0965	41.38dB	0.9816	0.0152	4.8723	38.785	7.8268s
Hyper-E2VID	38.16dB	0.9600	0.0363	34.94dB	0.9174	0.0955	41.83dB	0.9833	0.0154	4.9671	40.253	0.4184s
SSML	37.13dB	0.9528	0.0518	33.50dB	0.9030	0.1328	40.86dB	0.9794	0.0204	5.6987	37.128	3.0927s
WGSE	39.43dB	0.9667	0.0305	35.58dB	0.9209	0.0867	42.54dB	0.9847	0.0131	4.8783	38.165	0.8819s
Ours + WGSE	40.28dB	0.9728	0.0255	36.46dB	0.9308	0.0783	43.09dB	0.9868	0.0126	4.4911	**32.126**	0.3906s
Spk2ImgNet	40.13dB	0.9719	0.0259	36.28dB	0.9319	0.0786	42.92dB	0.9862	0.0133	5.5243	40.253	1.3363s
Ours + Spk2ImgNet	**40.37dB**	0.9737	0.0238	**36.53dB**	**0.9324**	**0.0771**	**43.12dB**	**0.9870**	0.0122	**4.4535**	34.166	0.4342s
STIR	39.78dB	0.9706	0.0241	35.33dB	0.9233	0.0917	42.75dB	0.9857	0.0128	4.8051	40.464	0.1225s
Ours + STIR	40.35dB	**0.9738**	**0.0230**	36.45dB	0.9320	0.0787	43.11dB	0.9869	**0.0119**	4.4959	35.531	0.3126s

Table 1: Quantitative comparison of reconstruction performance and running time on the synthetic and real-captured evaluation datasets. **Red** and blue indicate the best and the second-best performance, respectively.

Comparison with Existing Methods

Quantitative Results. Table 1 shows the quantitative comparison of reconstruction performance and running time. The values come from the average across all the clips (N_c= 10) of each sample. According to the table, our methods achieve the best PSNR/SSIM/LPIPS results on the three synthetic evaluation sets with different motion distributions, which verifies the performance advantages and the generalization ability. Compared with the corresponding SOTA methods, ours + WGSE, ours + Spk2ImgNet and ours + STIR bring performance gains on all the datasets, respectively. The performance gaps among our three methods with different initial reconstruction methods are small, which shows the stability of our methods. Among the non-learning-based methods, MAF presents the best performance.

We implement experiments on Real-SPK to evaluate the performance on real-captured spike streams without ground truth. As shown in Table 1, ours + Spk2ImgNet and ours + WGSE achieve the best NIQE and BRISQUE results, respectively. The phenomenon shows the generalization of our models trained on the synthetic dataset. Our methods also obtain performance gains over the three initial methods on the real-captured data, respectively. As a non-learning-based method, MAF achieves comparable NIQE/BRISQUE results to the learning-based methods.

According to the running time comparison, our methods achieve faster inference speeds than most of the learning-based methods. TFP and TFSTP are the fastest and slowest methods among all the methods. With better performance, ours + Spk2ImgNet/WGSE achieves more than twice the inference speed of Spk2ImgNet/WGSE.

Visual Results. Fig. 6 shows the visual comparison on real-captured samples. The samples involve bright (dense spikes) and dark (sparse spikes) scenes with high-speed motion, respectively. In the figure, our methods achieve the best texture details. Since TFP does not consider the underlying motion, its visual results present significant blur.

Case	Split	Align	SIA	LIM	PSNR(dB) of Ours +			
					WGSE	S2INet	STIR	*Avg.*
(1)					39.59	39.64	39.56	39.60
(2)	✓				39.89	39.95	39.88	39.90
(3)	✓	✓			39.95	40.06	39.99	40.00
(4)	✓		✓		40.05	40.14	40.07	40.09
(5)	✓		✓	✓	**40.28**	**40.37**	**40.35**	**40.34**

Table 2: Ablation study for reconstruction, where SIA means the spike interval-guided alignment module, S2INet means Spk2ImgNet, and LIM means the light intensity memory.

Ablation Study

We implement ablation studies to verify the effectiveness of the proposed strategies and modules. Table 2 presents the studies for reconstruction. Case (1) is a baseline version without spike clip splitting and light intensity memory (LIM) transfer. Compared with Case (1), we split the input spike clip in Case (2) to explore the impact. To verify the effectiveness of alignment, we introduce a DCN-based module to Case (2) to align features from the split parts, resulting in Case (3). In Case (4), we replace the module in Case (3) with the SIA module to show the impact of the alignment guidance of spike intervals. Compared with the final version Case (5), Case (4) only removes the LIM transfer.

Table 3 shows the studies for motion estimation. In Case (A), we remove the motion memory (MM) transfer and use the flow head of frozen parameters, which can be regarded as RAFT with spike interval input. Case (B) only employs the frozen flow head to explore the impact of the motion memory transfer by comparing it with Case (A). In Case (C), we replace the input of the final version from spike intervals to spike frames for motion estimation. In Case (D), we employ a commonly used RAFT-style spike camera motion estimation network with more parameters, Spike2Flow (Zhao et al. 2022). Case (E) is our final version. By comparing Case (B)

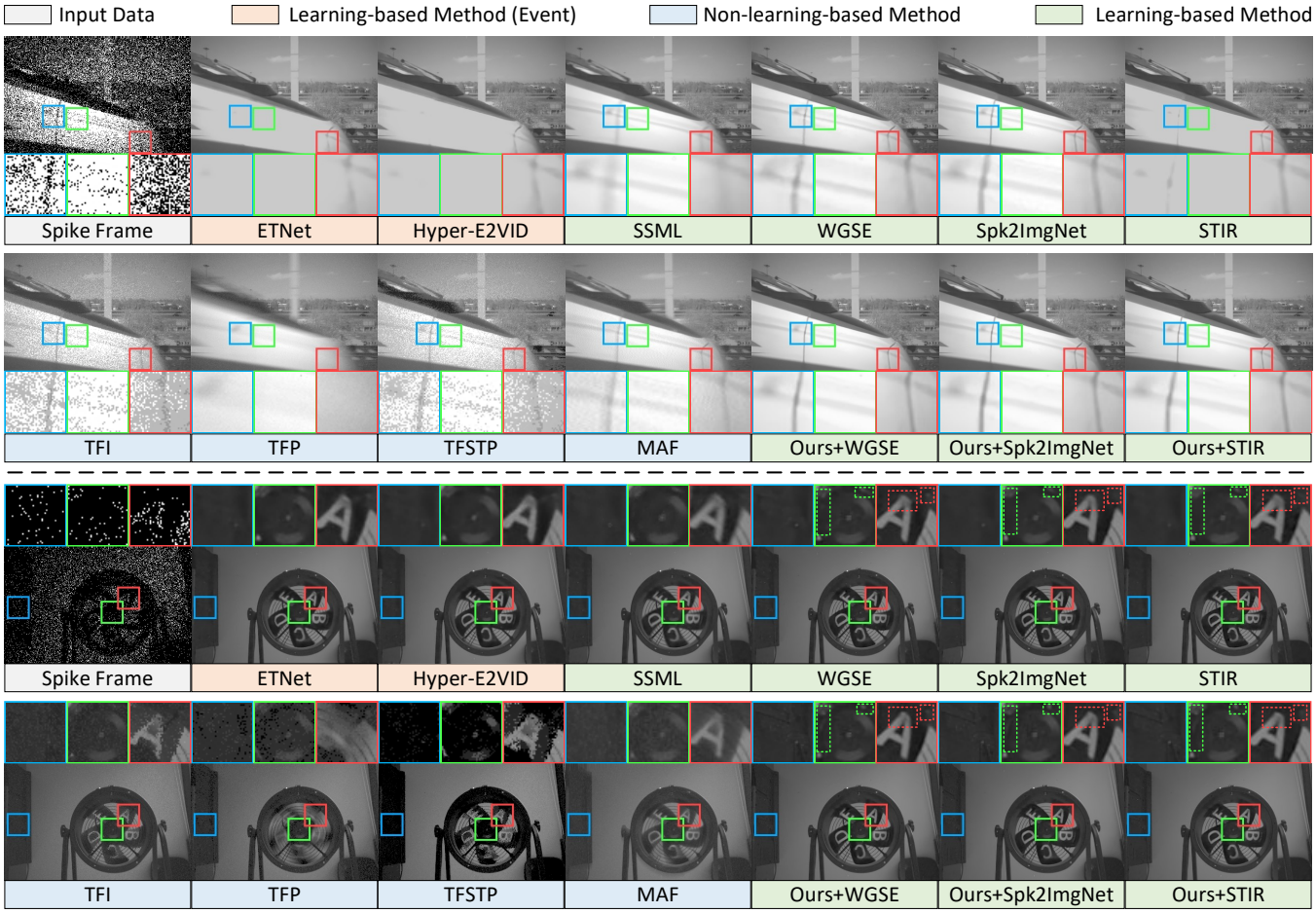

Figure 6: Visual comparison on the real-captured spike streams from Real-SPK. The first sample (bright scene) records a ***fast-driving*** train (350KM/h). The second one (dark scene) records a ***fast-rotating*** fan. Please zoom in for better visual comparison.

Case	MM	Inter.	LFH	S2F	PSNR(dB) of Ours +			
					WGSE	S2INet	STIR	*Avg.*
(A)		✓			40.14	40.23	40.11	40.16
(B)	✓	✓			40.19	40.32	40.25	40.26
(C)	✓		✓		40.24	40.26	40.29	40.27
(D)				✓	40.23	40.32	40.24	40.27
(E)	✓	✓	✓		**40.28**	**40.37**	**40.35**	**40.34**

Table 3: Ablation study for motion estimation, where MM means the motion memory, LFH means the learnable flow head, and S2F means motion estimation by Spike2Flow.

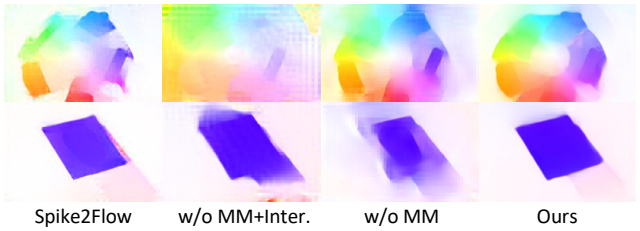

Figure 7: Visualization of ablation study for our motion estimation networks on two real-captured spike streams.

and Case (E), we can find the performance gains by LFH, without additional parameters. By comparing Case (D) and Case (E), we can verify the effectiveness of our SSMT-based motion estimation method. We also present optical flow visualization of different cases in Fig. 7.

Conclusions

Motivated by the human-like memory mechanism, we explore a framework of spike stream memory transfer to enhance image reconstruction of the retina-inspired spike cameras. By transferring previously inferred light intensity and motion, we mitigate the spike quantization effects and achieve more efficient motion estimation. Besides, we suppress the potential blur from inter- and intra-stream clips by the inferred motion and our spike interval-guided alignment module when aggregating long-term spike signals, respectively. Experiments on both synthetic and real-captured spike streams demonstrate our superior performance.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant 22127807, Grant 62461160310, and Grant 62072009, and by the Innovative Research Groups of the National Natural Science Foundation of China under Grant 62521007.

References

Chen, S.; Duan, C.; Yu, Z.; Xiong, R.; and Huang, T. 2022. Self-Supervised Mutual Learning for Dynamic Scene Reconstruction of Spiking Camera. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI)*, 2859–2866.

Chi, Y.; Gnanasambandam, A.; Koltun, V.; and Chan, S. H. 2020. Dynamic low-light imaging with quanta image sensors. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, 122–138. Springer.

Cho, K.; Van Merriënboer, B.; Bahdanau, D.; and Bengio, Y. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*.

Choi, J. H.; Elgendy, O. A.; and Chan, S. H. 2018. Image reconstruction for quanta image sensors using deep neural networks. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6543–6547. IEEE.

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

Ding, Z.; Zhao, R.; Zhang, J.; Gao, T.; Xiong, R.; Yu, Z.; and Huang, T. 2022. Spatio-temporal recurrent networks for event-based optical flow estimation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, 525–533.

Dong, Y.; Xiong, R.; Fan, X.; Yu, Z.; Tian, Y.; and Huang, T. 2025a. Self-Supervised Learning for Color Spike Camera Reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6231–6240.

Dong, Y.; Xiong, R.; Fan, X.; Zhu, S.; Wang, J.; and Huang, T. 2025b. Dynamic Scene Reconstruction for Color Spike Camera via Zero-Shot Learning. *IEEE Transactions on Computational Imaging*, 11: 129–141.

Dong, Y.; Xiong, R.; Zhang, J.; Yu, Z.; Fan, X.; Zhu, S.; and Huang, T. 2024a. Super-Resolution Reconstruction from Bayer-Pattern Spike Streams. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 24871–24880.

Dong, Y.; Xiong, R.; Zhao, J.; Fan, X.; Zhang, X.; and Huang, T. 2025c. Color Spike Camera Reconstruction via Long Short-Term Temporal Aggregation of Spike Signals. *IEEE Transactions on Image Processing*, 34: 5312–5324.

Dong, Y.; Xiong, R.; Zhao, J.; Zhang, J.; Fan, X.; Zhu, S.; and Huang, T. 2024b. Joint Demosaicing and Denoising for Spike Camera. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 1582–1590.

Dong, Y.; Xiong, R.; Zhao, J.; Zhang, J.; Fan, X.; Zhu, S.; and Huang, T. 2024c. Learning a Deep Demosaicing Network for Spike Camera With Color Filter Array. *IEEE Transactions on Image Processing*, 33: 3634–3647.

Dong, Y.; Zhao, J.; Xiong, R.; and Huang, T. 2022a. 3D Residual Interpolation for Spike Camera Demosaicing. In *2022 IEEE International Conference on Image Processing (ICIP)*, 1461–1465. IEEE.

Dong, Y.; Zhao, J.; Xiong, R.; and Huang, T. 2022b. High-speed scene reconstruction from low-light spike streams. In *2022 IEEE International Conference on Visual Communications and Image Processing (VCIP)*, 1–5. IEEE.

Dutton, N. A.; Gyongy, I.; Parmesan, L.; Gnecchi, S.; Calder, N.; Rae, B. R.; Pellegrini, S.; Grant, L. A.; and Henderson, R. K. 2015. A SPAD-based QVGA image sensor for single-photon counting and quanta imaging. *IEEE Transactions on Electron Devices*, 63(1): 189–196.

Dutton, N. A.; Parmesan, L.; Holmes, A. J.; Grant, L. A.; and Henderson, R. K. 2014. 320×240 oversampled digital single photon counting image sensor. In *2014 Symposium on VLSI Circuits Digest of Technical Papers*, 1–2. IEEE.

Ercan, B.; Eker, O.; Saglam, C.; Erdem, A.; and Erdem, E. 2024. HyperE2VID: Improving Event-Based Video Reconstruction via Hypernetworks. *IEEE Transactions on Image Processing*, 33: 1826–1837.

Fan, B.; Yin, J.; Dai, Y.; Xu, C.; Huang, T.; and Shi, B. 2024. Spatio-temporal interactive learning for efficient image reconstruction of spiking cameras. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 21401–21427.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

Hu, L.; Ding, Z.; Liu, M.; Ma, L.; and Huang, T. 2024a. Learning to Robustly Reconstruct Dynamic Scenes from Low-Light Spike Streams. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part XVII*, 88–105. Springer.

Hu, L.; Ma, L.; Guo, Y.; and Huang, T. 2024b. Scsim: a realistic spike cameras simulator. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, 1–6. IEEE.

Hu, L.; Zhao, R.; Ding, Z.; Ma, L.; Shi, B.; Xiong, R.; and Huang, T. 2022. Optical flow estimation for spiking camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 17844–17853.

Kai, D.; Lu, J.; Zhang, Y.; and Sun, X. 2024. EvTexture: Event-driven Texture Enhancement for Video Super-Resolution. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 22817–22839.

Kai, D.; Zhang, Y.; and Sun, X. 2023. Video super-resolution via event-driven temporal alignment. In *2023 IEEE International Conference on Image Processing (ICIP)*, 2950–2954. IEEE.

Kai, D.; Zhang, Y.; Wang, J.; Xiao, Z.; Xiong, Z.; and Sun, X. 2025. Event-Enhanced Blurry Video Super-Resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 4175–4183.

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*, 43: 566–576.

Nah, S.; Baik, S.; Hong, S.; Moon, G.; Son, S.; Timofte, R.; and Mu Lee, K. 2019. NTIRE 2019 challenge on video deblurring and super-resolution: Dataset and study. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*.

Nah, S.; Hyun Kim, T.; and Mu Lee, K. 2017. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3883–3891.

Pont-Tuset, J.; Perazzi, F.; Caelles, S.; Arbeláez, P.; Sorkine-Hornung, A.; and Van Gool, L. 2017. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*.

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

Teed, Z.; and Deng, J. 2020. RAFT: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, 402–419. Springer.

Weng, W.; Zhang, Y.; and Xiong, Z. 2021. Event-based video reconstruction using transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2563–2572.

Xia, L.; Ding, Z.; Zhao, R.; Zhang, J.; Ma, L.; Yu, Z.; Huang, T.; and Xiong, R. 2024. Unsupervised optical flow estimation with dynamic timing representation for spike camera. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36.

Xiang, X.; Zhu, L.; Li, J.; Wang, Y.; Huang, T.; and Tian, Y. 2021. Learning super-resolution reconstruction for high temporal resolution spike stream. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(1): 16–29.

Xiao, P.; Zhang, Y.; Kai, D.; Peng, Y.; Zhang, Z.; and Sun, X. 2024. ESTME: Event-driven Spatio-Temporal Motion Enhancement for Micro-Expression Recognition. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, 1–6. IEEE.

Zhang, J.; Jia, S.; Yu, Z.; and Huang, T. 2023. Learning temporal-ordered representation for spike streams based on discrete wavelet transforms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 137–147.

Zhao, J.; Xie, J.; Xiong, R.; Zhang, J.; Yu, Z.; and Huang, T. 2021a. Super resolve dynamic scene from continuous spike streams. In *Proceedings of the IEEE/CVF 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11996–12005.

Zhao, J.; Xiong, R.; Xie, J.; Shi, B.; Yu, Z.; Gao, W.; and Huang, T. 2021c. Reconstructing Clear Image for High-Speed Motion Scene With a Retina-Inspired Spike Camera. *IEEE Transactions on Computational Imaging*, 8: 12–27.

Zhao, J.; Xiong, R.; Zhang, J.; Zhao, R.; Liu, H.; and Huang, T. 2023a. Learning to super-resolve dynamic scenes for neuromorphic spike camera. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 3579–3587.

Zhao, R.; Xiong, R.; Zhang, J.; Yu, Z.; Zhu, S.; Ma, L.; and Huang, T. 2023b. Spike Camera Image Reconstruction Using Deep Spiking Neural Networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(6): 5207–5212.

Zhao, R.; Xiong, R.; Zhang, J.; Zhang, X.; Yu, Z.; and Huang, T. 2024a. Optical Flow for Spike Camera with Hierarchical Spatial-Temporal Spike Fusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 7496–7504.

Zhao, R.; Xiong, R.; Zhao, J.; Yu, Z.; Fan, X.; and Huang, T. 2022. Learning optical flow from continuous spike streams. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 7905–7920.

Zhao, R.; Xiong, R.; Zhao, J.; Zhang, J.; Fan, X.; Yu, Z.; and Huang, T. 2024b. Boosting Spike Camera Image Reconstruction from a Perspective of Dealing with Spike Fluctuations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 24955–24965.

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 *Proceedings of the IEEE/CVF 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 *2019 IEEE International Conference on Multimedia and Expo (ICME)*, 1432–1437. IEEE.

Zhu, Z.; Xiong, R.; Xie, J.; Wang, Y.; Zhang, X.; and Huang, T. 2025a. High Dynamic Range Imaging with Time-Encoding Spike Camera. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Zhu, Z.; Xiong, R.; Zhao, J.; Zhao, R.; Fan, X.; Zhu, S.; and Huang, T. 2025b. High Dynamic Range Imaging for Dynamic Scenes Based on Multi-Level Spike Camera. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(6): 5394–5406.