High-speed Image Reconstruction through Short-term Plasticity for Spiking Cameras

Yajing Zheng[1†], Lingxiao Zheng[1†], Zhaofei Yu[1,2*], Boxin Shi[1,2],
Yonghong Tian[1], Tiejun Huang[1,2]
[1]Department of Computer Science and Technology, Peking University
[2] Institute for Artificial Intelligence, Peking University

Abstract

Fovea, located in the centre of the retina, is specialized for high-acuity vision. Mimicking the sampling mechanism of the fovea, a retina-inspired camera, named spiking camera, is developed to record the external information with a sampling rate of 40,000 Hz, and outputs asynchronous binary spike streams. Although the temporal resolution of visual information is improved, how to reconstruct the scenes is still a challenging problem. In this paper, we present a novel high-speed image reconstruction model through the short-term plasticity (STP) mechanism of the brain. We derive the relationship between postsynaptic potential regulated by STP and the firing frequency of each pixel. By setting up the STP model at each pixel of the spiking camera, we can infer the scene radiance with the temporal regularity of the spike stream. Moreover, we show that STP can be used to distinguish the static and motion areas and further enhance the reconstruction results. The experimental results show that our methods achieve state-of-the-art performance in both image quality and computing time.

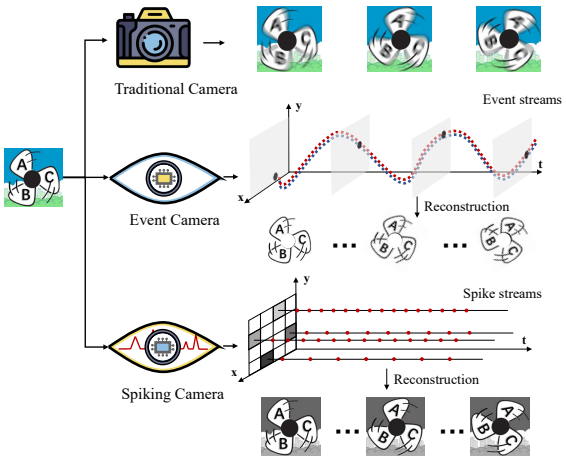

Figure 1. Illustration of working mechanism for traditional cameras, event cameras, and spiking cameras. Traditional cameras acquire images according to the constant frame rate, event cameras generate asynchronous events for all pixels when brightness changes exceed a certain threshold, and spiking cameras continuously capture photons and generate asynchronous spike for all pixels when the accumulated intensity reaches a predefined threshold.

1. Introduction

High-speed imaging is desired in scientific imaging and professional photography for clearly recording fast-changing processes in physics experiments, rapidly moving particles in chemical reactions, fleeting moment in competitive sports and so on. The traditional digital camera records scenes with a constant shutter speed (e.g., 30 fps), which loses much visual information and suffers motion blur. High-speed cameras can output images with a relatively high time sampling frequency of millions Hz or even hundreds of millions Hz [1, 13]. However, enormous memory demands are needed to record these images. Moreover, high-speed cameras require specialized sensors that are highly expensive, which cannot be widely used.

Over the past decades, neuromorphic vision sensors [12] have attracted extensive attention with its bio-inspired visual recording mechanism. Unlike conventional cameras, which sample external light with the same exposure time for all pixels, the neuromorphic vision sensors mimick the retina's sampling process and generate asynchronously binary outputs for all pixels based on the scene radiance change. A commonly used neuromorphic vision sensors are dynamic vision sensors (DVS) (also called event cameras), in which events are generated only when the brightness change exceeds a certain threshold [22, 5, 9]. Event cameras have distinctive advantages over traditional frame cameras such as low-latency, sparse output, low power consumption, and high dynamic range. Despite this, it is difficult for an event camera to reconstruct textures in scenes as visual in-

† These authors contributed equally to this work.
* Corresponding authors: yuzf12@pku.edu.cn.

formation of the static scenes is lost. Inspired by the sampling mechanism of primate fovea located in the retina center [34, 18], another retina-inspired camera named spiking camera has been developed in recent years [6, 7]. Each pixel of the spiking camera continuously captures photons and generates a spike when the accumulated intensity reaches a predefined threshold. An intuitive illustration for traditional cameras, event cameras, and spiking cameras is shown in Fig. 1. In spiking cameras, the pixels sensed different scene radiance will fire spikes with varying frequencies — the stronger the radiance is, the faster the spikes fire. Compared with event cameras, the spiking camera retains *high-speed spatio-temporal* information for both *moving and static* objects, which is ready-to-use for scene reconstruction.

Recently, some scene reconstruction methods have been proposed by estimating the firing frequency of each pixel, as the photosensitive units receive different scene radiance will trigger spikes with different frequencies [38, 39, 36]. However, these methods require a predefined length of the time window, which often suffer the problems of motion blur and low image contrast if the window length is inappropriate. Besides, complex optimization algorithms are utilized to separate the motion and static areas as to make it impossible to reconstruct images in real-time. Therefore, how to estimate each pixel's firing frequency without a predefined time window and reconstruct the texture with high image quality and low latency is still unclear.

In this paper, we develop a new high-speed image reconstruction approach through the short-term plasticity (STP) mechanism of the brain [31, 30]. By employing the output spiking streams as the input of spiking neural networks with STP [17], we derive the relationship between the time-varying firing frequency of each pixel and the dynamics of the postsynaptic neuron, and further infer the scene radiance and the pixel value of the reconstructed images. Moreover, as the dynamics of STP model will fluctuate around a stable value if the spike frequency changes, we introduce a motion extraction method with STP to enhance the reconstruction results. The experimental results show that the proposed methods are capable to reconstruct high-speed motion scenes with high quality in real-time, the performance of which outperform state-of-the-art approaches.

Contributions:

- We propose bio-inspired image reconstruction methods by implicitly estimating each pixel's firing frequency rather than utilizing time windows.
- We propose a novel motion extraction method based on the STP dynamics intuitively, without using much extra computation while rebuilding scenes.
- The proposed methods outperform previous works in both image quality and computing time, which leverage spiking cameras' low latency to realize high-speed image reconstruction.

2. Related Works

Event-based Imaging. It is different for event cameras to reconstruct textures in scenes as visual information of the static scenes is lost. Some hybrid sensors combining event cameras and conventional digital cameras, such as ATIS [24], DAVIS [2], RGB-DAVIS imaging system [33, 11] and Celex [10], were developed in recent years. Based on these sensors, the scene could be reconstructed with high frame rate by combining events and frames directly [3, 21, 20, 28, 10] or warping the events to images [15, 29]. However, the difficulty in achieving reliable temporal synchronization between events and low-rate frames from traditional sensors makes them inapplicable in capturing the high-speed scene.

In recent years, generating high-speed and high dynamic range videos with event cameras based on deep neural networks (DNNs) has become a mainstream trend. Inspired by [37, 27], Rebecq et al. [25, 26] trained a recurrent UNet architecture (E2VID) end-to-end with simulated data. These works are later improved by introducing a temporal consistency loss based on [14] and achieve the state of the art. Scheerlinck et al. [28] proposed a *FireNet* reducing the E2VID model complexity by 99% with minor trade-offs in reconstruction quality. Except for using the recurrent architecture, generative adversarial networks (GANs) were used in [23, 32] to generate frames from events. Nevertheless, the computational cost of DNNs is high and does not leverage the low-power and the low-latency of event cameras.

High-speed Imaging based on Spiking Cameras. Based on the temporal characteristic of spike streams generated by spiking cameras, some reconstruction methods have been proposed [38, 39, 36]. Zhu et al. [38] presented "texture from inter-spike-intervals (TFI)" and "texture from playback (TFP)" to rebuild the scenes according to the firing interval and firing rate, respectively. As there is a trade-off between removing the motion blur and improving the image contrast, the length of the window needs to be carefully defined, which will significantly influence the results. To solve this problem, Zhu et al. [39] proposed to extract the motion area and reconstruct the static and motion area with different methods. Nevertheless, the motion extraction based on the graph cut needs to optimize the motion mask iteratively. Such an energy-based optimized way is time-consuming that diminishes the advantage of the low latency of spiking cameras. Zhao et al. [36] improved the signal to noise ratio by utilizing temporal correlations of signals, but it only applied to the scenes with linear motion.

3. Preliminaries

3.1. Short-term Plasticity

Short-term plasticity refers to the short-term change of synaptic strength, which is usually between tens to thou-

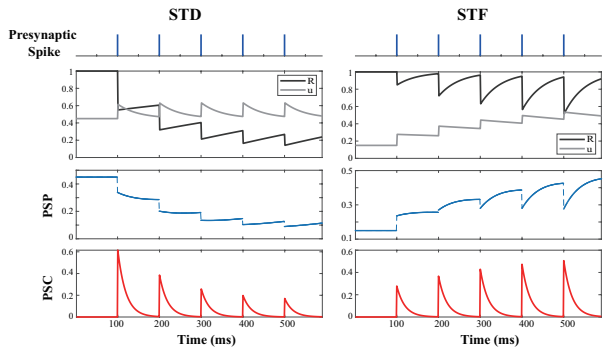

Figure 2. The postsynaptic potential and current generated by STP with received spike streams from a presynaptic neuron. Left: The short-term depression dominated model, the parameters are $\tau_D = 750\ ms,\ \tau_F = 50\ ms,\ U = 0.45,\ C = 0.3$. Right: The short-term facilitation dominate model, the parameters are $\tau_D = 50\ ms,\ \tau_F = 750\ ms,\ U = 0.15,\ C = 0.15$.

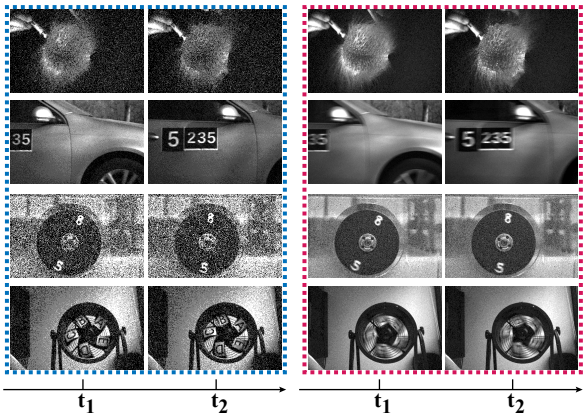

Figure 3. Reconstruction results of TFP with different length of time window. Images in the blue dotted box are recovered with $w = 8$, and images in the red dotted are recovered with $w = 32$.

sands of milliseconds. STP is sensitive to presynaptic spikes' temporal-distribution regularity and can transiently change postsynaptic potential (PSP) amplitude accordingly. When a postsynaptic neuron receives a sequence of action potentials (spikes) from a presynaptic neuron, the PSP changes according to:

$$\mathrm{PSP}(t) = A \cdot R(t) \cdot u(t), \tag{1}$$

where A denotes the maximum voltage value that an action potential can trigger on a postsynaptic neuron, $R(t)$ denotes the remaining number of available neurotransmitters in the axon at time t, and $u(t)$ denotes the release probability of neurotransmitter in the axon at time t. The following ordinary differential equations define the dynamics of $R(t)$ and $u(t)$:

$$\frac{dR(t)}{dt} = \frac{1 - R(t)}{\tau_D} - u(t^-)R(t^-)\delta(t - t_{sp}), \tag{2}$$

$$\frac{du(t)}{dt} = \frac{U - u(t)}{\tau_F} + C[1 - u(t^-)]\delta(t - t_{sp}). \tag{3}$$

Here $\delta(t)$ represents Dirac delta function, C is a constant parameter that influences the change of $u(t)$. Eq. 2 illustrates that the amount of neurotransmitters $R(t)$ decreases by $u(t^-)R(t^-)$ when a presynaptic spike releases at time t_{sp}, and recovers to 1 with a depression time constant τ_D. Note, the notation t^- denotes that these functions should be computed in the limit approaching the spike release time from below. Eq. 3 indicates that the release probability $u(t)$ increases by $C[1 - u(t^-)]$ once a presynaptic spike fires, and decays back to baseline release probability U with facilitation time constant τ_F. Similar to PSP, the postsynaptic current is formulated by:

$$\frac{d\mathrm{PSC}(t)}{dt} = -\frac{\mathrm{PSC}(t)}{\tau_s} + A \cdot R(t^-) \cdot u(t) \cdot \delta(t - t_{sp}). \tag{4}$$

Intuitively, the dynamics of $R(t)$ and $u(t)$ (Eq. 2 and Eq. 3) can be seen as two low-pass filters of the input spikes, and their cutoff frequencies are inversely proportional to time constants τ_D and τ_F. There are two types of STP named short-term depression and short-term facilitation, respectively. Short-term depression and short-term facilitation have opposite effects on synaptic efficacy, which can be seen in the middle and bottom of Fig. 2. Through adjusting the four parameters $\mathrm{STP}^{\boldsymbol{\theta}} = \{\tau_D, \tau_F, U, C\}$, STP can have forms being either short-term depression dominated or short-term facilitation dominated.

4. Methods

4.1. Overview of the method

The previous works to reconstruct the scene can be summarized as estimating the firing frequency of each pixel [36, 38, 39]. Fig. 3 illustrates the reconstruction results of TFP [38] with different time windows. One can find that a short time window leads to lower contrast and less motion blur, while the long one has higher contrast and more motion blur. Thus, it requires an appropriate predefined length of the time window to estimate the firing frequency accurately, so as to make the texture relatively high contrast and avoid the motion blur.

To mitigate the weakness of the setting of the time window, we set up the STP model at each pixel of the spiking cameras to record the temporal regularity of spikes implicitly. The dynamics of PSP regulated by STP is shown in Fig. 4a. We can find that PSP will converge to a steady value if the firing frequency of the input spike streams is fixed, no matter what type of STP is used.

Moreover, the steady value of PSP, the number of vesicles R, and the release probability u are all monotone increasing functions of firing frequency (shown in Fig. 4b). As mentioned above, the firing frequency of the spike

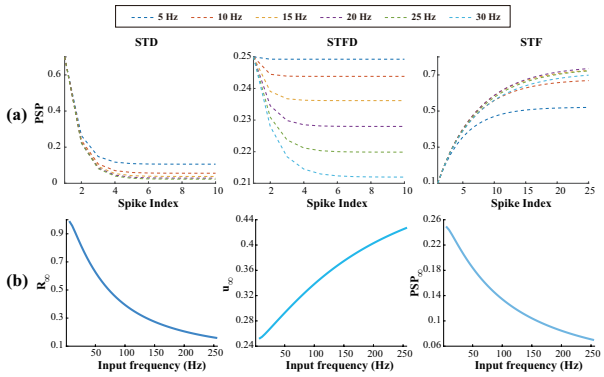

Figure 4. **(a):** The dynamic of PSP regulated by STP. The dotted lines with different colors refer to spike trains with different frequencies, from 5 Hz to 30 Hz. The short-term facilitation and depression model has a mixture of properties of both short-term depression and short-term facilitation. **(b):** the steady value of PSP, the number of neurotransmitters R, and the release probability u with respect to different input frequencies. The results are obtained with a short-term facilitation and depression model.

streams triggered by each photosensitive unit is proportional to the received scene radiance. Intuitively, we can estimate the scene radiance and the pixel value of the reconstructed images, as well as detect the motion area based on PSP. The details of our approach are presented in the following sections.

4.2. Texture Reconstruction through STP

By setting up the STP model at each pixel of the spiking cameras to record the output spike stream, we derive the equation between the time-varying firing frequency of each pixel and the dynamics of postsynaptic neuron. For the sake of derivation, and numerical implementation with simplicity and efficiency, the dynamics of $R(t)$ and $u(t)$ (Eq. 2 and Eq. 3) can be rewritten as the following difference equations by integrating between spikes n and $n+1$:

$$R_{n+1} = 1 - [1 - R_n(1-u_n)] \exp\left(-\frac{\Delta t_n}{\tau_D}\right), \tag{5}$$

$$u_{n+1} = U + [u_n + C(1-u_n) - U] \exp\left(-\frac{\Delta t_n}{\tau_F}\right), \tag{6}$$

where R_n and u_n denote the value of R and u between spikes n and $n+1$, Δt_n denotes the interval between spikes n and $n+1$. Similar to [4], we set $C = U$. If the spike rate ρ keeps constant, R and u will converge to their steady-state values $R_\infty(\rho)$ and $u_\infty(\rho)$:

$$R_\infty(\rho) = \frac{1 - \exp(-\frac{1}{\rho\tau_D})}{1 - [1 - u_\infty(\rho)]\exp(-\frac{1}{\rho\tau_D})}, \tag{7}$$

$$u_\infty(\rho) = \frac{U + (C-U)\exp(-\frac{1}{\rho\tau_F})}{1 - (1-C)\exp(-\frac{1}{\rho\tau_F})}. \tag{8}$$

Conversely, assuming that the spike rate ρ keeps constant and R and u have already converged to their steady-state values, we can estimate ρ from R and u separately through Eq. (7) and Eq. (8):

$$\rho_R = -\frac{1}{\tau_D \ln\left(\frac{1-R}{1-R(1-u)}\right)}, \tag{9}$$

$$\rho_u = -\frac{1}{\tau_F \ln\left(\frac{u-U}{C-U+u(1-C)}\right)}. \tag{10}$$

As the firing frequency of each pixel is proportional to the scene radiance, the estimated pixel value is a weighted average of ρ_R and ρ_u:

$$\hat{P}_{stp} \propto w_1 \cdot \rho_R + w_2 \cdot \rho_u. \tag{11}$$

By varying the weighted parameter $\boldsymbol{w} = \{w_1, w_2\}$, we can control the contribution of ρ_R and ρ_u to the constructed image. Here the parameters are set as $\text{STP}^{\boldsymbol{\theta}1}$: $\{\tau_D = 0.025\ ms, \tau_F = 0.25\ ms, C = U = 0.15\}$. The steps of this methods (texture from short-term plasticity, **TFSTP**) are summarized in Algorithm. 1.

Algorithm 1 Texture from STP (TFSTP)

Input: Spike streams $\mathcal{S}_{ij}$.
Output: Estimated pixel value $\hat{P}_{ij}$.

1: Initialize the parameters of STP, $\{\tau_D, \tau_F, U, C\}$, R and u, and the weight parameter w
2: Compute Δt_n using the spike data.
3: Update R_{n+1} and u_{n+1} using Eq. 5 and Eq. 6.
4: Estimated the firing frequency ρ_R and ρ_u using Eq. 9 and Eq. 10.
5: Estimate the pixel value using Eq. 11.

Fig. 5 compares the STP dynamics $\{\rho_R, \rho_u, R, u\}$ for dark and bright area with different scene radiance. In the dark area without moving objects (green circle in Fig. 5b), the spikes generated by a spiking camera have a nearly constant frequency. Hence the corresponding STP dynamics will converge to the steady value rapidly. When the car in the scenes moves across the bright areas (red circle in Fig. 5b) between 0 - 3.5 ms, there are some small-range fluctuations of the STP dynamics, but the overall trend is still converging towards the corresponding state of the bright place.

4.3. Brightness Change Detection

For most scenes, the TFSTP method proposed in Section 4.2 works pretty well. Nevertheless, for high-speed scenes with limited illumination, it may suffer from motion blur caused by rare spikes in the dark area, and the STP status is not updated in time. To solve this problem, we proposed another texture reconstruction method (texture from motion-based short-term plasticity, **TFMSTP**) based on detecting

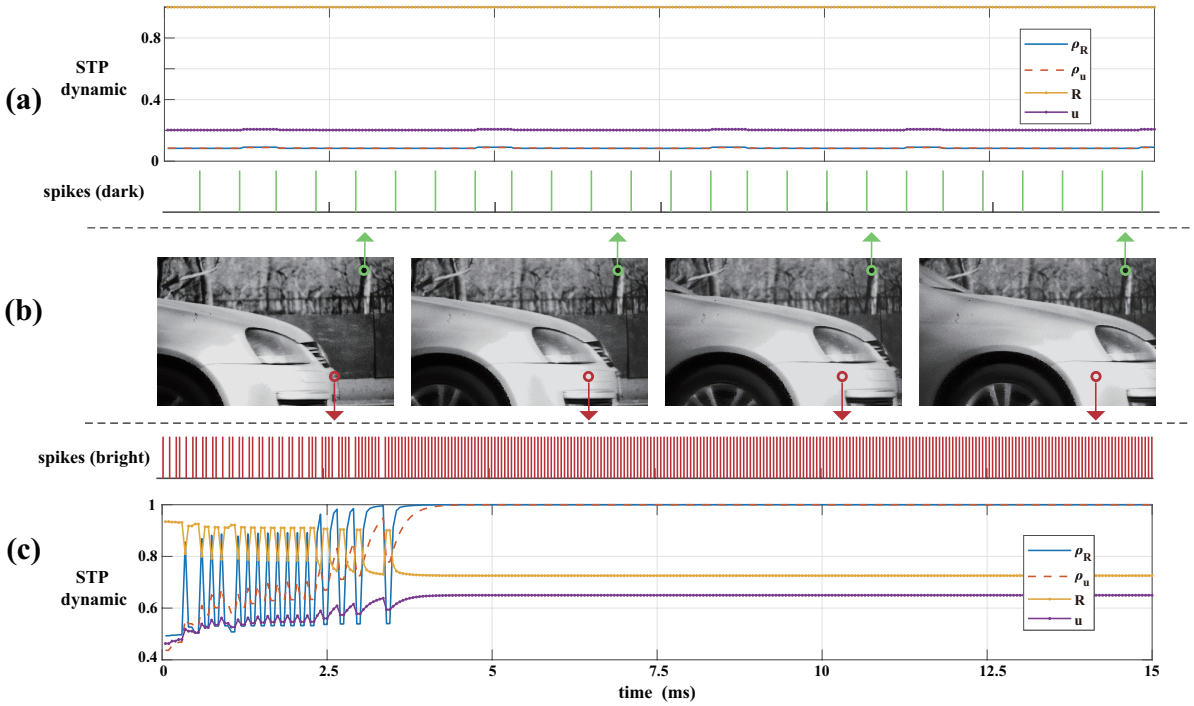

Figure 5. Difference between STP dynamics of dark and bright area. **(a):** Spike raster and the corresponding STP dynamics of dark area (green circle in **(b)**). **(b):** Scenes reconstruction via Algorithm 1. **(c):** Spike raster and STP dynamics of bright area (red circle in **(b)**).

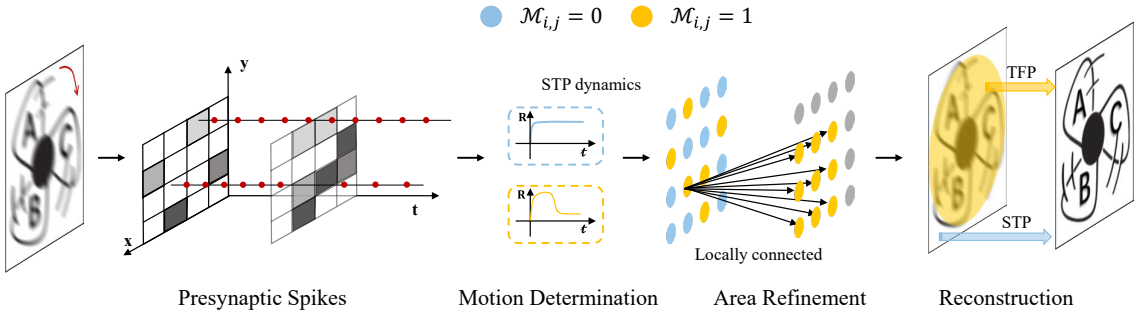

Figure 6. Illustration of the reconstruction process based on distinguishing the motion and static pixel through STP.

the motion area with STP. Specifically, we first detect the brightness change to extract the motion area and then reconstruct the motion and static areas via TFP and STP, separately.

Motion Determination. When there exists motion in the area, the STP value will vary around the steady value corresponding to the scene radiance (Fig. 5). Thus, we are able to detect the motion area by evaluating the STP dynamics, e.g. R, and u, which updates according to Eq. (5) and Eq. (6). The reconstruction process of TFMSTP is shown in Fig. 6, it begins with motion determination via STP. Since the number of vesicles R fluctuates with a larger range than the lease probability u in the short-term facilitation dominated model (shown in Fig. 2b), we detect pixels belonging to the motion area or not by using the difference between R_n and R_{n-1}:

$$\mathcal{M}_{i,j} = \begin{cases} 1, |R_n(i,j) - R_{n-1}(i,j)| \geq \theta \\ 0, |R_n(i,j) - R_{n-1}(i,j)| < \theta \end{cases}, \tag{12}$$

where $\mathcal{M}_{i,j}$ denotes whether pixel(i,j) belongs to the motion area, and θ is a predefined global constant.

Area Refinement. Except for finding the motion pixels, places along the moving trajectory of objects are also regarded as motion areas in our methods, which is achieved by feeding the $\mathcal{M}_{i,j}$ as the input voltage to a locally connected network consisted of leaky integrate-and-fire (LIF) neurons. The membrane potential $v(t)$ of LIF neuron

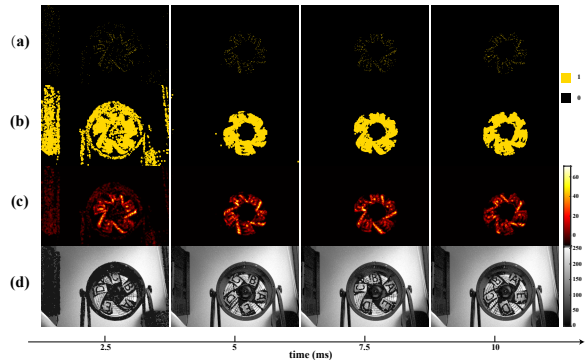

Figure 7. Example results of the Algorithm 2. **(a):** Motion mask $\mathcal{M}$ mentioned in Eq. 12. **(b):** The refined motion area χ obtained by Eq. 14. Yellow dots refer to states equal to 1, e.g. $\chi_{i,j} = 1$ or $\mathcal{M}_{i,j} = 1$, while the black ones refer to zero. **(c):** The voltage v of LIF neurons corresponding to the state of **(b)** (Eq. 13). **(d):** The estimated pixel value $\hat{P}'$.

changes according to:

$$\tau_m \frac{dv(t)}{dt} = -[v(t) - v_{rest}] + I_{i,j}(t), \tag{13}$$

where τ_m is the membrane time constant, and v_{rest} is the resting potential. The current $I_{i,j}$ of the neuron is the integrated result of neurons that locally connect to it, which is calculated as $I_{i,j} = \sum_{\boldsymbol{x}} \mathcal{M}_{\boldsymbol{x}}$ ($\boldsymbol{x}$ is the location of neurons connected to the LIF neuron at (i, j)). In our cases, τ_m equals the sampling frequency (25 μs) of the spiking cameras. The LIF neuron will release spikes when the membrane potential exceeds a certain threshold ϑ, and the membrane potential is reset to the resting potential. The state of the LIF neuron is changed as:

$$\chi_{i,j} = \begin{cases} 1, v \geq \vartheta \\ 0, v < \vartheta \end{cases} \tag{14}$$

After that, areas with $\chi = 0$ are regarded as static pixels while the ones with $\chi = 1$ refer to motion pixels. For static pixels, we estimate the pixel value using Eq. 11 derived by STP, and for motion pixels (i.e. $\chi_{i,j} = 1$), intensity of pixels are obtained by TFP with a small moving window $w = 8$. Therefore, we can achieve more texture details, high dynamic range, and low noise for the static area as well as low motion blur for the motion area. The proposed TFMSTP approach is summarized in Algorithm. 2. In Fig. 7, we show some example results obtained when detecting the brightness change with our proposed method. With continuous input of spike streams, the STP dynamic of each pixel gradually converges to a steady state, and only the motion area has state change (i.e. $\mathcal{M}_{i,j} = 1$ in Fig. 7a). To detect the motion area in time, the STP parameters, in this case, are set as $\text{STP}^{\boldsymbol{\theta}2} : \{\tau_D = \tau_F = 10ms, C = U = 0.15\}$. The segmented motion region have a clearer and smoother contour after **area refinement**. (Fig. 7b and Fig. 7c).

Algorithm 2 TFMSTP

Input: Spike streams $\mathcal{S}_{i,j}$.
Output: pixel value $\hat{P}'_{i,j}$, motion state $\chi_{i,j}$.

1: Initialize the parameters of STP, $\{\tau_D, \tau_F, U, C\}$, R and u, and the weighted parameter w
2: Compute Δt_n using the spike data.
3: Update $R(t)$ and $u(t)$ using Eq. 5 and Eq. 6.
4: Obtain $\mathcal{M}_{i,j}$ using Eq. 12.
5: Refine the area and get $\chi_{i,j}$ with Eq. 13 and Eq. 14.
6: Estimate pixel value $\hat{P}'_{\chi=1}$ with Algorithm 1.
7: Compute $\hat{P}'_{\chi=0}$ with TFP.

5. Evaluation

5.1. Datasets

To test the proposed image reconstruction algorithms, we use a publicly available dataset, including spike sequences captured by the spike camera with a sampling rate of 40,000 Hz, which is also used in [39]. This dataset contains eight sequences, which can be divided into two categories: high-speed scenes with the object's motion (Class A) and high-speed scenes with camera's ego-motion (Class B). Class A includes "Balloon", "Car", "Rotation1" and "Rotation2". Among them, "Balloon" records a balloon filled with water being punctured by a needle, "Car" describes a car traveling at a speed of 100 km/h, "Rotation1" describes a disk with 2000 rpm (revolutions per minute), and "Rotation2" depicts a rotating fan with 2600 rpm. Class B includes "Forest", "Railway", "Train" and "Viaduct-bridge"(V-b). These four sequences are recorded by a spiking camera in a high railway with a speed of 350 km/h.

5.2. Qualitative Measurement

As shown in Fig. 8, our method is less noisy than other methods, while maintaining low motion blur. For the scenes of high-speed rotation characters without enough scene radiance, there still exists motion blur on the edges of rotating characters (Fig. 9) with TFI, Graph-based method [39], OF-based [36] and the proposed method TFSTP. For TFMSTP method, the motion blur can be effectively eliminated through reconstruct the static and motion pixels separately. Moreover, except the TFMSTP, all the other recovered results on the back of a monitor are suffer from the low contrast, which is shown in the first row of Fig. 9. As the parameters ($\mathcal{STP}^{\boldsymbol{\theta}} : \{\tau_D, \tau_F, U, C\}$) setting are different in TFSTP and TFMSTP, the corresponding STP dynamics converged with different speed. When using the parameters $\text{STP}^{\boldsymbol{\theta}2}$ (Section. 4.3), the STP dynamics of each pixel converges to a steady state quickly, bringing about reconstructed results on the static region with higher contrast (Fig. 10) using a few timestamp. The reconstructed videos

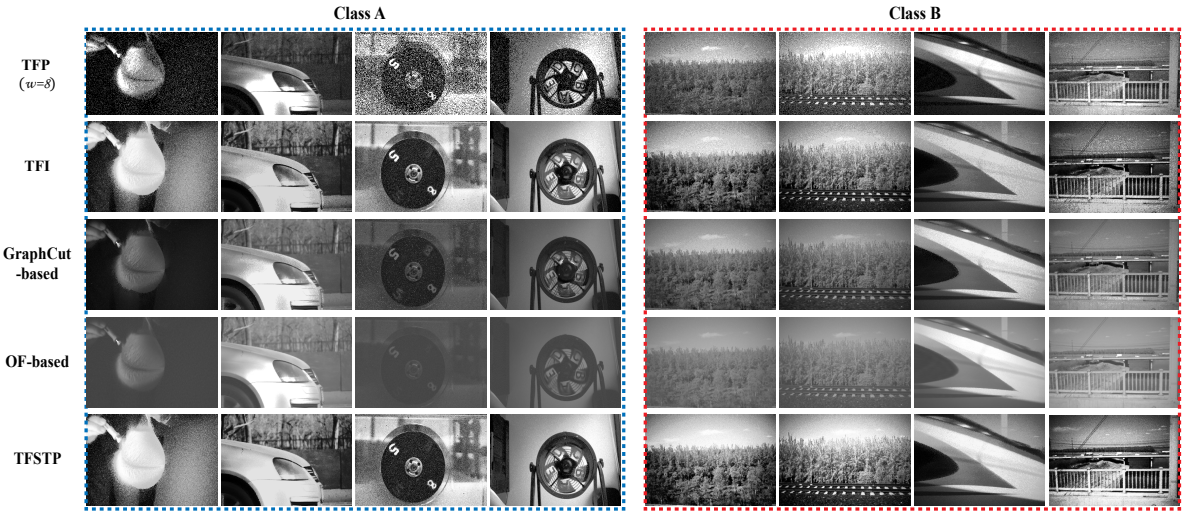

Figure 8. Reconstruction results of TFP ($w = 8$), TFI, GraphCut-based [39], OF-based [36], TFSTP (Ours).

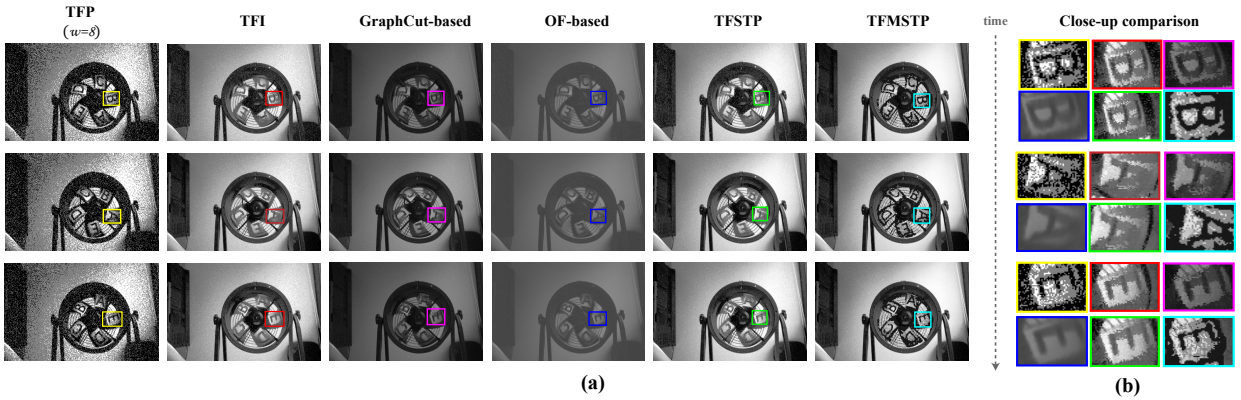

Figure 9. Comparison among different reconstruction methods on the scenes of "rotation2". **(b):** Closeups of the reconstructed results on character "B", "A" and "E" at different moment (1 ms, 4.5 ms, 11 ms), respectively.

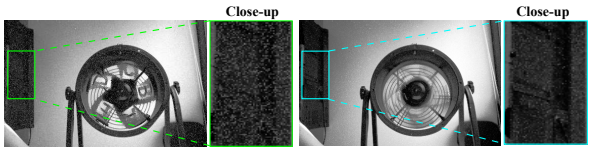

Figure 10. Comparison between reconstruction results through TFSTP with different parameters setting. The left image is reconstructed with the first set of STP parameters $\mathrm{STP}^{\boldsymbol{\theta}1}$, and the right one with $\mathrm{STP}^{\boldsymbol{\theta}2}$.

Metric	TFP(w=8)	TFI	GraphCut-based	OF-based	TFSTP
PSNR	19.96	16.92	19.03	16.46	**23.15**
SSIM	0.3776	0.6125	**0.7445**	0.7076	0.7300

Table 1. Referenced quantitative evaluation using simulated data

of all methods can be found in supplementary material.

5.3. Quantitative Evaluation

For quantitative evaluation, we first apply the reconstruction methods on a simulated dataset (supplementary material), and compare the TFP, TFI, GraphCut-based, OF-based, and TFSTP methods against the ground truth. As the synthesized dataset is generated only with the camera's ego-motion, it is unnecessary to distinguish the static and motion areas. Thus the TFMSTP method is not compared here. The results are reported in Tab. 1. The proposed TFMSP method achieves the best results in PSNR, and gets comparable results as GraphCut-based method in SSIM. The reconstruction results of our method are less noisy than GraphCut-based method (supplementary material).

Furthermore, to verify the effectiveness of the proposed methods on real-world data, we employ three no-reference image quality assessment metrics, namely two-dimensional (2-D) entropy [35], standard deviation and BIQI [19]. 2-D entropy uses both the gray value of a pixel and its local average gray value to evaluate the amount of information

Metric	Method	Class A (Object motion)				Class B (Camera's ego motion)				Average
		Balloon	Car	Rotation1	Rotation2	Forest	Railway	Train	V-b	
2-D entropy ↑	TFP [38]	3.88	5.53	4.00	4.80	5.30	5.33	6.03	6.01	5.11
	TFI [38]	12.05	7.50	11.38	10.21	8.48	9.02	8.24	7.49	9.30
	GraphCut-based [39]	9.75	9.58	11.10	11.05	11.35	11.29	10.60	10.46	10.65
	OF-based [36]	7.11	9.64	7.33	8.43	9.96	10.43	9.79	10.53	9.15
	TFSTP (ours)	**13.45**	**13.45**	**13.82**	**12.91**	**13.44**	**13.71**	**12.86**	**13.64**	**13.41**
	TFMSTP (ours)	13.01	12.01	13.12	12.48	12.62	12.86	11.28	12.89	12.53
standard deviation ↑	TFP [38]	46.41	46.09	**127.45**	**87.39**	51.77	68.45	59.82	56.99	68.04
	TFI [38]	72.91	68.99	71.03	69.31	68.44	66.05	66.38	63.72	68.35
	GraphCut-based [39]	31.83	58.34	27.04	37.74	46.94	37.82	66.49	36.31	42.81
	OF-based [36]	17.94	47.01	7.05	20.36	23.04	24.48	46.44	27.70	26.75
	TFSTP (ours)	73.94	74.18	72.91	73.15	**73.69**	73.44	73.60	73.52	73.55
	TFMSTP (ours)	**74.39**	**74.84**	74.29	77.23	73.31	**74.56**	**77.16**	**74.01**	**74.97**
BIQI ↓ (the lower, the better)	TFP [38]	61.61	53.76	75.93	85.61	62.19	66.17	58.23	55.11	64.83
	TFI [38]	71.33	61.27	81.64	39.33	83.97	78.85	64.48	66.13	68.38
	GraphCut-based [39]	**37.26**	28.40	52.40	23.89	45.51	36.62	34.30	29.46	35.98
	OF-based [36]	38.14	32.95	69.06	36.29	52.13	51.08	38.49	31.54	43.71
	TFSTP (ours)	53.01	41.30	54.80	**17.68**	**28.28**	28.75	26.52	39.97	36.29
	TFMSTP (ours)	45.07	**25.82**	**46.71**	17.83	32.81	**22.30**	**24.73**	**29.70**	**30.62**

Table 2. Comparison among different reconstruction methods.

carried by the image, larger 2-D entropy means more information. Standard deviation evaluates the contrast of the image, and larger standard deviation means higher contrast. BIQI considers JPEG quality, JP2K quality, noise, motion blur and fast fading of the image. Different from the former two metrics, a lower BIQI score indicates higher image quality.

As shown in Tab. 2, our methods achieve better results than other methods in almost all three metrics, which is consistent with subjective observation in Fig. 8 and Fig. 9. Note that TFP shows abnormally high standard deviation on dataset "Rotation1" and "Rotation2", which is caused by the high noise in the reconstruction image (see the first row of Fig. 8). By comparing our two methods, we find that TFSTP performs slightly better in 2-D entropy while TFMSTP is slightly better in standard deviation and BIQI.

5.4. Computational Complexity

Here we evaluate the computational complexity of our methods. For comparison, we consider the problem of reconstructing a K-frame video with a size $H \times W$. For each pixel, the TFSTP method only needs to update R, u, ρ_R, ρ_u when a spike generates at that pixel. Therefore, it needs at most K updates for each pixel. The time complexity of the TFSTP method is $O(HWK)$. Note that writing a $H \times W \times K$ video into the memory also takes $O(HWK)$ time, so our reconstruction method has reached the minimum asymptotic time complexity in theory.

Different to TFSTP, the TFMSTP method needs extra steps to determine whether a pixel belongs to the motion area or not with Eq. 12. However, this only takes constant time for each pixel. It does not affect the asymptotic time complexity.

In comparison, the GraphCut-based method in [39] takes at least $O(H^3W^3)$ time to implement graph cut for each frame, thus it takes at least $O(H^3W^3K)$ time to reconstruct a K-frame video. The OF-based method in [36] takes $O(H \cdot W \cdot K \cdot T \cdot iter)$ time to reconstruct a K-frame video. Therefore, our methods achieve a significantly lower time complexity than other methods.

6. Conclusions

In this paper, we propose a novel bio-inspired image reconstruction method for spiking cameras. Distinct from previous works, the spatiotemporal statics of spike streams is recorded implicitly by the short-term plasticity. By analyzing the dynamic characteristic of STP, we are able to infer the scene radiance and the pixel value of the reconstructed images. Besides, we show that STP can also be used to exact motion area to enhance reconstruction results. The theoretical analysis and experimental results show that our methods can reconstruct high-quality images with low computational complexity. Moreover, as the spiking cameras are one of the time-to-saturate cameras [8], our reconstruction methods can also be applied to other cameras developed based on similar principles (e.g. SPAD [16]).

7. Acknowledgement

This work is supported by National Natural Science Foundation of China (62027804, 62088102, 61825101, 61961130392), and Beijing Academy of Artificial Intelligence (BAAI). We thank Lin Zhu and Jing Zhao for providing the code and data.

T denotes the size of the time window used in their method, and $iter$ denotes the number of iteration in computing the optical flow.

References

[1] DK Bradley, PM Bell, OL Landen, JD Kilkenny, and J Oertel. Development and characterization of a pair of 30–40 ps x-ray framing cameras. *Review of Scientific Instruments*, 66(1):716–718, 1995. 1

[2] Christian Brandli, Raphael Berner, Minhao Yang, Shih-Chii Liu, and Tobi Delbruck. A 240× 180 130 db 3 μs latency global shutter spatiotemporal vision sensor. *IEEE Journal of Solid-State Circuits*, 49(10):2333–2341, 2014. 2

[3] Christian Brandli, Lorenz Muller, and Tobi Delbruck. Real-time, high-speed video decompression using a frame-and event-based davis sensor. In *2014 IEEE International Symposium on Circuits and Systems*, 2014. 2

[4] Rui P Costa, P Jesper Sjostrom, and Mark CW Van Rossum. Probabilistic inference of short-term synaptic plasticity in neocortical microcircuits. *Frontiers in Computational Neuroscience*, 7:75, 2013. 4

[5] Tobi Delbrück, Bernabe Linares-Barranco, Eugenio Culurciello, and Christoph Posch. Activity-driven, event-based vision sensors. In *Proceedings of 2010 IEEE International Symposium on Circuits and Systems*, 2010. 1

[6] Siwei Dong, Tiejun Huang, and Yonghong Tian. Spike camera and its coding methods. In *Data Compression Conference*, 2017. 2

[7] Siwei Dong, Lin Zhu, Daoyuan Xu, Yonghong Tian, and Tiejun Huang. An efficient coding method for spike camera using inter-spike intervals. In *Data Compression Conference*, 2019. 2

[8] Abbas El Gamal. High dynamic range image sensors. In *Tutorial at International Solid-State Circuits Conference*, volume 290, 2002. 8

[9] 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

[10] Menghan Guo, Jing Huang, and Shoushun Chen. Live demonstration: A 768× 640 pixels 200meps dynamic vision sensor. In *IEEE International Symposium on Circuits and Systems*, 2017. 2

[11] Jin Han, Chu Zhou, Peiqi Duan, Yehui Tang, Chang Xu, Chao Xu, Tiejun Huang, and Boxin Shi. Neuromorphic camera guided high dynamic range imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 2

[12] Giacomo Indiveri and Rodney Douglas. Neuromorphic vision sensors. *Science*, 288(5469):1189–1190, 2000. 1

[13] J Itatani, F Quéré, Gennady L Yudin, M Yu Ivanov, Ferenc Krausz, and Paul B Corkum. Attosecond streak camera. *Physical Review Letters*, 88(17):173903, 2002. 1

[14] Wei-Sheng Lai, Jia-Bin Huang, Oliver Wang, Eli Shechtman, Ersin Yumer, and Ming-Hsuan Yang. Learning blind video temporal consistency. In *Proceedings of the European Conference on Computer Vision*, 2018. 2

[15] Han-Chao Liu, Fang-Lue Zhang, David Marshall, Luping Shi, and Shi-Min Hu. High-speed video generation with an event camera. *The Visual Computer*, 33(6-8):749–759, 2017. 2

[16] Sizhuo Ma, Shantanu Gupta, Arin C Ulku, Claudio Bruschini, Edoardo Charbon, and Mohit Gupta. Quanta burst photography. *ACM Transactions on Graphics (TOG)*, 39(4):79–1, 2020. 8

[17] Wolfgang Maass. Networks of spiking neurons: The third generation of neural network models. *Neural Networks*, 10(9):1659–1671, 1997. 2

[18] Richard H Masland. The neuronal organization of the retina. *Neuron*, 76(2):266–280, 2012. 2

[19] AK Moorthy and AC Bovik. A modular framework for constructing blind universal quality indices. *IEEE Signal Processing Letters*, 17, 2009. 7

[20] Liyuan Pan, Richard Hartley, Cedric Scheerlinck, Miaomiao Liu, Xin Yu, and Yuchao Dai. High frame rate video reconstruction based on an event camera. *arXiv preprint arXiv:1903.06531*, 2019. 2

[21] Liyuan Pan, Cedric Scheerlinck, Xin Yu, Richard Hartley, Miaomiao Liu, and Yuchao Dai. Bringing a blurry frame alive at high frame-rate with an event camera. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2

[22] Lichtsteiner Patrick, Christoph Posch, and Tobi Delbruck. A 128x 128 120 db 15μ s latency asynchronous temporal contrast vision sensor. *IEEE Journal of Solid-state Circuits*, 43:566–576, 2008. 1

[23] Stefano Pini, Guido Borghi, and Roberto Vezzani. Learn to see by events: Color frame synthesis from event and rgb cameras. In *International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 2020. 2

[24] Christoph Posch, Daniel Matolin, and Rainer Wohlgenannt. An asynchronous time-based image sensor. In *2008 IEEE International Symposium on Circuits and Systems*, 2008. 2

[25] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. Events-to-video: Bringing modern computer vision to event cameras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2

[26] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. High speed and high dynamic range video with an event camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 2

[27] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-assisted Intervention*, 2015. 2

[28] Cedric Scheerlinck, Nick Barnes, and Robert Mahony. Continuous-time intensity estimation using event cameras. In *Asian Conference on Computer Vision*, 2018. 2

[29] Prasan Shedligeri and Kaushik Mitra. Photorealistic image reconstruction from hybrid intensity and event-based sensor. *Journal of Electronic Imaging*, 28(6):063012, 2019. 2

[30] Misha Tsodyks, Klaus Pawelzik, and Henry Markram. Neural networks with dynamic synapses. *Neural Computation*, 10(4):821–835, 1998. 2

[31] Misha V Tsodyks and Henry Markram. The neural code between neocortical pyramidal neurons depends on neurotransmitter release probability. *Proceedings of the National Academy of Sciences*, 94(2):719–723, 1997. 2

[32] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2

[33] Zihao W 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 2

[34] Heinz Wässle. Parallel processing in the mammalian retina. *Nature Reviews Neuroscience*, 5(10):747–757, 2004. 2

[35] Li Xi, Liu Guosui, and Jinlin Ni. Autofocusing of isar images based on entropy minimization. *IEEE Transactions on Aerospace and Electronic Systems*, 35(4):1240–1252, 1999. 7

[36] Jing Zhao, Ruiqin Xiong, and Tiejun Huang. High-speed motion scene reconstruction for spike camera via motion aligned filtering. In *2020 IEEE International Symposium on Circuits and Systems*, 2020. 2, 3, 6, 7, 8

[37] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Ev-flownet: Self-supervised optical flow estimation for event-based cameras. *arXiv preprint arXiv:1802.06898*, 2018. 2

[38] Lin Zhu, Siwei Dong, Tiejun Huang, and Yonghong Tian. A retina-inspired sampling method for visual texture reconstruction. In *IEEE International Conference on Multimedia and Expo*, 2019. 2, 3, 8

[39] Lin Zhu, Siwei Dong, Jianing Li, Tiejun Huang, and Yonghong Tian. Retina-like visual image reconstruction via spiking neural model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 2, 3, 6, 7, 8

Supplementary Material: High-speed Image Reconstruction through Short-term Plasticity for Spiking Cameras

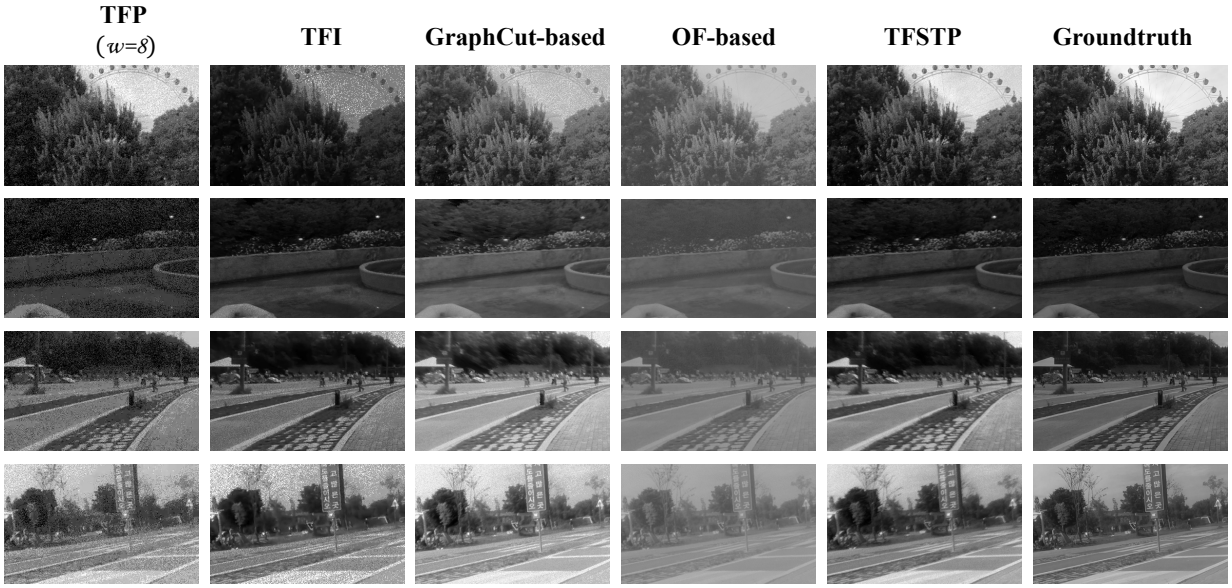

Figure S1. Comparison among different reconstruction methods on synthetic data.

A. Spiking Camera

Fovea-like Sampling Method. Inspired by the sampling mechanism of primate fovea [4, 3], spiking cameras take advantage of spike sequences to represent the brightness change in the spatial-temporal domain [1, 2]. Specifically, the photosensitive units continuously capture photons and increase the photodiode voltage. When the accumulated intensity exceeds a given threshold, a spike is generated and the photodiode voltage is reset to a predefined reset voltage. This process can be formulated as:

$$\text{A spike is generated at time } t^f \text{ if } \int_{t^{f-1}}^{t^f} I(t)dt \geq \phi, \quad \text{(S1)}$$

where $I(t)$ denotes the scene radiance, ϕ denotes the predefined threshold, and t^{f-1} represents the firing moment of the last spike. The spikes generated by spiking cameras can be represented by a 3-tuple $\mathcal{S} : \{x, y, t\}$, where $\{x, y\}$ denotes the spatial coordinates of the spikes in the photosensitive units, and t is the spike firing timestamp.

Texture Reconstruction from Inter-spike-interval (TFI). Based on the sampling mechanism of spiking cameras, the photosensitive units receive different scene radiance will trigger spikes with different frequencies. The inter-spike-interval (ISI) decreases as the scene radiance increases. Therefore, the pixel value (proportional to scene radiance) can be estimated by the interval between two neighboring spikes:

$$\hat{P}_{TFI} = \frac{C}{\Delta t}, \quad \text{(S2)}$$

where C refers to the maximum dynamic range of the spiking camera, and Δt represents the ISI.

Texture reconstruction from Playback (TFP). The TFP method infers the pixel value by collecting the spikes in a moving time window. By counting these spikes, we have

$$\hat{P}_{TFP} = \frac{N_w}{w} \cdot C, \quad \text{(S3)}$$

where C is the maximum dynamic range of the spiking camera, w is the size of the time window, and N_w is the total number of spikes collected in the time window.

B. Selection of the STP parameters

For the two reconstruction methods of TFSTP and TFMSTP, we have different preferences for parameter selection. The length of the time constants (τ_D and τ_F) and the relationship between them (e.g., $\tau_D < \tau_F$ or $\tau_D = \tau_F$) will

affect the convergent time of the STP dynamics, which implement some sort of adaptive time window. Larger time constants bring the higher contract, less noise, and more texture details in the static area but more motion blur in the motion area. Smaller time constants bring a smaller convergent rate, leading to the dark area with lower contrast in the early stage of reconstruction ($1\ ms$). This effect has been shown in Fig. 10 of the main text. In TFSTP, we need to balance the reconstruction quality between static and motion areas, so smaller time constants are chosen in $STP^{\theta 1}$. But in TFMSTP, we prioritize detecting the motion areas when choosing time constants, so larger time constants are chosen in $STP^{\theta 2}$.

C. Quantitative Evaluation on Simulated Data

In order to make referenced quantitative comparisons, we use the synthesized data[1] made by Zhao et al. [5]. Fig. S1 compares some reconstruction results of different methods on the simulated data. It can be noticed that although the GraphCut-based [6] method can obtain the best results in SSIM, its reconstruction images contain more noise than the results of TFSTP.

References

[1] Siwei Dong, Tiejun Huang, and Yonghong Tian. Spike camera and its coding methods. In *Data Compression Conference*, 2017. 1

[2] Siwei Dong, Lin Zhu, Daoyuan Xu, Yonghong Tian, and Tiejun Huang. An efficient coding method for spike camera using inter-spike intervals. In *Data Compression Conference*, 2019. 1

[3] Richard H Masland. The neuronal organization of the retina. *Neuron*, 76(2):266–280, 2012. 1

[4] Heinz Wässle. Parallel processing in the mammalian retina. *Nature Reviews Neuroscience*, 5(10):747–757, 2004. 1

[5] Jing Zhao, Ruiqin Xiong, Hangfan Liu, Jian Zhang, and Tiejun Huang. Spike2imgnet: Learning to reconstruct dynamic scene from continuousspike stream. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[6] Lin Zhu, Siwei Dong, Jianing Li, Tiejun Huang, and Yonghong Tian. Retina-like visual image reconstruction via spiking neural model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 2

[1] The simulated data is publicly available at https://cove.thecvf.com/datasets/517.