Reconstructing Clear Image for High-Speed Motion Scene With a Retina-Inspired Spike Camera

Jing Zhao, *Graduate Student Member, IEEE*, Ruiqin Xiong, *Senior Member, IEEE*, Jiyu Xie,
Boxin Shi, *Senior Member, IEEE*, Zhaofei Yu, *Member, IEEE*, Wen Gao, *Fellow, IEEE*,
and Tiejun Huang, *Senior Member, IEEE*

***Abstract*—Conventional digital cameras typically accumulate all the photons within an exposure period to form a snapshot image. It requires the scene to be quite still during the imaging time, otherwise it would result in blurry image for the moving objects. Recently, a retina-inspired spike camera has been proposed and shown great potential for recording high-speed motion scenes. Instead of capturing the visual scene by a single snapshot, the spike camera records the dynamic light intensity variation continuously. Each pixel on spike camera sensor accumulates the incoming photons independently and persistently, which fires a spike and restarts the photon accumulation immediately once the dispatch threshold is reached, producing a continuous stream of spikes recorded at very high temporal resolution. To recover the dynamic scene from captured spike stream, this paper presents an image reconstruction approach for spike camera. In order to generate high-quality reconstruction, we investigate the temporal correlation along motion trajectories and exploit it via adaptive temporal filtering. In particular, we present a hierarchical motion-aligned temporal filtering scheme, combining short-term filtering with long-term filtering to take advantage of long-term temporal correlation with low model complexity. Experimental results demonstrate that the proposed scheme outperforms the existing schemes significantly, producing much better objective and subjective qualities for spike camera image reconstruction.**

***Index Terms*—High-speed motion, image reconstruction, motion alignment, neuromorphic camera, spike camera, temporal correlation.**

I. Introduction

A. Motivation

Conventional digital cameras use an exposure window to accumulate all the incoming photons within that period to form a snapshot image. Such imaging mechanism can produce clear images with fine details for still scenes. However, for dynamic scenes with high speed motion, a single point on a moving object may be projected to different pixels on the image sensor, resulting in blurry image. To capture the motion process of dynamic scenes, frame-based high-speed cameras adopt a very short exposure time. With extremely reduced exposure, the moving distance of an object point projected on the image sensor becomes extremely short so that the captured image becomes less blurry. However, the reduced amount of incoming photons also leads to lower signal-to-noise ratio (SNR) in the formed images. With the recent prevalence of emerging computer vision applications such as autonomous driving and unmanned aerial vehicle, there is an increasing demand for high-speed motion scene imaging [1]. This makes the limitations of conventional cameras more evident.

To address the challenges in high-motion imaging, some biologic-inspired event cameras have been proposed [2]–[13], including DVS [3], ATIS [4], DAVIS [5] and CeleX [6]–[8], etc. Instead of recording the visual information by conventional image frames, event cameras monitor the light and send out events asynchronously to describe the light intensity changes. These events are recorded at very high temporal accuracy, e.g. on the order of micro-seconds or even nanoseconds [14], [15]. Such cameras are good at capturing high-speed motion and particularly suitable for motion detection and moving object tracking. However, they also have certain limitations. Probably the most critical one is that they can hardly reconstruct texture details of the visual scene.

Recently, a novel retina-inspired spike camera [16]–[19] has been proposed for capturing dynamic scenes with high temporal resolution. The spike camera no longer adopts the concept of image frames. Instead, each pixel on the spike camera sensor continuously accumulates the incoming light and fires a spike when a certain amount of photons is arrived. Different from the conventional cameras that use the same period for photon accumulation at every pixel, each pixel of spike camera works independently and fires a stream of spikes asynchronously, as shown in Fig. 1(a). The continuous spike stream provides a flexible representation to record the dynamic variation process of light intensity. In addition, different from the event cameras that only record relative light intensity changes, spike camera provides a more explicit format for recovering the absolute light intensity of the visual scenes.

Manuscript received June 9, 2021; revised September 12, 2021 and November 25, 2021; accepted December 8, 2021. Date of publication December 17, 2021; date of current version January 12, 2022. This work was supported in part by the National Natural Science Foundation of China under Grants 62072009, 22127807, and 61772041, and in part by the National Basic Research Program of China under Grant 2021YFF0900501. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Stanley H. Chan. *(Corresponding author: Ruiqin Xiong.)*

Jing Zhao, Ruiqin Xiong, Boxin Shi, Zhaofei Yu, Wen Gao, and Tiejun Huang are with the Institute of Digital Media, School of Electronic Engineering and Computer Science, Peking University, Beijing 100871, China (e-mail: jzhaopku@pku.edu.cn; rqxiong@pku.edu.cn; shiboxin@pku.edu.cn; yuzf12@pku.edu.cn; wgao@pku.edu.cn; tjhuang@pku.edu.cn).

Jiyu Xie is with the School of Information Science and Technology, University of Science and Technology of China, Hefei 230026, China (e-mail: xjy646@mail.ustc.edu.cn).

Digital Object Identifier 10.1109/TCI.2021.3136446

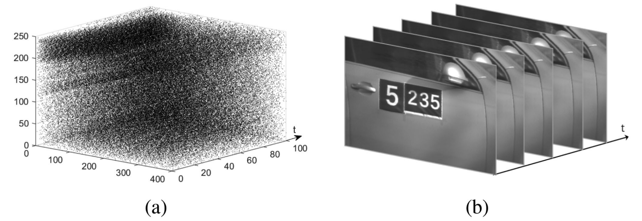


Fig. 1. Image reconstruction for spike camera. Here is an example of high-speed motion scene, where a car drives at a speed of 100 km/h. (a) The asynchronous spikes generated by spike camera. A black point represents a spike at a specific space-time location. (b) The reconstructed visual scene by our method.

B. Scope of the Paper

This paper aims to study the image reconstruction problem for spike cameras, as illustrated in Fig. 1. In order to recover the texture details from captured spike data, we may infer the instantaneous light intensity at any moment according to the inter-spike intervals (i.e. how long it takes the sensor pixel to accumulate the pre-specified amount of photons). This can provide a preliminary visual recovery. However, due to the existence of thermal noise and the Poisson effect of photon arrival, results of such simple light inference usually appear to be noisy, unstable, and spatio-temporally incoherent. Intuitively, in order to suppress the perturbation caused by noise, photon accumulation in a longer period should be considered, just as conventional cameras do using the exposure window. At the same time, special attention should be paid to object motion to avoid the mixing of lights from different object points, otherwise it will blur the image details.

In this paper, we propose an image reconstruction method for spike camera. In particular, we aim to handle the challenges brought by the conflict between high-speed motion and photon accumulation. For this purpose, we exploit the temporal correlation along motion trajectories. Considering that the variety in scene content leads to diversity in temporal correlation, we employ an adaptive temporal auto-regressive (TAR) model to formulate the temporal correlation. In order to generate low-noise reconstruction, it is preferred to exploit the signal along a long motion trajectory. However, as the motion trajectory becomes longer, more parameters are involved and the model would become complicated and unstable. To address the issue, we propose a motion-aligned hierarchical temporal filtering scheme, combining short-term filtering (STF) with long-term filtering (LTF). To be specific, we firstly employ STF to exploit local temporal correlation, and then establish a long-term temporal auto-regressive model based on the results of STF, so that long-term temporal correlation can be exploited with lower model complexity.

C. Related Works

1) Event Camera: In 1991, Mahowald *et al.* [2] published a moving cat on the cover of *Scientific American*, marking the birth of the first silicon retina. This pioneering work proves that a chip based on the neural architecture of eye can be a more powerful way to do computation, officially ignited the emerging field of neuromorphic vision sensors. Delbruck *et al.* [3] developed Dynamic Vision Sensor (DVS) to represent light intensity change with asynchronous sparse events. To recover light intensity, Posh *et al.* [4] proposed an Asynchronous Time-based Image Sensor (ATIS), which introduces the event-triggered light intensity measurement circuit to reconstruct the pixel at the changing pixel. Delbruck's *et al.* [5] developed the Dynamic and Active Pixel Vision Sensor (DAVIS) to make up for DVS texture imaging defect and was extended to color DAVIS346 [9]. Chen *et al.* [7], [8] increased the bit width of event to restore the scene texture. Different from these event cameras that focus on the variation of light intensity, the spike camera fires a positive signal to represent the arriving of a certain amount of photons, providing a more explicit input format to reconstruct the texture of the outer scenes.

2) Single-Photon Camera: Over the past decades, many efforts have been made to develop alternative sensors with photon-counting ability [20], [21], [24], [26]–[33]. Quanta Image Sensor (QIS) and Single-Photon Avalanche Diodes (SPAD) are two mainstream technologies. These emerging sensors are sensitive to single photoelectron, where the presence or absence of electron results in a logical binary output of "0" or "1" upon readout. Benefiting from the single-photon sensitivity, single photon cameras have shown potential for the applications under low illumination [34], [35]. Table I compares the SPAD, QIS and the spike camera [25].

Some image reconstruction algorithms have been proposed for SPAD and QIS cameras. Gyongy *et al.* [36] proposed to compensate for motion and spatially re-assign the photon detections to reconstruct the high-speed moving objects with minimal motion artifacts. Ma *et al.* [37] presented a quanta burst photography framework which can efficiently align and merge binary sequences into intensity images with minimal motion blur, high signal-to-noise ratio, and high dynamic range. Chi *et al.* [35] developed a student-teacher framework to handle noise and motion simultaneously. Seets *et al.* [38] and Iwabuchi *et al.* [39] proposed deblurring methods for single-photon imaging to suppress the motion blur and achieved competitive performance in dynamic scenes.

3) Image and Video Denoising: The problem we aim to solve in this paper is the estimation of a reliable visual signal from a stream of sensed light-intensity data perturbed by Possion effect and noise. Therefore it is highly related to image and video denoising. In the past decades, many methods have been proposed for image denoising and achieved great success [40]–[50]. Many works among them, such as BM3D [40], WNNM [41] and BAS [42], etc., consider that image signals can be sparsely represented in properly selected or learned domain. Many recent works, such as [44]–[47], [50]–[53], use deep neural networks (DNN) to learn the characteristics of images and noises so that the noises can be separated from the image signals. Besides, some burst denoising methods [49], [50], [54]–[57] merge a sequence of underexposed noisy images into a single clean image, so as to exploit the temporal correlation to handle the challenges brought by low light and motion.

4) Spike Camera Imaging: Spike camera is a retina-inspired neuromorphic camera proposed recently [16]–[19]. A few

TABLE I
COMPARISON WITH SPADS AND QISS

Camera	SPC SPAD [20]	SwissSPAD2 [21]	Time-gated SPAD [22]	QIS [23]	QIS [24]	Spike Camera [25]
Resolution	320×240	512×512	1024×1000	1376×768	1024× 1024	250×400
Pixel Pitch	8 μm	16.4 μm	9.4 μm	3.6 μm	1.1 μm	20 μm
Frames Per Second	16k	97.7k	24k	1k	1040	40k
Sensor Data Rate	1.54Gbps	10.24Gbps	25Gbps	1Gbps	1Gbps	4Gbps

texture reconstruction algorithms have been proposed for spike camera in recent years [18], [58]. The "texture from inter-spike interval" (TFI) [18] method infers the instantaneous light intensity according to inter-spike intervals. The "texture from playback" (TFP) method considers the number of spikes in a longer time window. The former one is usually quite noisy while the latter one is usually blurry when fast motion exists, so that a trade-off between motion blur and signal-to-noise ratio would become the problem. The TVS [58] method exploits a retina-like visual image reconstruction framework to improve the reconstruction quality. However, visually annoying noise and artifacts can still be observed in the reconstructed images.

The remainder of this paper is organized as follows. Section II gives an overview of the spike camera and the spike generation process. Section III discusses the preliminary light intensity inference. Section IV presents a motion-aligned image reconstruction framework for spike camera. Section V describes the proposed motion-aligned hierarchical temporal filtering method. Sections VI and VII report the experimental settings and results, respectively. Section VIII discusses the limitations and future works. Section IX concludes the paper.

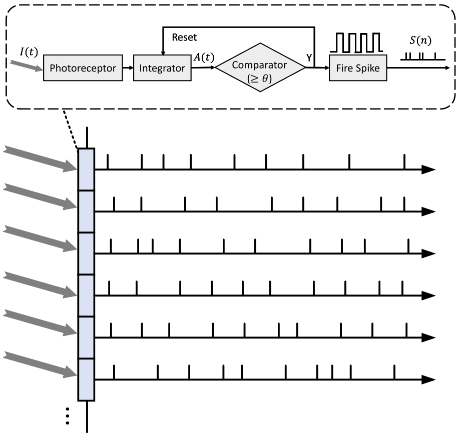


Fig. 2. The operation principle of spike camera. Each pixel on the sensor accumulates incoming photons persistently, and fires a stream of spikes to record the dynamic variation of light intensity. The sensor is illustrated by the blue column, and the spikes are illustrated by the short vertical lines along the horizontal time arrow.

II. Overview of Spike Camera

A. Idea of Spike Camera

Different from the conventional digital cameras that use a certain exposure time window to accumulate all the photoelectric information within that interval and compact them into a single snapshot image, the spike camera [16]–[18] abandons the concept of exposure window. Instead, it monitors the incoming light and fires a continuous stream of spikes to record the dynamic light intensity variation process, as illustrated in Fig. 2. The firing of each spike stands for the arrival of a very small number photons. The spike stream is recorded at a very high temporal resolution so that the dynamic light intensity variation process may be recovered from the spike data accurately.

B. Spike Generation

The mechanism of spike camera sensor is illustrated in Fig. 2. The sensor is composed of an array of pixels, each of which records the light intensity independently. Each pixel consists of three major components: photoreceptor, integrator, and comparator. The photoreceptor captures the incident light from the scenes and converts the instantaneous light intensity $I(t)$ into a voltage that can be recognized by the integrator. The integrator accumulates the electric charges from photoreceptor continuously, while the comparator checks the accumulated signal

$$A(t) = \int_0^t \eta I(x) dx \tag{1}$$

persistently. The constant η here denotes the photoelectric conversion rate. Once the accumulated signal reaches a dispatch threshold θ, the pixel sets up a flag signal for firing a spike (we call it spike flag). This signal is also sent to reset the integrator, immediately restarting a new "integrate-and-fire-spike" cycle. With such "compare-and-reset" operation, the signal accumulated by the integrator can be formulated as

$$A(t) = \int_0^t \eta I(x) dx \mod \theta. \tag{2}$$

Ideally, the aforementioned flag signal indicates the firing of a spike that should be immediately transmitted to all the following circuits. In actual hardware implementation, the spike is represented by a binary bit that will be read out under the control of a clock signal.

Fig. 3 illustrates the electric charge accumulation process and the generation of spike stream, with respect to a light intensity signal shown in Fig. 3(a). We can see that the time points in Fig. 3(b) where the signal $A(t)$ resets to zero are exactly the firing time of generated spikes. These points divide the working period into a set of intervals, in each of which the signal $I(t)$ integrates to a constant (i.e. θ/η), corresponding to the amount of photons represented by a spike. In current design, the θ is controlled by a reference voltage, which can be adjusted to accommodate different luminance conditions.

C. Spike Cycle and Inter-Spike Interval

Since each pixel works independently, we can restrict our discussion to a single pixel at this moment. Suppose $\{t_1, t_2, t_3, \cdots\}$ are the firing time of the generated spikes, it is straightforward

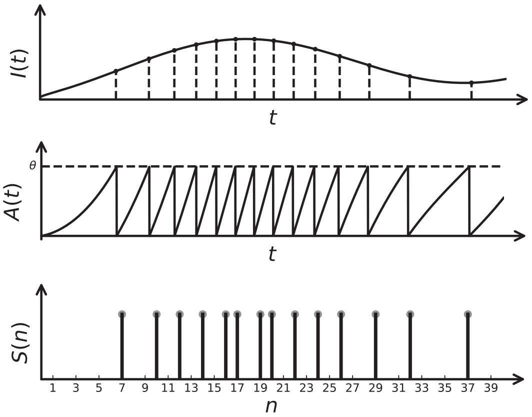


Fig. 3. The process of spike generation for one pixel. (a) Top: the incoming light intensity $I(t)$. It is worth mentioning that in practice the sensor can not observe the actual intensity directly. Instead, it receives a stream of incoming photons that follow a Poisson process. (b) Middle: the accumulated electric charges $A(t)$ with reset when $A(t)$ reaches the dispatch threshold θ. (c) Bottom: the spike array $S(n)$ read out by the camera. Each stick here stands for one spike. Note that the reading is synchronized by a clock signal – a spike is read out as “‘1” and no-spike is read out as “0”.

that the firing time of the k-th spike (i.e., t_k) satisfies

$$\int_{t_{k-1}}^{t_k} \eta I(x)dx = \theta. \tag{3}$$

Clearly, the photon accumulation for the k-th spike starts at t_{k-1} and ends at t_k. We call this period as the “life cycle” of the k-th spike, or the “inter-spike interval”. We can observe from Fig. 3 how the firing frequency of spikes varies with the instantaneous light intensity. When the incoming light is strong, it produces a dense spike stream with short inter-spike intervals. When the light becomes weak, it produces a sparse spike stream with longer inter-spike intervals.

D. Read Out of Spikes

A pixel on spike image sensor may fire spikes at arbitrary time, but the camera can only read out the spikes as a discrete-time binary signal $S(n)$. To be more specific, the camera checks the spike flag periodically, at the time $t = nT$, $n = 1, 2, \ldots$, with a fixed interval T. If the spike flag has been set up at the time $t = nT$, it reads out $S(n) = 1$ and clears the flag for this pixel immediately for the coming of next spike; otherwise, it reads out $S(n) = 0$. Apparently, the k-th spike fired at the time t_k will be read out as $S(n_k) = 1$ with

$$n_k = \lceil t_k / T \rceil. \tag{4}$$

Therefore, the index n of $S(n)$ is a discrete approximation of the continuous time t. To keep the time information of each spike as accurate as possible, the readout interval T should be small enough.[1]

E. Spike Data Format

The camera uses a high speed polling to periodically check the status of every pixel. To be more specific, in the current implementation, the camera checks 40000 times per second. Each time when it checks, it reads out the spike flag (“0” or “1”) of every pixel and forms a $H \times W$ spike frame.[2] This binary frame is compressed by a simple method and sent out via a high speed data interface. As the time lapses, the camera produces a three-dimensional $H \times W \times N$ spike cubic $S(x, y, n)$ – an array of spike frames, as illustrated in Fig. 1(a). For the convenience of later discussions, we use $S_n(\boldsymbol{z})$ instead to represent the spike frame array, where n is the time index and $\boldsymbol{z} = (x, y)$ denotes the pixel coordinate.

III. Light Intensity Inference for Spike Camera

The purpose of spike camera is to record the dynamic light-intensity variation for high-speed motion scenes. Once the spike frame array is captured, we aim to recover the instantaneous intensity at any time, denoted by $I_n(\boldsymbol{z})$.

A. Interval-Based Inference

A natural way to derive the light intensity is to consider the “life cycle” of each spike. A constant amount of electric charges is accumulated during this period. Therefore, we can infer the light intensity from the inter-spike intervals.

To be concrete, let us consider a specific pixel $\boldsymbol{z}$ and an arbitrary time index n. The start and end time index of the current spike cycle that covers the point $(\boldsymbol{z}, n)$ can be calculated by

$$P(\boldsymbol{z}, n) = \max \left\{ k \mid S_k(\boldsymbol{z}) = 1, k < n \right\}, \tag{5}$$

$$N(\boldsymbol{z}, n) = \min \left\{ k \mid S_k(\boldsymbol{z}) = 1, k \geqslant n \right\}. \tag{6}$$

Since a single spike cycle typically lasts an extremely short time, we can safely assume that the light intensity remains constant within this period. According to the spike generation model in (3), we have

$$\eta I_n(\boldsymbol{z}) \cdot [N(\boldsymbol{z}, n) - P(\boldsymbol{z}, n)] \cdot T + \epsilon_n(\boldsymbol{z}) \approx \theta \tag{7}$$

Here we introduced a small random perturbation $\epsilon_k(\boldsymbol{z})$ to account for the noise caused by dark current or the Poisson effect of incoming light. We use “$\approx$” in (7) because the discrete time index n_k is only an approximation of the actual firing time t_k, as formulated in (4). Based on (7), the instantaneous light intensity can be estimated as[3]

$$\widehat{I_n}(\boldsymbol{z}) = \frac{\theta}{\eta T \left[N(\boldsymbol{z}, n) - P(\boldsymbol{z}, n) \right]} \tag{8}$$

B. The Challenges

The above interval-based method essentially produces a visual reconstruction of the scene based on the instantaneous light intensity. Fig. 4(a) illustrates the image reconstructed according to (8). We can see the produced image appears to be quite noisy. This is because the instantaneous light intensity is actually difficult to measure accurately. Indeed, even under a constant luminance, the number of incoming photons in a very short interval is a random variable, which is typically Poisson distributed. Another factor that may contribute to the noisy reconstruction is

[1] The interval T is 25 μs in the current spike camera.

[2] The resolution in current implementation is 400×250.

[3] This method is called “texture-from-interval” (TFI) in [18].

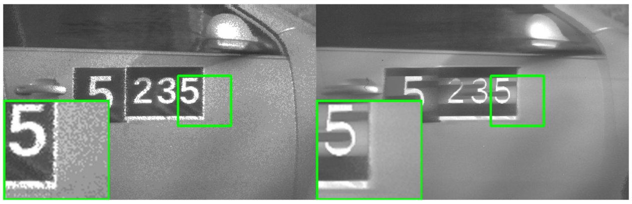


Fig. 4. Reconstruction images for the test scene *High-Speed Car*, created by interval based light intensity inference methods. (a) Left: the interval-based reconstruction based on single spike life cycle. (b) Right: the reconstruction based on multiple spike life cycles (with $m = 5$).

the inaccurate evaluation of the spike interval, due to the discrete approximation of t_k.

To suppress the effect of noise, an intuitive way is to jointly consider the photons received in multiple spike cycles. In this way, the estimation becomes

$$\widehat{I}_n(\boldsymbol{z}) = \frac{(2m-1)\cdot\theta}{\eta T\left[N^{(m)}(\boldsymbol{z},n) - P^{(m)}(\boldsymbol{z},n)\right]}, \tag{9}$$

with

$$P^{(m)}(\boldsymbol{z},n) = \max\left\{k \mid \sum_{i=k}^{n-1} S_i(\boldsymbol{z}) = m, k < n\right\}, \tag{10}$$

$$N^{(m)}(\boldsymbol{z},n) = \min\left\{k \mid \sum_{i=n}^{k} S_i(\boldsymbol{z}) = m, k \geqslant n\right\}. \tag{11}$$

Here 2 m is the number of spikes and $N^{(m)}(\boldsymbol{z},n) - P^{(m)}(\boldsymbol{z},n)$ is the total length of these $2m-1$ spike cycles. Fig. 4(b) illustrates an example of reconstructed image using (9), with $m = 5$. Indeed, by averaging the photons within multiple spike cycles, the influence of perturbation $\epsilon_k(\boldsymbol{z})$ can be remarkably suppressed for static scenes, producing much better recovery. However, for dynamic scene with high-speed motion, the movement of objects leads to undesired motion blur as shown in Fig. 4(b). As a conclusion, it is no longer appropriate to simply average the photons in the direction of temporal axis, when the objects move fast.

IV. Spike Camera Image Reconstruction With Motion Alignment

As we discussed, imaging for high-motion scene is very challenging. Photon accumulation is required to reduce the influence of sensor noise, but the existence of high-speed motion leads to the mixing of light from different object points. To address this challenges, we propose the idea of light-intensity inference with motion alignment. More specifically, we propose a motion aligned temporal filtering scheme to exploit the temporal correlation of light along motion trajectories.

A. Overall Framework

We aim to restore the true light intensity I at any moment from the recorded spike data S, with the best quality we may achieve. Considering the motion of objects, it would be beneficial to exploit the temporal correlation along motion trajectories, so that high-quality reconstruction can be achieved without incurring motion blur. To this end, we propose a motion-aligned reconstruction framework, as illustrated in Fig. 5. Suppose I_k is the key frame to reconstruct. We first infer the instantaneous light intensity at different moments via (8), producing a sequence of preliminary estimation $\widehat{I}_n$, $n = 1, 2, \ldots$. Then, we perform motion estimation based on these preliminary estimated reconstruction image, producing the displacement fields $\{\boldsymbol{u}_{k\to k+i}\}$ between the key frame and a set of neighboring frames. Finally, based on the displacement fields, motion-aligned temporal filtering is performed to regularize the preliminary estimations, generating the ultimate high-quality reconstruction image $\bar{I}_k$.

B. Motion Estimation

In order to exploit the temporal correlation to reduce the influence of noise without incurring motion blur, we need to find out the motion trajectories that go through each current pixel we aim to reconstruct, so that the pixels on the key frame $\widehat{I}_k$ can be mapped to reference frames $\{\widehat{I}_{k+i}\}$, $i = \pm 1, \pm 2, \ldots$ in the neighborhood. This problem can be solved via many optical flow algorithms [59]–[65]. In this paper, we followed the most widely used classical optical flow estimation strategy. To be specific, we assume that the light intensity along motion trajectories is consistent and the motion is spatially smooth, leading to the following optimization function

$$\min_{\boldsymbol{u}_{k\to k+i}} \left|\nabla \boldsymbol{u}_{k\to k+i}\right|_2^2 + \eta\left|\hat{I}_{k+i}\left(\boldsymbol{z} + \boldsymbol{u}_{k\to k+i}(\boldsymbol{z})\right) - \hat{I}_k(\boldsymbol{z})\right|_2^2 \tag{12}$$

Here η weighs between the light consistency term and the smoothness regularization term. Solving this optimization problem with Euler-Lagrange equations, we get the displacement field $\boldsymbol{u}_{k\to k+i}$ that maps the pixels in I_k to the pixels in I_{k+i}.

C. Motion-Aligned Temporal Filtering

It is worth noting that employing a fixed filter for the whole image cannot utilize the temporal correlation efficiently. On the one hand, due to the variety of scene content, the temporal correlations along motion trajectories may vary remarkably from one pixel location to another. On the other hand, due to the existence of object occlusion or illumination changes, the assumption of temporal consistency of light intensity along motion trajectories can be non-reliable in some cases. When outliers appear on the motion trajectories, they need be treated differently.

To handle the temporal correlation along motion trajectories adaptively, we propose to utilize an auto-regressive (AR) model [66], which has shown great potential for many image processing problems [67]–[72].

1) Temporal Auto-Regressive (TAR) Model: To characterize the temporal correlation among spike array, the light intensity can be modeled as an auto-regressive process along motion trajectories, as illustrated by Fig. 6. This can be formulated by

$$I_k(\boldsymbol{z}) = \sum_{i\in\phi} \alpha_i I_{k+i}\left(\boldsymbol{z} + \boldsymbol{u}_{k\to k+i}(\boldsymbol{z})\right) + \varepsilon. \tag{13}$$

Here ϕ is a template of time-index offset, representing the temporal dependency structure of the auto-regression model.

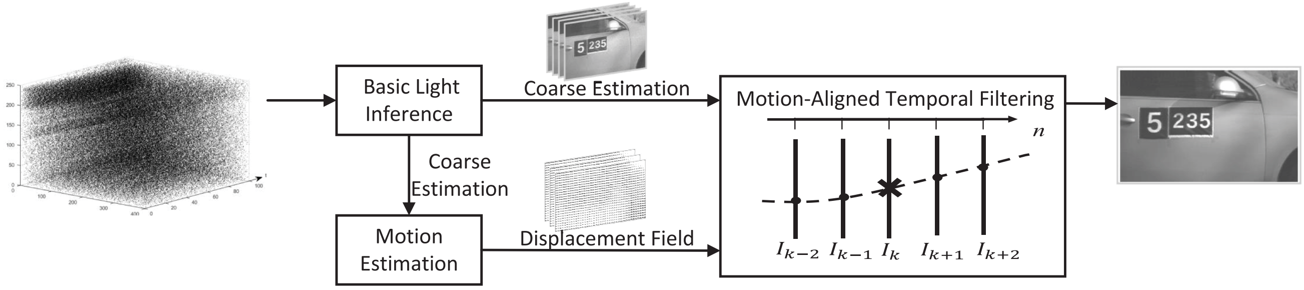


Fig. 5. The framework of motion-aligned image reconstruction for spike camera.

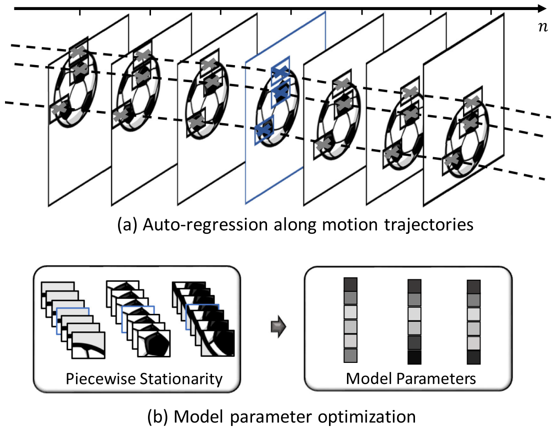

n
(a) Auto-regression along motion trajectories
Piecewise Stationarity
Model Parameters
(b) Model parameter optimization

Fig. 6. Motion-aligned temporal auto-regressive model. The light intensity of key pixels (represented by blue signs) can be inferred according to the corresponding supporting pixels (represented by the grey signs) along the same motion trajectories. The model parameters are adaptively determined by the local temporal structures.

A typical choice for ϕ is $\{\pm 1, \pm 2, \ldots, \pm K\}$. And $\{\alpha_i\}$ is a set of parameters for the TAR model to control the weighting in the linear combination (13). The term ε is a perturbation independent of the spatial-temporal location, and it accounts for both the fine details of image signal and random noise.

The validity of the TAR model hinges on a mechanism that adaptively adjusts the model parameters α_i to reflect the local temporal correlation structure of the visual signal. The assumption that motion and light intensity changes are locally smooth suggests piecewise stationarity. In other words, the parameters α_i remain nearly constant in a small locality, although they may vary significantly for different region. Such piecewise stationarity makes it possible to learn the signal structure by fitting the light intensity samples within a local window to the TAR model. Based on the learned structure, we can exploit the temporal correlation to generate reconstruction images with better quality.

2) Spatially-Adaptive Temporal Filtering: Suppose I_k is the image that we aim to reconstruct and $\boldsymbol{z}$ is an arbitrary pixel. With the displacement fields $\{\boldsymbol{u}_{k\to k+i}\}, i = \pm 1, \pm 2, \ldots,$ we filter the preliminarily estimated luminance along motion trajectory using the learned TAR model, producing a more stable reconstruction

$$\bar{I}_k(\boldsymbol{z}) = \sum_{i\in\phi} \alpha_i \widehat{I}_{k+i}\left(\boldsymbol{z} + \boldsymbol{u}_{k\to k+i}(\boldsymbol{z})\right). \tag{14}$$

It is worth noting that $\boldsymbol{u}_{k\to k+i}(\boldsymbol{z})$ in (14) can be sub-pixel displacement. To fetch the light intensity at sub-pixel location $\widehat{I}_{k+i}(\boldsymbol{z} + \boldsymbol{u}_{k\to k+i}(\boldsymbol{z}))$, interpolation methods such as bicubic interpolation [73], or the more sophisticated content adaptive methods such as NEDI [74] and SAI [68], can be used.

Based on the assumption of piecewise stationarity, the parameters $\boldsymbol{\alpha}$ of TAR model for pixel location $\boldsymbol{z}$ in I_k can be adapted to the signal local structure by solving the following least-square optimization problem:

$$\arg\min_{\boldsymbol{\alpha}} \sum_{\boldsymbol{z}'\in\Omega_{\boldsymbol{z}}} \left(I_k(\boldsymbol{z}') - \sum_{i\in\phi} \alpha_i I_{k+i}\left(\boldsymbol{z}' + \boldsymbol{u}_{k\to k+i}(\boldsymbol{z}')\right)\right)^2. \tag{15}$$

Of course, the true signal I_k is not available and we use $\widehat{I}_k$ instead to solve (15). Here $\Omega_{\boldsymbol{z}}$ is a two-dimensional local window around the pixel $\boldsymbol{z}$. In general, the size of $\Omega_{\boldsymbol{z}}$, i.e., the number of neighboring pixels the window contains, is set to be much bigger than the length of ϕ so that the above problem can be reliably solved and avoid over-fitting.

D. Limitation

In order to fully exploit the temporal correlation to improve reconstruction quality, a long-term TAR model combining the signal along a long motion trajectory, should be employed. This aspect is particularly important for the case of our high-speed spike camera since its sampling rate is up to 40000 Hz. However, with the increase in the size of ϕ, the complexity of TAR model increases accordingly, introducing a large set of parameters $\{\alpha_i\}$. As discussed in Section IV-C, to obtain a reliable estimation of $\{\alpha_i\}$, the window Ω should be increased accordingly. However, the validity of TAR model relies on the piecewise stationarity that the temporal correlation structure within local spatial window are near constant. If the size of Ω becomes too large, the stationarity assumption no longer holds and it will affect the accuracy of TAR model.

V. Motion-Aligned Hierarchical Temporal Filtering

To address the difficulty of TAR model learning for long motion trajectories, we propose a hierarchical temporal AR model. Based on this, we develop a **M**otion-**A**ligned **H**ierarchical **T**emporal **F**iltering (MAHTF) scheme, as shown in Fig. 7. It adopts a hierarchical filtering as illustrated in Fig. 8, which first utilizes a short-term filtering (STF) and then a long-term filtering (LTF) with reduced model degree of freedom. In this way, the long-term temporal correlation can be exploited without over-fitting.

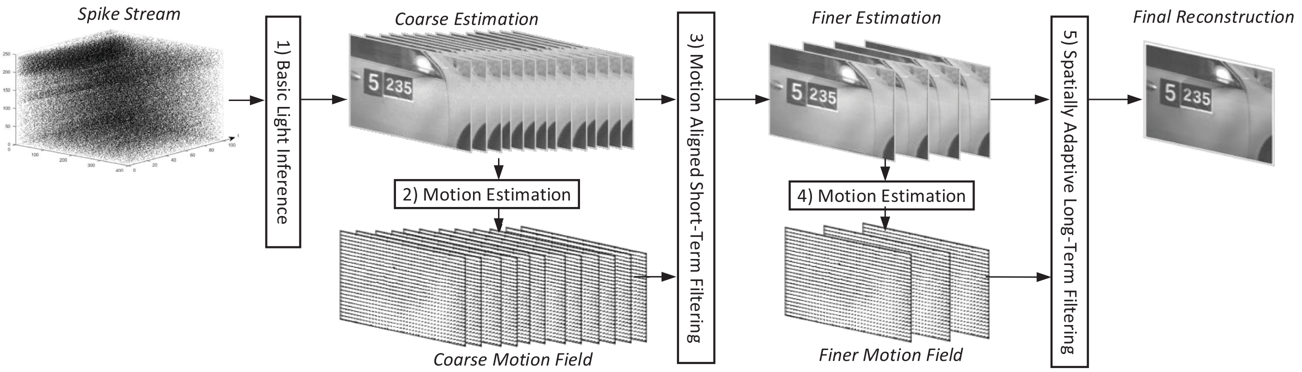


Fig. 7. The proposed Motion-Aligned Hierarchical Temporal Filtering (MAHTF) framework for spike camera image reconstruction.

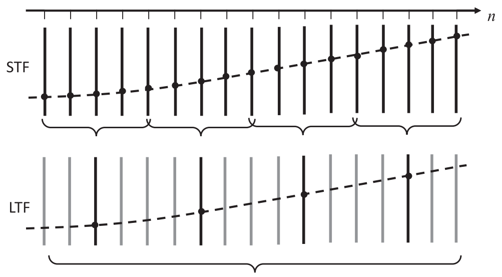


Fig. 8. Hierarchical temporal filtering along motion trajectories. Short-term filtering (STF) is first employed to exploit local temporal correlation to stabilize the initial estimation. Then, a long-term filtering (LTF) is applied to the sub-sampled results, so that long-term temporal correlation can be exploited with lower model complexity.

A. Overview

The overall framework is illustrated in Fig. 7. To achieve high reconstruction quality, we recover the visual scene via three stages in a coarse-to-fine manner. Suppose I_k is the frame that we aim to reconstruct. Firstly, an initial estimation of $\{I_{k+i}\}$, denoted by $\{\widehat{I}_{k+i}\}$, $i = 0, \pm 1, \pm 2, \ldots$, is inferred by (8). Then, the motion between $\widehat{I}_k$ and $\widehat{I}_{k+i}$ is estimated, resulting in the displacement fields $\boldsymbol{u}_{k\to k+i}$, $i = \pm 1, \pm 2, \ldots$. Assuming that a visual scene generally do not change much in very short time, a motion-aligned STF with a fixed filter is performed on $\{\widehat{I}_{k\pm i}\}$, so as to exploit the short-term temporal correlation and reduce the degree of freedom of the subsequent long-term TAR model. The STF produces a sequence of finer estimations $\{\widetilde{I}_{k\pm i}\}$. Finally, to further refine the reconstruction, motion estimation is performed to refine the displacement fields, and a motion-aligned LTF is conducted on $\{\widetilde{I}_{k\pm i}\}$ using a hierarchical long-term TAR model. The TAR model adaptively adjust the model parameters according to the local content structure. In particular, since the short-term correlation has already been exploited by STF, we establish the long-term TAR model based on temporal subsampling. This design helps to reduce the complexity of TAR model, but it does not affect the utilization of long-term temporal correlation.

B. Short-Term Filtering

We use short-term filtering to exploit the short-time correlation and reduce the freedom of TAR model for long-term correlation. In this paper, a short term is defined as a very short time, e.g. a few spike polling points ($0.1 \sim 0.2$ ms). In general, the natural image signal typically exhibits very strong temporal correlation in short term along motion trajectories, and the correlation structure tends to be the same, a fixed filter can be utilized for STF, resulting in an estimation of $I_k(\boldsymbol{z})$ formulated by

$$\widetilde{I}_k(\boldsymbol{z}) = \frac{1}{C}\sum_{i=-r_s}^{r_s} \omega_i \cdot \widehat{I}_{k+i}\left(\boldsymbol{z} + \boldsymbol{u}_{k\to k+i}(\boldsymbol{z})\right). \quad (16)$$

Here, r_s is the radius of short-term filtering and $C = \sum \omega_i$ is normalization factor. Under a widely used Markov model, the strength of temporal correlation in visual signal decays with the temporal distance. Therefore, we adopt a relatively simple filter for STF, in which the filter weight is formulated by

$$\omega_i = e^{-\frac{i^2}{2\sigma^2}}. \quad (17)$$

Here, σ is determined by the radius parameter r_s.

C. Long-Term Filtering

We employ a long-term filtering based on $\{\widetilde{I}_n\}$ to take advantage of temporal correlation in a long time. In order to exploit the correlation adaptively, we establish a TAR model along the motion trajectories. Specifically, since short-term correlation is already exploited via STF, we establish the TAR model with temporal subsampling, as shown in Fig. 8. The model produces a final estimation of $I_k(\boldsymbol{z})$ from several estimated frames with fixed frame intervals:

$$\bar{I}_k(\boldsymbol{z}) = \sum_{i=-r_l}^{r_l} \alpha_i \cdot \widetilde{I}_{k+i\cdot\bar{T}}\left(\boldsymbol{z} + \boldsymbol{u}_{k\to k+i\cdot\bar{T}}(\boldsymbol{z})\right). \quad (18)$$

Here, r_l represents the radius of TAR model and $\{\alpha_i\}$ is a set of filter weights, which can be adjusted adaptively according to the signal structures. $\bar{T}$ denotes the sampling interval for LTF, which is typically no larger than the length of STF, so that the correlation along the whole motion trajectories can be exploited.

Compared with the original TAR model formulated by (14), the freedom of the above TAR model is remarkably reduced with the aid of STF, so that the parameters of long-term TAR model can be adapted to its correlation structure. In addition, to avoid overfitting, we add a regularization term to constrain the

TABLE II
DETAIL INFORMATION OF REAL CAPTURED SPIKE DATA

Name	Resolution	Length (s)	Sampling rate
Train-350km/h	400×250	0.2	20000 HZ
Car-100km/h	400×250	0.2	20000 HZ
Falling Doll	400×250	0.01	20000 HZ
Fan-2600rpm	400×250	0.22	20000 HZ

parameters. Therefore, the filter weights is calculated by:

$$\arg\min_{\boldsymbol{\alpha}} \eta\|\boldsymbol{\alpha}\|_2^2 + \sum_{\boldsymbol{z}'\in\Omega_{\boldsymbol{z}}} \left(\widetilde{I}_k(\boldsymbol{z}') - \sum_{i=-r_l}^{r_l} \alpha_i \widetilde{I}_{k+i\cdot\bar{T}}\left(\boldsymbol{z}' + \boldsymbol{u}_{k\to k+i\cdot\bar{T}}(\boldsymbol{z}')\right)\right)^2, \quad (19)$$

Here $\Omega_{\boldsymbol{z}}$ is a spatial local window around pixel $\boldsymbol{z}$.

D. Discussion of Complexity

To speed up the algorithm, we use STF as a pre-processing and the complexity of the pre-processing is $O(n)$. Then, the major computational complexity is on motion alignment and the parameter estimation of LTF filtering weights $\{\alpha_i\}$. The complexity of motion alignment is proportional to the number of frames $2r_l$ used in LTF. The complexity of filtering weight calculation is related to both the window size and the degree of spatial overlap. Since LTF determines the filtering wights of one block at a time by solving (19), the larger the block and the smaller the overlap, the faster the algorithm runs. However, a large block size may reduce the adaptability of the TAR model. Thus, there needs a trade-off between algorithm complexity and model accuracy. Suppose each image is divided to m blocks, the time complexity to reconstruct n images is $O(n \times m + n \times 2r_l)$.

VI. EXPERIMENT SETTINGS

A. Implementation

We evaluate the performance of the proposed method on both synthesized data and real captured data. For synthesized data experiments, we develop a spike camera simulator to simulate the working mechanism of spike camera (as illustrated in Section II), so that we can generate spikes from some virtual scenes. The simulator outputs both the synthesized spike sequences and the corresponding ground truth image at each sampling time point. In this way, we can evaluate the reconstruction performance with objective image quality metrics, such as Peak Signal-to-Noise Ratio (PSNR) and Structural SIMilarity (SSIM). For real data experiments, we not only use the PKU-Spike-High-Speed dataset[4] but also capture additional spike streams using the PKU FSM spike camera system as shown in Fig. 9. The detail of the sequences is shown in Table II.

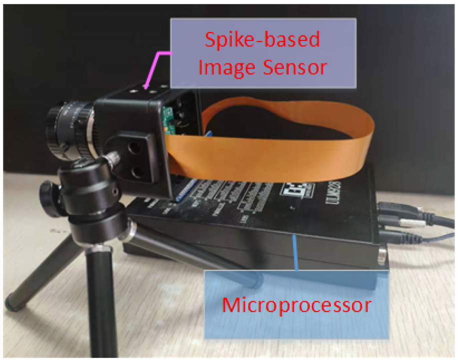


Fig. 9. The PKU FSM spike camera system, which consists of a spike-based image sensor and a microprocessor.

In this paper, the temporal window of MAHTF is set to 40, in which the radius of the STF is set to 6 and the interval of long-term TAR model is set to 8. The size of two-dimensional local window is set to 25×25.

B. Spike Data Synthesis

In order to evaluate the reconstruction method objectively, we develop a spike camera simulator to generate synthesized spike sequences from image based virtual scene.

To be specific, we regard each selected image as the scene to record and suppose that there is a relative motion between the scene and the spike camera sensor, which results in that the sensor captures different region of the scene at different moment. To simulate the spike camera working mechanism, each pixel of the sensor accumulates the light intensity (i.e., the pixel value of image) continuously, while the sensor checks the accumulated value of all the pixels periodically, producing a sequence of $H \times W$ spike frames. To be more specific, for the n-th polling, if the accumulated value at the coordinate **z** reaches the predefined threshold, it outputs $S_n(\boldsymbol{z}) = 1$ and clears the value of pixel $\boldsymbol{z}$. Otherwise, it outputs $S_n(\boldsymbol{z}) = 0$. At the time of each polling, we also extract the image region that the sensor currently monitors, producing a sequence of $H \times W$ ground truth.

VII. EXPERIMENTAL RESULTS

In this section, we evaluate the performance of the proposed method and compare it with the existing previous schemes. We also conduct some ablation studies to analyze the factors that influence the performance of spike camera image reconstruction. Supplementary material is available at IEEE DataPort [75] to demonstrate the spike data and reconstruction results.

A. Comparison With Previous Methods

We compare our method with the existing spike camera reconstruction methods, i.e. "Texture from inter-spike interval (TFI)"[18], "Texture from playback (TFP)"[18] and "Texture via Spiking Neural Model (TVS)"[58] objectively and subjectively. Here both the temporal window of TFP and the search window of TFI are set to 40.

[4] A dataset real captured with spike camera, which is publicly available at [Online]. Available: https://www.pkuml.org/resources/pku-spike-high-speed.html.

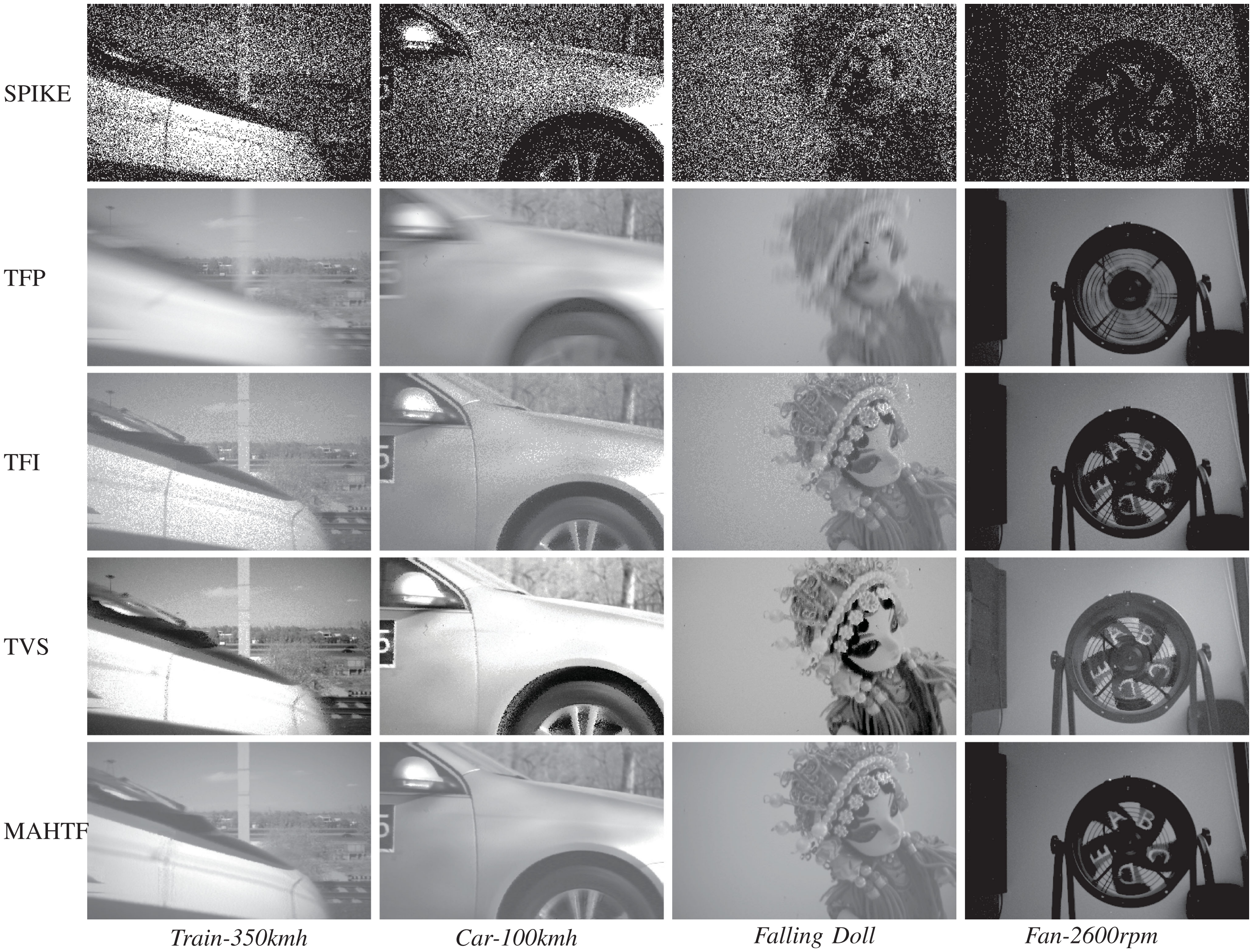


Fig. 10. Comparison of different reconstruction methods on real captured spike data. (Please enlarge the figure to observe details.) The proposed MAHTF achieves the best visual quality.

1) Visual Quality Evaluation: Figs. 10 and 11 show the images reconstructed from real spike data and synthesized spike data, respectively. We note that the visual quality performance of our proposed method clearly outperforms other reconstruction methods. For the TFI method, it can recover the outline of fast moving objects well, but the signal-to-noise ratio (SNR) is generally unsatisfactory. For example, the reconstructed image intensity fluctuates drastically with time on many pixels. This severely influences the perceptual quality. The TFP method, on the other hand, leads to apparent undesired motion blur on the reconstructions, especially for the regions with fast motion. The TVS method achieves better visual quality than TFI and TFP, but noticeable noise remain on the reconstruction, especially for the bright regions with short spike intervals. In contrast, our proposed method achieves better reconstruction that is more temporally stable, and it restores clear textures with fine details, even for the regions with high speed motion.

TABLE III
THE PSNR RESULTS OF DIFFERENT METHODS (UNIT: DB)

Sequences	TFI	TFP	Proposed
Airplane	20.36	22.32	**30.80**
Beacon	22.64	22.43	**31.02**
Cap	23.97	26.70	**32.89**
Drifting	23.05	25.38	**32.64**
Lena	22.04	24.14	**34.28**
House	24.81	18.92	**26.38**
Motorcycle	24.15	20.37	**28.82**
Parrot	23.44	24.43	**34.29**
Window	24.15	24.69	**34.32**
Pepper	21.38	26.30	**34.25**
Average	23.00	23.56	**31.97**

2) Objective Quality Evaluation: Tables III and IV show the average PSNR and SSIM results on the synthesized spike sequences, respectively. We note that the proposed MAHTF reconstruction method significantly outperforms the previous spike camera reconstruction methods. We believe the reason is that the proposed method achieves long-term photon accumulation for the novel camera model via motion-aligned temporal filtering, which can effectively handle the conflict between high-speed motion and light intensity stabilization. Such motion alignment has not been consider in all the previous image reconstruction methods for spike camera.

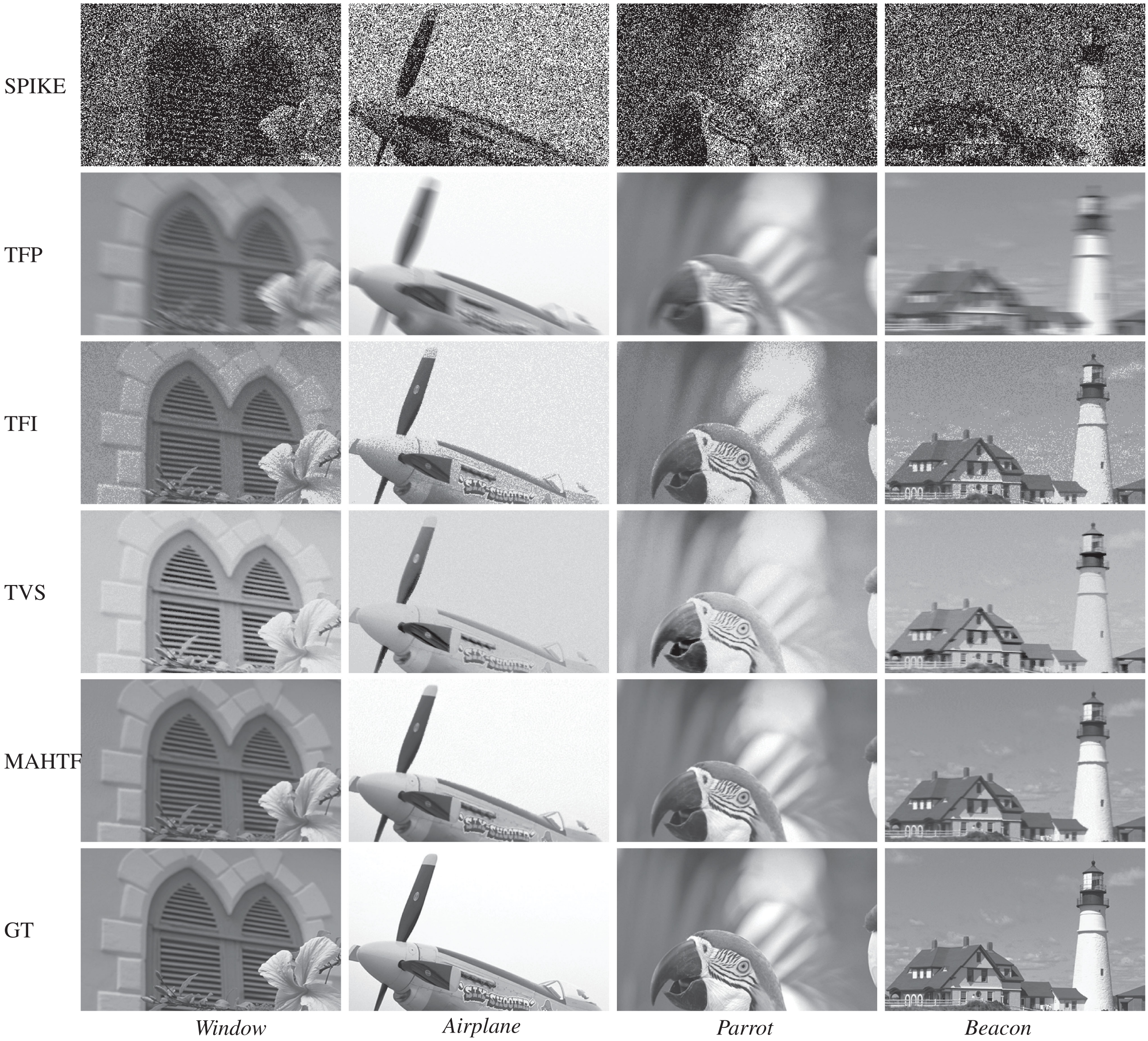


Fig. 11. Comparison of different reconstruction methods on synthesized spike data. (Please enlarge the figure to observe details.) The proposed MAHTF achieves the best visual quality.

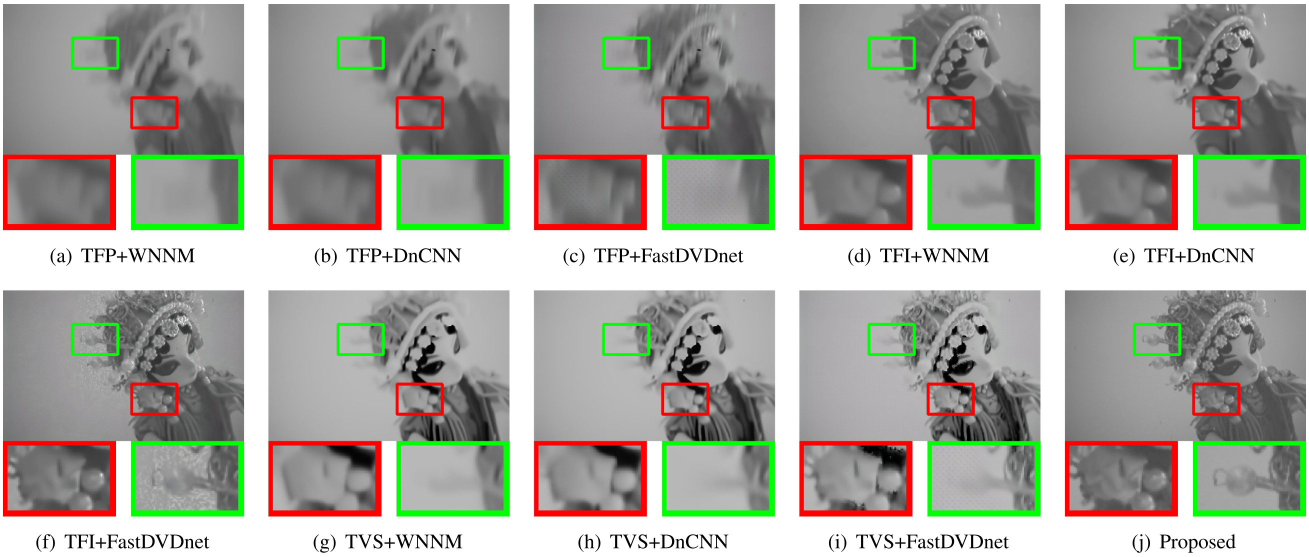
B. Comparison With Denoising Methods

We compare the proposed method with several alternative schemes combining existing spike camera image reconstruction algorithms with denoising. For denoising algorithms, we consider WNNM [41], DnCNN [47] and FastDVDnet [52]. For existing spike camera reconstruction methods, we use TFP [18], TFI [18] and TVS [58].

Table V presents the PSNR and SSIM results of the compared schemes. We note that the proposed MAHTF scheme achieves the best performance among them. Fig. 12 shows the visual quality of reconstruction image. We can observe that the proposed MAHTF outperforms other pipelined reconstruction methods. For TFP-based methods, the added denoising post-processing can not remove the blur effect. For TFI-based and TVS-based methods, the denoising operation is helpful, but it cannot provide a quality comparable to MAHTF. It removes part of image noise but also degrade some of the texture details. The reason may be that these denoising algorithms have not taken the characteristics of spike camera data into consideration.

C. Comparison With Conventional Camera

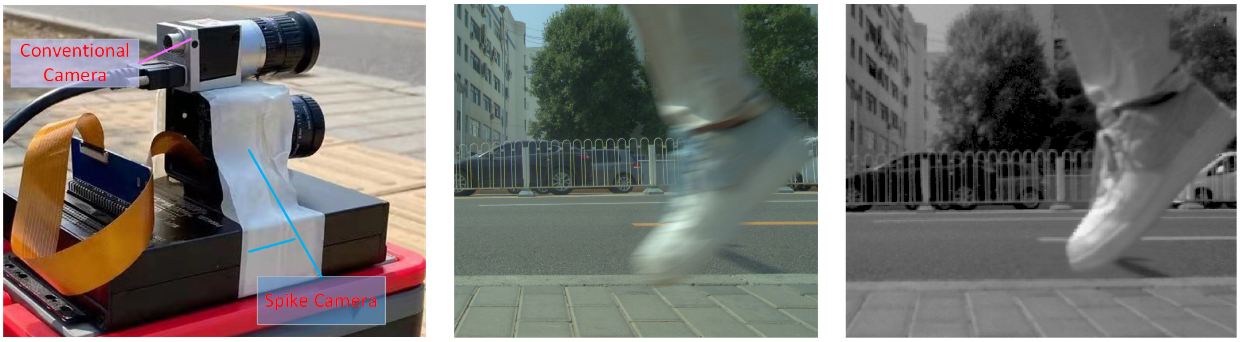
We set up a hybrid camera system to compare the proposed method with conventional camera. As shown in Fig. 13(a), the hybrid camera system is composed of a conventional camera and a spike camera. The conventional camera adopts the auto-exposure mode. Fig. 13 shows the visual comparison for a high-speed motion scene. We note that the image captured by a conventional camera at 120 fps is blurry. The reason is that conventional cameras accumulate all the photons within a exposure window to form a snapshot, ignoring the object



(a) TFP+WNNM (b) TFP+DnCNN (c) TFP+FastDVDnet (d) TFI+WNNM (e) TFI+DnCNN

(f) TFI+FastDVDnet (g) TVS+WNNM (h) TVS+DnCNN (i) TVS+FastDVDnet (j) Proposed

Fig. 12. Comparison with denoising methods on the “falling doll”.



(a) Hybrid camera system (b) Result of conventional camera (c) Result of spike camera

Fig. 13. Visual comparison with a conventional camera for a high-speed motion scene, where the leg is moving very fast. Our MAHTF can reconstruct the high-speed moving scene without motion blur.

TABLE IV
THE SSIM RESULTS OF DIFFERENT METHODS

Sequences	TFI	TFP	Proposed
Airplane	0.4207	0.8062	**0.8709**
Beacon	0.4219	0.7319	**0.9131**
Cap	0.5214	0.7442	**0.9172**
Drifting	0.6324	0.7085	**0.9347**
Lena	0.5436	0.6554	**0.9284**
House	0.8057	0.4050	**0.8717**
Motorcycle	0.8055	0.6138	**0.9445**
Parrot	0.4803	0.7987	**0.9467**
Window	0.5482	0.7143	**0.9518**
Pepper	0.4665	0.7493	**0.9043**
Average	0.5646	0.6927	**0.9183**

motion in that interval. The spike camera produce a continuous spike streams to record the high-speed dynamic scene at a much higher temporal resolution. By properly modeling the motion and temporal correlation, we can reconstruct a clear image (as shown in Fig. 13(c)) for each time point.

D. *Comparison With Single-Photon Camera Imaging Method*

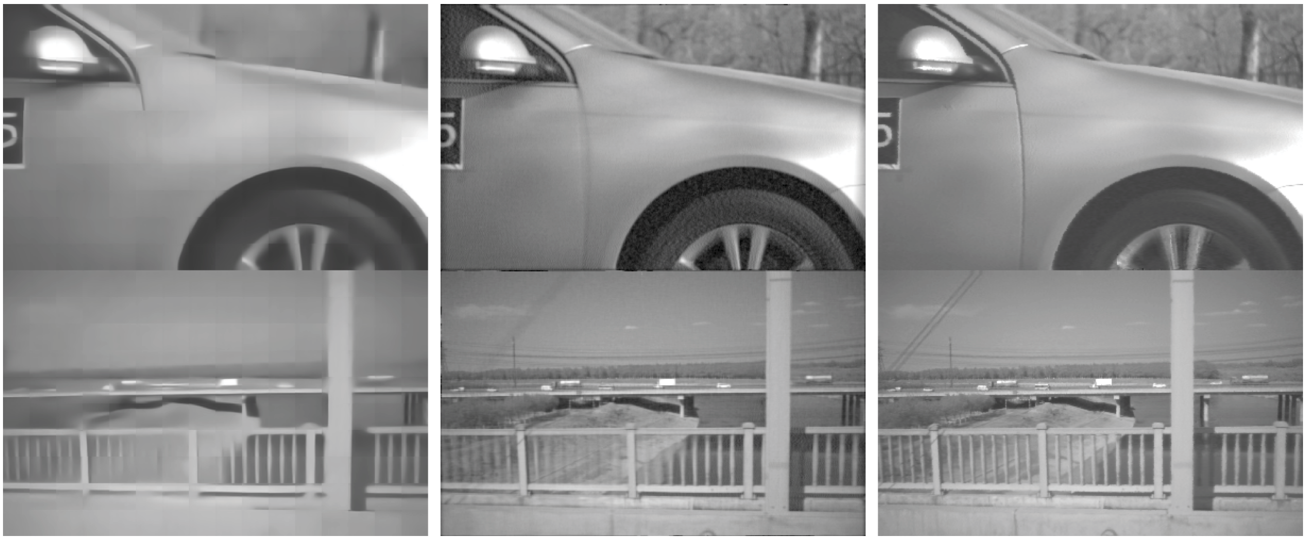
To evaluate the performance of our proposed framework on reconstructing high-speed moving scenes, we compare the proposed framework with the representative single-photon imaging methods, i.e., [37] and [35]. For [37], the temporal window of each block is set to 40. For [35], we use the code and pretrained model, which are publicly available at https://github.itap.purdue.edu/StanleyChanGroup/ECCV2020_Dynamic. Fig. 14 shows the reconstruction results. We note that the proposed method achieves the best performance on spike camera reconstruction. In fact, the working mechanism difference between single-photon cameras and the spike camera results in the difference of output data meaning, such that single-photon imaging methods are not well-matching for spike camera imaging.

E. *Visualization of Intermediate Results*

In order to better evaluate the proposed scheme, we visualize the intermediate results in Fig. 15, including the preliminary reconstruction, the finer reconstruction, the preliminary motion

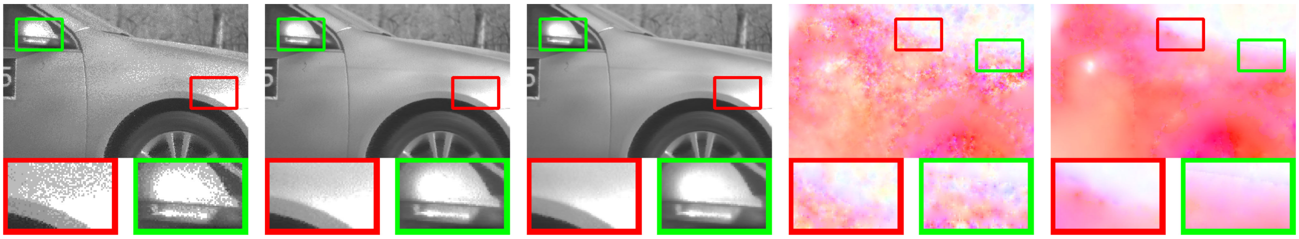
TABLE V
COMPARISON WITH DENOISING METHODS

Metric	Sequences	TFI+WNNM	TFI+DnCNN	TFI+FastDVDnet	TFP+WNNM	TFP+DnCNN	TFP+FastDVDnet	Proposed
PSNR	Airplane	18.97	19.62	19.00	22.29	22.34	22.28	**30.80**
	Beacon	26.76	27.53	26.16	22.37	22.40	22.46	**31.02**
	Cap	28.82	29.57	28.17	26.07	26.06	26.44	**32.89**
	Drifting	25.85	26.79	26.64	23.64	23.95	25.16	**32.64**
	Lena	28.38	28.48	24.92	23.85	23.97	24.29	**34.28**
	House	24.15	24.75	26.77	18.84	18.86	18.96	**26.38**
	Motorcycle	24.63	26.01	27.20	19.78	20.03	20.49	**28.82**
	Parrot	28.52	29.02	26.63	24.34	24.42	24.50	**34.29**
	Window	30.27	30.93	28.80	24.32	24.17	24.68	**34.32**
	Pepper	29.10	28.69	23.94	25.94	26.07	26.46	**34.25**
	Average	*26.55*	*27.14*	*25.82*	*23.14*	*23.22*	*23.57*	***31.97***
SSIM	Airplane	0.9167	**0.9224**	0.6312	0.8454	0.8447	0.7873	0.8709
	Beacon	0.8583	0.8705	0.7473	0.7631	0.7590	0.7487	**0.9131**
	Cap	0.8144	0.8373	0.8404	0.7319	0.7259	0.7219	**0.9172**
	Drifting	0.6395	0.6805	0.7809	0.5260	0.5317	0.6580	**0.9347**
	Lena	0.7504	0.7443	0.6882	0.6668	0.6671	0.6847	**0.9284**
	House	0.7235	0.6762	0.8345	0.3966	0.3926	0.4144	**0.8717**
	Motorcycle	0.7747	0.8164	0.8936	0.5405	0.5573	0.6240	**0.9445**
	Parrot	0.8689	0.8679	0.7983	0.8389	0.8389	0.8309	**0.9467**
	Window	0.8930	0.8981	0.7965	0.7226	0.7020	0.7385	**0.9518**
	Pepper	0.7396	0.7210	0.6241	0.7500	0.7508	0.7605	**0.9043**
	Average	*0.7979*	*0.8035*	*0.7635*	*0.6782*	*0.6770*	*0.7069*	***0.9183***



(a) The work [35] (b) The work [37] (c) Proposed

Fig. 14. Comparison with the competitive single-photon image sensor reconstruction methods.



(a) Preliminary reconstruction (b) Finer reconstruction (c) Final reconstruction (d) Preliminary motion (e) Refined motion

Fig. 15. Visualization of intermediate results. The quality is improved progressively.

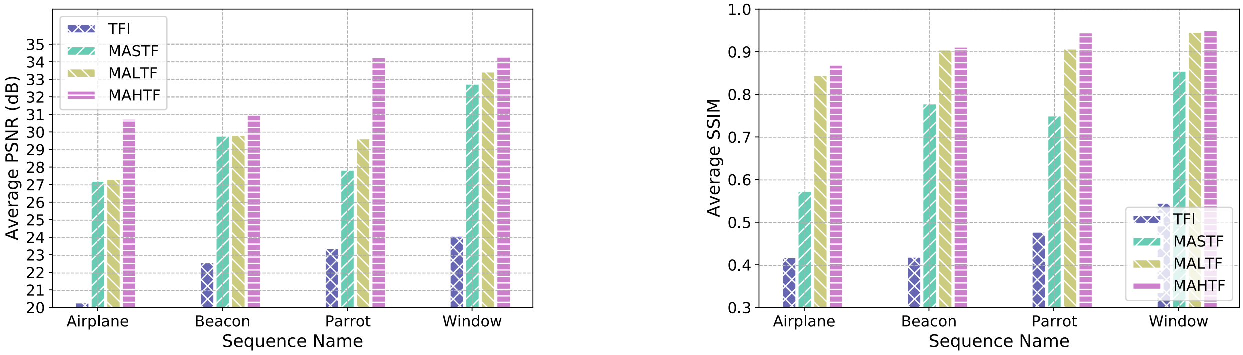


Fig. 16. Improvement of hierarchical temporal filtering. We compare the proposed MAHTF with motion-aligned short-term filtering (with a window radius of 6) and motion-aligned long-term filtering (with a window radius of 40).

field and the refined motion field. We note that the quality of reconstruction is improved progressively. As for motion estimation, we note that the refined motion field estimated based on finer reconstruction can describe the movement of objects more accurately than the preliminary motion field.

F. *Ablation Study for Hierarchical Temporal Filtering*

We look into the the benefit of hierarchical temporal filtering. We compare the proposed MAHTF method with two alternative methods: (1) motion-aligned short-term filtering (MASTF), (2) motion-aligned long-term filtering (MALTF). The MALTF method here uses the same strategy for choosing $\{\alpha_i\}$ as MASTF. We also compare them with the basic light inference method without any temporal filtering, i.e. TFI. Fig. 16 shows the PSNR and SSIM results. We can observe that long term filtering generally achieves better performance than short-term filtering, which confirms the necessity of utilizing long-term temporal correlation for spike camera image reconstruction. Moreover, we note that MAHTF outperforms MALTF, confirming that the hierarchical filtering structure plays an important role in improving the reconstruction quality.

VIII. Discussion and Limitation

A. *Super-Resolution*

As described above, spike camera uses a high-speed “integrate-and-fire-spike” mechanism to record the visual scenes at extremely high temporal resolution. In fact, the high-speed spiking mechanism also enables the spike camera to capture finer texture details beyond the pixel resolution of sensor itself. Due to the relative motion between the camera and the objects, the sensor may sample different points of an objects at different moments. By properly exploiting the motion information, an higher-resolution reconstruction with finer details may be recovered. We will explore this aspect in our future works.

B. *Limitations*

The proposed framework assumes that the motion can be well solved by optical flow algorithms, which is usually appropriate for camera motion and rigid object motion. When there exists complex motion and the assumption does not well hold, the auto-regressive model can mitigate the mismatch by adaptively exploiting the temporal correlation. However, when the scene contains several small objects or undergoes nonrigid motion, the reconstruction may suffer in undesired artifacts. In addition, the effectiveness of our proposed method is also based on the assumption of the brightness constancy along motion trajectories during a short-term interval (about $0.1 \sim 1$ ms). For the scenes where the light intensity along motion trajectories changes at very high frequency (e.g. $10\ k$ Hz or higher), the motion-aligned temporal filtering is inferior. Despite of this, the proposed method can still handle most of the high-speed scenes.

IX. Conclusion

The recently-invented spike camera has demonstrated its great potential for recording high-speed motion scenes. This paper addresses the image reconstruction problem for spike camera and proposes an effective scheme to reconstruct high-quality images from the captured three-dimensional spike data array. Taking the object movement into consideration, we employ motion-aligned filtering to exploit the temporal correlation along motion trajectories. This allows us to exploit the photo-electronic information in a relatively large time window, without mixing the light from different object points, so that high quality reconstruction can be achieved. In particular, we propose a hierarchical temporal filtering structure, combining short-term filtering with long-term temporal auto-regressive model to take advantage of long-term correlation with reduced model complexity. Experiments on both real captured spike data and synthesized spike data demonstrate that the proposed scheme achieves significantly improved reconstruction performance for spike camera.

Acknowledgment

The authors would like to thank Dr. Siwei Dong from Peking University for providing the spike camera dataset and the spike player software.

References

[1] T. Moeslund, A. Hilton, and V. Krger, “A survey of advances in vision-based human motion capture and analysis,” *Comput. Vis. Image Understanding*, vol. 104, no. 2, pp. 90–126, 2006.

[2] M. A. Mahowald and C. Mead, “The silicon retina,” *Sci. Amer.*, vol. 264, no. 5, pp. 76–83, 1991.

[3] P. Lichtsteiner, C. Posch, and T. Delbruck, “A 128 120 dB 15 μs latency asynchronous temporal contrast vision sensor,” *IEEE J. Solid-State Circuits*, vol. 43, no. 2, pp. 566–576, Feb. 2008.

[4] C. Posch, D. Matolin, and R. Wohlgenannt, “An asynchronous time-based image sensor,” in *Proc. IEEE Int. Symp. Circuits Syst.*, 2008, pp. 2130–2133.

[5] C. Brandli, R. Berner, M. Yang, S. Liu, and T. Delbruck, “A 240 130 dB 3 μs latency global shutter spatiotemporal vision sensor,” *IEEE J. Solid-State Circuits*, vol. 49, no. 10, pp. 2333–2341, Oct. 2014.

[6] M. Guo, J. Huang, and S. Chen, “Live demonstration: A 768 × 640 pixels 200meps dynamic vision sensor,” in *Proc. IEEE Int. Symp. Circuits Syst.*, 2017, pp. 1–1.

[7] M. Guo, R. Ding, and S. Chen, “Live demonstration: A dynamic vision sensor with direct logarithmic output and full-frame picture-on-demand,” in *Proc. IEEE Int. Symp. Circuits Syst.*, 2016, pp. 456–456.

[8] J. Huang, M. Guo, and S. Chen, “A dynamic vision sensor with direct logarithmic output and full-frame picture-on-demand,” in *Proc. IEEE Int. Symp. Circuits Syst.*, 2017, pp. 1–4.

[9] D. P. Moeys *et al.*, “A sensitive dynamic and active pixel vision sensor for color or neural imaging applications,” *IEEE Trans. Biomed. Circuits Syst.*, vol. 12, no. 1, pp. 123–136, Feb. 2018.

[10] C. Posch, T. Serrano-Gotarredona, B. Linares-Barranco, and T. Delbruck, “Retinomorphic event-based vision sensors: Bio-inspired cameras with spiking output,” *Proc. IEEE*, vol. 102, no. 10, pp. 1470–1484, Oct. 2014.

[11] T. Delbrck, B. Linares-Barranco, E. Culurciello, and C. Posch, “Activity-driven, event-based vision sensors,” in *Proc. IEEE Int. Symp. Circuits Syst.*, 2010, pp. 2426–2429.

[12] L. Pan, C. Scheerlinck, X. Yu, R. Hartley, M. Liu, and Y. Dai, “Bringing a blurry frame alive at high frame-rate with an event camera,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 6820–6829.

[13] H. Rebecq, R. Ranftl, V. Koltun, and D. Scaramuzza, “High speed and high dynamic range video with an event camera,” in *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 6, pp. 1964–1980, Jun. 2021.

[14] M.V. Srinivasan, R. J. Moore, S. Thurrowgood, D. Soccol, and D. Bland, “From biology to engineering: Insect vision and applications to robotics,” in *Frontiers in Sensing*, Berlin, Germany: Springer, 2012, pp. 19–39.

[15] G. Gallego *et al.*, “Event-based vision: A survey,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 154–180, 2020.

[16] S. Dong, T. Huang, and Y. Tian, “Spike camera and its coding methods,” in *Proc. Data Compression Conf.*, 2017, Art. no. 437.

[17] S. Dong, L. Zhu, D. Xu, Y. Tian, and T. Huang, “An efficient coding method for spike camera using inter-spike intervals,” in *Proc. Data Compression Conf.*, 2019, Art. no. 568.

[18] L. Zhu, S. Dong, T. Huang, and Y. Tian, “A retina-inspired sampling method for visual texture reconstruction,” in *Proc. IEEE Int. Conf. Multimedia Expo*, 2019, pp. 1432–1437.

[19] J. Han *et al.*, “Neuromorphic camera guided high dynamic range imaging,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 1730–1739.

[20] N. A. Dutton *et al.*, “A SPAD-based QVGA image sensor for single-photon counting and quanta imaging,” *IEEE Trans. Electron Devices*, vol. 63, no. 1, pp. 189–196, Jan. 2016.

[21] C. Bruschini *et al.*, “Monolithic SPAD arrays for high-performance, time-resolved single-photon imaging,” in *Proc. Int. Conf. Opt. MEMS Nanophoton.*, 2018, pp. 1–5.

[22] K. Morimoto *et al.*, “Megapixel time-gated SPAD image sensor for scientific imaging applications,” *Proc. High-Speed Biomed. Imag. Spectroscopy VI*, vol. 11654, 2021, Art. no. 116540U.

[23] S. Masoodian, A. Rao, J. Ma, K. Odame, and E. R. Fossum, “A 2.5 pj/b binary image sensor as a pathfinder for quanta image sensors,” *IEEE Trans. Electron Devices*, vol. 63, no. 1, pp. 100–105, Jan. 2016.

[24] A. Gnanasambandam, O. Elgendy, J. Ma, and S. H. Chan, “Megapixel photon-counting color imaging using quanta image sensor,” *Opt. Exp.*, vol. 27, no. 12, pp. 17298–17310, 2019.

[25] J. Gao, Y. Wang, K. Nie, Z. Gao, and J. Xu, “The analysis and suppressing of non-uniformity in a high-speed spike-based image sensor,” *Sensors*, vol. 18, no. 12, 2018, Art. no. 4232.

[26] N. A. Dutton, L. Parmesan, A. J. Holmes, L. A. Grant, and R. K. Henderson, “320× 240 oversampled digital single photon counting image sensor,” in *Proc. Symp. VLSI Circuits Dig. Tech. Papers*, 2014, pp. 1–2.

[27] I. Gyongy, N. Dutton, P. Luca, and R. Henderson, “Bit-plane processing techniques for low-light, high speed imaging with a SPAD-based QIS,” in *Proc. Int. Image Sensor Workshop*, 2015, pp. 284–287.

[28] N. A. Dutton12 *et al.*, “Oversampled ITOF imaging techniques using SPAD-based quanta image sensors,” in *Proc. Int. Image Sensor Workshop*, pp. 170–173, 2015.

[29] C. Niclass, A. Rochas, P.-A. Besse, R. S. Popovic, and E. Charbon, “CMOS imager based on single photon avalanche diodes,” *Proc. 13th Int. Conf. Solid-State Sensors*, vol. 1, pp. 1030–1034, 2005.

[30] E. Charbon, “Will avalanche photodiode arrays ever reach 1 megapixel,” in *Proc. Int. Image Sensor Workshop*, 2007, pp. 246–249.

[31] L. Sbaiz, F. Yang, E. Charbon, S. Susstrunk, and M. Vetterli, “The gigavision camera,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2009, pp. 1093–1096.

[32] F. Yang, L. Sbaiz, E. Charbon, S. Süsstrunk, and M. Vetterli, “On pixel detection threshold in the gigavision camera,” in *Proc. Digit. Photography VI*, 2010, p. 7537.

[33] F. Yang, Y. M. Lu, L. Sbaiz, and M. Vetterli, “Bits from photons: Oversampled image acquisition using binary poisson statistics,” *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 1421–1436, Apr. 2012.

[34] I. Gyongy, T. Al Abbas, N. A. Dutton, and R. K. Henderson, “Object tracking and reconstruction with a quanta image sensor,” in *Proc. Int. Image Sensor Workshop*, 2017, pp. 242–245.

[35] Y. Chi, A. Gnanasambandam, V. Koltun, and S. H. Chan, “Dynamic low-light imaging with quanta image sensors,” in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 122–138.

[36] I. Gyongy, N. A. Dutton, and R. K. Henderson, “Single-photon tracking for high-speed vision,” *Sensors*, vol. 18, no. 2, 2018, Art. no. 323.

[37] S. Ma, S. Gupta, A. C. Ulku, C. Bruschini, E. Charbon, and M. Gupta, “Quanta burst photography,” *ACM Trans. Graph.*, vol. 39, no. 4, pp. 79–1, 2020.

[38] T. Seets, A. Ingle, M. Laurenzis, and A. Velten, “Motion adaptive deblurring with single-photon cameras,” in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2021, pp. 1945–1954.

[39] K. Iwabuchi, Y. Kameda, and T. Hamamoto, “Image quality improvements based on motion-based deblurring for single-photon imaging,” *IEEE Access*, vol. 9, pp. 30080–30094, 2021.

[40] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, “Image denoising by sparse 3-D transform-domain collaborative filtering,” *IEEE Trans. Image Process.*, vol. 16, no. 8, pp. 2080–2095, Aug. 2007.

[41] S. Gu, L. Zhang, W. Zuo, and X. Feng, “Weighted nuclear norm minimization with application to image denoising,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 2862–2869.

[42] R. Xiong *et al.*, “Image denoising via bandwise adaptive modeling and regularization exploiting nonlocal similarity,” *IEEE Trans. Image Process.*, vol. 25, no. 12, pp. 5793–5805, Dec. 2016.

[43] S. H. Chan, X. Wang, and O. A. Elgendy, “Plug-and-play ADMM for image restoration: Fixed-point convergence and applications,” *IEEE Trans. Comput. Imag.*, vol. 3, no. 1, pp. 84–98, Mar. 2017.

[44] H. C. Burger, C. J. Schuler, and S. Harmeling, “Image denoising: Can plain neural networks compete with BM3D?,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 2392–2399.

[45] V. Jain and S. Seung, “Natural image denoising with convolutional networks,” *Proc. 21st Int. Conf. Neural Inf. Process. Syst.*, vol. 21, pp. 769–776, 2008.

[46] X. Mao, C. Shen, and Y.-B. Yang, “Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections,” *Adv. Neural Inf. Process. Syst.*, vol. 29, pp. 2802–2810, 2016.

[47] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, “Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising,” *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3142–3155, Jul. 2017.

[48] O. A. Elgendy, A. Gnanasambandam, S. H. Chan, and J. Ma, “Low-light demosaicking and denoising for small pixels using learned frequency selection,” *IEEE Trans. Comput. Imag.*, vol. 7, pp. 137–150, Jan. 2021.

[49] S. W. Hasinoff *et al.*, “Burst photography for high dynamic range and low-light imaging on mobile cameras,” *ACM Trans. Graph.*, vol. 35, no. 6, pp. 1–12, 2016.

[50] B. Mildenhall, J. T. Barron, J. Chen, D. Sharlet, R. Ng, and R. Carroll, “Burst denoising with kernel prediction networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2502–2510.

[51] M. Maggioni, G. Boracchi, A. Foi, and K. Egiazarian, “Video denoising, deblocking, and enhancement through separable 4-D nonlocal spatiotemporal transforms,” *IEEE Trans. Image Process.*, vol. 21, no. 9, pp. 3952–3966, Sep. 2012.

[52] M. Tassano, J. Delon, and T. Veit, “FastDVDnet: Towards real-time deep video denoising without flow estimation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 1354–1363.

[53] A. Davy, T. Ehret, J.-M. Morel, P. Arias, and G. Facciolo, “A non-local CNN for video denoising,” in *Proc. IEEE Int. Conf. Image Process.*, pp. 2409–2413, 2019.

[54] Z. Liu, L. Yuan, X. Tang, M. Uyttendaele, and J. Sun, “Fast burst images denoising,” *ACM Trans. Graph.*, vol. 33, no. 6, pp. 1–9, 2014.

[55] O. Liba *et al.*, “Handheld mobile photography in very low light,” *ACM Trans. Graph.*, vol. 38, no. 6, pp. 1–16, 2019.

[56] C. Godard, K. Matzen, and M. Uyttendaele, “Deep burst denoising,” in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 538–554.

[57] M. Aittala and F. Durand, “Burst image deblurring using permutation invariant convolutional neural networks,” in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 731–747.

[58] L. Zhu, S. Dong, J. Li, T. Huang, and Y. Tian, “Retina-like visual image reconstruction via spiking neural model,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 1438–1446.

[59] B. K. Horn and B. G. Schunck, “Determining optical flow,” *Techn. Appl. Image Understanding*, vol. 281, pp. 319–331, 1981.

[60] D. Sun, S. Roth, and M. J. Black, “Secrets of optical flow estimation and their principles,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 2432–2439.

[61] D. Sun, S. Roth, J. Lewis, and M. J. Black, “Learning optical flow,” in *Proc. Eur. Conf. Comput. Vis.*, 2008, pp. 83–97.

[62] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert, “High accuracy optical flow estimation based on a theory for warping,” in *Proc. Eur. Conf. Comput. Vis.*, 2004, pp. 25–36.

[63] A. Dosovitskiy *et al.*, “Learning optical flow with convolutional networks,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 2758–2766.

[64] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz, “PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8934–8943.

[65] Z. Teed and J. Deng, “RAFT: Recurrent all-pairs field transforms for optical flow,” in *Proc. Eur. Conf. Comput. Vision*, 2020, pp. 402–419.

[66] X. Wu, E. Barthel, and W. Zhang, “Piecewise 2D auto-regression for predictive image coding,” in *Proc. IEEE Int. Conf. Image Process.*, 1998, pp. 901–904.

[67] Y. Zhang, D. Zhao, X. Ji, R. Wang, and W. Gao, “A spatio-temporal auto regressive model for frame rate upconversion,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19, no. 9, pp. 1289–1301, Sep. 2009.

[68] X. Zhang and X. Wu, “Image interpolation by adaptive 2-D autoregressive modeling and soft-decision estimation,” *IEEE Trans. Image Process.*, vol. 17, no. 6, pp. 887–896, Jun. 2008.

[69] X. Zhang, R. Xiong, W. Lin, S. Ma, J. Liu, and W. Gao, “Video compression artifact reduction via spatio-temporal multi-hypothesis prediction,” *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 6048–6061, Dec. 2015.

[70] M. Li, J. Liu, X. Sun, and Z. Xiong, “Image/video restoration via multiplanar autoregressive model and low-rank optimization,” *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 15, no. 4, pp. 1–23, 2019.

[71] W. Dong, L. Zhang, R. Lukac, and G. Shi, “Sparse representation based image interpolation with nonlocal autoregressive modeling,” *IEEE Trans. Image Process.*, vol. 22, no. 4, pp. 1382–1394, Apr. 2013.

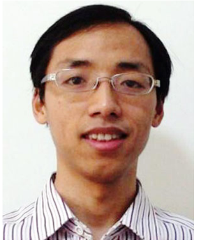
[72] D. Minnen, J. Ballé, and G. D. Toderici, “Joint autoregressive and hierarchical priors for learned image compression,” *Proc. 32nd Int. Conf. Neural Inf. Process. Syst.*, vol. 31, pp. 10771–10780, 2018.

[73] H. Hou and H. Andrews, “Cubic splines for image interpolation and digital filtering,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 26, no. 6, pp. 508–517, Dec. 1978.

[74] X. Li and M. T. Orchard, “New edge-directed interpolation,” *IEEE Trans. Image Process.*, vol. 10, no. 10, pp. 1521–1527, Oct. 2001.

[75] J. Zhao and R. Xiong, “Reconstruct clear image for high-speed motion scene with retina-inspired spike camera - supplementary material,” 2020. [Online]. Available: http://dx.doi.org/10.21227/3e24-mt42

Jing Zhao (Graduate Student Member, IEEE) received the B.S. degree from Wuhan University, Wuhan, China, in 2017. She is currently working toward the Ph.D. degree with Peking University, Beijing, China. Her research interests include computational imaging, neuromorphic camera, and image processing. She was the recipient of the Best Paper Award at the 2019 CVPR UG2+ Workshop.



Ruiqin Xiong (Senior Member, IEEE) received the B.S. degree from the University of Science and Technology of China, Hefei, China, in 2001, and the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2007. From 2007 to 2009, he was a Senior Research Associate with the University of New South Wales, Sydney, NSW, Australia. In 2010, he joined the School of Electronic Engineering and Computer Science, Peking University, Beijing, China, where he is currently a Professor. He has authored or coauthored more than 140 technical articles in referred international journals and conferences. His research interests include image and video processing, statistical image modeling, deep learning, neuromorphic camera, and computational imaging. He was the recipient of the Best Student Paper Award from SPIE Conference on Visual Communications and Image Processing in 2005. He was also the co-recipient of the Best Paper Award from IEEE Conference on Visual Communications and Image Processing in 2011 and the Best Student Paper Award from IEEE Conference on Visual Communications and Image Processing in 2017.

Jiyu Xie received the B.S. degree in electronic information engineering in 2020 from the University of Science and Technology of China, Hefei, China, where he is currently working toward the M.A. degree. His research interests include image/video coding and processing.

Boxin Shi (Senior Member, IEEE) received the B.E. degree from the Beijing University of Posts and Telecommunications, Beijing, China, in 2007, the M.E. degree from Peking University, Beijing, China, in 2010, and the Ph.D. degree from the University of Tokyo, Tokyo, Japan, in 2013. He is currently a Boya Young Fellow Assistant Professor and Research Professor with Peking University, where he leads the Camera Intelligence Lab. Before joining PKU, he did postdoctoral research with the MIT Media Lab, Singapore University of Technology and Design, Singapore, Nanyang Technological University, Singapore, from 2013 to 2016, and was a Researcher with the National Institute of Advanced Industrial Science and Technology from 2016 to 2017. He won the Best Paper Runner-up Award at International Conference on Computational Photography 2015. He was an Editorial Board Member of the *International Journal of Computer Vision* and an Area Chair of CVPR/ICCV.

Zhaofei Yu (Member, IEEE) received the B.S. degree from the Hong Shen Honors School, College of Optoelectronic Engineering, Chongqing University, Chongqing, China, in 2012, and the Ph.D. degree from Automation Department, Tsinghua University, Beijing, China, in 2017. He is currently an Assistant Professor with the Institute for Artificial Intelligence, Peking University, Beijing, China. His current research interests include artificial intelligence, brain-inspired computing, and computational neuroscience.

Wen Gao (Fellow, IEEE) received the Ph.D. degree in electronics engineering from the University of Tokyo, Tokyo, Japan, in 1991. He is currently a Professor of computer science with the School of Electronic Engineering and Computer Science, Institute of Digital Media, Peking University, Beijing, China. Before joining Peking University, he was a Professor of computer science with the Harbin Institute of Technology, Harbin, China, from 1991 to 1995, and a Professor with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. He has authored or coauthored extensively, including five books and more than 600 technical articles in refereed journals and conference proceedings in the areas of image processing, video coding and communication, pattern recognition, multimedia information retrieval, multimodal interfaces, and bioinformatics. He is a Member of the China Engineering Academy. He is the Chair of a number of prestigious international conferences on multimedia and video signal processing, such as the IEEE International Conference on Multimedia and Expo and ACM Multimedia, and served on the Advisory and Technical Committees of numerous professional organizations. He served or serves on the Editorial Board of several journals, such as the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON AUTONOMOUS MENTAL DEVELOPMENT, *EURASIP Journal of Image Communications*, and *Journal of Visual Communication and Image Representation*.

Tiejun Huang (Senior Member, IEEE) is currently a Professor with the School of Computer Science, Peking University, Beijing, China, and the Director of the Beijing Academy for Artificial Intelligence. He authored or coauthored more than 300 peer-reviewed papers on leading journals and conferences, and he is also a co-editor of four ISO/IEC standards, five National standards, and four IEEE standards. He holds more than 50 granted patents. His research interests include visual information processing and neuromorphic computing. He is a Fellow of CAAI, CCF, the Vice Chair of the China National General Group on AI Standardization. Professor Huang was the recipient of the National Award for Science and Technology of China (Tier-2) for three times (2010, 2012 & 2017).