Super Resolve Dynamic Scene from Continuous Spike Streams

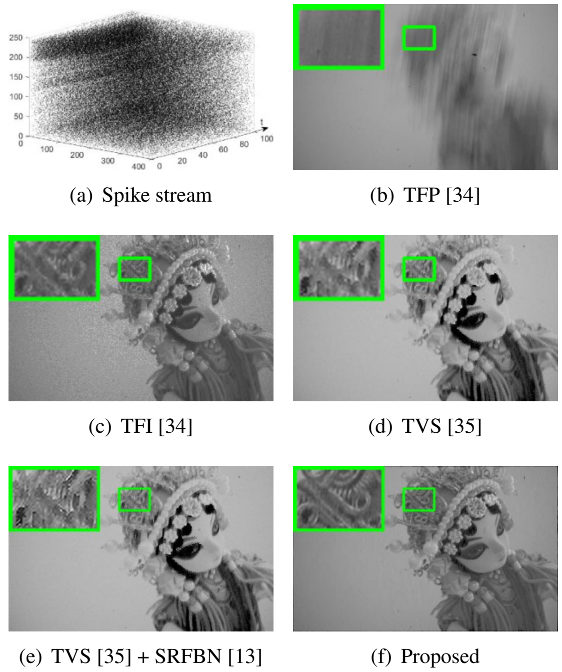
Jing Zhao[1], Jiyu Xie[2], Ruiqin Xiong[1]*, Jian Zhang[3], Zhaofei Yu[1], Tiejun Huang[1]
[1]School of Electronic Engineering and Computer Science, Peking University, Beijing, China
[2]University of Science and Technology of China, Hefei, China
[3]Shenzhen Graduate School, Peking University, Shenzhen, China
{jzhaopku, rqxiong, zhangjian.sz, yuzf12, tjhuang}@pku.edu.cn, xjy646@mail.ustc.edu.cn

Abstract

Recently, a novel retina-inspired camera, namely spike camera, has shown great potential for recording high-speed dynamic scenes. Unlike conventional digital cameras that compact the visual information within an exposure interval into a single snapshot, the spike camera continuously outputs binary spike streams to record the dynamic scenes, yielding a very high temporal resolution. Most of the existing reconstruction methods for spike camera focus on reconstructing images with the same resolution as spike camera. However, as a trade-off of high temporal resolution, the spatial resolution of spike camera is limited, resulting in inferior details of the reconstruction. To address this issue, we develop a spike camera super-resolution framework, aiming to super resolve high-resolution intensity images from the low-resolution binary spike streams. Due to the relative motion between the camera and the objects to capture, the spikes fired by the same sensor pixel no longer describes the same points in the external scene. In this paper, we exploit the relative motion and derive the relationship between light intensity and each spike, so as to recover the external scene with both high temporal and high spatial resolution. Experimental results demonstrate that the proposed method can reconstruct pleasant high-resolution images from low-resolution spike streams.

1. Introduction

With the development of real-time computer vision applications, such as unmanned aerial vehicle, autonomous driving and robotics, the inherent limitations of conventional digital cameras become increasingly evident. Conventional cameras generally accumulate the photoelectric information in a certain exposure window to form a snapshot frame. Such an 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 can be projected onto different pixels on the sensor, resulting in blurry artifacts for the moving objects.



(a) Spike stream (b) TFP [34]

(c) TFI [34] (d) TVS [35]

(e) TVS [35] + SRFBN [13] (f) Proposed

Figure 1. The image for a doll falling from a height. (a) Spike streams captured by the spike camera [6]. (b)-(d) Spike camera reconstruction with different methods [34, 35], which suffers low resolution with poor details. (e) x2 super resolution result by combining the state-of-the-art spike camera reconstruction method [35] with competitive image super-resolution method [13]. (f) x2 reconstruction by super-resolving the intensity image from spike streams with the proposed method, which can reconstruct the textures and details.

To address this issue, a novel neuromorphic camera, namely spike camera, has been developed to record dynamic scenes [6, 7]. Unlike the conventional cameras that capture the visual scene by a snapshot, the spike camera abandons the concept of exposure window. Instead, it mon-

[*] Ruiqin Xiong (rqxiong@pku.edu.cn) is the corresponding author.

itors the incoming light persistently and fires continuous spike streams to record dynamic scenes at very high temporal resolution. In addition, different from the bio-inspired event cameras that send events to record the *relative* light intensity changes, spike cameras fire spikes to record the arrival of a very small amount of photons, which provides more explicit information for recovering the *absolute* intensity.

Despite the great potential of spike camera for capturing dynamic scenes, recovering high-quality images from binary spike streams still remains an important and challenging issue, which has gained increasing attention in recent years [33–35]. Some works [33, 34] exploit the characteristics of spike streams and infer the instantaneous light intensity by estimating the firing frequency of each pixel. In addition, Zhu *et al.* [35] design retina-like visual image reconstruction frameworks to solve the problem. However, these methods mainly focus on suppressing the noises and blurry artifacts of reconstruction, ignoring the issue of low resolution. In fact, as a trade-off of low latency and low power consumption, current spike cameras have a relatively low spatial resolution. To generate high resolution images, an intuitive approach is to combine the spike camera reconstruction methods with image super-resolution algorithms [13, 16, 31, 32]. However, such pipelined schemes usually cannot achieve promising reconstruction, as the scene details have already been lost in the first reconstruction stage.

In this paper, we develop a novel image reconstruction framework to super-resolve high-quality intensity images from continuous spike streams. Due to the relative motion between the camera and the objects to capture, the spikes fired by the same sensor pixel no longer describes the same points in the external scene and each spike can be mapped to different locations in the scene. By exploiting the relative motion, we can recover the scene with a resolution much higher than that directly provided by the spike streams. To this end, we carefully analyze the working mechanism of spike camera. Based on the spike camera imaging principle, we formulate the relationship between the image intensity and each spike, so as to derive super-resolved intensity from the spike streams. The main contribution of this paper are summarized as follows:

- We present a super-resolution framework for spike camera. To the best of our knowledge, we are the first to super-resolve low-resolution (LR) spike streams to high-resolution (HR) intensity images.
- Instead of simply applying image super-resolution algorithms to LR reconstructions of spike camera, we derive the relationship between light intensity and each spike, so as to estimate pixel-wise super-resolved intensity from spike streams.
- Experimental results demonstrate that the proposed method can reconstruct pleasant HR intensity images from LR binary spike streams and recover fine details, which cannot be reconstructed with the state-of-the-art methods.

2. Related Work

Bio-inspired event camera Event cameras, also known as neuromorphic cameras, have gained growing attention with their distinctive advantages such as high speed, high dynamic range and low power consumption [1, 15]. Instead of recording the visual information in the whole exposure interval by a snapshot, event cameras monitor the variation of light intensity persistently, with each pixel generating a event stream to describe the changes of light intensity. Thus far, event cameras have shown great potential for capturing motion information in dynamic scenes, and have been applied to many computer vision applications, such as object detection and tracking [20, 21]. However, as only the relative light intensity changes are recorded, event cameras can hardly reconstruct the texture details of the visual scenes, and many efforts [2, 19, 23] have been made to solve this problem. Different from these event cameras, spike camera fires a positive signal to represent the arriving of a certain amount of photons, which provides a more explicit input format for reconstructing absolute light intensity.

Spikes to intensity images To reconstruct dynamic scenes from the asynchronous spike streams, many spike camera reconstruction methods have been proposed in recent years [33–35]. Inspired by the conventional imaging model, TFP [34] recovered light intensity by averaging the spikes in a virtual exposure window. This model is suitable for static scenes. However, for dynamic scenes, a single point on the moving objects can be projected onto different pixels on the sensor, leading to motion blur as shown in Fig. 1(b). To address this issue, TFI [34] inferred the instantaneous light intensity according to inter-spike intervals, which can provide a primary visual recovery of dynamic scenes, even for the regions with high-speed motion. However, affected by thermal noise, the reconstruction are visually unpleasant as shown in Fig. 1(c). To simultaneously handle the challenges brought by high-speed motion and noises, Zhao *et al.* [33] proposed to reconstruct dynamic scenes via motion-aligned filtering and Zhu *et al.* [35] developed a retina-like visual image reconstruction framework, which achieves state-of-the-art performance as shown Fig. 1(d). However, all of these methods ignored the issue of low-resolution.

Image and video super-resolution In the past decades, various image super-resolution (SR) methods have been proposed to recover a high-resolution image from its low-resolution counterpart [5, 8, 13, 29, 30]. Early works used interpolation techniques based on sampling theory like linear or bicubic. These methods run fast, but can not re-

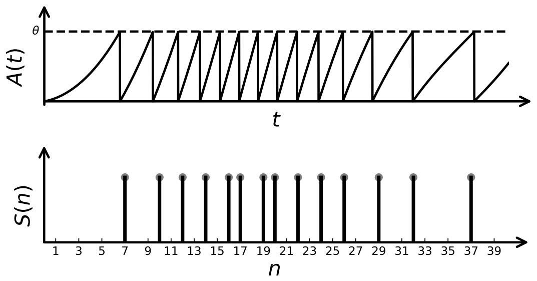


Figure 2. Example of spike generation process at the pixel p. Top: the instantaneous electric charges $A(t)$ with reset when $A(t)$ reaches the dispatch threshold θ. Bottom: the spike stream $S(n)$ read out by the pixel.

build realistic textures. To address these issues, improved approaches devoted to establishing complex mapping functions between LR and HR images based on neighbor embedding or sparse coding [8]. More recently, convolutional neural network (CNN) based image SR have achieved impressive results. Dong *et al.* [5] first proposed a shallow CNN for image SR. Inspired by the pioneering work, many new architectures, such as EDSR [16], RDN [32], DEPN [10], and SRFBN [13], have been proposed and achieved promising performance. In addition, based on the image SR methods and further to grasp the temporal consistency, many video SR algorithms [4,11,12,14,27] were developed to handle spatio-temporal information simultaneously. To improve the spatial resolution of spike camera reconstruction, we can first reconstruct intensity images from spike streams and then apply these image SR methods to the reconstruction. However, such pipelined schemes can not achieve promising results, as the details have already been lost in the first reconstruction stage.

3. Preliminary

In this section, we first formulate the working mechanism of spike camera, and then present the spike camera super-resolution problem.

3.1. Spike camera working mechanism

Spike camera is composed of an array of pixels, each of which records light intensity independently. Each pixel consists of three major components: photoreceptor, integrator, and comparator. The photoreceptor captures the incident light from the outer scenes and converts the light intensity into a voltage that can be recognized by the integrator. The integrator accumulates the electric charges from the photoreceptor continuously, while the comparator checks the accumulated signal persistently. Once the dispatch threshold θ is reached, a spike is fired and the integrator is reset, restarting a new “integrate-and-fire” cycle.

Since each pixel works independently, we can restrict our discussion to a single pixel $p = (r, c)$. The instantaneous electric charge amount of pixel p at the time t can be formulated as:

$$A(t) = \int_{\Omega_p} \int_0^t \alpha \cdot I(z,x) dx dz \mod \theta. \qquad (1)$$

Here Ω_p denotes the spatial region that pixel p covers, $I(z,t)$ denotes the light intensity of position $z = (x,y)$ at time t and α is the photoelectric conversion rate. Spikes may be fired at arbitrary time t, but the camera can only read out the spikes as a discrete-time binary signal $S(n)$ (as shown in Fig. 2). To be more specific, the camera checks the spike flag with a fixed short interval T. If a spike flag has been set up at the time t, with $(n-1)T < t \leq nT$, it reads out $S(n) = 1$. Otherwise, it reads out $S(n) = 0$. As the light comes in continuously, all the pixels on the sensor work simultaneously and independently, firing spikes to represent the arrival of every certain amount of photons. With the time going on, the camera would produce a binary spike array $S \in \{0,1\}^{H \times W \times N}$ as shown in Fig. 1(a).

3.2. Problem statement of spike camera SR

The purpose of spike camera is to record the dynamic light intensity variation process for high-speed motion scenes. Once the spike array is captured, we aim to recover the instantaneous intensity at any time. In particular, considering the limited spatial resolution of spike camera, we aim to super resolve high-quality intensity images with fine details. Instead of using a pipelined method that simply combines spike camera reconstruction algorithms with existing image SR methods, we propose to directly estimate pixel-wise super-resolved intensity. It is an ill-posed inverse problem, which can be described as follows. Given the binary spike array $S \in \{0,1\}^{H \times W \times N}$, the objective is to restore high-quality HR intensity images $I^{\text{HR}} \in [0,255]^{cH \times cW \times N}$ from the LR spike array, where c is the upscale factor.

4. Approach

As shown in Fig. 3, due to the relative motion between the camera and the object, the spikes fired by the same sensor pixel no longer describes the same points on the object in the outer scene. Instead, it records the light intensity at various locations. That is, each spike can be mapped to different locations on the scene. By properly exploiting the relative motion between the camera sensor and the scene, it is possible to recover the scene with a resolution much higher than that directly provided by the pixels of spike camera. To this end, we develop a motion-guided spike camera super-resolution (MGSR) framework in order to super-resolve HR intensity images from the LR spike streams.

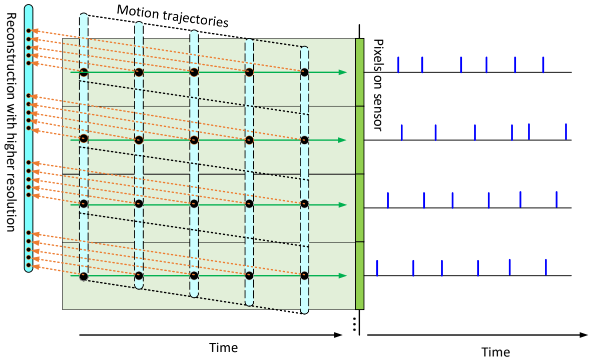


Figure 3. The principle of spike camera SR imaging. Left: the dynamic scene. Right: the spike streams read out by the sensor. Due to the relative motion between the camera and the objects, the spikes fired by the same sensor pixel can be mapped to various locations on the scene to reconstruct.

4.1. Intensity-spike relationship

As introduced in Section 3.1, each spike corresponds to a certain amount of photons, which can be denoted as a tuple $s:(p,t_s,t_e)$ with $p=(r,c)$ representing the pixel coordinate. Here t_s and t_e represent the start and end time of the current spike cycle, respectively. Based on Eq. (1), the relationship of the spike $s:(p,t_s,t_e)$ and intensity I can be formulated as

$$\theta = \int_{\Omega_p}\int_{t_s}^{t_e} \alpha \cdot I(z,t)dtdz. \tag{2}$$

Suppose that we aim to reconstruct the scene at time k. Based on the assumption of brightness constancy, every $I(z,t)$ can be represented with the intensity of the corresponding point at time k as $I(z+u_{t\rightarrow k}(z),k)$, where $u_{t\rightarrow k}(z)$ denotes the displacement that maps the point z in the scene at time t to the corresponding point in the key scene at time k. Thus, we can model the relationship between the intensity of the key scene and arbitrary spike $s:(p,t_s,t_e)$ as

$$\theta = \int_{\Omega}\int_{t_s}^{t_e} \alpha \cdot I(z,k)\cdot \mathcal{M}_s(z,t)dtdz. \tag{3}$$

Here Ω denotes the field of view of the camera sensor, $I(z,k)$ denotes the light intensity of position z at time k and $\mathcal{M}_s(z,t)$ is the binary mask that denotes whether the intensity $I(z,k)$ contributes to the spike s at the time t. That is, if z's corresponding point $z+u_{k\rightarrow t}(z)$ locates in the spatial region that the pixel p covers, $I(z,k)$ contributes to the pixel and $\mathcal{M}_s(z,t)$ is set to 1. Otherwise, $\mathcal{M}_s(z,t)$ is set to 0. Apparently, the mask $\mathcal{M}_s(z,t)$ can be calculated by

$$\mathcal{M}_s(z,t) = \begin{cases} 1, & z+u_{k\rightarrow t}(z)\in\Omega_p \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

where Ω_p denotes the spatial region that the pixel p covers. For simplicity, we use I_k to denote the light intensity of the scene at time k. Considering that $I_k(z)$ is constant in time, the Eq. (3) can be reformulated as

$$\begin{aligned}\theta &= \int_{\Omega}\int_{t_s}^{t_e} \alpha \cdot I_k(z)\cdot \mathcal{M}_s(z,t)dtdz \\ &= \int_{\Omega} \alpha \cdot I_k(z)\left(\int_{t_s}^{t_e} \mathcal{M}_s(z,t)dt\right)dz \\ &= \int_{\Omega} \alpha \cdot I_k(z)\cdot \mathcal{W}_s(z)dz,\end{aligned} \tag{5}$$

with $\mathcal{W}_s(z)=\int_{t_s}^{t_e}\mathcal{M}_s(z,t)dt$ representing how much that $I_k(z)$ contributes to the spike $s:(p,t_s,t_e)$.

4.2. Spike camera SR

Based on the above analysis, the relationship between arbitrary $I_k(z)$ and spike $s:(p,t_s,t_e)$ can be modelled. In order to super-resolve intensity images, we can resample the reconstruction plane into finer grids, and model the relationship between I_k^{HR} and the spike s as:

$$\theta = \sum_q \alpha \cdot I_k^{\text{HR}}(q)\cdot \mathcal{W}_s(q). \tag{6}$$

Here $q=(m,n)$ denotes the coordinate in I_k^{HR} and $\mathcal{W}_s(q)$ denotes how much that $I_k^{\text{HR}}(q)$ contributes to the spike $s:(p,t_s,t_e)$. Once enough spikes are accumulated in a short time interval around k, we can super-resolve the key intensity image I_k^{HR} by minimizing the following loss function $J(I_k^{\text{HR}})$:

$$J(I_k^{\text{HR}}) = \sum_{s=1}^{N}\|\alpha\cdot\mathcal{W}_s I_k^{\text{HR}}-\theta\|_2^2, \tag{7}$$

where N denotes the number of spikes in a selected temporal window. $\mathcal{W}_s\in\mathbb{R}^{1\times M}$ with $M=cH\times cW$ denoting the number of sub-pixels to reconstruct.

To address this problem, we develop a motion-guided spike camera SR (MGSR) framework (as illustrated in Fig. 4) . Firstly, a fundamental light inference algorithm is applied to the spike streams S, producing a sequence of basic intensity images $\{I_t^{\text{LR}}\}, t\in\phi_k$. A typical choice for ϕ_k is $\{k,k\pm1,k\pm2,\cdots\}$. With the basic reconstruction, we can estimate the displacement fields of different frames and map the points in I_k^{HR} to other frames. Then we can further calculate how much that each pixel $I_k^{\text{HR}}(q)$ contributes to each spike $s:(p,t_s,t_e)$, producing a sequence of contribution map $\{\mathcal{W}_s\}$. Based on $\{\mathcal{W}_s\}$, the high resolution intensity image I^{HR} can be easily derived by solving Eq. (7).

Light inference Recall Section 3.1, each spike $s:(p,t_s,t_e)$ corresponds to a certain amount of photons. Assuming that the light intensity in a short spike interval is

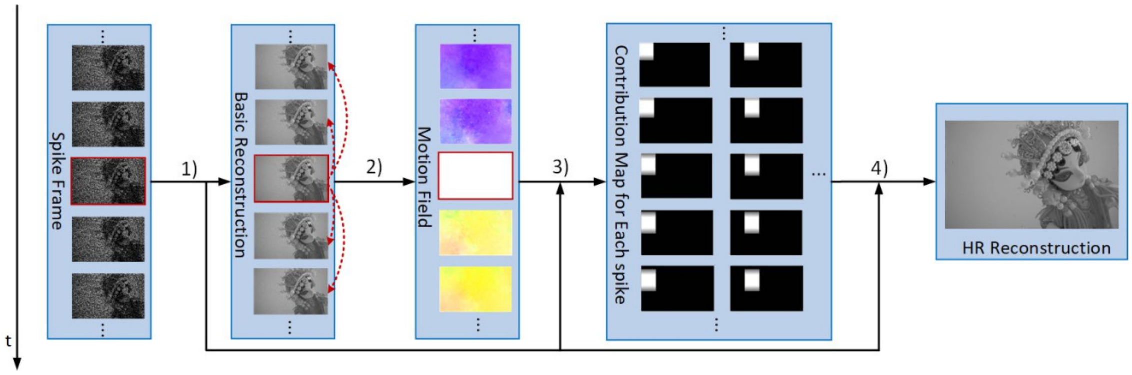


Figure 4. Illustration of the proposed motion-guided spike camera super-resolution (MGSR). The MGSR consists of four main processes: 1) Light inference 2) Motion estimation 3) Contribution weight calculation and 4) Super-resolution Imaging.

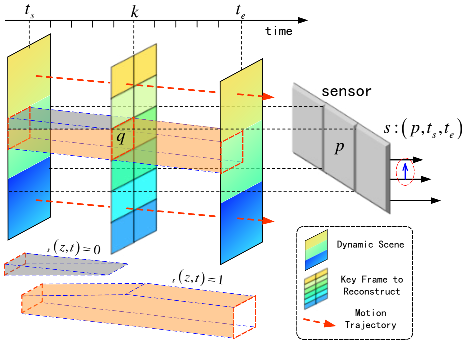


Figure 5. Illustration of contribution weight calculation. The slanted tube passing the image pixel q along the motion trajectory denotes the regions related to $I_k^{\text{HR}}(q)$. Given arbitrary (z,t), if it satisfies that $z + u_{k\to t}(z) \in \Omega_p, t_e < t \le t_s$, it denotes that $I_k^{\text{HR}}(z)$ contributes to the spike $s : (p, t_s, t_e)$ at time t, namely $\mathcal{M}_s(z,t) = 1$. Otherwise, $\mathcal{M}_s(z,t) = 0$. Thus, the volume of the orange tube corresponds the weight that $I_k^{\text{HR}}(q)$ contributes to the spike $s : (p, t_s, t_e)$.

Algorithm 1 MGSR

```
Input: Spike stream frames $\{S_t\}, t \in \phi_k$
Output: HR images $I_k^{\text{HR}}$
1: Infer the fundamental light intensity $\{I_t\}, t \in \phi_k$ using Eq. (8)
2: Estimate relative motion $\{u_{k\to t}\}$, using Eq. (9)
3: Calculate contribution map for each spike according to Eq. (4) and Eq. (10), producing $\{\mathcal{W}_s\}$.
4: Reconstruct high resolution intensity image $I_k^{\text{HR}}$ using Eq. (7) and Eq. (11).
```

stable, we can roughly infer instantaneous light intensity by

$$I_t^{\text{LR}}(p) = \frac{\theta}{\alpha \cdot (t_e - t_s)}, \tag{8}$$

with $t_e < t \le t_s$. It is worth noting that these basic reconstructions are only used for estimating the relative motion.

Motion estimation In the past decades, optical flow algorithms [3, 9, 24–26] have shown great potential for estimating dense motion fields. To achieve the following contribution weight calculation, we employ an optical flow algorithm to the coarse estimates, so as to infer the displacement field between key frame I_k^{LR} and reference frame I_t^{LR}, which is expressed as:

$$u_{k\to t} = \mathcal{F}(I_k^{\text{LR}}, I_t^{\text{LR}}). \tag{9}$$

Here $\mathcal{F}(\cdot)$ denotes the optical flow algorithm. $u_{k\to t} = (u_{k\to t}^h, u_{k\to t}^v)$ denotes the displacement field that maps the pixels in I_k^{LR} to the pixels in I_t^{LR}.

Calculation of contribution weight With the displacement field $u_{k\to t}$, given arbitrary point z, we can easily infer whether $I_k^{\text{HR}}(z)$ contributes to the spike $s : (p, t_s, t_e)$ at the time t $(t_s < t \le t_e)$ via Eq. (4), producing contribution mask $\mathcal{M}_s$. Then we can calculate the weight that each image pixel $I_k^{\text{HR}}(q)$ contributes to the spike $s : (p, t_s, t_e)$ as

$$\mathcal{W}_s(q) = \int_{z\in\Omega_q} \int_{t_s}^{t_e} \mathcal{M}_s(z,t)dtdz, \tag{10}$$

producing the corresponding contribution map $\mathcal{W}_s$. Here Ω_q denotes the spatial region that pixel q covers in I_k^{HR}. Due to the relative motion between the camera sensor and the scene, a spike is generally related to multiple pixels in the I_k^{HR}. The number of related pixels increases with the increase of motion speed and spike life cycle, i.e., $t_e - t_s$. Fig. 5 illustrates the contribution weight calculation and Fig. 6 shows examples of the contribution weight maps with different relative motion.

Super-resolution imaging Once enough spikes accumulated, we can super-resolve a $cH \times cW$ intensity image via solving the optimization problem depicted in Eq.(7). In this paper, we use gradient descent method [22] to solve the problem, which can be formulated as:

$$I_k^{\text{HR}} := I_k^{\text{HR}} - \gamma \cdot \nabla_{I_k^{\text{HR}}} J(I_k^{\text{HR}}; \mathcal{W}_s), \tag{11}$$

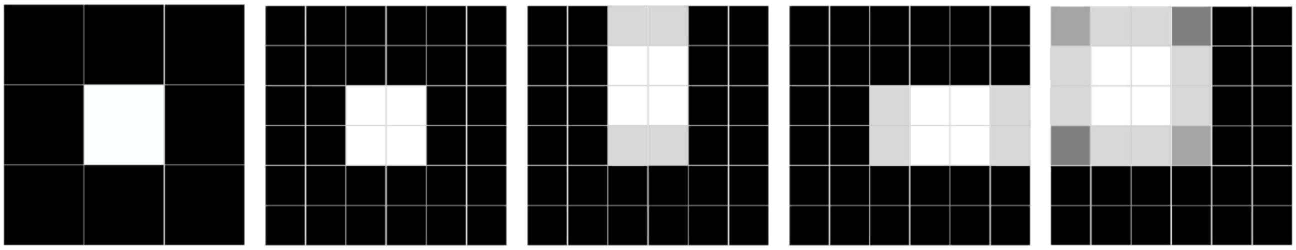

(a) S_t (b) $\mathcal{W}_s^1$ (c) $\mathcal{W}_s^2$ (d) $\mathcal{W}_s^3$ (e) $\mathcal{W}_s^4$

Figure 6. Illustration of contribution weight maps with different motion. For simplicity, we use global motion as an example. Due to the resample, the weight map grid is finer than the spike frame (sensor) grid. (a) The spike frame at time t with a spike fired at center coordinate p. (b) The contribution weight map of I_k^{HR} to spike $S_t(p)$ with relative motion: $u_{k\to t}^h = 0, u_{k\to t}^v = 0$. That is, the scene is static. The spike only corresponds to the region equal to the spike coordinate. (c) The contribution weight map to spike $S_t(p)$ with relative motion: $u_{k\to t}^h = 0, u_{k\to t}^v = v$. Due to vertical movement, the related region expands in vertical direction. The weight in the top and bottom is smaller than the ones in center as the points in the two ends do not always contribute to the spike during the whole spike life cycle. (d) The contribution weight map to spike $S_t(p)$ with relative motion: $u_{k\to t}^h = -v, u_{k\to t}^v = 0$. (e) The contribution weight map to spike $S_t(p)$ with relative motion$u_{k\to t}^h = v, u_{k\to t}^v = v$.

where γ is the update step. In particular, we can also use this algorithm as a general reconstruction algorithm. We can set c as 1 to reconstruct an image with the same spatial resolution to that provided by the spike streams. The proposed MGSR approach is summarized in Algorithm 1.

5. Experimental results

5.1. Setting

Datasets To evaluate the performance of the proposed method, we not only use the real-world PKU-Spike-High-Speed Dataset [1] but also capture several additional spike sequences for the scenes with textures using the FSM spike camera [34]. These spike sequences are captured with 20000HZ or 40000HZ. The spatial resolution of them is 400×250. They 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 "Doll", "Car", "Train". Among them, "Doll" records a doll falling from a height, "Car" describes a car traveling at a speed of 100 km/h and "Train" describes a car traveling at a speed of 350 km/h. Class B includes "Fruits", "Keyboard", "Clock" , "Railway" and "Viaduct-bridge" (V-b). These five sequences are recorded by a spike camera with a very high-speed motion.

Implementation details The number of spikes should increase with the spatial resolution of reconstruction. As discussed above, we use the spikes in a neighbouring time window of the key image. In this paper, the time window size is set to $20 \times c$. We employ the optical flow estimation algorithm proposed by Sun *et. al.* [24] to infer the relative motion. The update step γ in Eq. (11) is initialized to 5×10^{-4} and decreases during the optimization.

5.2. Comparison with the state-of-the-art methods

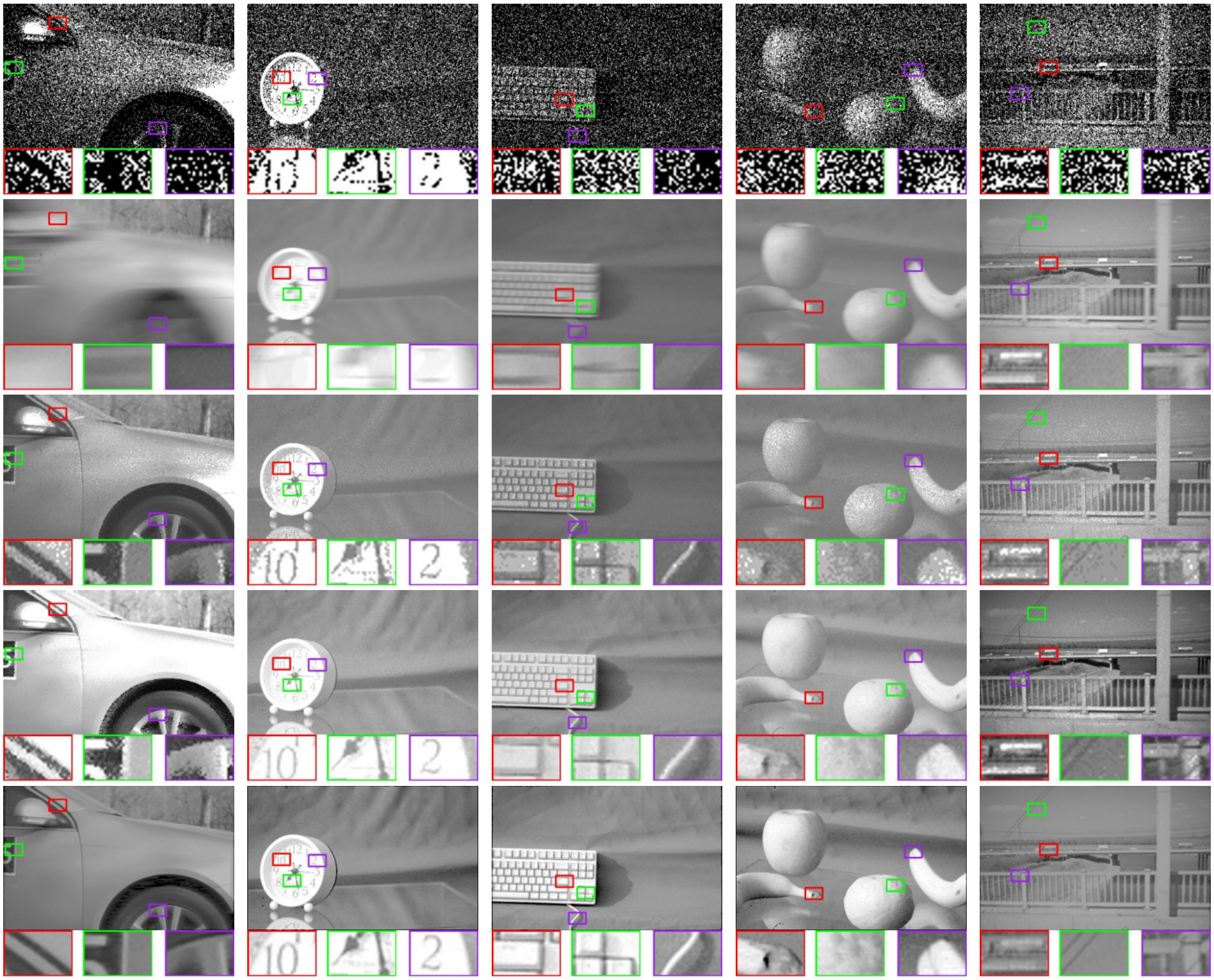
Image reconstruction To evaluate the proposed method, we first compare it with three recent spike camera reconstruction methods, i.e., texture from playback (TFP) [34], texture from interval (TFI) [34] and texture via spiking neural model (TVS) [35].

1) Qualitative evaluation. Fig. 7 shows the reconstruction results of different methods for dynamic scenes. The visual quality of the reconstructions produced by our proposed method is evidently better than the competing methods. Note that there are severe undesired blurry artifacts in the reconstruction of TFP, especially for the regions with high-speed motion. Although the TFI and TVS can well reconstruct the outlines of fast-moving objects, the reconstruction typically appears to be noisy. In contrast, our proposed method achieves stable reconstruction without blurry artifacts or obvious noise.

2) Quantitative evaluation. For quantitative evaluation, we employ two widely used no-reference image quality assessment metrics, namely blind/referenceless image spatial quality evaluator (BRISQUE) [17] and naturalness image quality evaluator (NIQE) [18]. A lower score indicates higher image quality. As illustrated in Table 1, the proposed MGSR outperforms other reconstruction methods in both two metrics.

Super-resolution To evaluate the SR performance, we combine two representative image super-resolution algorithms, i.e. ANR [28] and SRFBN [13], with the competing spike camera reconstruction methods, i.e., TFI and TVS, to super resolve intensity images from spike streams, and compare our proposed MGSR with them. Fig. 8 shows the x4 SR results. We note that the proposed MGSR can reconstruct more details than other reconstruction methods. For instance, we can even reconstruct the letters on the

[1] The dataset is publicly available at https://www.pkuml.org/resources/pku-spike-high-speed.html.



(a) Car (b) Clock (c) Keyboard (d) Fruits (e) V-B

Figure 7. Comparison of different reconstruction methods on real captured spike data. From top to bottom: Spike, TFP [34], TFI [34], TVS [35], MGSR (x1). Please enlarge for more details.

Metric	Method	Class A			Class B					Average
		Doll	Car	Train	Railway	Clock	Keyboard	Fruits	V-B	
BRISQUE (↓)	TFP	**21.91**	29.63	**12.74**	27.33	31.32	38.92	28.47	35.46	28.22
	TFI	43.24	43.39	43.45	37.99	43.46	43.15	43.45	43.10	42.65
	TVS	28.38	41.19	42.02	27.47	42.36	34.81	36.34	32.72	35.66
	MGSR (x1)	30.55	**22.27**	13.05	**21.42**	**28.69**	**23.85**	**21.09**	**25.89**	**23.35**
NIQE (↓)	TFP	8.20	7.64	10.62	6.57	7.84	8.61	9.00	8.15	8.32
	TFI	7.96	13.02	6.49	8.14	22.60	14.73	24.05	10.18	13.42
	TVS	7.48	9.31	6.78	7.01	13.42	11.36	12.43	8.23	9.50
	MGSR (x1)	**4.51**	**4.38**	**3.14**	**4.80**	**6.28**	**6.14**	**6.22**	**6.51**	**5.25**

Table 1. Comparison among different reconstruction methods.

keyboard. This is because that for the pipelined methods, the information of scene details have already been lost in the first stage and the following image super-resolution can hardly restore it. Different from these methods, we directly super resolve HR images from the spike streams, which can exploit the photoelectric information more efficiently.

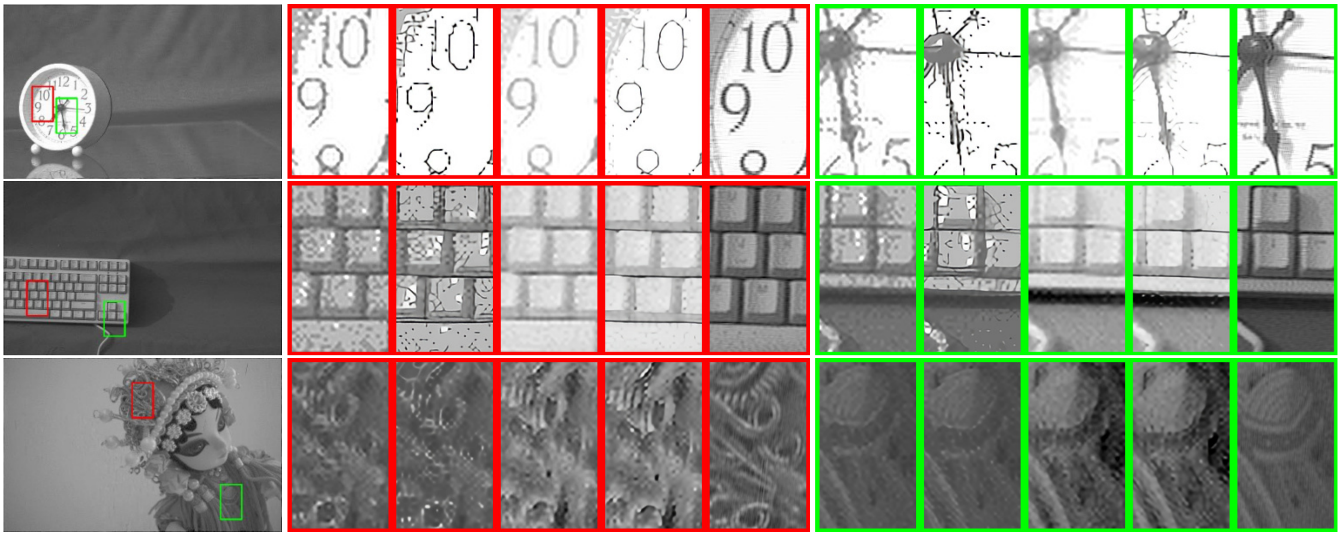


Figure 8. x4 SR results in comparison to the state-of-the-art methods. From left to right: TFI [34]+ANR [28], TFI [34]+SRFBN [13], TVS [35]+ANR [28], TVS [35]+SRFBN [13] and the proposed MGSR.

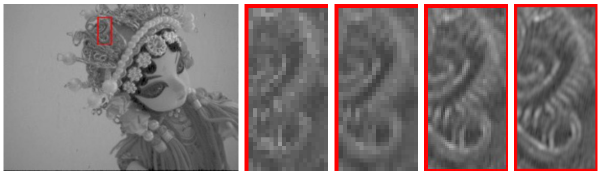


Figure 9. Illustration of spatial resolution gain. From left to right: Full image, TFI, MGSR(x1), MGSR(x2), MGSR(x4)

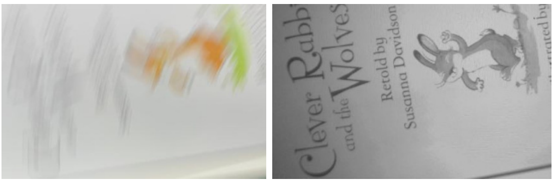


Figure 10. Visual comparison for a high-speed motion scene, where a book fell from a height. Left: image captured by iPhone 11. Right: spike camera reconstruction using MGSR (x2).

In addition, to further show the spatial resolution gain, we use the proposed MGSR to reconstruct the images with different upscale factors, i.e., x1, x2, x4. As illustrated in Fig. 9, with the increase of upscale factor, the reconstructed image contains more details, which demonstrates the effectiveness of our proposed MGSR.

5.3. Comparison with the conventional camera

We also compare the bio-inspired spike camera with the conventional CMOS camera. Fig. 10 shows the visual comparison for a high-speed motion scene, where a book fell from a height. We note that the image captured by the iPhone 11 is blurred. The reason may be that the conventional camera accumulated the photoelectric within the whole exposure window to form a snapshot, ignoring the object motion. Different from the conventional camera, the spike camera produce continuous spike streams to record the high-speed dynamic scene. By properly modeling the motion and temporal correlation, we can reconstruct a clear image (as shown in Fig. 10) for arbitrary time.

6. Conclusion

In this paper, we develop the first framework to super resolve high-speed dynamic scenes from LR spike streams. Due to the relative motion between camera sensor and the objects, each spike can be mapped to multiple points in the external scene, which provides an opportunity to recover the scene with a much higher resolution than that provided by spike streams. By analyzing the working mechanism of spike camera, we derive the relationship between light intensity and each spike, so as to super resolve high-quality intensity images from the spike streams. Experiments on real-life captured spike data show that the proposed framework reconstructs high-quality images with fine details in comparison to the-state-of-the-art methods in both the same size image reconstruction and super-resolution.

Acknowledge

This work was supported in part by the National Key Research and Development Program of China under Grant 2020AAA0130401 and the National Natural Science Foundation of China under Grant 62072009, 62088102, 61772041, and was also supported by the Cloud Brain II of Peng Cheng Lab, Shenzhen, China.

References

[1] Christian Brandli, Raphael Berner, Minhao Yang, Shih-Chii Liu, and Tobi Delbruck. A 240x180 130db 3us latency global shutter spatiotemporal vision sensor. *IEEE Journal of Solid-State Circuits*, 49(10):2333–2341, 2014.

[2] Christian Brandli, Lorenz Muller, and Tobi Delbruck. Real-time, high-speed video decompression using a frame-and event-based davis sensor. In *IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 686–689, 2014.

[3] Thomas Brox, Andrés Bruhn, Nils Papenberg, and Joachim Weickert. High accuracy optical flow estimation based on a theory for warping. In *European Conference on Computer Vision (ECCV)*, pages 25–36, 2004.

[4] Jose Caballero, Christian Ledig, Andrew Aitken, Alejandro Acosta, Johannes Totz, Zehan Wang, and Wenzhe Shi. Real-time video super-resolution with spatio-temporal networks and motion compensation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4778–4787, 2017.

[5] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *European Conference on Computer Vision (ECCV)*, pages 184–199, 2014.

[6] Siwei Dong, Tiejun Huang, and Yonghong Tian. Spike camera and its coding methods. In *Data Compression Conference (DCC)*, pages 437–437, 2017.

[7] Siwei Dong, Lin Zhu, Daoyuan Xu, Yonghong Tian, and Tiejun Huang. An efficient coding method for spike camera using inter-spike intervals. *arXiv preprint arXiv:1912.09669*, 2019.

[8] Weisheng Dong, Lei Zhang, Rastislav Lukac, and Guangming Shi. Sparse representation based image interpolation with nonlocal autoregressive modeling. *IEEE Transactions on Image Processing*, 22(4):1382–1394, 2013.

[9] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2758–2766, 2015.

[10] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Deep back-projection networks for super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1664–1673, 2018.

[11] Younghyun Jo, Seoung Wug Oh, Jaeyeon Kang, and Seon Joo Kim. Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3224–3232, 2018.

[12] Armin Kappeler, Seunghwan Yoo, Qiqin Dai, and Aggelos K Katsaggelos. Video super-resolution with convolutional neural networks. *IEEE transactions on computational imaging*, 2(2):109–122, 2016.

[13] Zhen Li, Jinglei Yang, Zheng Liu, Xiaomin Yang, Gwanggil Jeon, and Wei Wu. Feedback network for image super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3867–3876, 2019.

[14] Renjie Liao, Xin Tao, Ruiyu Li, Ziyang Ma, and Jiaya Jia. Video super-resolution via deep draft-ensemble learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 531–539, 2015.

[15] Patrick Lichtsteiner, Christoph Posch, and Tobi Delbruck. A 128×128 120db 15us latency asynchronous temporal contrast vision sensor. *IEEE Journal of Solid-State Circuits*, 43(2):566–576, 2008.

[16] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 136–144, 2017.

[17] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing*, 21(12):4695–4708, 2012.

[18] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a “completely blind” image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012.

[19] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6820–6829, 2019.

[20] Bharath Ramesh, Andrés Ussa, Luca Della Vedova, Hong Yang, and Garrick Orchard. Low-power dynamic object detection and classification with freely moving event cameras. *Frontiers in Neuroscience*, 14:135, 2020.

[21] Bharath Ramesh and Hong Yang. Boosted kernelized correlation filters for event-based face detection. In *IEEE Winter Conference on Applications of Computer Vision Workshops (WACVW)*, pages 155–159, 2020.

[22] Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016.

[23] Cedric Scheerlinck, Nick Barnes, and Robert Mahony. Continuous-time intensity estimation using event cameras. In *Asian Conference on Computer Vision (ACCV)*, pages 308–324, 2018.

[24] Deqing Sun, Stefan Roth, and Michael J Black. Secrets of optical flow estimation and their principles. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2432–2439, 2010.

[25] Deqing Sun, Stefan Roth, JP Lewis, and Michael J Black. Learning optical flow. In *European Conference on Computer Vision (ECCV)*, pages 83–97, 2008.

[26] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8934–8943, 2018.

[27] Xin Tao, Hongyun Gao, Renjie Liao, Jue Wang, and Jiaya Jia. Detail-revealing deep video super-resolution. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4472–4480, 2017.

[28] Radu Timofte, Vincent Desmet, and Luc Vangool. A+: Adjusted anchored neighborhood regression for fast super-resolution. In *Asian Conference on Computer Vision (ACCV)*, 2014.

[29] Jun Xie, Rogerio Schmidt Feris, and Ming-Ting Sun. Edge-guided single depth image super resolution. *IEEE Transactions on Image Processing*, 25(1):428–438, 2015.

[30] Xiangjun Zhang and Xiaolin Wu. Image interpolation by adaptive 2-d autoregressive modeling and soft-decision estimation. *IEEE Transactions on Image Processing*, 17(6):887–896, 2008.

[31] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *European Conference on Computer Vision (ECCV)*, pages 286–301, 2018.

[32] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2472–2481, 2018.

[33] Jing Zhao, Ruiqin Xiong, and Tiejun Huang. High-speed motion scene reconstruction for spike camera via motion aligned filtering. In *IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–5, 2020.

[34] 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 (ICME)*, pages 1432–1437, 2019.

[35] Lin Zhu, Siwei Dong, Jianing Li, Tiejun Huang, and Yonghong Tian. Retina-like visual image reconstruction via spiking neural model. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1438–1446, 2020.