Learning to Robustly Reconstruct Dynamic Scenes from Low-light Spike Streams

Liwen Hu[1], Ziluo Ding[2], Mianzhi Liu[3], Lei Ma[1,3⋆], and Tiejun Huang[1]

[1] State Key Laboratory of Multimedia Information Processing, School of Computer Science, Peking University {huliwen, lei-ma, tjhuang}@pku.edu.cn
[2] Beijing Academy of Artificial Intelligence
{ziluoding}@baai.ac.cn
[3] College of Future Technology, Peking University
{liumianzhi}@stu.pku.edu.cn

Abstract. Spike camera with high temporal resolution can fire continuous binary spike streams to record per-pixel light intensity. By using reconstruction methods, the scene details in high-speed scenes can be restored from spike streams. However, existing methods struggle to perform well in low-light environments due to insufficient information in spike streams. To this end, we propose a bidirectional recurrent-based reconstruction framework to better handle such extreme conditions. In more detail, a **l**ight-**r**obust **rep**resentation (LR-Rep) is designed to aggregate temporal information in spike streams. Moreover, a fusion module is used to extract temporal features. Besides, we synthesize a reconstruction dataset for high-speed low-light scenes where light sources are carefully designed to be consistent with reality. The experiment shows the superiority of our method. Importantly, our method also generalizes well to real spike streams. Our project is: https://github.com/Acnext/Learning-to-Robustly-Reconstruct-Dynamic-Scenes-from-Low-light-Spike-Streams/.

Keywords: Spike camera · Reconstruction

1 Introduction

As a neuromorphic sensor with high temporal resolution (40,000 Hz), spike camera [14, 41] has shown enormous potential for high-speed visual tasks, such as reconstruction [3, 5–7, 36, 37, 40, 42, 43], optical flow estimation [13, 33, 39], and depth estimation [19, 21, 35]. Different from event cameras [1, 4, 20], it can record per-pixel light intensity by accumulating photons and firing continuous binary spike streams. Correspondingly, high-speed dynamic scenes can be reconstructed from spike streams. Recently, many deep learning methods have advanced this field and shown great success in reconstructing more detailed scenes. However, existing methods struggle to perform well in low-light high-speed scenes due to insufficient information in spike streams.

⋆ Corresponding author.

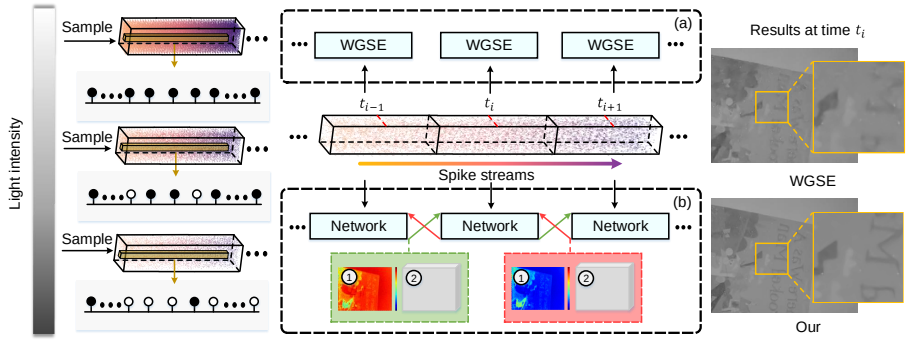

Fig. 1: Overview of reconstruction for high-speed spike streams. **Left**: with decreasing light intensity, more sparse spike streams are difficult to extract features. A black circle is a spike. **Middle**: (a) The state-of-the-art method, WGSE [34]. The arrow with a gradient color is the timeline. (b) Our reconstruction method. Green (red) lines denote the forward (backward) data flow. ① (②) is the release time of spikes (temporal features). ① (②) in forward and backward data flow is independent. **Right**: reconstructed results from WGSE and our method.

A dilemma arises for visual sensors, that is, the quality of sampled data can greatly decrease in a low-light environment [10, 11, 17, 18, 39]. Low-quality data creates many difficulties for all kinds of vision tasks. Similarly, the reconstruction for the spike camera also suffers from this problem. To improve the performance of reconstruction in low-light high-speed scenes, two non-trivial matters should be carefully considered. First, constructing a low-light high-speed scene dataset for spike camera is crucial to evaluating different methods. However, due to the frame rate limitations of traditional cameras, it is difficult to capture images clearly in real high-speed scenes as supervised signals. Instead of it, a reasonable way is to synthesize datasets for spike camera [13, 35, 37, 43]. To ensure the reliability of the reconstruction dataset, synthetic low-light high-speed scenes should be as consistent as possible with the real world, *e.g.* light source. Second, as shown in Fig. 1, with the decrease of illuminance in the environment, the total number of spikes in spike streams decreases greatly which means the valid information in spike streams can greatly decrease. Fig. 1(a) shows that the state-of-the-art method often fail under low-light conditions since they have no choice but to rely on inadequate information.

In this work, we aim to address all two issues above-mentioned. In more detail, a reconstruction dataset for high-speed low-light scenes is proposed. We carefully design the scene by controlling the type and power of the light source and generating noisy spike streams based on [38]. Besides, we propose a light-robust reconstruction method as shown in Fig. 1(b). Specifically, to compensate for information deficiencies in low-light spike streams, we propose a **l**ight-**r**obust **rep**resentation (LR-Rep). In LR-Rep, the release time of forward and backward spikes is used to update a **g**lobal **i**nter-**s**pike **i**nterval (GISI). Then, to further

excavate temporal information in spike streams, LR-Rep is fused with forward (backward) temporal features. During the feature fusion process, we add alignment information to avoid the misalignment of motion from different timestamps. Finally, the scene is clearly reconstructed from fused features.

Empirically, we show the superiority of our reconstruction method. Importantly, our method also generalizes well to real spike streams. In addition, extensive ablation studies demonstrate the effectiveness of each component. The main contributions of this paper can be summarized as follows:

• A reconstruction dataset for high-speed low-light scenes is proposed. We carefully construct varied low-light scenes that are close to reality.

• We propose a bidirectional recurrent-based reconstruction framework where a light-robust representation, LR-Rep, and fusion module can effectively compensate for information deficiencies in low-light spike streams.

• Experimental results on real and synthetic datasets have shown our method can more effectively handle spike streams in high-speed low-light scenes than previous methods.

2 Related Work

2.1 Low-light Vision

Low-light environment has always been a challenge not only for human perception but also for computer vision methods. For **traditional cameras**, some works [2, 9, 11, 15, 25, 27, 29, 31] mainly concern the enhancement of low-light images. Wei *et al.* [27] propose the LOL dataset containing low/normal-light image pairs and propose a deep Retinex-Net including a Decom-Net for decomposition and an Enhance-Net for illumination adjustment. Guo *et al.* [11] proposes Zero-DCE which formulates light enhancement as a task of image-specific curve estimation with a deep network. Retinexformer [2] formulates a simple yet principled One-stage Retinex-based Framework to light up low-light images. Besides, some work focuses on the robustness of vision tasks to low-light, *e.g.* object detection. Wang *et al.* [24] combines with the image enhancement algorithm to improve the accuracy of object detection. For **spike camera**, it is also affected by low-light environments. Dong *et al.* [8] propose a real low-light high-speed dataset for reconstruction. However, it lacks corresponding image sequences as ground truth. Besides, the concurrent work [44] synthesizes a low-light spike stream dataset. However, it only contains static scenes and cannot evaluate the performance of reconstruction methods during motion.

2.2 Reconstruction for Spike Camera

The reconstruction of high-speed dynamic scenes has been a popular topic for spike camera. Based on the statistical characteristics of spike stream, Zhu *et al.* [41] first reconstruct high-speed scenes. Zhao *et al.* [36] improved the smoothness of reconstructed scenes by introducing motion aligned filter. Zhu *et*

al. [42] construct a dynamic neuron extraction model to distinguish the dynamic and static scenes. With the rise of spiking neural networks [22, 23, 30, 45], for enhancing reconstruction results, Zheng *et al.* [40] uses short-term plasticity mechanism to exact motion area. Zhao *et al.* [37] first proposes a deep learning-based reconstruction framework, Spk2ImgNet (S2I), to handle the challenges brought by both noise and high-speed motion. Chen *et al.* [3] build a self-supervised reconstruction framework by introducing blind-spot networks. It achieves desirable results compared with S2I. The reconstruction method [34] presents a novel Wavelet Guided Spike Enhancing (WGSE) paradigm. By using multi-level wavelet transform, the noise in the reconstructed results can be effectively suppressed. Besides, we would like to mention the concurrent work, RSIR [44]. In RSIR, the AST representation is used to adaptively extract the number of spike in a spike stream under different illuminations. Then, a multi-scale wavelet recurrent network can reconstruct images from the AST representation. However, AST compresses a spike stream into a spike number map which ignores dynamic information, resulting in more motion blur for high-speed low-light scenes. This greatly limits the contribution of RSIR to spike camera reconstruction, as the original intention of spike camera is to handle high-speed dynamic scenes. Unlike AST, our proposed representation, LR-Rep, first calculates global inter-spike interval map (GISI). It can better preserve dynamic information while aggregating temporal information.

2.3 Spike Camera Simulation

Spike camera simulation is a popular way to generate spike streams and accurate labels. Zhao *et al.* [37] first convert interpolated image sequences with high frame rate into spike stream. Based on [37], the simulators [16, 38, 43] add some random noise to generate spike streams more accurately. To avoid motion artifacts caused by interpolation, Hu *et al.* [13] presents the spike camera simulator (SPCS) combining simulation function and rendering engine tightly. Then, based on SPCS, optical flow datasets for spike camera are first proposed. Zhang *et al.* [35] generate the first spike-based depth dataset by the spike camera simulation. Zhang *et al.* [34] generate the first semantic segmentation spike streams dataset by the spike camera simulation.

3 Reconstruction Datasets

In order to train and evaluate the performance of reconstruction methods in low-light high-speed scenes, we propose two low-light spike stream datasets, **R**and **L**ow-**L**ight **R**econstruction (RLLR) and **L**ow-**L**ight **R**econstruction (LLR) based on spike camera model. RLLR is used as our train dataset and LLR is carefully designed to evaluate the performance of different reconstruction methods as test dataset. We first introduce the spike camera model, and then introduce our datasets where noisy spike streams are generated by the spike camera model.

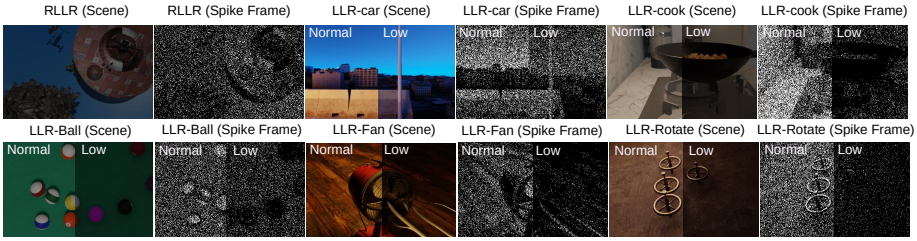

Fig. 2: Proposed datasets, RLLR and LLR. RLLR includes random scenes and LLR includes carefully designed scenes. A Spike Frame is a slice of generated spike streams on a temporal axis.

Spike camera model Each pixel on the spike camera model converts light signal into the current signal and accumulates the input current. For pixel $\mathbf{x} = (x, y)$, if the accumulation of input current reaches a fixed threshold ϕ, a spike is fired and then the accumulation can be reset as,

$$A(\mathbf{x}, t) = A_{\mathbf{x}}(t) \bmod \phi = \int_0^t I_{tot}(\mathbf{x}, \tau) d\tau \bmod \phi, \tag{1}$$

$$I_{tot}(\mathbf{x}, \tau) = I_{in}(\mathbf{x}, \tau) + I_{dark}(\mathbf{x}, \tau), \tag{2}$$

where $A(\mathbf{x}, t)$ is the accumulation at time t, $A_{\mathbf{x}}(t)$ is the accumulation without reset before time t, $I_{in}(\mathbf{x}, \tau)$ is the input current at time τ (proportional to light intensity) and $I_{dark}(\mathbf{x}, \tau)$ is the main fixed pattern noise in spike camera, *i.e.* dark current [12, 38, 43]. Further, due to limitations of circuits, each spike is read out at discrete time $nT, n \in \mathbb{N}$ (T is a micro-second level). Thus, the output of the spike camera is a spatial-temporal binary stream S with $H \times W \times N$ size. The H and W are the height and width of the sensor, respectively, and N is the temporal window size of the spike stream. According to the spike camera model, it is natural that the spikes (or information) in low-light spike streams are sparse because reaching the threshold is lengthy.

RLLR As shown in Fig. 2, RLLR includes 100 random low-light high-speed scenes where high-speed scenes are first generated by SPCS [13] and then the light intensity of all pixels in each scene is darkened by multiplying a random constant (0-1). Each scene in RLLR continuously records a spike stream with $400 \times 250 \times 1000$ size and corresponding image sequence. Then, for each image, we clip a spike stream with $400 \times 250 \times 41$ size from the spike stream as input.

LLR As shown in Fig. 2, LLR includes 5×2 carefully designed high-speed scenes where we use the scenes with five kinds of motion (named Ball, Car, Cook, Fan, and Rotate) and each scene corresponds to two light sources (normal and low). To ensure the reliability of our scenes, different light sources are used, and the power of the light source is consistent with the real world. Besides, the motion in Ball, Cook, Fan, and Rotate is from [13] while the motion in Car is created based on vehicle speed in the real world. Hence, the motion of objects is close to the real world. Each scene in LLR continuously records 21 spike streams with

$400 \times 250 \times 41$ size and 21 corresponding images. In the proposed datasets, we consider the noise of spike camera based on [38].

4 Method

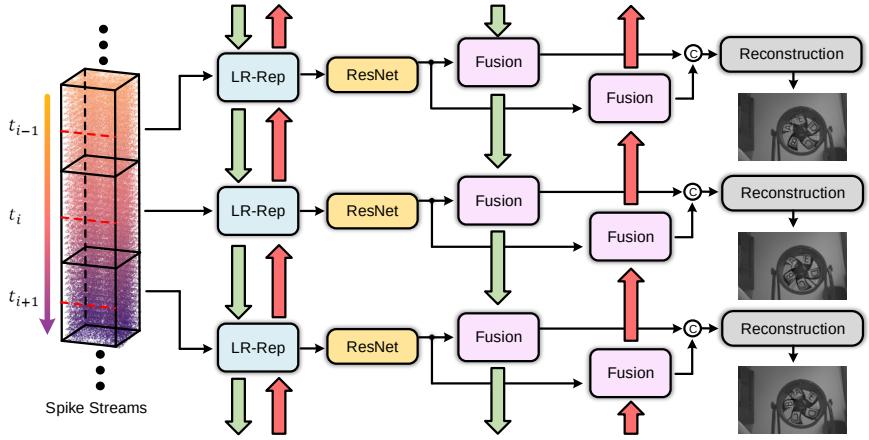

Fig. 3: Illustration of the proposed bidirectional recurrent-based reconstruction framework. It includes a light-robust representation, feature extractor (ResNet), fusion, and reconstruction. The green and red lines represent the forward and backward data flow. The two kinds of data flow are independent.

4.1 Problem Statement

For simplicity, we write $\mathbf{S}_t \in \{0,1\}^{H\times W\times(2\Delta t+1)}$ to denote a spike stream from time $t-\Delta t$ to $t+\Delta t$ ($2\Delta t+1$ is the fixed temporal window) and write $\mathbf{Y}_t \in \mathbb{R}^{H\times W}$ to denote the instantaneous light intensity received in spike camera at time t. Reconstruction is to use continuous spike streams, $\{\mathbf{S}_{t_i}, t_i = i*(2\Delta t+1)|i = 1,2,3...K\}$ to restore the light intensity information at different time, $\{\mathbf{Y}_{t_i}, t_i = i*(2\Delta t+1)|i = 1,2,3...K\}$. Generally, the temporal window $2\Delta t+1$ is set as 41 which is the same with [3, 34, 37].

4.2 Overview

To overcome the challenge of low-light spike streams, *i.e.* the recorded information is sparse (see Fig.1), we propose a light-robust reconstruction method that can fully utilize temporal information of spike streams. It is beneficial from two modules: 1. A light-robust representation, LR-Rep. 2. A fusion module. As shown

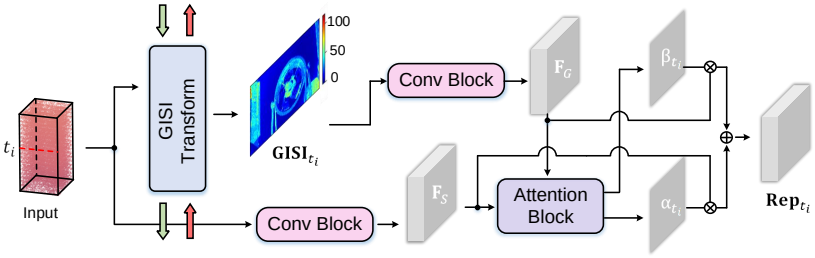

Fig. 4: Illustration of the proposed light-robust representation. We use convolution blocks to extract shallow features from input spike stream and GISI, respectively. Then they are fused by an attention block.

in Fig. 3, to recover the light intensity information at time t_i, $\mathbf{Y}_{t_i}$, we first calculate the light-robust representation at time t_i, written as $\mathbf{Rep}_{t_i}$. Then, we use a ResNet module to extract deep features, $\mathbf{F}_{t_i}$, from $\mathbf{Rep}_{t_i}$. $\mathbf{F}_{t_i}$ is fused with forward (backward) temporal features as $\mathbf{F}^f_{t_i}$ ($\mathbf{F}^b_{t_i}$). Finally, we reconstruct the image at time t_i, $\hat{\mathbf{Y}}_{t_i}$ with $\mathbf{F}^f_{t_i}$ and $\mathbf{F}^b_{t_i}$.

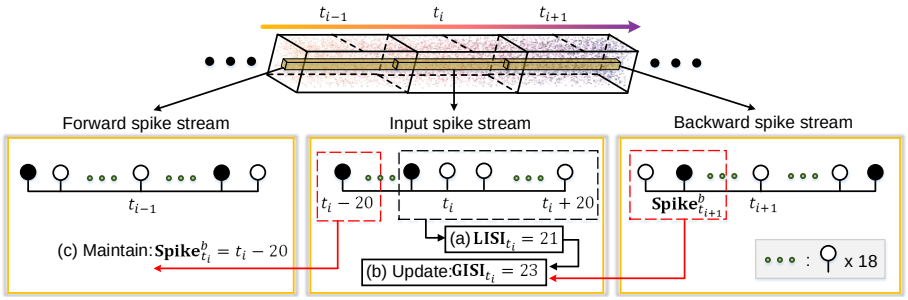

Fig. 5: Illustration of GISI transform for backward in a pixel. (a). Calculate the local inter-spike interval, $\mathbf{LISI}_{t_i}$ from the input spike stream [3, 39]. (b). Update global inter-spike interval, $\mathbf{GISI}_{t_i}$ based on the release time of backward spike, $\mathbf{Spike}^b_{t_{i+1}}$ and $\mathbf{LISI}_{t_i}$. (c). Maintain and transmit the release time of backward spike, $\mathbf{Spike}^b_{t_i}$. Black (white) circle is a (no) spike and the red line is backward data flow.

4.3 Light-robust Representation

As shown in Fig. 4, a light-robust representation, LR-Rep, is proposed to aggregate the information in low-light spike streams. LR-Rep mainly consists of two parts, GISI transform and feature extraction.

GISI transform Calculating the local inter-spike interval from the input spike stream is a common operation [3, 39] and we call it as LISI transform. Different

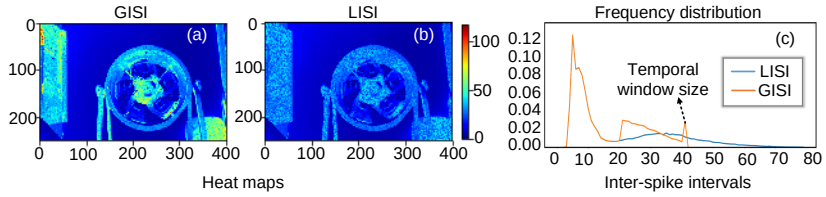

Fig. 6: (a) and (b) show the visualizations of $\mathbf{GISI}_{t_i}$ and $\mathbf{LISI}_{t_i}$ in a real spike stream. (c) shows the distribution of pixel-wise values in $\mathbf{GISI}_{t_i}$ and $\mathbf{LISI}_{t_i}$.

from LISI transform, we propose a GISI transform that can utilize the release time of forward and backward spikes to obtain the global inter-spike interval $\mathbf{GISI}_{t_i}$. It needs to be performed twice, i.e. once forward and once backward respectively. Taking GISI transform backward as an example, it can be summarized as three steps as shown in Fig. 5. GISI transform can extract more temporal information from spike streams than LISI transform as shown in Fig. 6.

Feature extraction After GISI transform, we separately extract shallow features of $\mathbf{GISI}_{t_i}$ and input spike stream, $\mathbf{F}_G$ and $\mathbf{F}_S$ through convolution block. Finally, $\mathbf{Rep}_{t_i}$ is obtained by an attention module where $\mathbf{F}_G$ and $\mathbf{F}_S$ are integrated, *i.e.*

$$[\beta_{t_i}, \alpha_{t_i}] = Att([\mathbf{F}_G, \mathbf{F}_S]), \tag{3}$$

$$\mathbf{Rep}_{t_i} = \beta_{t_i}\mathbf{F}_G + \alpha_{t_i}\mathbf{F}_S, \tag{4}$$

where $Att(\cdot)$ denotes an attention block including 3-layer convolution with 3-layer activation function and $\mathbf{Rep}_{t_i}$ is our LR-Rep at time t_i.

4.4 Fusion and Reconstruction

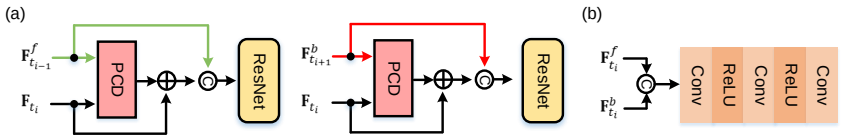

Fig. 7: Illustration of fusion module and reconstruction module. (a) denotes forward (green line) and backward (red line) fusion modules. (b) denotes the reconstruction module.

We first extract the deep feature $\mathbf{F}_{t_i}$ of $\mathbf{Rep}_{t_i}$ through a ResNet with 16 layers. Then, as shown in Fig. 7(a), for forward, temporal features $\mathbf{F}^f_{t_{i-1}}$ and $\mathbf{F}_{t_i}$ are fused as temporal features of the input spike stream $\mathbf{F}^f_{t_i}$. For backward,

temporal features $\mathbf{F}^b_{t_{i+1}}$ and $\mathbf{F}_{t_i}$ are fused as temporal features of the input spike stream $\mathbf{F}^b_{t_i}$. To avoid the misalignment of motion from different timestamps, we use a Pyramid Cascading and Deformable convolution (PCD) [26] to add alignment information to $\mathbf{F}_{t_i}$. The above process can be written as,

$$\mathbf{F}_{t_i} = f(\mathbf{Rep}_{t_i}), \tag{5}$$

$$\mathbf{F}^f_{t_i} = f([\mathbf{F}_{t_i} + a(\mathbf{F}^f_{t_{i-1}}, \mathbf{F}_{t_i}), \mathbf{F}^f_{t_{i-1}}]), \tag{6}$$

$$\mathbf{F}^b_{t_i} = f([\mathbf{F}_{t_i} + a(\mathbf{F}^b_{t_{i+1}}, \mathbf{F}_{t_i}), \mathbf{F}^b_{t_{i+1}}]), \tag{7}$$

where $f(\cdot)$ denotes the feature extraction and $a(\cdot,\cdot)$ denotes the PCD module. Finally, as shown in Fig. 7(b), we use forward and backward temporal features ($\mathbf{F}^b_{t_i}$ and $\mathbf{F}^f_{t_i}$) to reconstruct the current scene at time t_i, *i.e.*

$$\hat{\mathbf{Y}}_{t_i} = c([\mathbf{F}^b_{t_i}, \mathbf{F}^f_{t_i}]), \tag{8}$$

$$\mathcal{L} = \sum_{i=1}^{K} \|\hat{\mathbf{Y}}_{t_i} - \mathbf{Y}_{t_i}\|_1 \tag{9}$$

where $c(\cdot)$ denotes 3-layer convolution with 2-layer ReLU, $\mathcal{L}$ is the loss function, $\|\cdot\|_1$ denotes 1-norm and K is the number of continuous spike streams.

5 Experiment

5.1 Implementation Details

We train our method in the proposed dataset, RLLR. Consistent with previous work [3,34,37], the temporal window of each input spike stream is 41. The spatial resolution of input spike streams is randomly cropped the spike stream to 64×64 during the training procedure and the batch size is set as 8. Besides, forward (backward) temporal features and the release time of spikes in our method are maintained from 21 continuous spike streams. We use Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.99$. The learning rate is initially set as 1e-4 and scaled by 0.1 after 70 epochs. The model is trained for 100 epochs on 1 NVIDIA A100-SXM4-80GB GPU.

5.2 Results

We compare our method with traditional reconstruction methods, *i.e.* TFI [41], STP [40], SNM [42] and deep learning-based reconstruction methods, *i.e.* SSML [3], Spk2ImgNet (S2I) [37], WGSE [34], concurrent work RSIR [44]. The supervised learning methods, S2I, WGSE and RSIR, are trained on RLLR. We evaluate methods on two kinds of data:
(1) The carefully designed **synthetic** dataset, LLR.
(2) The **real** spike streams dataset, PKU-Spike-High-Speed [37] and low-light high-speed spike streams dataset [8].

Table 1: PSNR and SSIM of reconstruction results on synthetic dataset, LLR. The best performance is bolded. Note that high PSNR (S2I, WGSE, and Ours are above 40) is a normal occurrence due to low-light scenes (see appendix).

Metric	TFI ICME'19	RSIR MM'23	SSML IJCAI'22	S2I CVPR'21	STP CVPR'21	SNM PAMI'22	WGSE AAAI'23	Ours This paper
PSNR	31.409	34.121	38.432	40.883	24.882	25.741	42.959	**45.075**
SSIM	0.72312	0.88337	0.89942	0.95915	0.55537	0.80281	0.97066	**0.98681**

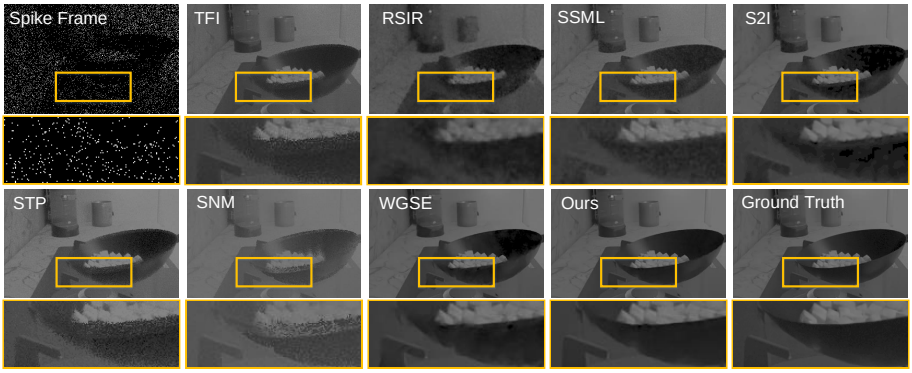

Fig. 8: An example of different methods on the LLR. Spike Frame is a slice of input spike streams on temporal axis. Please enlarge for details. More results are in appendix.

The **reproduction** of these methods is from their official source codes.

Results on our synthetic dataset As shown in Table. 1, we use the two reference image quality assessment (IQA) metrics, *i.e.* PSNR and SSIM to evaluate the performance of different methods on LLR. We can find that our method achieves the best reconstruction performance and has a PSNR gain over 2dB than the state-of-the-art reconstruction method, WGSE, which demonstrates its effectiveness. Fig. 8 shows the visualization results from different reconstruction methods. We can find that our method can better restore motion details in low-light motion regions than other methods. Besides, RSIR is designed to handle spike streams in static scenes and we find that it can suffer from large motion blur in low-light high-speed scenes.

Results on real datasets For real data, we test different methods on two spike stream datasets, PKU-Spike-High-Speed [37] and low-light spike streams [8]. PKU-Spike-High-Speed includes 4 high-speed scenes under normal-light conditions and [8] includes 5 high-speed scenes under low-light conditions. Fig. 9 shows the reconstruction results. Note that we apply the traditional enhancement method [32] to reconstruction results on [8] because scenes are too dark. We can find that, for high-speed scenes under normal-light conditions, deep learning-based methods (SSML, RSIR, S2I, WGSE, and Ours) can reconstruct scene details well. However, for high-speed scenes under low-light conditions, SSML and RSIR introduce a large amount of motion blur while S2I and WGSE may

introduce some artifacts in dark backgrounds. Our method can more effectively restore the information in scenes, i.e., clear texture.

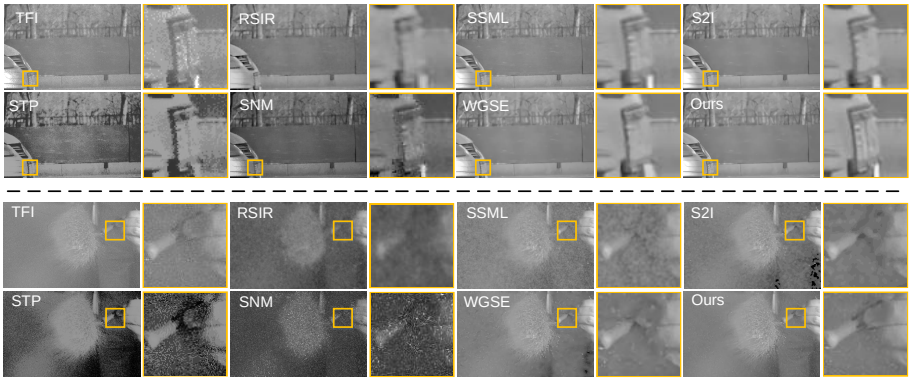

Fig. 9: Results from different reconstruction methods on the real datasets, PKU-Spike-High-Speed [37] (Top) and low-light high-speed spike streams dataset [8] (Bottom). For low-light high-speed spike streams dataset, we apply the traditional enhancement method [32] to reconstruction results because the scene is too dark. More results are in our appendix.

As shown in Fig. 10, we perform a user study written as US [15,28] to quantify the visual quality of different methods. For each scene in datasets, we randomly select reconstructed images at the same time from different methods and display them on the screen (the image order is randomly shuffled). 20 human subjects (university degree or above) are invited to independently score the visual quality of the reconstructed image. The scores of visual quality range from 1 to 8 (worst to best quality). The average subjective scores for each spike stream dataset are shown in Fig. 10 and our method reaches the highest US score in all methods.

Temporal consistency of reconstructed results Our reconstruction method is stable to spike stream at different moments. Fig. 11 shows the continuous reconstructed results in a real high-speed low-light scene. We find that our method can recover scene details at different moments, while the state-of-the-art WGSE introduces temporal-varying artifacts. Besides, we also provide a reconstruction video in our supplementary material.

5.3 Ablation

Proposed modules To investigate the effect of the proposed light-robust representation LR-Rep, the **a**djacent (forward and backward) **d**eep temporal **f**eatures (ADF), *i.e.* $\mathbf{F}^b_{t_i}$ and $\mathbf{F}^f_{t_i}$ in our fusion module, the **a**lignment **i**nformation in our **f**usion module (AIF) and GISI transform in LR-Rep, we compare 5 baseline methods with our final method. (A) is the basic baseline without LR-Rep, ADF, and AIF. Table. 2 shows ablation results on the proposed dataset,

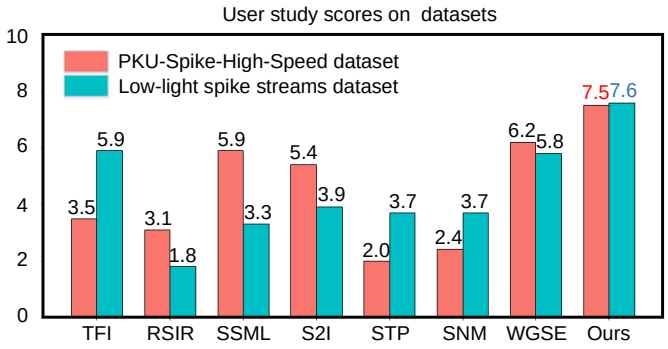

Fig. 10: User study scores (↑) of reconstructed images from different methods. The max (min) score is 8 (1). Red or blue color is the highest score on the dataset [37] or [8].

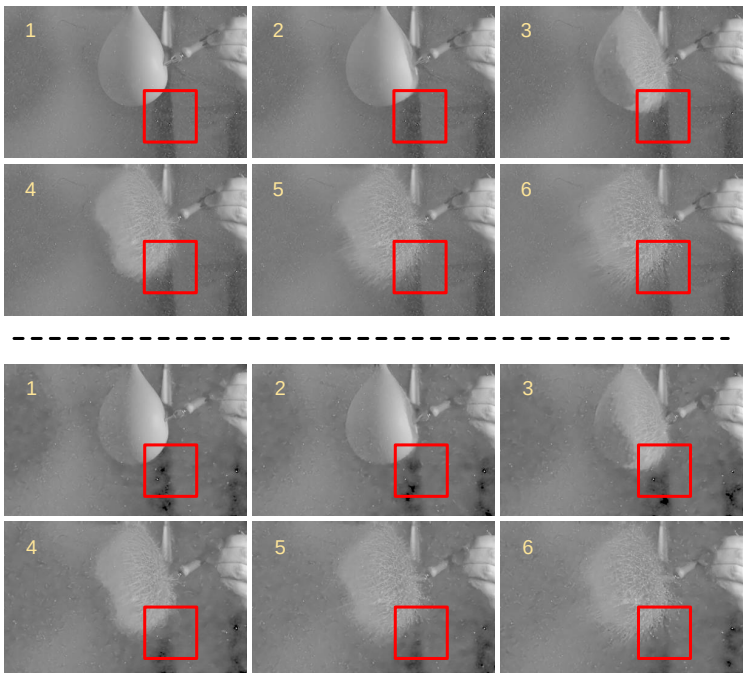

Fig. 11: A water polo bursting at high speed in a low-light indoor. We selected the reconstruction results under 6 sampling moments, and the interval between two adjacent sampling moments is 41/40000 s. The **top** is our method and the **bottom** is the state-of-the-art reconstruction method [34]. We apply the traditional enhancement method [32] to reconstruction results because the scene is too dark. Reconstructed videos are provided in supplementary materials.

LLR. The comparison between (A) and (C) ((B) and (D)) proves the effectiveness of LR-Rep. The comparison between (A) and (B) ((C) and (D)) proves the

effectiveness of ADF. Further, by adding the alignment information in the fusion module *i.e.* AIF, our final method (E) appropriately reduces the misalignment of motion from different timestamps and can reconstruct high-speed scenes more accurately than (D). Besides, the comparison between (E) and (F) shows GISI has better performance than LISI. This is because GISI can extract more temporal information than LISI (see Fig. 6). More importantly, the cost of using GISI instead of LISI is negligible (we only need to use two 400×250 matrices to store the time of the forward spike and the backward spike, respectively), which does not affect the parameter and efficiency of the network.

Table 2: Abltion results on the synthetic dataset, LLR. The best performance is bolded.

Index	Effect of different network structures	PSNR	SSIM
(A)	Basic baseline	42.743	0.97403
(B)	Adding ADF to (A)	44.151	0.98514
(C)	Adding LR-Rep to (A)	44.739	0.98636
(D)	Adding ADF & LR-Rep to (A)	44.956	0.98678
(E)	Adding ADF & LR-Rep & AIF	**45.075**	**0.98681**
(F)	Replacing GISI with LISI in (E)	44.997	0.98676

Comparison with other representation We compare the performance of different representations in our framework, *i.e.* (1) General representation of spike stream: TFI and TFP [41] (2) Tailored representation for reconstruction networks: AST in RSIR [44], AMIM [3] in SSML, SALI [37] in S2I and WGSE-1d [34] in WGSE. We replace LR-Rep in our method as the above representation. They are trained on the dataset, RLLR, and implementation details are the same as our method. As shown in Table. 3, our LR-Rep achieves the best performance which means LR-Rep can better adapt to our framework.

Table 3: Performance of different representation methods in our framework. All methods are trained on RLLR and are tested on LLR. The best performance is bolded.

Rep.	TFP ICME'19	TFI ICME'19	AST MM'23	AMIM IJCAI'22	SALI CVPR'21	WGSE-1d AAAI'23	LR-Rep Ours
PSNR	38.615	37.617	37.997	41.950	43.314	42.302	**45.075**
SSIM	0.96641	0.93632	0.95463	0.97493	0.98304	0.97438	**0.98681**

Train dataset size. The size of train datasets has an impact on the performance of our network. A larger train dataset typically provides more samples and a wider range of variations. In fact, proposed RLLR is enough for the reconstruction task of low-light spike streams. As shown in Table. 4, we find that as the dataset size increase, the performance of the model also improves. However, it is observed that the performance improvement becomes less significant after the

Table 4: Evaluation results on LLR. We train our network where 20%, 40%, 60%, and 80% of RLLR data are used as training set respectively. The best performance is bolded.

Metric	20%	40%	60%	80%	100%
PSNR	35.001	38.618	44.415	44.753	**45.075**
SSIM	0.93411	0.97113	0.98459	0.98581	**0.98681**

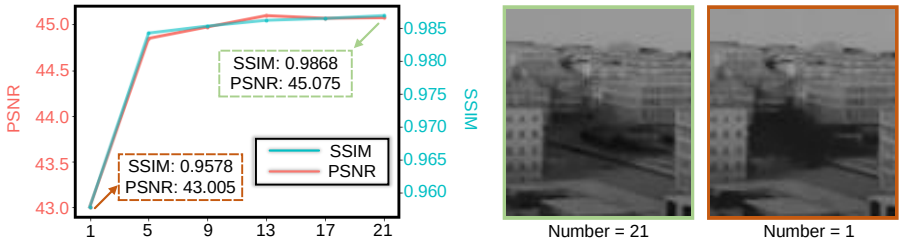

Fig. 12: Effect of the number of continuous spike streams on the performance. We test on the dataset, LLR. **Left**: PSNR and SSIM of the method under the different number of continuous spike streams. **Right**: Comparison of reconstruction images.

dataset size reaches 60% of RLLR. It shows that the proposed RLLR is sufficient for training our network.

The number of continuous spike streams For solving the reconstruction difficulty caused by inadequate information in low-light scenes, the release time of spike in LR-Rep and temporal features in fusion module are maintained forward and backward in a recurrent manner. The number of continuous spike streams has an impact on our method performance. Fig. 12 shows its effect on the performance. We can find that, as the number increases, the performance of our method can greatly increase until convergence. This is because, as the number increases, our method can utilize more temporal information until sufficient. The reconstrued image from 21 continuous spike streams has more details in a shaded area.

6 Conclusion

We propose a bidirectional recurrent-based reconstruction framework for spike camera to better handle different light conditions. In our framework, a light-robust representation (LR-Rep) is designed to aggregate temporal information in spike streams. Moreover, a fusion module is used to extract temporal features. To evaluate the performance of different methods in low-light high-speed scenes, we synthesize a reconstruction dataset where light sources are carefully designed to be consistent with reality. The experiment on both synthetic data and real data shows the superiority of our method.

Acknowledgement This work was supported by the National Science and Technology Major Project (Grant No. 2022ZD0116305), the Beijing Natural Science Foundation (Grant No. JQ24023), and the Beijing Municipal Science & Technology Commission Project (No.Z231100006623010).

References

1. Brandli, C., Berner, R., Yang, M., Liu, S.C., Delbruck, T.: A 240 × 180 130 db 3 μs latency global shutter spatiotemporal vision sensor. IEEE Journal of Solid-State Circuits (JSSC) **49**(10), 2333–2341 (2014)
2. Cai, Y., Bian, H., Lin, J., Wang, H., Timofte, R., Zhang, Y.: Retinexformer: One-stage retinex-based transformer for low-light image enhancement. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 12504–12513 (2023)
3. chen, S., Duan, C., Yu, Z., Xiong, R., Huang, T.: Self-supervised mutual learning for dynamic scene reconstruction of spiking camera. In: Proceedings of the International Joint Conference on Artificial Intelligence, IJCAI. pp. 2859–2866 (2022)
4. Delbrück, T., Linares-Barranco, B., Culurciello, E., Posch, C.: Activity-driven, event-based vision sensors. IEEE International Symposium on Circuits and Systems (ISCAS) pp. 2426–2429 (2010)
5. Dong, Y., Xiong, R., Zhang, J., Yu, Z., Fan, X., Zhu, S., Huang, T.: Super-resolution reconstruction from bayer-pattern spike streams. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24871–24880 (2024)
6. Dong, Y., Xiong, R., Zhao, J., Zhang, J., Fan, X., Zhu, S., Huang, T.: Joint demosaicing and denoising for spike camera. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 1582–1590 (2024)
7. Dong, Y., Xiong, R., Zhao, J., Zhang, J., Fan, X., Zhu, S., Huang, T.: Learning a deep demosaicing network for spike camera with color filter array. IEEE Transactions on Image Processing **33**, 3634–3647 (2024)
8. Dong, Y., Zhao, J., Xiong, R., Huang, T.: High-speed scene reconstruction from low-light spike streams. In: 2022 IEEE International Conference on Visual Communications and Image Processing (VCIP). pp. 1–5. IEEE (2022)
9. Fu, Z., Yang, Y., Tu, X., Huang, Y., Ding, X., Ma, K.K.: Learning a simple low-light image enhancer from paired low-light instances. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 22252–22261 (June 2023)
10. Graca, R., McReynolds, B., Delbruck, T.: Optimal biasing and physical limits of dvs event noise. arXiv preprint arXiv:2304.04019 (2023)
11. Guo, C., Li, C., Guo, J., Loy, C.C., Hou, J., Kwong, S., Cong, R.: Zero-reference deep curve estimation for low-light image enhancement. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1780–1789 (2020)
12. Hu, L., Ma, L., Guo, Y., Huang, T.: Scsim: A realistic spike cameras simulator. arXiv preprint arXiv:2405.16790 (2024)
13. Hu, L., Zhao, R., Ding, Z., Ma, L., Shi, B., Xiong, R., Huang, T.: Optical flow estimation for spiking camera. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 17844–17853 (2022)
14. Huang, T., Zheng, Y., Yu, Z., Chen, R., Li, Y., Xiong, R., Ma, L., Zhao, J., Dong, S., Zhu, L., et al.: 1000× faster camera and machine vision with ordinary devices. Engineering (2022)

15. Jiang, Y., Gong, X., Liu, D., Cheng, Y., Fang, C., Shen, X., Yang, J., Zhou, P., Wang, Z.: Enlightengan: Deep light enhancement without paired supervision. IEEE transactions on image processing **30**, 2340–2349 (2021)
16. Kang, Z., Li, J., Zhu, L., Tian, Y.: Retinomorphic sensing: A novel paradigm for future multimedia computing. In: Proceedings of the ACM International Conference on Multimedia (ACMMM). p. 144–152 (2021)
17. Li, C., Guo, C., Han, L., Jiang, J., Cheng, M.M., Gu, J., Loy, C.C.: Low-light image and video enhancement using deep learning: A survey. IEEE transactions on pattern analysis and machine intelligence **44**(12), 9396–9416 (2021)
18. Li, C., Guo, C., Loy, C.C.: Learning to enhance low-light image via zero-reference deep curve estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(8), 4225–4238 (2021)
19. Li, J., Liu, J., Wei, X., Zhang, J., Lu, M., Ma, L., Du, L., Huang, T., Zhang, S.: Uncertainty guided depth fusion for spike camera. arXiv preprint arXiv:2208.12653 (2022)
20. Lichtsteiner, P., Posch, C., Delbruck, T.: A 128 $\times$ 128 120 db 15 μs latency asynchronous temporal contrast vision sensor. IEEE Journal of Solid-State Circuits (JSSC) **43**(2), 566–576 (2008)
21. Liu, J., Zhang, Q., Li, J., Lu, M., Huang, T., Zhang, S.: Unsupervised spike depth estimation via cross-modality cross-domain knowledge transfer. arXiv preprint arXiv:2208.12527 (2022)
22. Shen, J., Ni, W., Xu, Q., Tang, H.: Efficient spiking neural networks with sparse selective activation for continual learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 611–619 (2024)
23. Shen, J., Xu, Q., Liu, J.K., Wang, Y., Pan, G., Tang, H.: Esl-snns: An evolutionary structure learning strategy for spiking neural networks. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 86–93 (2023)
24. Wang, J., Yang, P., Liu, Y., Shang, D., Hui, X., Song, J., Chen, X.: Research on improved yolov5 for low-light environment object detection. Electronics **12**(14), 3089 (2023)
25. Wang, T., Zhang, K., Shen, T., Luo, W., Stenger, B., Lu, T.: Ultra-high-definition low-light image enhancement: A benchmark and transformer-based method. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 2654–2662 (2023)
26. Wang, X., Chan, K.C., Yu, K., Dong, C., Change Loy, C.: Edvr: Video restoration with enhanced deformable convolutional networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (June 2019)
27. Wei, C., Wang, W., Yang, W., Liu, J.: Deep retinex decomposition for low-light enhancement. arXiv preprint arXiv:1808.04560 (2018)
28. Wilson, T.D.: On user studies and information needs. Journal of documentation **37**(1), 3–15 (1981)
29. Wu, Y., Pan, C., Wang, G., Yang, Y., Wei, J., Li, C., Shen, H.T.: Learning semantic-aware knowledge guidance for low-light image enhancement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1662–1671 (2023)
30. Xu, Q., Gao, Y., Shen, J., Li, Y., Ran, X., Tang, H., Pan, G.: Enhancing adaptive history reserving by spiking convolutional block attention module in recurrent neural networks. Advances in Neural Information Processing Systems **36** (2024)

31. Xu, X., Wang, R., Lu, J.: Low-light image enhancement via structure modeling and guidance. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9893–9903 (2023)
32. Ying, Z., Li, G., Gao, W.: A bio-inspired multi-exposure fusion framework for low-light image enhancement. arXiv preprint arXiv:1711.00591 (2017)
33. Zhai, M., Ni, K., Xie, J., Gao, H.: Spike-based optical flow estimation via contrastive learning. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2023)
34. Zhang, J., Jia, S., Yu, Z., Huang, T.: Learning temporal-ordered representation for spike streams based on discrete wavelet transforms. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 137–147 (2023)
35. Zhang, J., Tang, L., Yu, Z., Lu, J., Huang, T.: Spike transformer: Monocular depth estimation for spiking camera. In: European Conference on Computer Vision (ECCV) (2022)
36. Zhao, J., Xiong, R., Huang, T.: High-speed motion scene reconstruction for spike camera via motion aligned filtering. In: International Symposium on Circuits and Systems (ISCAS). pp. 1–5 (2020)
37. Zhao, J., Xiong, R., Liu, H., Zhang, J., Huang, T.: Spk2imgnet: Learning to reconstruct dynamic scene from continuous spike stream. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11996–12005 (2021)
38. Zhao, J., Zhang, S., Ma, L., Yu, Z., Huang, T.: Spikingsim: A bio-inspired spiking simulator. In: 2022 IEEE International Symposium on Circuits and Systems (ISCAS). pp. 3003–3007. IEEE (2022)
39. Zhao, R., Xiong, R., Zhao, J., Yu, Z., Fan, X., Huang, T.: Learninng optical flow from continuous spike streams. In: Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS) (2022)
40. Zheng, Y., Zheng, L., Yu, Z., Shi, B., Tian, Y., Huang, T.: High-speed image reconstruction through short-term plasticity for spiking cameras. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6358–6367 (2021)
41. Zhu, L., Dong, S., Huang, T., Tian, Y.: A retina-inspired sampling method for visual texture reconstruction. In: IEEE International Conference on Multimedia and Expo (ICME). pp. 1432–1437 (2019)
42. Zhu, L., Dong, S., Li, J., Huang, T., Tian, Y.: Ultra-high temporal resolution visual reconstruction from a fovea-like spike camera via spiking neuron model. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(1), 1233–1249 (2022)
43. Zhu, L., Li, J., Wang, X., Huang, T., Tian, Y.: Neuspike-net: High speed video reconstruction via bio-inspired neuromorphic cameras. In: IEEE International Conference on Computer Vision (ICCV). pp. 2400–2409 (2021)
44. Zhu, L., Zheng, Y., Geng, M., Wang, L., Huang, H.: Recurrent spike-based image restoration under general illumination. In: Proceedings of the ACM International Conference on Multimedia (ACMMM). pp. 8251—-8260 (2023)
45. Zhu, Y., Fang, W., Xie, X., Huang, T., Yu, Z.: Exploring loss functions for time-based training strategy in spiking neural networks. Advances in Neural Information Processing Systems **36** (2024)