Boosting Spike Camera Image Reconstruction from a Perspective of Dealing with Spike Fluctuations

Rui Zhao[1,2] Ruiqin Xiong[1,2*] Jing Zhao[1,2] Jian Zhang[3] Xiaopeng Fan[4] Zhaofei Yu[1] Tiejun Huang[1,2]
[1]School of Computer Science, Peking University
[2]National Key Laboratory for Multimedia Information Processing, Peking University
[3]School of Electronic and Computer Engineering, Peking University
[4]School of Computer Science and Technology, Harbin Institute of Technology
ruizhao@stu.pku.edu.cn
{rqxiong, jzhaopku, zhangjian.sz, yuzf12, tjhuang}@pku.edu.cn fxp@hit.edu.cn

Abstract

As a bio-inspired vision sensor with ultra-high speed, spike cameras exhibit great potential in recording dynamic scenes with high-speed motion or drastic light changes. Different from traditional cameras, each pixel in spike cameras records the arrival of photons continuously by firing binary spikes at an ultra-fine temporal granularity. In this process, multiple factors impact the imaging, including the photons' Poisson arrival, thermal noises from circuits, and quantization effects in spike readout. These factors introduce fluctuations to spikes, making the recorded spike intervals unstable and unable to reflect accurate light intensities. In this paper, we present an approach to deal with spike fluctuations and boost spike camera image reconstruction. We first analyze the quantization effects and reveal the unbiased estimation attribute of the reciprocal of differential of spike firing time (DSFT). Based on this, we propose a spike representation module to use DSFT with multiple orders for fluctuation suppression, where DSFT with higher orders indicates spike integration duration between multiple spikes. We also propose a module for inter-moment feature alignment at multiple granularities. The coarser alignment is based on patch-level cross-attention with a local search strategy, and the finer alignment is based on deformable convolution at the pixel level. Experimental results demonstrate the effectiveness of our method on both synthetic and real-captured data. The source code and dataset are available at https://github.com/ruizhao26/BSF.

1. Introduction

Vision technology has undergone remarkable advancements recently. Machine vision in scenes with high-speed motion or drastic light changes is a key challenge in emerging applications such as autonomous driving [14], unmanned aerial vehicles [72], and assistant referees in sports [18]. Traditional digital cameras typically record scenes with a frame rate of 30 Hz $\sim$ 120 Hz, which is inadequate to fulfill the demands of these applications.

*Corresponding author.

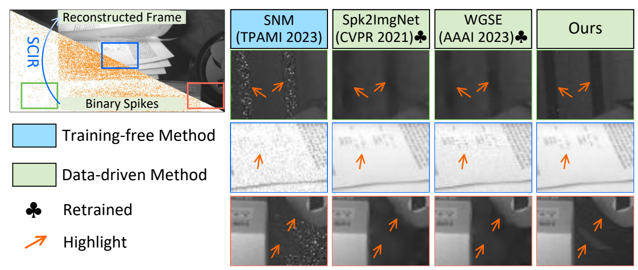

Figure 1. Illustration of spike camera image reconstruction (SCIR) and comparison among recent methods [58, 64, 77] and ours. On the top-left, an orange point means a spike. Our method can better recover the textures. Please zoom in for more details.

Neuromorphic cameras (NeurCams) are a new kind of bio-inspired vision sensor designed to handle the above-mentioned challenges. NeurCams record light intensity at an ultra-high temporal resolution. One kind of NeurCams are event cameras [22, 29, 35]. They employ a differential imaging model, in which each pixel records the scene by outputting events whenever the *change* in light intensity in the logarithmic domain surpasses a certain threshold.

Different from event cameras, spike cameras [8, 18] employ an integral imaging model. Each pixel *accumulates* photons from scenes independently. Whenever the accumulation of a pixel reaches a predefined threshold, it fires a spike and restarts the accumulation. The reading out of spikes is at a high temporal resolution of 40 kHz. Thus, spike cameras can describe light intensities in a very fine temporal granularity by reporting each pixel's status of receiving photons at a very high frequency. Based on these characteristics, spike cameras can handle scenes with high-speed motion or drastic light changes. Recently, many tasks have been researched for spike cameras, including image reconstruction [9, 10, 60, 64, 69, 73], coding [8, 12, 74], object tracking [21, 70, 79], optical flow estimation [17, 48, 55, 62, 67], and depth estimation [44, 57].

Reconstructing clear images from spikes at an arbitrary time is a key task for spike cameras. However, the recording of photons is affected by multiple factors. First, the arrival of photons follows a Poisson process. Second, the circuits introduce thermal noise. Third, the spike reading is controlled by a clock signal, which introduces quantization effects. These factors introduce *fluctuations and randomness* to spikes, i.e., even when the light intensity is constant, the integration period of each spike changes over time. Thus, reconstructing clear images from spikes is challenging.

Several spike camera image reconstruction (SCIR) methods have been proposed from different perspectives. These methods can be roughly categorized into filtering-based method (FM) [61, 65, 73], neuronic-models-based method (NMM) [69, 71, 75, 77], and deep-learning-based method (DLM) [5, 64]. Many of them consider spatial-temporal information aggregation to obtain clear images. However, the fluctuations in spikes are not fully explored.

In this paper, we propose a deep neural network that is boosted from a perspective of dealing with spike fluctuations to obtain clear images. We first analyze the characteristics of the quantization effects of spikes. The differential of spike firing time (DSFT) [67] measures the duration of spike integral to which a spatial-temporal point belongs. We reveal the reciprocal of DSFT can offer an unbiased estimation of light intensities under quantization effects when the input is stable. Based on this attribute, we propose a multi-order DSFT fusion (MODF) module for spike representation. The MODF extracts features from the reciprocal of DSFT to pursue unbiased light information. Further considering photons' Poisson arrival, we extend DSFT to higher orders. Higher-order DSFT represents intervals of multiple spikes, describing stabilized light intensities in a longer temporal scale. The MODF uses multi-order DSFT to approximate light intensities with higher fidelity. It extracts features from the reciprocal of DSFT with multiple orders and fuses them in pursuit of reducing the influence of fluctuations in spikes.

Further considering the impact of the dynamic changes of light and motions. we propose a multi-granularity alignment (MGA). It aligns features in a pyramidal fashion. At each level, the feature is aligned by coarse-granularity alignment (CA) and fine-granularity alignment (FA). The CA aligns the reference features towards the key features using a patch-level cross-attention with a local search strategy. Based on the initial patch-level alignment of CA, the FA aligns features at pixel level using deformable convolutions. We synthesize a benchmark dataset that incorporates multiple light intensities and Poisson effects to evaluate SCIR methods. Our method achieves state-of-the-art performance on both synthetic and real-captured data. Our key contributions can be summarized as follows.

(1) We analyze the characteristics of spikes under the quantization effects and reveal the unbiased estimation attributes of the reciprocal of DSFT.

(2) We propose a network with multi-order DSFT fusion (MODF) and multi-granularity alignment (MGA) modules. These two modules contribute to our model for obtaining more clear images from spikes

(3) Experiments conducted on synthetic and real-captured data demonstrate that the proposed method achieves state-of-the-art performance for SCIR.

2. Related Work

Image reconstruction for spike cameras.

Filtering-based methods. TFP and TFI [73] use the firing rate in a temporal window and the firing interval of a spike to represent the light intensity, respectively. Zhao et al. [61, 65] propose to align the preliminary reconstructed images with optical flow [3] and fuse aligned images using a temporal auto-regressive model [47] for long-term filtering. MGSR [63] predicts super-resolved images from spikes based on the mapping of coordinates with different scales based on optical flow.

Neuronic-model-based methods. SNM [75, 77] is a neuronic model with three-layer leaky integrate and fire (LIF) neurons. The model is trained based on the spike-timing-dependent plasticity (STDP) mechanism [2]. TFSTP [69, 71] uses the short-term plasticity (STP) mechanism [27] to construct the relationship between binary spikes and light intensities based on postsynaptic potentials.

Deep-learning-based methods. Spk2ImgNet [64] aligns features at different moments with pyramidal deformable convolutional networks [7, 43, 78] and fuses them for SCIR. SSML [5] employs a blind-spot network [20] for SCIR. Zhang et al. [58] use wavelet to enhance spike representation for SCIR. Zhao et al. [68] use deep spiking neural networks for SCIR. Zhao et al. [66] propose a deep unfolding network for obtaining super-resolved SCIR. Xiang et al. [50] propose a spike-based super-resolution network using optical flow for alignment. Methods with hybrid cameras based on traditional [4, 15, 16, 49] and event [76] cameras for improving the imaging performance of the spike camera are also proposed recently.

Image reconstruction for other emerging cameras.

Event cameras record changes in light intensities in the logarithmic domain using polar events. There are optimization-based [1, 30] and deep-learning-based methods such as E2Vid [36, 37], FireNet [39], and ETNet [46]

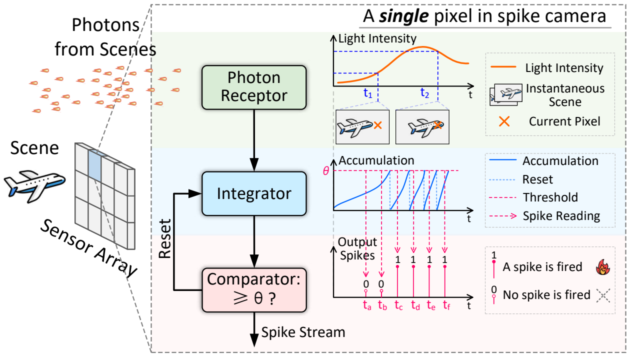

Figure 2. Working mechanism of spike cameras.

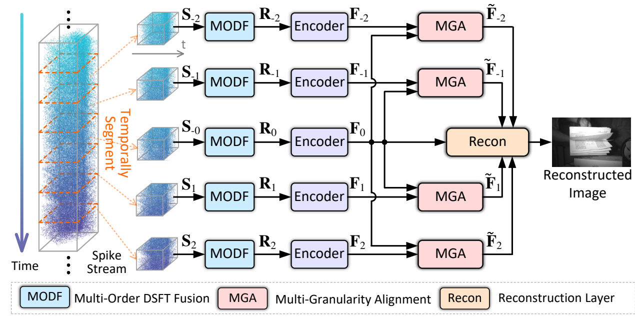

Figure 3. The overall architecture of the proposed BSF network for spike camera image reconstruction.

proposed for event-based image reconstruction. Since event cameras mainly record moving objects, reconstructing images with only events may cause errors of estimation for global contrast. Recently, more works [23, 32, 33, 40, 41] use traditional images to offer light information of static regions to help event cameras reconstruct clear images.

Quanta image sensor (QIS) [26] is a kind of photon-counting sensor. it can be divided into CCD / CMOS (CIS) QIS [25] and single-photon avalanche detector (SPAD) QIS [54]. QIS aims at photon-level detection and it can work well in ultra-low-light scenes with very few photons [6, 11, 13, 34, 38].

3. Working Mechanism of Spike Camera

The working mechanism of spike cameras is shown in Fig. 2. Each pixel of the spike camera comprises three main components: a photon receptor, an integrator, and a comparator. The incoming photons are captured by the photon receptor and accumulated by the integrator. Whenever the number of accumulated photons reaches a predefined threshold θ, a spike is fired, and the integrator is reset. Suppose $\mathbf{L} = \mathbf{L}(\mathbf{x}, t)$ is the expected number of arrival photons at a pixel area per unit time, where $\mathbf{x} = (x, y)$ is spatial coordinate and t is time stamp. The accumulation in the integrator can be formulated as:

$$\mathbf{A}(\mathbf{x}, t) = \int_0^t \alpha \mathcal{P}\left(\mathbf{L}(\mathbf{x}, \tau)\right) \mathrm{d}\tau \mod \theta, \tag{1}$$

where $\mathbf{A}$ is the accumulation, α is the quantum conversion coefficient of photons, and $\mathcal{P}$ means Poisson sampling:

$$P(X = k;\ J) = \frac{J^k}{k!} e^{-J},\ k \in \mathbb{Q}, \tag{2}$$

where P means probability, and J is the expected number of arrival photons.$\mathbb{Q}$ means the natural number set. The thermal noises are omitted here.

spike cameras read spike arrays out at an ultra-high speed of up to 40 kHz. Their outputs can be formulated as $\mathbf{S} \in \mathbb{B}^{H \times W \times T}$, where $\mathbb{B}$ means the binary domain. The threshold θ is configured to ensure that *no more than one* spike can be fired within any spike-reading interval.

4. Methods

4.1. Overall Architecture

The overall architecture of the proposed method is shown in Fig. 3. We segment 5 spike sub-streams $\{\mathbf{S}_i\}_{i=-2}^{2}$ in the temporal axis from the spike stream $\mathbf{S} \in \mathbb{B}^{H \times W \times T}$ to reconstruct the scene at moment t_0, where i is the index indicating the time. $\mathbf{S}_i$ is centered at moment t_i: $\mathbf{S}_i(\mathbf{x}) = \{\mathbf{S}(\mathbf{x}, t)\}_{t=t_i - w_h}^{t_i + w_h}$, where w_h is the window radius. We refer to t_0 as *key* moment and refer to $\{t_i\}_{|i|=\{1,2\}}$ as *referecen* moments. These five spike sub-streams are extracted into representations $\{\mathbf{R}_i\}_{i=-2}^{2}$ through the multi-order DSFT fusion (MODF) module, where DSFT is differential of the spike firing time [67]. Then the representations are encoded into features $\{\mathbf{F}_i\}_{i=-2}^{2}$. Features at reference moments are then aligned to the key moment by the multi-granularity alignment (MGA) module. The reconstructed image is obtained by fusing the key feature and aligned reference features using several convolutional layers.

4.2. Multi-order DSFT Fusion

The purpose of spike representation is to extract initial light information from spikes. In the imaging process of spike cameras, the recording of photons has randomness due to the Poisson process of photons' arrival, making the time for the number of photons to reach the threshold exhibit randomness. Besides, since the spike readout is controlled by a clock signal, the reading time and firing time are usually slightly different, introducing quantization effects. In short, the periods of spikes are unstable and do not directly reflect the light intensity, i.e., the spikes are fluctuating.

To extract stable light intensities under the impacts of the above-mentioned fluctuation factors of spikes, we design a multi-order DSFT fusion module based on three propositions as follows. Note that we focus on the processing of a single spike sub-stream in this subsection, we omit the subscript index i that indicates the time.

Proposition 1: Using DSFT as input. The concept of differential of spike firing time (DSFT) is proposed in Spike2Flow [67]. As shown in Fig. 4, the DSFT of each

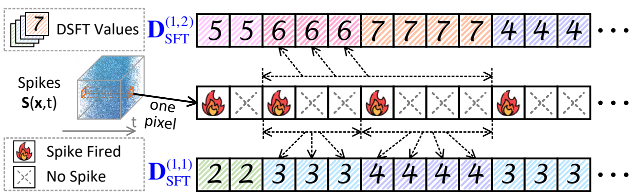

Figure 4. Illustration of differential of spike firing time (DSFT).

point in the 3D coordinate represents the duration of the spike integration period it belongs to. In other words, the DSFT measures the time interval between the previous and the next spike. In this paper, the version of DSFT mentioned above is named (1,1)-order DSFT $\mathbf{D}_{\mathrm{SFT}}^{(1,1)}$.

In binary spikes, "1" represents light intensities in a time process but not a time point. Thus, the binary spikes cannot reflect light intensities at the reading moment in a simple way. The light intensity at the reading moment of each "1" can be different. In DSFT, the value represents the current light intensity of each point in the 3D coordinate. Thus, DSFT offers more relative information about light than binary spikes, and we use DSFT to pursue light intensities contained in physical reality.

Proposition 2: Processing DSFT in reciprocal domain. In this part, we analyze the DSFT and quantization effects. In the analyses of this proposition, we omit the thermal noise and assume the photons' arrival is constant. Suppose the firing threshold of the spike camera is θ, the spike-reading time interval is T_{r}, and the incoming rate of photons is ζ. During a spike-reading time interval, the number of photons reaching a pixel area is $L = \zeta T_{\mathrm{r}}$. We can infer the light intensity through the ratio of θ and $\mathbf{D}_{\mathrm{SFT}}^{(1,1)}$: $\widetilde{L} = \theta / \mathbf{D}_{\mathrm{SFT}}^{(1,1)}$. When $\theta \mod L = 0$, the $\mathbf{D}_{\mathrm{SFT}}^{(1,1)}$ is stable. However, as shown in Fig. 5, when $\theta \mod L \neq 0$, even when the photon's arrival is constant, the $\mathbf{D}_{\mathrm{SFT}}^{(1,1)}$ has different values go up and down around θ/L. This is the *value instability attribute* of DSFT, which contributes to the fluctuations of spikes. To handle such issues caused by quantization effects, we propose the following theorem and design a spike representation module based on the theorem.

Theorem 4.1. *Suppose the symbolic definition is the same as above and* $\theta \mod L \neq 0$. *When the photons' arrival is constant, The (1,1)-order DSFT has only two values* $\{\lfloor \theta/L \rfloor, \lceil \theta/L \rceil\}$ *and its distribution is as follows:*

$$\begin{cases} \Pr\left\{\mathbf{D}_{SFT}^{(1,1)} = \lfloor \theta/L \rfloor\right\} = p_1 = (\lceil \theta/L \rceil - \theta/L) \cdot \dfrac{\lfloor \theta/L \rfloor}{\theta/L} \\ \Pr\left\{\mathbf{D}_{SFT}^{(1,1)} = \lceil \theta/L \rceil\right\} = p_2 = (\theta/L - \lfloor \theta/L \rfloor) \cdot \dfrac{\lceil \theta/L \rceil}{\theta/L} \end{cases} \tag{3}$$

where $\Pr\{\cdot\}$ *means probability.*

The proof of the above theorem is in Sec. 7 of the supplementary material (abbreviated as *supp* hereafter). Eq. (3)

Time / T_{r}	1	2	3	4	5	6	7	8	9	10	···
Accumulation	0.4θ	0.8θ	0.2θ	0.6θ	0	0.4θ	0.8θ	0.2θ	0.6θ	0	···
Spike	☒	☒	🔥	☒	🔥	☒	☒	🔥	☒	🔥	···
$\mathbf{D}_{\mathrm{SFT}}^{(1,1)}$	3	3	2	2	3	3	3	2	2	3	···

Figure 5. Value instability attribute of DSFT caused by quantization effects when photons' arrival is stable. Here $L \equiv 0.4\theta$.

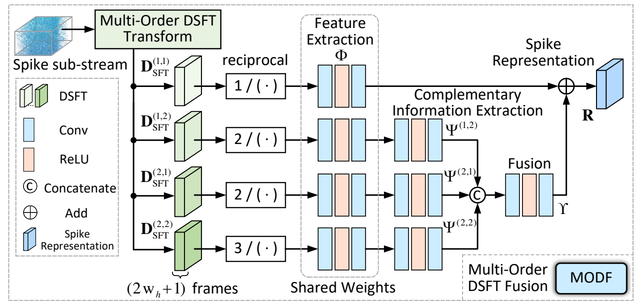

Figure 6. Illustration of the multi-order DSFT fusion (MODF).

shows that when photons' arrival is constant, the $\mathbf{D}_{\mathrm{SFT}}^{(1,1)}$ has only two values: the ceiling value and the floor value of θ/L. This distribution indicates the estimated $\widetilde{L} = \theta/\mathbf{D}_{\mathrm{SFT}}^{(1,1)}$ fluctuate. Although the $\mathbf{D}_{\mathrm{SFT}}^{(1,1)}$ is unstable when $\theta \mod L \neq 0$, we find that $1/\mathbf{D}_{\mathrm{SFT}}^{(1,1)}$ is an *unbiased* estimation of L/θ according to Theorem 4.1:

$$\mathbb{E}\left(\frac{1}{\mathbf{D}_{\mathrm{SFT}}^{(1,1)}}\right) = \frac{1}{\lfloor \theta/L \rfloor} \cdot p_1 + \frac{1}{\lceil \theta/L \rceil} \cdot p_2 = \frac{L}{\theta}. \tag{4}$$

Through the reciprocal of $\mathbf{D}_{\mathrm{SFT}}^{(1,1)}$, the unbiased estimation of L can be obtained. Note that since $\mathbf{D}_{\mathrm{SFT}}^{(1,1)}$ is not constant, $\mathbb{E}(\mathbf{D}_{\mathrm{SFT}}^{(1,1)})$ is a biased estimation for θ/L according to *harmonic mean inequality* (introduced in Sec. 7 of *supp*). Thus, we process the DSFT in its reciprocal domain to pursue an unbiased estimation of light intensities.

Proposition 3: Fusing DSFT with multiple orders. Considering Poisson noises, motion, and light change, the constant photon-arrival assumption is practically limited in reality. In pursuit of stable light information under these factors that contribute to spike fluctuations, we expanded the (1,1)-order DSFT to (n_1, n_2)-order DSFT:

$$\begin{aligned} \mathbf{D}_{\mathrm{SFT}}^{(n_1,n_2)}(\mathbf{x},t) &= \mathbf{T}_{\mathrm{next}}^{(n_2)}(\mathbf{x},t) - \mathbf{T}_{\mathrm{prev}}^{(n_1)}(\mathbf{x},t) \\ &= \min\left\{\tau \,\middle|\, \sum\nolimits_{k=t+1}^{\tau} \mathbf{S}(\mathbf{x},k) = n_2,\ \tau > t\right\} \\ &\quad - \max\left\{\tau \,\middle|\, \sum\nolimits_{k=\tau}^{t} \mathbf{S}(\mathbf{x},k) = n_1,\ \tau \le t\right\}, \end{aligned} \tag{5}$$

where $\mathbf{T}_{\mathrm{next}}^{(n_2)}(\mathbf{x},t)$ is the time stamp of the n_2-th next spike at pixel $\mathbf{x}$ and time stamp t. $\mathbf{T}_{\mathrm{prev}}^{n_1}(\mathbf{x},t)$ is the time stamp of the n_1-th previous spike, where spike at $(\mathbf{x},t)$ is counted into $\mathbf{T}_{\mathrm{prev}}^{n_1}$ if $\mathbf{S}(\mathbf{x},t) = 1$. The $\mathbf{D}_{\mathrm{SFT}}^{(n_1,n_2)}(\mathbf{x},t)$ represents the spike interval between the n_1-th previous spike and the n_2-th next spike. DSFT with higher order can suppress the Poisson effects to stabilize fluctuations. Thus, we use DSFT with multiple orders to obtain stable light information

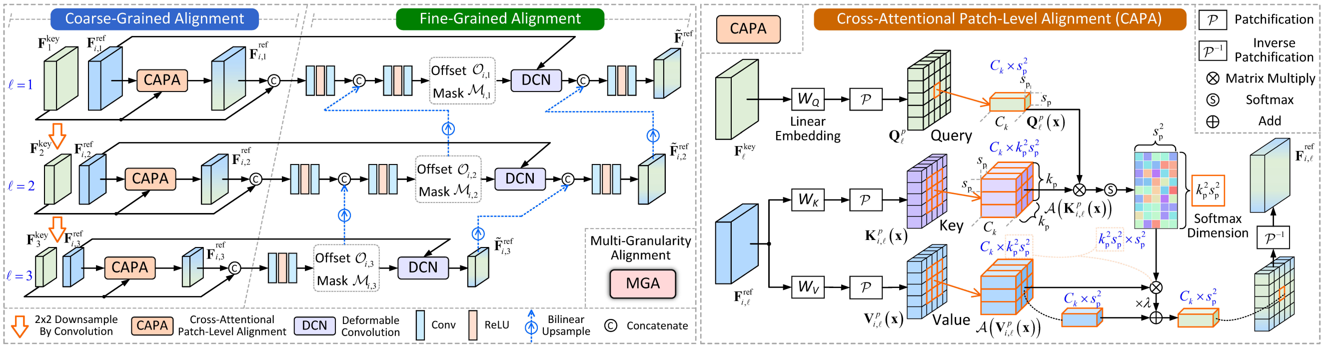

Figure 7. Illustration of the multi-granularity alignment (MGA) module for aligning $\mathbf{F}_i^{\text{ref}}$ to $\mathbf{F}^{\text{key}}$. On the left is the architecture of the MGA module. On the right is the architecture of the cross-attentional patch-level alignment (CAPA) module in the MGA.

Based on the above discussion, we propose a multi-order DSFT fusion (MODF) module. As shown in Fig. 6. We use $\{\mathbf{D}_{\text{SFT}}^{(n_1,n_2)}\}_{n_1,n_2\in\{1,2\}}$ as input. Similar to spike sub-streams, all the DSFT streams have a length of $(2w_h+1)$. All the DSFT streams are taken reciprocal and normalized by multiplying with interval numbers $\chi_i = n_1+n_2-1$ of the corresponding order. Then the reciprocal DSFT are extracted to be features through a weight-shared feature extractor Φ. Compared with DSFT with higher orders, $\mathbf{D}_{\text{SFT}}^{(1,1)}$ undergoes less motion blur since it contains the fewest spike intervals. Thus, we use features from $\mathbf{D}_{\text{SFT}}^{(1,1)}$ as foundation and use features from higher-order DSFT to enhance the information in $\mathbf{D}_{\text{SFT}}^{(1,1)}$. This process can be formulated as:

$$\mathbf{R} = \Phi\Big(\frac{1}{\mathbf{D}_{\text{SFT}}^{(1,1)}}\Big) + \Upsilon\Big(\text{Cat}\Big(\Big\{\Psi^{\xi_i}\Big(\Phi\Big(\frac{\chi_i}{\mathbf{D}_{\text{SFT}}^{\xi_i}}\Big)\Big)\Big\}_{i=1}^{3}\Big)\Big),$$
$$\text{where}\begin{cases}\{\xi_i\}_{i=1}^3 = \{\xi_1,\xi_2,\xi_3\} = \{(1,2),\ (2,1),\ (2,2)\},\\ \{\chi_i\}_{i=1}^3 = \Big\{\sum(\xi_i)-1\Big\}_{i=1}^3 = \{2,2,3\},\end{cases} \tag{6}$$

where Φ is the feature extractor, $\Psi^{(n_1,n_2)}$ is for extracting information from $\mathbf{D}_{\text{SFT}}^{(n_1,n_2)}$, ξ_i is the index of (n_1,n_2), Υ means fusion operation through convolution, and Cat is channel-wise concatenation. The representations $\{\mathbf{R}\}_{i=-2}^{2}$ are then encoded to be features $\{\mathbf{F}\}_{i=-2}^{2}$, where the encoder consists of 4 residual blocks.

4.3. Multi-granularity Alignment

The utilization of long-term information is key to reconstructing high-quality images from spike streams. We achieve this objective by aligning information from spike sub-streams at different moments: we align the information from features at moments $\{t_i\}_{|i|=1,2}$ (reference features) to the t_0 moment (key feature) and then fuse all these features.

The MODF module aims to extract stable representations of light intensities from binary spikes. However, multiple factors contribute to the fluctuations of spikes in practice. These factors make corresponding areas in features at different moments have different values, improving the matching error in the alignment process.

To realize robust alignment, we propose a multi-granularity alignment (MGA) module with a pyramid structure. In each pyramidal level, the alignment is from coarse granularity to fine granularity. As shown in the left of Fig. 7, the input features are downsampled by convolutions to construct a pyramid. Suppose $\mathbf{F}_\ell^{\text{key}}$ is key feature at t_0 moment at the ℓ-th pyramidal level, and $\mathbf{F}_{i,\ell}^{\text{ref}}$ is reference feature at t_i moment at the ℓ-th pyramidal level. $\mathbf{F}_{i,1}^{\text{key}} = \mathbf{F}_i^{\text{key}}$ and $\mathbf{F}_1^{\text{ref}} = \mathbf{F}^{\text{ref}}$ are initial input of MGA. In each pyramidal level, the reference feature is first aligned by coarse-grained alignment (CA) and then aligned by fine-grained alignment (FA). The CA locally aligns features at a patch level, and the FA further aligns features at the pixel level. In CA, we propose a Cross-Attentional Patch-level Alignment (CAPA) with a local search strategy for initial coarse alignment. The design propositions and details of the CAPA are as follows.

The aim of alignment is to align the reference features $\mathbf{F}_{i,\ell}^{\text{ref}}$ to spatial coordinates of the key feature $\mathbf{F}_\ell^{\text{key}}$, i.e., searching for corresponding contents of $\mathbf{F}_\ell^{\text{key}}$ in $\mathbf{F}_{i,\ell}^{\text{ref}}$. Considering the above analysis, we design a ***cross-attentional operation*** in CAPA. We use $\mathbf{F}_\ell^{\text{key}}$ to construct the Query, and use $\mathbf{F}_{i,\ell}^{\text{ref}}$ to construct the Key and Value. Patch-level operation is a classic auxiliary strategy in pixel-level tasks [24, 52, 53]. Considering the light intensity information is more stable in a local region than in a pixel, we design CA to be ***patch-level*** to provide a foundation for subsequent pixel-level alignment. Thus, the embedding procedure in the CAPA can be formulated as:

$$\mathbf{Q}_\ell^{\text{p}} = \mathcal{Z}[\mathbf{Q}_\ell] = \mathcal{Z}[W_Q\mathbf{F}_\ell^{\text{key}}], \tag{7}$$
$$\mathbf{K}_{i,\ell}^{\text{p}} = \mathcal{Z}[\mathbf{K}_{i,\ell}] = \mathcal{Z}[W_K\mathbf{F}_{i,\ell}^{\text{ref}}], \tag{8}$$
$$\mathbf{V}_{i,\ell}^{\text{p}} = \mathcal{Z}[\mathbf{V}_{i,\ell}] = \mathcal{Z}[W_V\mathbf{F}_{i,\ell}^{\text{ref}}], \tag{9}$$

where $\mathcal{Z}$ is the patchification operation with $s_{\text{p}}\times s_{\text{p}}$ size. Suppose after padding for patchification, the spatial resolution of feature $\mathbf{F}_\ell^{\text{key}}$ and $\mathbf{F}_{i,\ell}^{\text{ref}}$ at the ℓ-th level is $H_\ell\times W_\ell$, the spatial resolution of $\mathbf{Q}_\ell^{\text{p}}, \mathbf{K}_{i,\ell}^{\text{p}}$ and $\mathbf{V}_{i,\ell}^{\text{p}}$ is $\widehat{H}_\ell\times\widehat{W}_\ell$, where $\widehat{H}_\ell = H_\ell/s_{\text{p}}$ and $\widehat{W}_\ell = W_\ell/s_{\text{p}}$. In this way, we can realize transforming $\mathbf{F}^{\text{ref}}_{i,\ell}$ to approximate $\mathbf{F}^{\text{key}}_{\ell}$ according to the relationship of "using $\mathbf{F}^{\text{key}}_{\ell}$ to query $\mathbf{F}^{\text{ref}}_{i,\ell}$". In other words, the cross-attention implements alignment from $\mathbf{F}^{\text{ref}}_{i,\ell}$ towards $\mathbf{F}^{\text{key}}_{\ell}$. Based on the above discussion, we design a ***patch-level local search strategy*** in CAPA. The attention operation can be formulated as follows:

$$\widehat{\mathbf{V}}^{\text{p}}_{i,\ell}(\mathbf{x}) = \mathcal{A}(\mathbf{V}^{\text{p}}_{i,\ell}(\mathbf{x}))\sigma\left(\frac{(\mathbf{Q}^{\text{p}}_{\ell})^{\top}(\mathbf{x})\ \mathcal{A}(\mathbf{K}^{\text{p}}_{i,\ell}(\mathbf{x}))}{\sqrt{C_k}}\right), \quad (10)$$

where $\mathbf{x} \in \mathbb{Q}^{\widehat{H}_\ell \times \widehat{W}_\ell}$. $\mathcal{A}(\cdot)$ is the local sampling operator with $k_{\text{p}} \times k_{\text{p}}$ size. σ means softmax on the dimension with $k_{\text{p}}^2 s_{\text{p}}^2$ channels as shown in the right of Fig. 7. C_k is the channel number of the query, key, and value. The $\mathcal{A}(\cdot)$ can be formulated as:

$$\mathcal{A}(\mathbf{V}^{\text{p}}_{i}(\mathbf{x})) = \left\{\mathbf{V}^{\text{p}}_{i}(\mathbf{x}+\boldsymbol{\delta})\right\}_{\boldsymbol{\delta}\in\mathcal{N}(\mathbf{x};k_{\text{p}})}, \quad (11)$$

where $\mathcal{N}(\mathbf{x}; k_{\text{p}})$ is a $k_{\text{p}} \times k_{\text{p}}$ area centered on $\mathbf{x}$. Through $\mathcal{A}$, we construct key vectors over a larger range for each query vector, which realizes search operation in $\mathbf{F}^{\text{ref}}_{i,\ell}$ for alignment towards $\mathbf{F}^{\text{key}}_{\ell}$. Based on the patch-level operation and the local search strategy, CAPA implements alignment in a coarse granularity with more stable light information than global pixel-level alignment. The illustration of CAPA and the size of each tensor are shown on the right of Fig. 7. The coarse-grained alignment $\widehat{\mathbf{F}}^{\text{p}}_{i,\ell}$ is obtained through inverse patchification of $\widehat{\mathbf{V}}^{\text{p}}_{i,\ell}$:

$$\widehat{\mathbf{F}}^{\text{ref}}_{i,\ell}(\mathbf{x}) = \mathbf{F}^{\text{ref}}_{i,\ell}(\mathbf{x}) + \lambda\mathcal{Z}^{-1}\left[\widehat{\mathbf{V}}^{\text{p}}_{i}\right](\mathbf{x}),\ \ \mathbf{x}\in\mathbb{Q}^{H_\ell\times W_\ell}, \quad (12)$$

where $\mathcal{Z}^{-1}$ means inverse patchification operation and λ is a learnable parameter. Through CAPA, each pixel in $\mathbf{F}^{\text{ref}}_{i,\ell}$ is aligned to $\mathbf{F}^{\text{key}}_{\ell}$ based on $k_p \times k_p$ patches centered at $\mathbf{F}^{\text{key}}_{\ell}$.

The coarse aligned $\widehat{\mathbf{F}}^{\text{ref}}_{i,\ell}$ is then aligned in fine-granularity by deformable convolutions (DCN) [7, 78]:

$$\widetilde{\mathbf{F}}^{\text{ref}}_{i}(\mathbf{x}) = \sum_{\boldsymbol{\delta}\in\mathcal{N}(\mathbf{x};k_{\text{d}})} K(\boldsymbol{\delta})\widehat{\mathbf{F}}^{\text{ref}}_{i}(\mathbf{x}+\boldsymbol{\delta}+\mathcal{O}_i(\mathbf{x},\boldsymbol{\delta}))\mathcal{M}_i(\mathbf{x},\boldsymbol{\delta}), \quad (13)$$

where the subscript ℓ is omitted. $\widetilde{\mathbf{F}}^{\text{ref}}$ is the fine-grained aligned feature. k_d is the kernel size of the DCN. $\mathcal{O}$ and $\mathcal{M}$ are the offset and mask of the DCN, respectively. They are obtained from the initial aligned $\widehat{\mathbf{F}}^{\text{ref}}_{i}$ and $\mathbf{F}^{\text{key}}$, and they are passed and fused between different pyramidal levels by upsampling as shown in the left of Fig. 7. Besides, the aligned $\widetilde{\mathbf{F}}_i$ is also passed and fused between different levels by upsampling. In short, the MGA aligns features at different scales and granularities. The CAPA uses patch-level light information to reduce matching errors caused by spike fluctuations, aiming to provide reliable foundations for fine-grained alignment.

The key feature and aligned reference features are then fused through the reconstruction layer, which is composed of several layers of convolutions and ReLU:

$$\widetilde{I}(t_0) = \text{Recon}\left(\text{Cat}\left(\{\widetilde{\mathbf{F}}\}_{i=-2}^{-1},\ F_0,\ \{\widetilde{\mathbf{F}}\}_{i=1}^{2}\right)\right). \quad (14)$$

5. Experiments

5.1. Data Preparation

To synthesize spike data, we refer to the simulation procedure proposed in literature [64] and extend the simulator. First, we use an advanced video frame interpolation method EMA-VFI [56] to achieve high-fidelity continuous scene generation. Second, based on the temporally interpolated continuous frames, we set a parameter η to simulate different levels of light intensity. Third, we simulate the Poisson process of photons' arrival.

Since the aperture size and spike firing threshold are adjustable, the settings of parameters in the simulation pipeline are *without loss of generality*. Based on the continuous scenes, we establish 3 different light intensity factors $\eta = \{1.00, 0.75, 0.50\}$. γ is the conversion of pixel values to the expected number of arriving photons within a single pixel during one readout interval T_{r}, and we set γ as 60. We set the quantum conversion factor α as 0.7. Suppose a pixel value at moment t is $I(t)$. During interval $(t, t+T_{\text{r}})$, the integral in the accumulator is $\Delta\mathbf{A}(\mathbf{x}) = \alpha\mathcal{P}(\eta I(\mathbf{x}, t))$, where $\mathcal{P}$ is Poisson sampling. We set the firing threshold as $\theta = \max(I)\cdot\gamma$. Since $\alpha < 1$, the number of fired spikes within T_{r} will not exceed one.

We use the REDS dataset [31] at 120 FPS and 1280×720 resolution to generate the REDS-SCIR dataset. In REDS, there are 240 scenes for training and 30 scenes for evaluation. For each training scene, we crop it to 12 scenes at 256×256 resolution. For each evaluation scene, we crop it to 4 scenes at 384×512 resolution. Given that we have 3 light intensity factors $\eta = \{1.00, 0.75, 0.50\}$, there are $240\times12\times3 = 8640$ and $30\times4\times3 = 360$ scenes for training and evaluation, respectively. For each scene, we use 40 frames to generate 400 spike frames. We use the high-quality gray images from the REDS as ground truths.

5.2. Implementation Details

In the experiments, we set the patchification size s_{p} in CAPA as 3. Since CAPA is used for initial alignment, we set kernel size k_{p} of local sampling operation as 3 for simplicity. During training, we randomly crop the spikes to 96×96 spatially, and we use random horizontal and vertical flips as well as random rotation for data augmentation. The network is trained for 60 epochs with a batch size of 8. We use Adam optimizer [19] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The learning rate is initially set as 1e-4 and scaled by 0.5 every 10 epoch. The network is trained based on ℓ_1 loss between the normalized estimated $\widetilde{I}(t_0)$ and its ground truth $I_{\text{gt}}(t_0)$:

$$\mathcal{L} = \left\|\widetilde{I}(t_0)/(\alpha\cdot\eta) - I_{\text{gt}}(t_0)\right\|_1. \quad (15)$$

5.3. Comparison with Existing Methods

We divide existing methods for comparison into 4 parts: (A) traditional training-free methods, (B) event-based image re-

Part	Method	$\eta = 1.00$			$\eta = 0.75$			$\eta = 0.50$			Params (M)
		PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	
(A)	TFP [73]	27.27	0.711	0.265	26.73	0.669	0.300	25.62	0.581	0.370	—
	TFI [73]	23.55	0.634	0.329	24.77	0.673	0.293	26.77	0.713	0.249	—
	TFSTP [69]	20.35	0.678	0.270	19.62	0.685	0.252	21.10	0.707	0.247	—
	MAHTF [65]	29.57	0.879	0.112	30.07	0.884	0.113	29.65	0.869	0.136	—
(B)	FireNet [39] ♣	34.38	0.922	0.077	33.87	0.911	0.084	32.62	0.884	0.105	0.038
	ETNet [46] ♣	33.24	0.918	0.082	32.85	0.909	0.089	31.96	0.889	0.109	22.179
(C)	SSML [5]	32.60	0.920	0.088	32.09	0.907	0.097	31.00	0.879	0.122	2.385
	SSML [5] ♣	33.94	0.923	0.075	33.27	0.909	0.088	32.01	0.883	0.116	2.385
(D)	Spk2ImgNet [64]	35.21	0.953	0.036	34.70	0.945	0.044	33.75	0.926	0.064	3.904
	Spk2ImgNet [64] ♣	39.16	0.966	0.024	38.27	0.958	0.032	36.59	0.940	0.051	3.904
	WGSE [58]	35.21	0.950	0.039	34.98	0.947	0.042	34.11	0.931	0.057	3.806
	WGSE [58] ♣	38.97	0.964	0.027	38.23	0.957	0.034	36.75	0.940	0.049	3.806
	BSF (Ours)	**39.76**	**0.970**	**0.021**	**39.09**	**0.964**	**0.027**	**37.76**	**0.951**	**0.040**	2.477

Table 1. Quantitative results on the evaluation set on REDS-SCIR with full-reference metrics. ♣ means the network is retrained on REDS-SCIR with the *same settings* as ours. Best in **red bold** and second best in blue. ↑ and ↓ means larger and smaller is better, respectively.

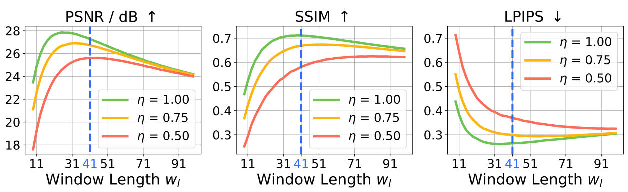

Figure 8. The impact of the window length on TFP [73].

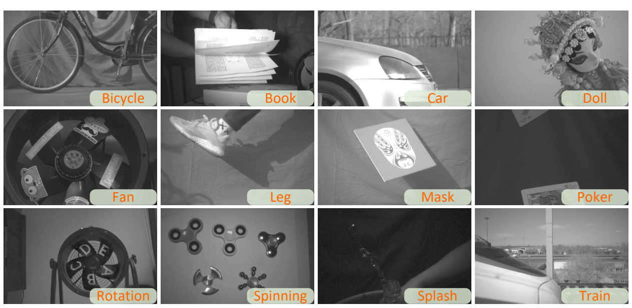

Figure 9. The real-captured scene used for evaluation. Each scene is shown through a frame reconstructed by our method.

construction networks, (C) self-supervised SCIR methods, and (D) supervised SCIR methods. We do not compare SNM [77] on REDS-SCIR since the official program will be unresponsive with 384×512 resolution input. Although part (B) is initially proposed for event data, it is designed for processing streaming data rather than specifically for event data. Given that spikes are also streaming data, we select them for comparison. We use three full-reference metrics: PSNR, SSIM [45], and LPIPS [59] (Alex version). All these three metrics measure the distance between predicted images and their corresponding ground truths. The evaluation on REDS-SCIR is based on predicted images normalized by α and η in the way like Eq. (15): $I_{\text{norm}} = I_{\text{pred}}/(\alpha \cdot \eta)$.

Methods in part (A) are training-free. TFP [73] is the temporal mean of a segment of spikes, which has a hyperparameter of window length w_l. As shown in Fig. 8, we test the w_l in a range of $\{2n+1\}_{n=4}^{49}$ on the REDS-SCIR. We select a $w_l = 41$ that performs well on different η. For part (B), we clip the 60 frames centered at the moment to be reconstructed into 10 segments for recurrent inputting. For parts (C) and (D), they are originally trained on spike data synthesized from REDS [64], thus, we preserve both the original and retrained version. Besides, when retraining SSML [5], we use its original self-supervised loss. As shown in Table 1, our method achieves the best performance across all the η on the three metrics. Note that part (A) has no parameters since they are not deep learning methods.

Part	Method	BRISQUE ↓	PIQE ↓	HOSA ↓
(A)	TFP [73]	37.502	45.956	35.436
	TFI [73]	37.708	45.148	30.892
	TFSTP [69]	37.585	38.714	29.173
	SNM [75]	32.089	41.927	29.334
	MAHTF [65]	30.910	30.068	26.757
(B)	FireNet [39] ♣	25.545	25.076	35.305
	ETNet [46] ♣	33.403	46.682	36.482
(C)	SSML [5]	29.240	25.491	35.399
	SSML [5] ♣	32.234	26.981	35.554
(D)	Spk2ImgNet [64]	29.351	26.745	25.761
	Spk2ImgNet [64] ♣	29.180	39.593	31.287
	WGSE [58]	24.637	27.831	25.657
	WGSE [58] ♣	23.429	30.673	27.434
	BSF (Ours)	**18.529**	**23.477**	**25.523**

Table 2. Quantitative results on real-captured data.

Besides synthetic data, we also compare the above methods using data captured by spike cameras in the real world. We use spikes of 12 scenes, which are shown in Fig. 9. For quantitative comparison, since there are no ground truths, we employ three blind image quality assessment metrics, namely BRISQUE [28], PIQE [42], and HOSA [51]. BRISQUE uses statistics of locally normalized luminance to quantify possible losses of naturalness. PIQE estimates quality only from perceptually significant spatial regions with local features. HOSA uses a small codebook based on high-order statistics aggregation to build the global quality-

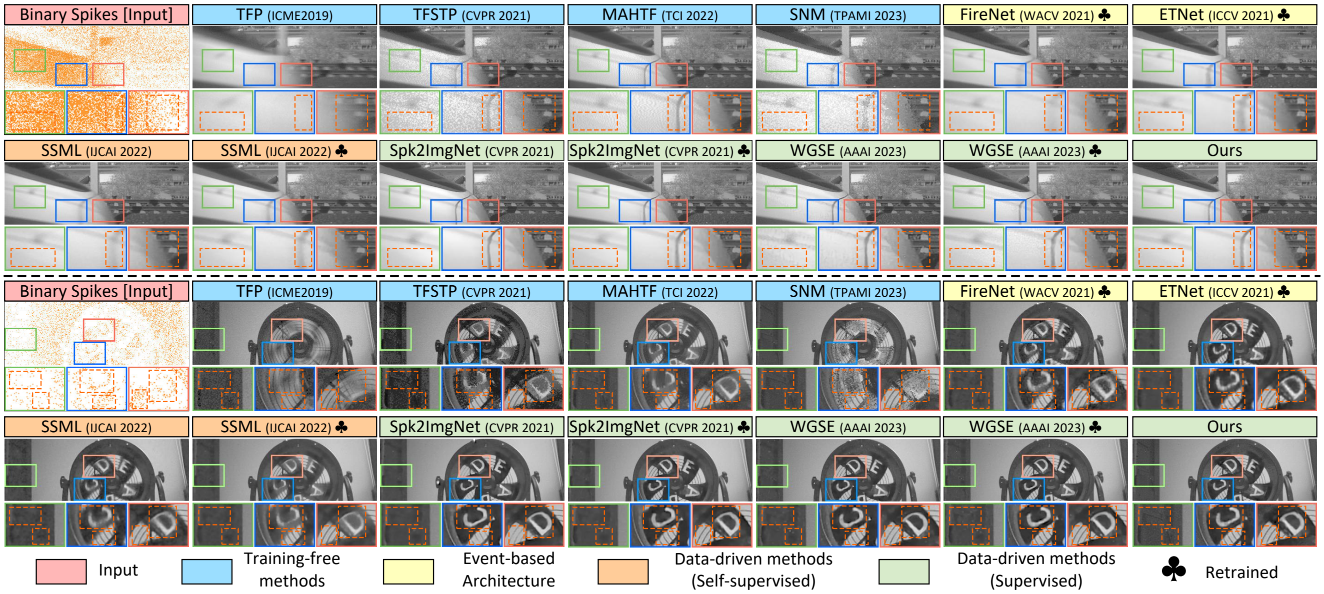

Figure 10. Visual comparison on real-captured data. In the visualization of spikes, an orange point means a spike. Gamma transformation with parameter 2.2 is used for visualization.

Case	MODF	MGA			PSNR ↑		
		DCN	CAPA	Pym	$\eta = 1.00$	$\eta = 0.75$	$\eta = 0.50$
(1)	Spike				38.44	37.69	36.29
(2)	D (1,1)				38.78	38.10	36.80
(3)	✓				38.99	38.32	37.00
(4)	✓	✓		1	38.91	38.26	36.96
(5)	✓	✓	✓	1	39.06	38.39	37.06
(6)	✓	✓		2	39.02	38.39	37.11
(7)	✓	✓	✓	2	39.39	38.73	37.41
(8)	✓	✓		3	39.38	38.77	37.53
(9)	✓	✓	✓	3	**39.76**	**39.09**	**37.76**

Table 3. Ablation studies of the proposed network. Best in bold.

aware image representation. As shown in Table. 2, our method outperforms other methods on all three metrics.

5.4. Ablation Studies

We implement a series of ablation studies to verify the effectiveness of the proposed modules. We first focus on the proposed multi-order DSFT fusion (MODF) and multi-granularity alignment (MGA) as shown in Table. 3. Cases (1–3) are about the MODF module. “Spike” means there is only one branch with binary spikes as input, and “D (1,1)” means there is only one branch with $\mathbf{D}_{\mathrm{SFT}}^{(1,1)}$ as input. Cases (4–9) explore the impact of pyramidal levels and cross-attentional patch-level alignment (CAPA) on our network. Cases (1–3) show the effectiveness of using multi-order DSFT. Cases (4–9) show the effectiveness of pyramidal alignment and using CAPA for initialization.

We also study the impact of the number of input frames on the network. As shown in Table. 4, we select $\{1,3,5,7\}$ as the number of sub-streams, i.e., the number of input frames

N_{IF}	$\eta = 1.00$			$\eta = 0.75$			$\eta = 0.50$		
	P ↑	S ↑	L ↓	P ↑	S ↑	L ↓	P ↑	S ↑	L ↓
21	38.14	0.959	0.032	37.44	0.952	0.039	36.08	0.935	0.055
41	39.35	0.969	**0.021**	38.67	0.963	**0.027**	37.34	0.949	0.041
61	**39.76**	**0.970**	**0.021**	**39.09**	**0.964**	**0.027**	**37.76**	**0.951**	**0.040**
81	39.70	0.969	0.022	39.03	0.963	0.028	37.69	0.950	0.042

Table 4. Ablation studies on the number of input frames N_{IF}.

N_{IF} is $\{21,41,61,81\}$. Since the space is limited, we use the initial letters to represent the 3 reference metrics. When N_{IF} is small, the performance grows with N_{IF} grows, but this growth tends to converge when N_{IF} is large. We set 61 as the N_{IF} since the performance is converged.

6. Conclusions

We propose a method for reconstructing clear images from spike streams with boosted approaches for dealing with spike fluctuations. We reveal the unbiased estimation attribute of the reciprocal of DSFT and design a multi-order DSFT fusion (MODF) module. We also propose a pyramidal multi-granularity alignment (MGA) module. The MGA uses a cross-attentional patch-level operation with a local search strategy for initialization and uses deformable convolution for pixel-level alignment. Experimental results show that the proposed method achieves state-of-the-art performance on both synthetic and real-captured data.

Acknowledgements

This work was supported in part by the National Key R&D Program of China under Grant 2021YFF0900501, and in part by the National Natural Science Foundation of China under Grants 62072009, 22127807.

References

[1] Patrick Bardow, Andrew J Davison, and Stefan Leutenegger. Simultaneous optical flow and intensity estimation from an event camera. In *CVPR*, pages 884–892, 2016. 2

[2] Guo-qiang Bi and Mu-ming Poo. Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type. *Journal of Neuroscience*, 18(24):10464–10472, 1998. 2

[3] Thomas Brox, Andrés Bruhn, Nils Papenberg, and Joachim Weickert. High accuracy optical flow estimation based on a theory for warping. In *ECCV*, pages 25–36, 2004. 2

[4] Yakun Chang, Chu Zhou, Yuchen Hong, Liwen Hu, Chao Xu, Tiejun Huang, and Boxin Shi. 1000 fps hdr video with a spike-rgb hybrid camera. In *CVPR*, pages 22180–22190, 2023. 2

[5] Shiyan Chen, Chaoteng Duan, Zhaofei Yu, Ruiqin Xiong, and Tiejun Huang. Self-supervised mutual learning for dynamic scene reconstruction of spiking camera. In *IJCAI*, pages 2859–2866, 2022. 2, 7

[6] Yiheng Chi, Abhiram Gnanasambandam, Vladlen Koltun, and Stanley H Chan. Dynamic low-light imaging with quanta image sensors. In *ECCV*, pages 122–138, 2020. 3

[7] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *CVPR*, pages 764–773, 2017. 2, 6

[8] Siwei Dong, Tiejun Huang, and Yonghong Tian. Spike camera and its coding methods. In *DCC*, pages 437–437, 2017. 1, 2

[9] Yanchen Dong, Jing Zhao, Ruiqin Xiong, and Tiejun Huang. 3d residual interpolation for spike camera demosaicing. In *ICIP*, pages 1461–1465, 2022. 2

[10] Yanchen Dong, Jing Zhao, Ruiqin Xiong, and Tiejun Huang. High-speed scene reconstruction from low-light spike streams. In *VCIP*, pages 1–5, 2022. 2

[11] Omar A Elgendy, Abhiram Gnanasambandam, Stanley H Chan, and Jiaju Ma. Low-light demosaicking and denoising for small pixels using learned frequency selection. *IEEE TCI*, 7:137–150, 2021. 3

[12] Kexiang Feng, Chuanmin Jia, Siwei Ma, and Wen Gao. Spikecodec: An end-to-end learned compression framework for spiking camera. *arxiv*, 2023. 2

[13] Abhiram Gnanasambandam and Stanley H Chan. Image classification in the dark using quanta image sensors. In *ECCV*, pages 484–501, 2020. 3

[14] Sorin Grigorescu, Bogdan Trasnea, Tiberiu Cocias, and Gigel Macesanu. A survey of deep learning techniques for autonomous driving. *Journal of Field Robotics*, 37(3):362–386, 2020. 1

[15] Jin Han, Chu Zhou, Peiqi Duan, Yehui Tang, Chang Xu, Chao Xu, Tiejun Huang, and Boxin Shi. Neuromorphic camera guided high dynamic range imaging. In *CVPR*, pages 1730–1739, 2020. 2

[16] Jin Han, Yixin Yang, Peiqi Duan, Chu Zhou, Lei Ma, Chao Xu, Tiejun Huang, Imari Sato, and Boxin Shi. Hybrid high dynamic range imaging fusing neuromorphic and conventional images. *IEEE TPAMI*, 2023. 2

[17] Liwen Hu, Rui Zhao, Ziluo Ding, Lei Ma, Boxin Shi, Ruiqin Xiong, and Tiejun Huang. Optical flow estimation for spiking camera. In *CVPR*, pages 17844–17853, 2022. 2

[18] Tiejun Huang, Yajing Zheng, Zhaofei Yu, Rui Chen, Yuan Li, Ruiqin Xiong, Lei Ma, Junwei Zhao, Siwei Dong, Lin Zhu, et al. 1000× faster camera and machine vision with ordinary devices. *Engineering*, 2022. 1

[19] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2014. 6

[20] Jaakko Lehtinen, Jacob Munkberg, Jon Hasselgren, Samuli Laine, Tero Karras, Miika Aittala, and Timo Aila. Noise2noise: Learning image restoration without clean data. In *ICML*, pages 2965–2974, 2018. 2

[21] Jianing Li, Xiao Wang, Lin Zhu, Jia Li, Tiejun Huang, and Yonghong Tian. Retinomorphic object detection in asynchronous visual streams. In *AAAI*, pages 1332–1340, 2022. 2

[22] Patrick Lichtsteiner, Christoph Posch, and Tobi Delbruck. A 128×128 120 db 15μ s latency asynchronous temporal contrast vision sensor. *IEEE JSSC*, 43(2):566–576, 2008. 1

[23] Songnan Lin, Jiawei Zhang, Jinshan Pan, Zhe Jiang, Dongqing Zou, Yongtian Wang, Jing Chen, and Jimmy Ren. Learning event-driven video deblurring and interpolation. In *ECCV*, pages 695–710. Springer, 2020. 3

[24] Ao Luo, Fan Yang, Xin Li, and Shuaicheng Liu. Learning optical flow with kernel patch attention. In *CVPR*, pages 8906–8915, 2022. 5

[25] Jiaju Ma and Eric R Fossum. Quanta image sensor jot with sub 0.3 e-rms read noise and photon counting capability. *IEEE EDL*, 36(9):926–928, 2015. 3

[26] Jiaju Ma, Stanley Chan, and Eric R Fossum. Review of quanta image sensors for ultralow-light imaging. *IEEE TED*, 69(6):2824–2839, 2022. 3

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

[28] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE TIP*, 21(12):4695–4708, 2012. 7

[29] Diederik Paul Moeys, Federico Corradi, Chenghan Li, Simeon A Bamford, Luca Longinotti, Fabian F Voigt, Stewart Berry, Gemma Taverni, Fritjof Helmchen, and Tobi Delbruck. A sensitive dynamic and active pixel vision sensor for color or neural imaging applications. *IEEE TBCS*, 12(1): 123–136, 2017. 1

[30] Gottfried Munda, Christian Reinbacher, and Thomas Pock. Real-time intensity-image reconstruction for event cameras using manifold regularisation. *IJCV*, 126:1381–1393, 2018. 2

[31] Seungjun Nah, Sungyong Baik, Seokil Hong, Gyeongsik Moon, Sanghyun Son, Radu Timofte, and Kyoung Mu Lee. Ntire 2019 challenge on video deblurring and super-resolution: Dataset and study. In *CVPRW*, 2019. 6

[32] 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 *CVPR*, pages 6820–6829, 2019. 3

[33] Liyuan Pan, Richard Hartley, Cedric Scheerlinck, Miaomiao Liu, Xin Yu, and Yuchao Dai. High frame rate video reconstruction based on an event camera. *IEEE TPAMI*, 44(5): 2519–2533, 2020. 3

[34] Jiayong Peng, Zhiwei Xiong, Hao Tan, Xin Huang, Zheng-Ping Li, and Feihu Xu. Boosting photon-efficient image reconstruction with a unified deep neural network. *IEEE TPAMI*, 45(4):4180–4197, 2022. 3

[35] Christoph Posch, Daniel Matolin, and Rainer Wohlgenannt. A qvga 143 db dynamic range frame-free pwm image sensor with lossless pixel-level video compression and time-domain cds. *IEEE JSSC*, 46(1):259–275, 2010. 1

[36] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. Events-to-video: Bringing modern computer vision to event cameras. In *CVPR*, pages 3857–3866, 2019. 2

[37] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. High speed and high dynamic range video with an event camera. *IEEE TPAMI*, 43(6):1964–1980, 2019. 2

[38] Yash Sanghvi, Abhiram Gnanasambandam, and Stanley H Chan. Photon limited non-blind deblurring using algorithm unrolling. *IEEE TCI*, 8:851–864, 2022. 3

[39] Cedric Scheerlinck, Henri Rebecq, Daniel Gehrig, Nick Barnes, Robert Mahony, and Davide Scaramuzza. Fast image reconstruction with an event camera. In *WACV*, pages 156–163, 2020. 2, 7

[40] Stepan Tulyakov, Daniel Gehrig, Stamatios Georgoulis, Julius Erbach, Mathias Gehrig, Yuanyou Li, and Davide Scaramuzza. Time lens: Event-based video frame interpolation. In *CVPR*, pages 16155–16164, 2021. 3

[41] Stepan Tulyakov, Alfredo Bochicchio, Daniel Gehrig, Stamatios Georgoulis, Yuanyou Li, and Davide Scaramuzza. Time lens++: Event-based frame interpolation with parametric non-linear flow and multi-scale fusion. In *CVPR*, pages 17755–17764, 2022. 3

[42] N Venkatanath, D Praneeth, Maruthi Chandrasekhar Bh, Sumohana S Channappayya, and Swarup S Medasani. Blind image quality evaluation using perception based features. In *NCC*, pages 1–6, 2015. 7

[43] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *CVPRW*, 2019. 2

[44] Yixuan Wang, Jianing Li, Lin Zhu, Xijie Xiang, Tiejun Huang, and Yonghong Tian. Learning stereo depth estimation with bio-inspired spike cameras. In *ICME*, pages 1–6, 2022. 2

[45] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 13(4):600–612, 2004. 7

[46] Wenming Weng, Yueyi Zhang, and Zhiwei Xiong. Event-based video reconstruction using transformer. In *ICCV*, pages 2563–2572, 2021. 2, 7

[47] Xiaolin Wu, EU Barthel, and Wenhan Zhang. Piecewise 2d autoregression for predictive image coding. In *ICIP*, pages 901–904, 1998. 2

[48] Lujie Xia, Ziluo Ding, Rui Zhao, Jiyuan Zhang, Lei Ma, Zhaofei Yu, Tiejun Huang, and Ruiqin Xiong. Unsupervised optical flow estimation with dynamic timing representation for spike camera. In *NeurIPS*, 2023. 2

[49] Lujie Xia, Jing Zhao, Ruiqin Xiong, and Tiejun Huang. Svfi: spiking-based video frame interpolation for high-speed motion. In *AAAI*, pages 2910–2918, 2023. 2

[50] Xijie Xiang, Lin Zhu, Jianing Li, Yixuan Wang, Tiejun Huang, and Yonghong Tian. Learning super-resolution reconstruction for high temporal resolution spike stream. *IEEE TCSVT*, 2021. 2

[51] Jingtao Xu, Peng Ye, Qiaohong Li, Haiqing Du, Yong Liu, and David Doermann. Blind image quality assessment based on high order statistics aggregation. *IEEE TIP*, 25(9):4444–4457, 2016. 7

[52] Qingsen Yan, Weiye Chen, Song Zhang, Yu Zhu, Jinqiu Sun, and Yanning Zhang. A unified hdr imaging method with pixel and patch level. In *CVPR*, pages 22211–22220, 2023. 5

[53] Sukmin Yun, Hankook Lee, Jaehyung Kim, and Jinwoo Shin. Patch-level representation learning for self-supervised vision transformers. In *CVPR*, pages 8354–8363, 2022. 5

[54] Franco Zappa, Andrea L Lacaita, Sergio D Cova, and Piergiorgio G Lovati. Solid-state single-photon detectors. *Optical Engineering*, 35(4):938–945, 1996. 3

[55] Mingliang Zhai, Kang Ni, Jiucheng Xie, and Hao Gao. Spike-based optical flow estimation via contrastive learning. In *ICASSP*, pages 1–5, 2023. 2

[56] Guozhen Zhang, Yuhan Zhu, Haonan Wang, Youxin Chen, Gangshan Wu, and Limin Wang. Extracting motion and appearance via inter-frame attention for efficient video frame interpolation. In *CVPR*, pages 5682–5692, 2023. 6

[57] Jiyuan Zhang, Lulu Tang, Zhaofei Yu, Jiwen Lu, and Tiejun Huang. Spike transformer: Monocular depth estimation for spiking camera. In *ECCV*, pages 34–52, 2022. 2

[58] Jiyuan Zhang, Shanshan Jia, Zhaofei Yu, and Tiejun Huang. Learning temporal-ordered representation for spike streams based on discrete wavelet transforms. In *AAAI*, pages 137–147, 2023. 1, 2, 7

[59] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018. 7

[60] Yiyang Zhang, Ruiqin Xiong, and Tiejun Huang. Spike signal reconstruction based on inter-spike similarity. In *VCIP*, pages 1–5, 2022. 2

[61] Jing Zhao, Ruiqin Xiong, and Tiejun Huang. High-speed motion scene reconstruction for spike camera via motion aligned filtering. In *ISCAS*, pages 1–5, 2020. 2

[62] Jing Zhao, Ruiqin Xiong, Rui Zhao, Jin Wang, Siwei Ma, and Tiejun Huang. Motion estimation for spike camera data sequence via spike interval analysis. In *VCIP*, pages 371–374, 2020. 2

[63] Jing Zhao, Jiyu Xie, Ruiqin Xiong, Jian Zhang, Zhaofei Yu, and Tiejun Huang. Super resolve dynamic scene from continuous spike streams. In *ICCV*, pages 2533–2542, 2021. 2

[64] Jing Zhao, Ruiqin Xiong, Hangfan Liu, Jian Zhang, and Tiejun Huang. Spk2ImgNet: Learning to reconstruct dynamic scene from continuous spike stream. In *CVPR*, pages 11996–12005, 2021. 1, 2, 6, 7

[65] Jing Zhao, Ruiqin Xiong, Jiyu Xie, Boxin Shi, Zhaofei Yu, Wen Gao, and Tiejun Huang. Reconstructing clear image for high-speed motion scene with a retina-inspired spike camera. *IEEE TCI*, 8:12–27, 2021. 2, 7

[66] Jing Zhao, Ruiqin Xiong, Jian Zhang, Rui Zhao, Hangfan Liu, and Tiejun Huang. Learning to super-resolve dynamic scenes for neuromorphic spike camera. In *AAAI*, pages 3579–3587, 2023. 2

[67] Rui Zhao, Ruiqin Xiong, Jing Zhao, Zhaofei Yu, Xiaopeng Fan, and Tiejun Huang. Learning optical flow from continuous spike streams. In *NeurIPS*, pages 7905–7920, 2022. 2, 3

[68] Rui Zhao, Ruiqin Xiong, Jian Zhang, Zhaofei Yu, Shuyuan Zhu, Lei Ma, and Tiejun Huang. Spike camera image reconstruction using deep spiking neural networks. *IEEE TCSVT*, 2023. 2

[69] Yajing Zheng, Lingxiao Zheng, Zhaofei Yu, Boxin Shi, Yonghong Tian, and Tiejun Huang. High-speed image reconstruction through short-term plasticity for spiking cameras. In *CVPR*, pages 6358–6367, 2021. 2, 7

[70] Yajing Zheng, Zhaofei Yu, Song Wang, and Tiejun Huang. Spike-based motion estimation for object tracking through bio-inspired unsupervised learning. *IEEE TIP*, 32:335–349, 2022. 2

[71] Yajing Zheng, Lingxiao Zheng, Zhaofei Yu, Tiejun Huang, and Song Wang. Capture the moment: High-speed imaging with spiking cameras through short-term plasticity. *IEEE TPAMI*, 45(7):8127–8142, 2023. 2

[72] Alex Zihao Zhu, Dinesh Thakur, Tolga Özaslan, Bernd Pfrommer, Vijay Kumar, and Kostas Daniilidis. The multivehicle stereo event camera dataset: An event camera dataset for 3d perception. *IEEE RAL*, 3(3):2032–2039, 2018. 1

[73] Lin Zhu, Siwei Dong, Tiejun Huang, and Yonghong Tian. A retina-inspired sampling method for visual texture reconstruction. In *ICME*, pages 1432–1437, 2019. 2, 7

[74] Lin Zhu, Siwei Dong, Tiejun Huang, and Yonghong Tian. Hybrid coding of spatiotemporal spike data for a bio-inspired camera. *IEEE TCSVT*, 31(7):2837–2851, 2020. 2

[75] Lin Zhu, Siwei Dong, Jianing Li, Tiejun Huang, and Yonghong Tian. Retina-like visual image reconstruction via spiking neural model. In *CVPR*, pages 1438–1446, 2020. 2, 7

[76] Lin Zhu, Jianing Li, Xiao Wang, Tiejun Huang, and Yonghong Tian. Neuspike-net: High speed video reconstruction via bio-inspired neuromorphic cameras. In *ICCV*, pages 2400–2409, 2021. 2

[77] Lin Zhu, Siwei Dong, Jianing Li, Tiejun Huang, and Yonghong Tian. Ultra-high temporal resolution visual reconstruction from a fovea-like spike camera via spiking neuron model. *IEEE TPAMI*, 2022. 1, 2, 7

[78] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *CVPR*, pages 9308–9316, 2019. 2, 6

[79] Yaoyu Zhu, Yu Zhang, Xiaodong Xie, and Tiejun Huang. An fpga accelerator for high-speed moving objects detection and tracking with a spike camera. *Neural Computation*, 34 (8):1812–1839, 2022. 2