Recognizing Ultra-High-Speed Moving Objects with Bio-Inspired Spike Camera

Junwei Zhao[1,2], **Shiliang Zhang**[1 †], **Zhaofei Yu**[1,2 †], **Tiejun Huang**[1,2]

[1]National Key Laboratory for Multimedia Information Processing, School of Computer Science, Peking University
[2]Institute for Artificial Intelligence, Peking University
{jwz, slzhang.jdl, yuzf12, tjhuang}@pku.edu.cn

Abstract

Bio-inspired spike camera mimics the sampling principle of primate fovea. It presents high temporal resolution and dynamic range, showing great promise in fast-moving object recognition. However, the physical limit of CMOS technology in spike cameras still hinders their capability of recognizing ultra-high-speed moving objects, e.g., extremely fast motions cause blur during the imaging process of spike cameras. This paper presents the first theoretical analysis for the causes of spiking motion blur and proposes a robust representation that addresses this issue through temporal-spatial context learning. The proposed method leverages multi-span feature aggregation to capture temporal cues and employs residual deformable convolution to model spatial correlation among neighbouring pixels. Additionally, this paper contributes an original real-captured spiking recognition dataset consisting of 12,000 ultra-high-speed (equivalent speed >500 km/h) moving objects. Experimental results show that the proposed method achieves 73.2% accuracy in recognizing 10 classes of ultra-high-speed moving objects, outperforming existing spike-based recognition methods. Resources will be available at https://github.com/Evin-X/UHSR.

Introduction

Conventional frame cameras suffer from visual cue loss and severe motion blur in high-speed scenarios due to their limited frame rate (e.g., 30 fps) and single-exposure imaging principle, as shown in Fig. 1 (a). In contrast, bio-inspired spike cameras simulate the sensing principle of retinal photosensitive cells, where each pixel perceives light independently and generates spikes asynchronously (Zheng et al. 2023c). This unique imaging principle allows spike cameras to achieve a sampling frequency that is $1000\times$ higher than human vision (Huang et al. 2022), enhancing their capability in capturing high-speed moving objects, as shown in Fig. 1 (b). Previous efforts (e.g., Zhao et al. (2021b), Hu et al. (2022), Zhang et al. (2022a), Zhao et al. (2022)) have shown promising advantages of spike cameras in recording and recognizing high-speed objects over frame cameras.

Current spike cameras are implemented using CMOS technology. The photo-electric conversion time of each pixel sets a limit on the maximum sampling frequency of spike signals (El-Desouki et al. 2009). If the speed of a moving object exceeds a theoretical upper threshold (e.g., 500 km/h), the spike signals will become distorted, resulting in information loss during the imaging process. Consequently, the brightness intensity maps generated from spike streams will present blur, as depicted in Fig. 1 (c). We designate this phenomenon as spiking motion blur. In spiking vision community, research on spiking motion blur is still in its early stage. This work is hence motivated to explore the capability of spike cameras in recording ultra-high-speed motions that exceed their physical limit.

Moreover, the lack of datasets has hindered the research on ultra-high-speed moving object recognition. Existing spiking datasets, as shown in Table 1, mainly consist of synthetic data generated from video frames rather than real data. The speeds of moving objects in synthetic data are constrained by video frame rates, which are much lower than the sampling frequency of spike cameras. Additionally, the majority of these datasets are designed for pixel-level vision tasks (Zheng et al. 2023b) such as depth estimation, optical flow estimation, and reconstruction, making them unsuitable for instance-level tasks. Although Zhao et al. (2023a) and Zhao et al. (2023b) introduced datasets for neuromorphic recognition, the speeds of moving objects in these datasets are lower than ultra-high speed. Therefore, a dataset featuring ultra-high-speed motions is required.

We theoretically analyze the causes of motion blur in the spike camera imaging process. First, we establish the relationship between moving objects and spike signal generation. Then, we analyze the distortion conditions of spike signal sampling based on the Shannon sampling theorem and find that spiking motion blur is caused by temporal undersampling and spatial misalignment. Based on findings of the analysis, we propose a robust representation learning method that utilizes the temporal-spatial contexts of spike streams to address the issues of spiking motion blur. We employ multi-span dilated convolution in the temporal domain to extract temporal features and perform re-weighting aggregation on the extracted temporal features from different spans. In the spatial domain, we utilize cascaded residual deformable convolution to capture the correlation among neighbouring pixels. Finally, we integrate the temporal and spatial features through cross-attention fusion.

† Corresponding authors

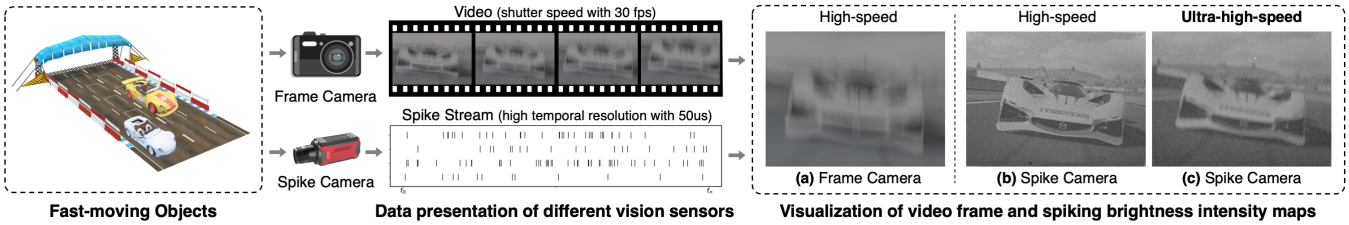


Figure 1: Comparison of fast-motion recordings. (a) is captured by a frame camera, suffering from significant motion blur. (b)-(c) are generated from spike streams recorded by a spike camera, which presents stronger capability in capturing high-speed motions, but is still hindered by the physical limit in capturing ultra-high speeds. (Details in the Experiment Section)

Spiking Dataset	Vision Task	Reference	Year	Sim or Real Data?	High-Level Vision Task?	Ultra-High Speed?
S-DENSE (Zhang et al. 2022a)	Depth Estimation	ECCV	2022	Sim	×	×
S-KITTI (Wang et al. 2022)	Depth Estimation	ICME	2022	Sim	×	×
RSSF (Zhao et al. 2022)	Flow Estimation	NeurIPS	2022	Sim	×	×
SPIFT (Hu et al. 2022)	Flow Estimation	CVPR	2022	Sim	×	×
Spk-Vimeo (Xiang et al. 2021)	Reconstruction	T-CSVT	2021	Sim	×	×
Spk-REDS (Zhao et al. 2021a)	Reconstruction	CVPR	2021	Sim	×	×
PKU-Recon (Zhu et al. 2020)	Reconstruction	CVPR	2020	Real	×	×
SpiReco (Zhao et al. 2023b)	Recognition	T-CSVT	2023	Real	✓	×
HSSR (Zhao et al. 2023a)	Recognition	ACM MM	2023	Real	✓	×
UHSR (Ours)	Recognition	AAAI	2024	Real	✓	✓

Table 1: Differences between the UHSR and other existing datasets.

Besides above methods, this paper contributes a spiking dataset for Ultra-High-Speed object Recognition, named UHSR dataset. Specifically, we construct the UHSR dataset using a motion platform similar to (Zhao et al. 2023b), which allows us to capture ultra-high-speed objects in a laboratory environment. The platform provides motions with equivalent speeds exceeding 500 km/h, significantly faster than existing datasets. Experimental results show that, our method boosts the baseline by 8.3% accuracy on recognizing 101 classes of ultra-high-speed moving objects. Besides, our method achieves state-of-the-art performances on the high-speed SpiReco and ultra-high-speed UHSR datasets.

Our contributions can be summarized as follows:

- To the best of our knowledge, we present the first theoretical analysis for underlying causes of spiking motion blur, which reveals the physical limit of current spike cameras in recording high-speed motions.
- We propose an original method for recognizing ultra-high-speed moving objects. Our method effectively addresses the issue of spiking motion blur through temporal-spatial context learning.
- We contribute a new spiking recognition dataset featuring ultra-high-speed motions, where objects are recorded at equivalent speeds exceeding 500 km/h at a camera-object distance of 10 meters.

Related Work

Overview of Spike Camera Model

This section briefly introduces the working principle of the spike camera. Each pixel consists of a photon-receptor, integrator, and comparator. The photon-receptor records photons, which are accumulated by the integrator and converted into voltage. The comparator continuously compares the voltage with a threshold θ. When the voltage exceeds θ, a spike is emitted and the accumulation is reset. The spike generation process on a pixel can be expressed as,

$$\int_{t}^{t+nT_r} \lambda I(t)dt \geq \theta, \tag{1}$$

where $I(t)$ is the brightness intensity of a pixel at time t, λ denotes photoelectric conversion rate, and T_r is the temporal resolution (e.g., $50\mu s$). Spike stream s is a binary array with the dimension of $\mathbb{R}^{T\times H\times W}$, where T is the time length, H and W are the height and width of the sensor. More details of spike cameras can be found in (Huang et al. 2022).

Neuromorphic Recognition Methods

In recent years, Spiking Neural Networks (SNNs) have gained attention for processing neuromorphic data, with various proposed models such as (Zheng et al. 2021; Fang et al. 2021; Meng et al. 2022; Wang et al. 2023b). However, there is a lack of research evaluating the performance of SNNs on spiking data recorded by spike cameras.

Currently, several efforts have been proposed for processing spike camera data. Zheng et al. (2023c) recovered brightness using spike signal intervals, while Zhao et al. (2022) introduced a spiking representation based on spike firing time differentials. Moreover, Zhao et al. (2023b) developed a denoised and motion-enhanced framework for recognizing high-speed moving objects. However, none of these approaches tackled the challenge of spiking motion blur resulting from ultra-high-speed motions.

Unlike spike cameras inspired by the foveal retina, event cameras are inspired by the peripheral retina (Gallego et al.

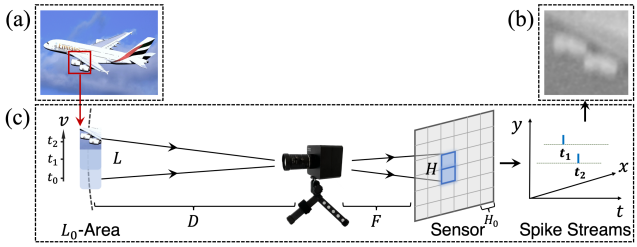


Figure 2: (a) ultra-high-speed motion results in (b) spiking motion blur. (c) illustrates the relationship between moving objects and their correlated spike streams.

2020). Each pixel in event cameras operates asynchronously and generates events when brightness change exceeds a threshold (Zheng et al. 2023a). Several event-based processing methods including (Wang et al. 2023a; Peng et al. 2023; Sun et al. 2023) rely on precise timestamps and polarity information, which are not available in spike streams. This makes them unsuitable for processing spiking data.

Analysis of Spiking Motion Blur

This section establishes the relationship between a moving object and its corresponding spike signals to analyze the cause of spiking motion blur. Based on the spike generation principle in Eq. (1), the time Δt required for emitting a spike can be calculated as,

$$\Delta t = nT_r \geq \frac{\theta}{\lambda \bar{I}}\,, \tag{2}$$

where $\bar{I}$ is the average brightness intensity in a short time interval. The perceived $\bar{I}$ by each pixel can be estimated as,

$$\bar{I} = \frac{\theta}{\lambda \Delta t_s}\,, \tag{3}$$

where Δt_s is the time interval between two adjacent spikes.

As shown in Fig. 2, assuming an object at distance D moves at speed v, and projects an inverted image on the sensing chip through the lens. The projected pixels keep recording the brightness independently and emit a spike once the voltage exceeds a threshold. According to the law of convex lens imaging, the relationship between the length of the object L and its image H can be represented as,

$$\frac{F}{H} = \frac{D}{L}\,, \tag{4}$$

where F denotes the focal lens. Given that, a pixel of size $H_0 \times H_0$ captures the brightness of a high-speed moving area $L_0 \times L_0$ (denoted as L_0-area) located at a distance D_0 from the lens. Following the Nyquist–Shannon sampling theorem, a pixel needs to emit a minimum of two spikes to ensure sufficient information sampling for the L_0-area. Hence, the Eq. (4) can be rewritten as,

$$\frac{F}{H_0} = \frac{D}{v(2\Delta t)}\,. \tag{5}$$

Substituting Eq. (2) into Eq. (5) yields the formula,

$$v \leq \eta \bar{I} D/F\,, \quad (\eta = \lambda H_0/2\theta)\,, \tag{6}$$

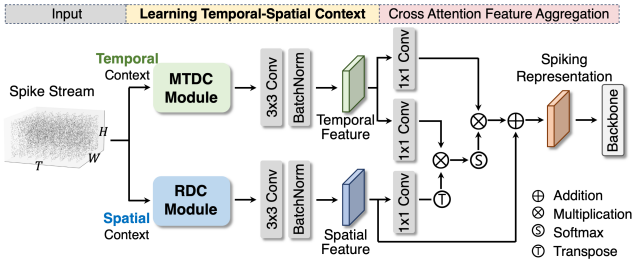


Figure 3: The overall architecture of the proposed TSC method. The MTDC module and RDC module learn the temporal and spatial context of spike streams, respectively. The generated features are fused by the CAFA module.

where η is a constant determined by sensor parameters of λ, H_0 and θ. Eq. (6) states that in a given scene with parameters such as $\bar{I}$, D, and F, if a spike camera is capable of capturing the moving objects clearly, there exists an upper bound on the moving speed. We denote this speed upper bound as,

$$V^* = \eta \bar{I} D/F. \tag{7}$$

When the object speed exceeds V^*, it causes temporal undersampling of brightness intensity, leading to spatial misalignment between the object and its spike signals. Additionally, we find that V^* is influenced by scene parameters, especially the brightness intensity $\bar{I}$, when D and F are fixed. A higher $\bar{I}$ results in a larger V^*, suggesting a correlation between spike camera performance and scene brightness intensity. In summary, given a camera and scene, if the object speed exceeds the upper bound or the scene brightness is too weak, motion blur may occur in the recorded spiking data.

Proposed Method

Given a spike stream $s \in \mathbb{R}^{T \times H \times W}$, our goal is to accurately recognize objects from s, which can be formulated as,

$$\Omega^* = \arg\min_{\Omega} \mathcal{L}(\mathcal{F}(S, \Omega), Y)\,, \tag{8}$$

where Ω represents parameters of the recognition model $\mathcal{F}$, $\mathcal{L}$ denotes the loss function of $\mathcal{F}$, and $S = \{s_0, s_1, ..., s_N\}$ denotes the spiking dataset with labels Y. Based on the spiking motion blur analysis, we propose a learning-based spiking representation $\mathcal{R}$ to tackle the issue of spiking motion blur by considering both the Temporal and Spatial Context (TSC), as illustrated in Fig 3. Therefore, the model $\mathcal{F}$ can be represented as $\mathcal{F} = \mathcal{B}(\mathcal{R}(\cdot))$, where $\mathcal{B}$ denotes the Backbone such as ResNet (He et al. 2016) and VGGNet (Simonyan and Zisserman 2014).

In the temporal domain, considering the continuity of motion process, errors in the recorded brightness intensity at time t_i can be compensated by incorporating the brightness information from neighbouring time steps. However, determining appropriate length of the neighboring time window is challenging due to various motion speeds of objects. To address this, we introduce the Multi-span Temporal-Dilated Convolution (MTDC) module that extracts temporal features at multi-scales and performs learnable re-weighted fusion of these features.

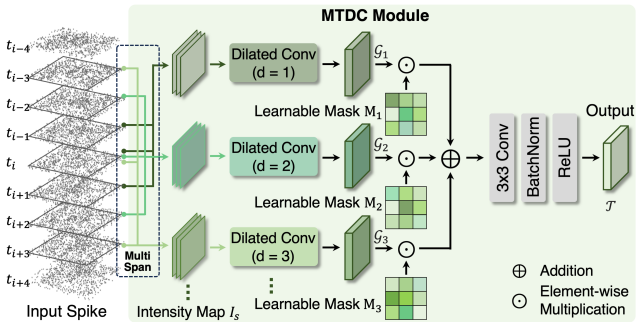


Figure 4: Structure of the MTDC module.

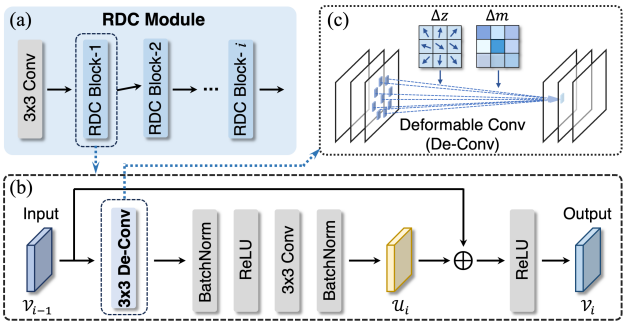


Figure 5: Structure of the RDC module. (a) RDC module. (b) RDC block. (c) Deformable convolution.

In the spatial domain, the high speed of moving objects leads to spatial misalignment of spikes, causing the brightness information of moving objects to be encoded in a neighbourhood of pixels rather than a single pixel. Determining the appropriate spatial neighbourhood is difficult due to diverse motion patterns. To tackle this, we introduce a spatial feature extractor based on Residual Deformable Convolution (RDC) blocks, and employ a cascaded architecture to enhance the feature extraction capability of this module.

To model the feature-level correlation between temporal and spatial domains, we employ Cross-Attention Feature Aggregation (CAFA) module to integrate temporal-spatial features, generating a robust spiking representation for improving recognition accuracy.

The structures of MTDC, RDC, and CAFA modules are detailed in the following sub-sections. We optimize the weights of representation learning modules in an end-to-end manner, jointly training the backbone and representation modules using cross-entropy loss.

Multi-span Temporal-Dilated Convolution

We estimate brightness intensity from spike streams at each time step t using Eq. (3). The intensity maps I_s are fed into multiple dilated convolution blocks, shown in Fig. 4, where each block conducts dilated convolutions with different dilated rates d to capture features at multiple temporal scales. Notably, the dilation is only applied along the temporal axis. In contrast to 3D convolution kernels, our algorithm mitigates the information overlap in the temporal domain and reduces parameter size, which facilitates network training. The generated feature map $\mathcal{G}$ is calculated as,

$$\mathcal{G}_d(t,h,w) = \sum_{\tau}\sum_{i,j} \kappa_d(\hat{\varepsilon}+\tau, \hat{\varepsilon}+i, \hat{\varepsilon}+j) \odot I_s(t+d\tau, h+i, w+j), \tag{9}$$

where $\kappa_d \in \mathbb{R}^{\varepsilon\times\varepsilon\times\varepsilon}$ are weights of a dilated convolution kernel with the dilated rate d, $\hat{\varepsilon} = \lfloor \frac{\varepsilon}{2} \rfloor$, $-\hat{\varepsilon} \le \tau, i, j \le \hat{\varepsilon}$, and $t = 1, 2, .., T$ is the time length of a input spike stream.

Experimental results illustrated in Fig. 9 show that the temporal correlation in spike streams weakens as the temporal span increases. Therefore, the feature maps $\mathcal{G}_d$, corresponding to different dilated rates d, are combined using element-wise multiplication $\odot$ with learnable masks $\mathbf{M}_d$. This enables the aggregation of features across various temporal spans with adaptive weightings. In this way, the temporal feature map $\mathcal{T}$ can be obtained by,

$$\mathcal{T} = f_a(\sum_d \mathbf{M}_d \odot \mathcal{G}_d), \tag{10}$$

where $f_a(\cdot)$ represents the operation pipeline consisting of 3×3 convolution, Batch Normalization (BN), and ReLU.

Residual Deformable Convolution

The RDC module takes brightness intensity maps I_s as input and extracts spatial context from neighbouring pixels. The input I_s is passed through a 3×3 convolutional layer and then processed by several cascaded RDC blocks as illustrated in Fig. 5 (a). The cascaded structure is adopted to enhance the capability of local feature extraction. Each RDC block, as shown in Fig. 5 (b), introduces deformable convolution (Zhu et al. 2019) to achieve a flexible perception range for modelling various motion cues. For a deformable convolution kernel with K sampling locations, as depicted in Fig. 5 (c), the weight and offset for the k-th location are denoted as ω_k and $\boldsymbol{z}_k$, respectively. The output feature map $\mathcal{U}$ at each position $\boldsymbol{z} = (h, w)$ can be computed as,

$$\mathcal{U}(t,\boldsymbol{z}) = \sum_k \omega_k \cdot I_s(t, \boldsymbol{z}+\boldsymbol{z}_k+\Delta\boldsymbol{z}_k)\cdot\Delta m_k, \tag{11}$$

where $\Delta\boldsymbol{z}_k$ and Δm_k are the learnable offsets and modulation scalar for the k-th location, respectively. Accordingly, the feature map $\mathcal{V}$ produced by the i-th RDC block can be obtained by,

$$\mathcal{V}_i = ReLU(\mathcal{V}_{i-1} \oplus f_b(\mathcal{U}_i)), \tag{12}$$

where $f_b(\cdot)$ represents the operation pipeline consisting of BN, ReLU, 3×3 convolution, and another BN. The feature map $\mathcal{V}_{i-1}$ generated by the previous RDC block is added $\oplus$ via a residual connection.

Cross-Attention Feature Aggregation

We employ cross-attention to model the relation between temporal and spatial domains. As shown in Fig. 3, the robust spiking representation $\mathcal{R}$ is generated by aggregating the temporal features $\mathcal{T}$ and spatial features $\mathcal{V}$. The computation process can be formulated as,

$$\mathcal{R} = Softmax(f_c(\mathcal{T}) \otimes f_c(\mathcal{V})^T) \otimes f_c(\mathcal{T}) \oplus f_d(\mathcal{V}), \tag{13}$$

where $f_c(\cdot)$ represents the operation pipeline consisting of 3×3 convolution, BN, 1×1 convolution, and $f_d(\cdot)$ represents the operation pipeline consisting of 3×3 convolution, BN.

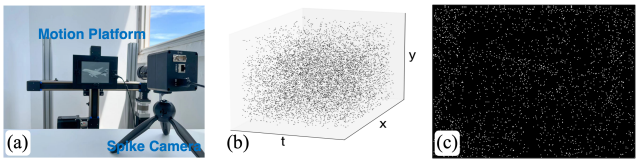


Figure 6: (a) High-speed motion platform for constructing the UHSR. (b)-(c) Visualization of spikes in 3D and 2D.

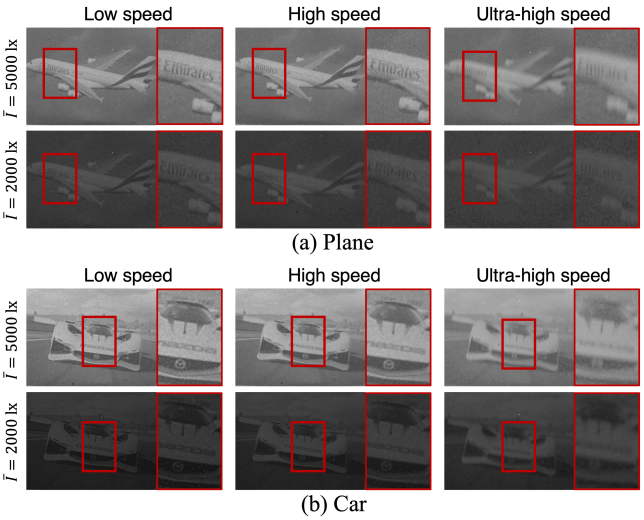


Figure 7: Visualization to the brightness intensity maps under each experimental setting recorded in Table 2.

Experiment

Dataset

(i) the SpiReco (Zhao et al. 2023b) is a collection of high-speed moving object recognition datasets captured using a spike camera. It consists of various motion patterns with different speeds. The sub-datasets of SpiReco include S-CIFAR (S-CIF) with 10 classes of 10,000 samples and S-CALTECH (S-CAL) with 101 classes of 8,710 samples. The test sets for each sub-dataset contain 1,500 samples.
(ii) the UHSR proposed in this study is a pioneering dataset for ultra-high-speed spiking recognition, as there is currently no dedicated dataset in this field. The related dataset, i.e., SpiReco, only covers motion speeds lower than 500 km/h, making it challenging to evaluate methods for ultra-high-speed motions. To collect UHSR, we employ a similar data collection platform to as shown in Fig. 6 (a), which simulates objects moving at speeds equivalent to 500∼700 km/h. The speed calculation follows the approach described in (Zhao et al. 2023b). We randomly select 6,000 images from CIFAR-10 and 6,000 images from CALTECH-101 for annotation, creating the Ultra-high-speed CIFAR (U-CIF) and Ultra-high-speed CALTECH (U-CAL) datasets, respectively. U-CIF contains 10 classes, and U-CAL consists of 101 classes. Training and testing sets are spilt as a ratio of 5:1. UHSR facilitates the evaluation of methods designed for ultra-high-speed spiking recognition.

Experimental Settings				Evaluation Results	
No.	Scene	$\bar{I}$ (lx)	v (km/h)	TDE↑	BIQI↓
1-1	Plane	∼5000	7.9 ($0.6V^*$)	**11.26**	**52.01**
1-2			13.0 ($1.0V^*$)	11.19	53.94
1-3			18.0 ($1.4V^*$)	10.53	67.40
1-4		∼2000	3.3 ($0.6V^*$)	**10.34**	**59.96**
1-5			5.4 ($1.0V^*$)	10.31	61.70
1-6			7.6 ($1.4V^*$)	9.84	76.35
2-1	Car	∼5000	7.9 ($0.6V^*$)	**11.14**	**56.33**
2-2			13.0 ($1.0V^*$)	11.09	58.82
2-3			18.0 ($1.4V^*$)	10.23	74.36
2-4		∼2000	3.3 ($0.6V^*$)	**10.17**	**61.47**
2-5			5.4 ($1.0V^*$)	10.03	63.95
2-6			7.6 ($1.4V^*$)	9.35	79.84

Table 2: Experimental settings and evaluation results.

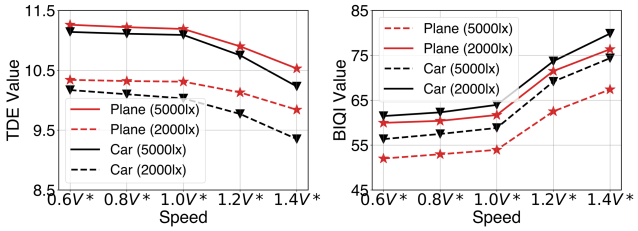


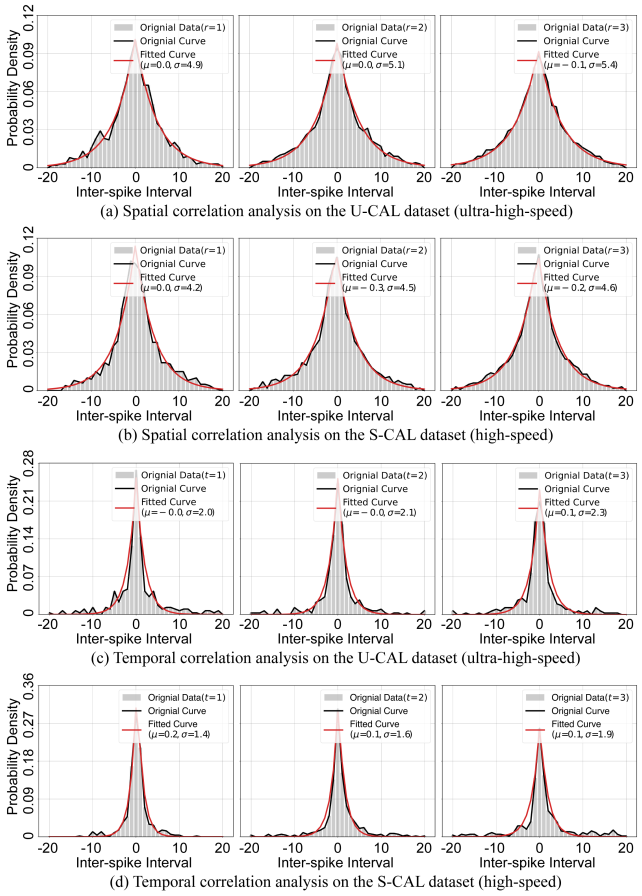
Figure 8: Evaluation results of spiking motion blur.

Implementation

We adopt the ImageNet pre-trained ResNet-18 as the backbone. Model parameters are trained using the SGD optimizer without data augmentation. The initial learning rate is set to 2e-4 with the LambdaLR scheduler. The training process runs for 30 epochs for each dataset with a batch size of 16. Input samples are downsampled to 124×124 using average pooling. θ/λ is set to 1, constraining the value of brightness intensity maps within [0, 1]. The framework is implemented in PyTorch and trained on NVIDIA RTX 4090 GPUs.

Validation of Spiking Motion Blur Analysis

We experimentally validate the analysis of spiking motion blur. High-speed motions are provided by the motion platform shown in Fig. 6 (a), driving the monitor at speeds ranging from 0 to 18 km/h. The camera-monitor distance is $D = 0.35m$, and the focal length is $F = 8mm$. The constant $\eta \approx 1.67e\text{-}5$ in Eq. (7) is calculated based on the spiking sensor parameters from (Huang et al. 2022). Experiments are conducted under two luminance conditions with two moving objects, i.e., an airplane and a racing car. The upper bound of moving speed V^* is calculated for each setup according to Eq. (7). We set the monitor motion speeds from 0.6 V^* to 1.4 V^* and record the spike streams. Brightness intensity maps are generated using Eq. (3), as depicted in Fig. 7. This approach estimates brightness intensity without optimization or reference information, and the quality of the reconstructed maps reflects the quality of spike streams, assuming the impact of spike noise remains stable.



(a) Spatial correlation analysis on the U-CAL dataset (ultra-high-speed)

(b) Spatial correlation analysis on the S-CAL dataset (high-speed)

(c) Temporal correlation analysis on the U-CAL dataset (ultra-high-speed)

(d) Temporal correlation analysis on the S-CAL dataset (high-speed)

Figure 9: Spatial and temporal correlation analysis of spiking data in the UHSR and SpiReco datasets.

We quantitatively evaluate the reconstructed maps using Two-Dimensional Entropy (TDE) (Xi, Guosui, and Ni 1999) and Blind Image Quality Index (BIQI) (Moorthy and Bovik 2009) as metrics. TDE indicates better image quality with higher values, while BIQI indicates better image quality with lower values. The evaluation results shown in Fig. 8 and summarized in Table 2 indicate that, brightness intensity maps generated from scenes with $v \leq V^*$ exhibit higher quality compared to those from scenes with $v > V^*$, and the quality decreases with increasing of v. Because spike signals become distorted when the motion speed exceeds V^*. Additionally, scenes with lower luminance, such as 2000 lx, manifest more prominent motion blur as the spike camera is sensitive to scene brightness. Fig. 7 visualizes those reconstructed maps, illustrating that scenes with motion speed $v > V^*$ present motion blur. These experimental results validate the effectiveness of the spiking motion blur analysis.

Ablation Study of Proposed Method

We experimentally investigate the temporal-spatial correlation of spike streams by randomly selecting 300 samples from UHSR and SpiReco, respectively. For each sample, we compute the similarity of inter-spike intervals within local pixel regions in the spatial domain, and within a certain time window for the same pixel in the temporal domain. Results are shown in Fig. 9, where each sub-figure displays the original data, the original data curve, and the curve fitted using the Laplace distribution. The Probability Density Function (PDF) of Laplace distribution can be expressed as,

$$f(x|\mu,\sigma) = \frac{1}{2\sigma}\exp\left(-\frac{|x-\mu|}{\sigma}\right), \tag{14}$$

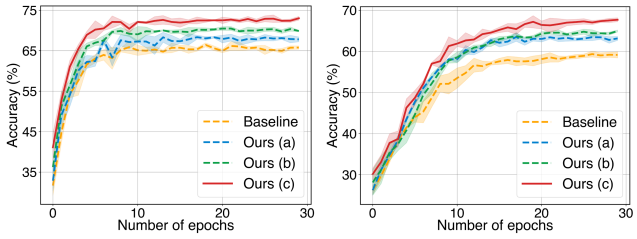
where x denotes the difference in inter-spike intervals, μ is the mean, and σ is the scale parameter. When σ is small, the PDF presents a steeper peak, leading to a more concentrated distribution and smaller deviation of variables around the μ.

Experimental results in Fig. 9 exhibit significant correlations in spike interval distributions within certain spatial (e.g., $r \leq 3$) and temporal (e.g., $t \leq 3$) ranges, with r indicating spatial region radius and t denoting temporal span. Specifically, the correlation decreases with the increasing of r and t. For instance, as shown in Fig. 9 (a), when $r = 1$, $\sigma = 4.9$, and when $r = 3$, $\sigma = 5.4$. Similarly, in Fig. 9 (c), when $t = 1$, $\sigma = 2.0$, and when $t = 3$, $\sigma = 2.3$. In our model design, the RDC module employs a 3-level cascaded structure, and the MTDC module adopts a dilated rate of $d = 3$. Moreover, the results also demonstrate that the spatial and temporal correlation on U-CAL is weaker than S-CAL, due to objects in U-CAL moving at higher speeds.

We then validate the effectiveness of MTDC and RDC modules, and summarize results in Table 3. Experimental setups are described in the Implementation Section. The baseline model is ResNet-18 and input streams are uniformly 3 ms. Fig. 10 visualizes the ablation experimental results on the UHSR dataset averaged by 3 runs. The results demonstrate that both MTDC and RDC modules effectively improve recognition accuracy. Notably, our method achieves a more substantial performance improvement on the UHSR, with an 8.3% increase on U-CAL and a 5.3% increase on S-CAL compared to the baseline. These results demonstrate that temporal-spatial context learning effectively improves the accuracy of recognizing ultra-high-speed objects.

Model	Module		SpiReco		UHSR	
	M1	M2	S-CIF	S-CAL	U-CIF	U-CAL
Baseline	-	-	55.1%	69.3%	65.7%	59.2%
Ours (a)	✓	-	57.6%	71.7%	68.3%	63.4%
Ours (b)	-	✓	58.9%	73.6%	70.5%	65.1%
Ours (c)	✓	✓	**60.8%**	**74.6%**	**73.2%**	**67.5%**

Table 3: Recognition accuracy of ablation experiments on MTDC (M1) and RDC (M2) modules.



(a) On the U-CIF dataset (b) On the U-CAL dataset

Figure 10: Visualization to ablation experimental results on the UHSR dataset.

Type	Method	Reference	Backbone Structure	SpiReco		UHSR	
				S-CIF	S-CAL	U-CIF	U-CAL
SNN	TDBN (Zheng et al. 2021)	AAAI'21	LIF Res-19	52.8%	62.2%	60.3%	53.1%
	SEWR (Fang et al. 2021)	NeurIPS'21	SEW Res-18	53.6%	63.7%	61.5%	53.9%
	DSR (Meng et al. 2022)	CVPR'22	LIF Res-18	53.2%	64.1%	61.2%	54.5%
Event	EtoF (Ahmad et al. 2022)	WACV'22	ResNet-18	53.9%	66.3%	61.8%	55.7%
	BEI (Cohen et al. 2018)	T-NNLS'18	ResNet-18	50.7%	61.8%	56.6%	50.3%
	SBNE (Zhang et al. 2022b)	CVPR'22	ResNet-18	54.2%	67.4%	62.3%	56.5%
Spike	TSR (Zhao et al. 2022)	NeurIPS'22	ResNet-18	56.3%	70.4%	65.9%	60.4%
	TFP (Zheng et al. 2023c)	T-PAMI'23	ResNet-18	55.6%	69.8%	63.1%	56.7%
	ISI (Zheng et al. 2023c)	T-PAMI'23	ResNet-18	55.1%	69.3%	65.7%	59.2%
	DMER (Zhao et al. 2023b)	T-CSVT'23	ResNet-18	57.9%	71.2%	67.8%	62.4%
	TSC (Ours)	AAAI'24	ResNet-18	**60.8%**	**74.6%**	**73.2%**	**67.5%**

Table 4: Comparison with SoTA methods on SpiReco (high-speed) and UHSR (ultra-high-speed) datasets.

Spike	U-CIF			U-CAL		
Methods	1 ms	3 ms	5 ms	1 ms	3 ms	5 ms
TSR	51.9%	65.9%	69.3%	43.2%	60.4%	64.5%
TFP	40.5%	63.1%	64.2%	30.4%	56.7%	61.9%
ISI	29.2%	65.7%	68.7%	17.6%	59.2%	63.3%
DMER	53.7%	67.8%	70.5%	46.3%	62.4%	64.7%
Ours	**62.4%**	**73.2%**	**73.9%**	**51.6%**	**67.5%**	**68.4%**

Table 5: Evaluation to the robustness of different methods.

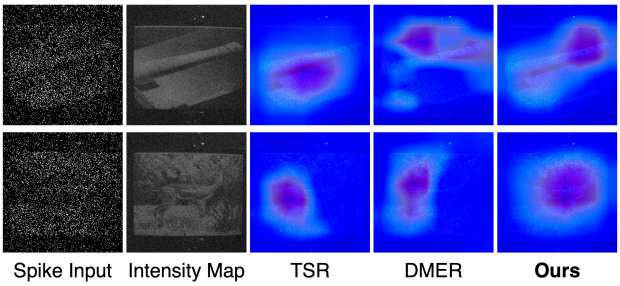


Figure 11: Visual inspection of TSR, DMER, and Ours using Grad-CAM (Selvaraju et al. 2017). Best viewed in color.

Comparison with State-of-The-Art Methods

As summarized in Table 4, we compare with SoTA spike-based methods including TSR, TFP, ISI and DMER. We also compare with SNNs and event-based methods. SNNs include TDBN, SEWR and DSR. For SNNs, spike streams are processed as sequential inputs with T time steps. Many event-based methods require precise timestamps for processing (Kim et al. 2021), which are not available in spiking data. Hence, we compare with those methods not requiring precise timestamps, including EtoF, BEI, and SBNE. To ensure a fair comparison, the input spike streams have a uniform time length of 3 ms and a spatial size of 124×124. Training is done for 30 epochs using ResNet-18 as the backbone for each method. Learning rates and optimizers follow recommendations from the official code of each method.

Training SNNs on spiking data poses optimization challenges, limiting their performances. Additionally, spike and event streams have different data representations, which may degrade the performance of event-based methods on spiking datasets even under an identical experimental setup. Compared to SoTA spike-based processing methods like TSR, TFP, ISI, and DMER, our method achieves substantial improvements in recognition accuracy, with a 5.4% and 5.1% increase on UHSR, and a 3.4% and 2.9% increase on SpiReco, respectively. These results demonstrate the effectiveness of our proposed method in addressing spiking motion blur, leading to enhanced recognition accuracy for ultra-high-speed moving objects.

We conduct robustness testing on spike-based methods to evaluate their performances under different input lengths of spike streams, as the length of captured data is different due to diverse motion speeds and limited camera view angles in real scenarios. We compare with others using 1 ms, 3 ms, and 5 ms spike streams as input, while keeping the other experimental settings consistent with those in Table 4. The results in Table 5 demonstrate that our method consistently outperforms other methods, showcasing its robustness.

We further apply Grad-CAM on spiking feature maps to show the interest area of models. We compare with learning-based spike processing methods, i.e., TSR and DMEM. The visualization in Fig. 11 shows that our method pays more attention to distinctive regions of moving objects compared with other methods. Additionally, results also show that the models trained on UHSR focus on the features of moving objects rather than the moving screen.

Conclusion

In this paper, we present the first theoretical analysis of spiking motion blur caused by ultra-high-speed motions, and validate the analysis through extensive experiments. Based on the analysis, we propose a robust spiking representation that learns the temporal-spatial context of spike streams, effectively improving recognition accuracy for ultra-high-speed objects. Experimental results demonstrate the superior accuracy and enhanced robustness of our method. Additionally, we construct an original spiking dataset for ultra-high-speed object recognition to facilitate further research.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China under Grant No. 62088102, U20B2052, 61936011, in part by the Okawa Foundation Research Award.

References

Ahmad, S.; Scarpellini, G.; Morerio, P.; and Del Bue, A. 2022. Event-driven Re-Id: A New Benchmark and Method Towards Privacy-Preserving Person Re-Identification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 459–468.

Cohen, G.; Afshar, S.; Orchard, G.; Tapson, J.; Benosman, R.; and van Schaik, A. 2018. Spatial and temporal downsampling in event-based visual classification. *IEEE Transactions on Neural Networks and Learning Systems*, 29(10): 5030–5044.

El-Desouki, M.; Jamal Deen, M.; Fang, Q.; Liu, L.; Tse, F.; and Armstrong, D. 2009. CMOS image sensors for high speed applications. *Sensors*, 9(1): 430–444.

Fang, W.; Yu, Z.; Chen, Y.; Huang, T.; Masquelier, T.; and Tian, Y. 2021. Deep residual learning in spiking neural networks. In *Advances in Neural Information Processing Systems*, volume 34, 21056–21069.

Gallego, G.; Delbrück, T.; Orchard, G.; Bartolozzi, C.; Taba, B.; Censi, A.; Leutenegger, S.; Davison, A. J.; Conradt, J.; Daniilidis, K.; et al. 2020. Event-based vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1): 154–180.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Hu, L.; Zhao, R.; Ding, Z.; Ma, L.; Shi, B.; Xiong, R.; and Huang, T. 2022. Optical Flow Estimation for Spiking Camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17844–17853.

Huang, T.; Zheng, Y.; Yu, Z.; Chen, R.; Li, Y.; Xiong, R.; Ma, L.; Zhao, J.; Dong, S.; Zhu, L.; et al. 2022. $1000\times$ faster camera and machine vision with ordinary devices. *Engineering*.

Kim, J.; Bae, J.; Park, G.; Zhang, D.; and Kim, Y. M. 2021. N-ImageNet: Towards robust, fine-grained object recognition with event cameras. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2146–2156.

Meng, Q.; Xiao, M.; Yan, S.; Wang, Y.; Lin, Z.; and Luo, Z.-Q. 2022. Training High-Performance Low-Latency Spiking Neural Networks by Differentiation on Spike Representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12444–12453.

Moorthy, A.; and Bovik, A. 2009. A modular framework for constructing blind universal quality indices. *IEEE Signal Processing Letters*, 17: 7.

Peng, Y.; Zhang, Y.; Xiao, P.; Sun, X.; and Wu, F. 2023. Better and Faster: Adaptive Event Conversion for Event-Based Object Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 2056–2064.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, 618–626.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Sun, L.; Zhang, Y.; Cheng, J.; and Lu, H. 2023. Asynchronous Event Processing with Local-Shift Graph Convolutional Network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 2402–2410.

Wang, D.; Jia, X.; Zhang, Y.; Zhang, X.; Wang, Y.; Zhang, Z.; Wang, D.; and Lu, H. 2023a. Dual Memory Aggregation Network for Event-Based Object Detection with Learnable Representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 2492–2500.

Wang, Q.; Zhang, T.; Han, M.; Wang, Y.; Zhang, D.; and Xu, B. 2023b. Complex dynamic neurons improved spiking transformer network for efficient automatic speech recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 102–109.

Wang, Y.; Li, J.; Zhu, L.; Xiang, X.; Huang, T.; and Tian, Y. 2022. Learning stereo depth estimation with bio-inspired spike cameras. In *2022 IEEE International Conference on Multimedia and Expo*, 1–6. IEEE.

Xi, L.; Guosui, L.; and Ni, J. 1999. Autofocusing of ISAR images based on entropy minimization. *IEEE Transactions on Aerospace and Electronic Systems*, 35(4): 1240–1252.

Xiang, X.; Zhu, L.; Li, J.; Wang, Y.; Huang, T.; and Tian, Y. 2021. Learning super-resolution reconstruction for high temporal resolution spike stream. *IEEE Transactions on Circuits and Systems for Video Technology*.

Zhang, J.; Tang, L.; Yu, Z.; Lu, J.; and Huang, T. 2022a. Spike Transformer: Monocular Depth Estimation for Spiking Camera. In *European Conference on Computer Vision*, 34–52. Springer.

Zhang, K.; Che, K.; Zhang, J.; Cheng, J.; Zhang, Z.; Guo, Q.; and Leng, L. 2022b. Discrete time convolution for fast event-based stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8676–8686.

Zhao, J.; Xiong, R.; Liu, H.; Zhang, J.; and Huang, T. 2021a. Spk2ImgNet: Learning to Reconstruct Dynamic Scene from Continuous Spike Stream. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11996–12005.

Zhao, J.; Xiong, R.; Xie, J.; Shi, B.; Yu, Z.; Gao, W.; and Huang, T. 2021b. Reconstructing clear image for high-speed motion scene with a retina-inspired spike camera. *IEEE Transactions on Computational Imaging*, 8: 12–27.

Zhao, J.; Ye, J.; Zhang, S.; Yu, Z.; and Huang, T. 2023a. Recognizing High-Speed Moving Objects with Spike Camera. In *Proceedings of the ACM International Conference on Multimedia*, 7657–7665.

Zhao, J.; Zhang, S.; Yu, Z.; and Huang, T. 2023b. SpiReco: Fast and Efficient Recognition of High-Speed Moving Objects with Spike Cameras. *IEEE Transactions on Circuits and Systems for Video Technology*.

Zhao, R.; Xiong, R.; Zhao, J.; Yu, Z.; Fan, X.; and Huang, T. 2022. Learning Optical Flow from Continuous Spike Streams. In *Advances in Neural Information Processing Systems*, volume 35, 7905–7920.

Zheng, H.; Wu, Y.; Deng, L.; Hu, Y.; and Li, G. 2021. Going deeper with directly-trained larger spiking neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 11062–11070.

Zheng, X.; Liu, Y.; Lu, Y.; Hua, T.; Pan, T.; Zhang, W.; Tao, D.; and Wang, L. 2023a. Deep learning for event-based vision: A comprehensive survey and benchmarks. *arXiv preprint arXiv:2302.08890*.

Zheng, Y.; Zhang, J.; Zhao, R.; Ding, J.; Chen, S.; Xiong, R.; Yu, Z.; and Huang, T. 2023b. SpikeCV: Open a Continuous Computer Vision Era. *arXiv preprint arXiv:2303.11684*.

Zheng, Y.; Zheng, L.; Yu, Z.; Huang, T.; and Wang, S. 2023c. Capture the Moment: High-speed Imaging with Spiking Cameras through Short-term Plasticity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhu, L.; Dong, S.; Li, J.; Huang, T.; and Tian, Y. 2020. Retina-like visual image reconstruction via spiking neural model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1438–1446.

Zhu, X.; Hu, H.; Lin, S.; and Dai, J. 2019. Deformable Convnets v2: More deformable, better results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9308–9316.