SOTA: Spike-Navigated Optimal TrAnsport Saliency Region Detection in Composite-bias Videos

Wenxuan Liu[1,2], **Yao Deng**[2], **Kang Chen**[1], **Xian Zhong**[2,*], **Zhaofei Yu**[3,1] and **Tiejun Huang**[1]
[1]State Key Laboratory for Multimedia Information Processing, Peking University
[2]Hubei Key Laboratory of Transportation Internet of Things, Wuhan University of Technology
[3]Institute for Artificial Intelligence, Peking University
liuwx66@pku.edu.cn, 361248@whut.edu.cn, mrchenkang@stu.pku.edu.cn,
zhongx@whut.edu.cn, {yuzf12, tjhuang}@pku.edu.cn

Abstract

Existing saliency detection methods struggle in real-world scenarios due to motion blur and occlusions. In contrast, spike cameras, with their high temporal resolution, significantly enhance visual saliency maps. However, the composite noise inherent to spike camera imaging introduces discontinuities in saliency detection. Low-quality samples further distort model predictions, leading to saliency bias. To address these challenges, we propose **S**pike-navigated **O**ptimal **Tr**A**n**sport Saliency Region Detection (SOTA), a framework that leverages the strengths of spike cameras while mitigating biases in both spatial and temporal dimensions. Our method introduces **S**pike-based **M**icro-debias (SM) to capture subtle frame-to-frame variations and preserve critical details, even under minimal scene or lighting changes. Additionally, **S**pike-based **G**lobal-debias (SG) refines predictions by reducing inconsistencies across diverse conditions. Extensive experiments on real and synthetic datasets demonstrate that SOTA outperforms existing methods by eliminating composite noise bias. Our code and dataset will be released at https://github.com/lwxfight/sota.

1 Introduction

Video saliency detection is essential for isolating objects from backgrounds across consistent frames [Zhao *et al.*, 2024c]. With the rapid advancement of digital media, it has become increasingly effective for surveillance applications, particularly in artificial intelligence [Liu *et al.*, 2022]. However, RGB cameras, constrained by short-exposure shutters, struggle with fast-moving objects and occlusions [Hu *et al.*, 2022], making continuous and accurate motion capture a significant challenge.

Human visual perception processes motion as continuous and uninterrupted [Sinha *et al.*, 2017]. Inspired by this, neuromorphic cameras abandon the conventional "frame" concept and output asynchronous sparse event streams [Liu *et al.*, 2024b], enabling high-speed object capture independent of shutter speed and frame rate [Zhou *et al.*, 2024]. However, they struggle to capture static scenes due to reliance on differential computation [Zhao *et al.*, 2024b]. In contrast, spike cameras [Dong *et al.*, 2017] employ a fovea-like sampling method (FSM), simulating the structure and function of the retina. This fusion of motion sensitivity and visual reconstruction allows effective tracking of continuously moving targets.

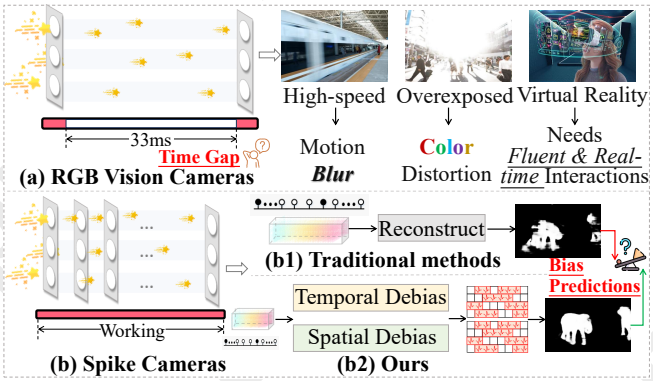


Figure 1: **Imaging Comparison.** (a) RGB vision cameras, which are affected by time gaps. (b) Spike camera imaging, where (b1) shows the traditional reconstruction method, influenced by environmental noise, leading to inaccurate saliency maps. (b2) presents the optimized reconstruction from a spatiotemporal perspective, resulting in a more accurate saliency map.

Despite their advantages, both traditional and neuromorphic cameras have limitations in detecting salient regions, as shown in Fig. 1(a) and (b). Traditional cameras suffer from motion blur, overexposed scenes, and real-time interaction challenges, often losing salient targets. While spike cameras mitigate some of these issues, their reliance on light imaging introduces new challenges. Extended motion over time can cause lighting variations, creating a domain gap in spiking data. A naive temporal transfer approach can introduce noise, amplifying negative effects. This raises the critical question: *How can we restore consistency in target regions while suppressing noise interference?*

To address this, we propose **S**pike-navigated **O**ptimal **Tr**A**n**sport Saliency Region Detection (SOTA) in Fig. 1(b2),

*Corresponding author.

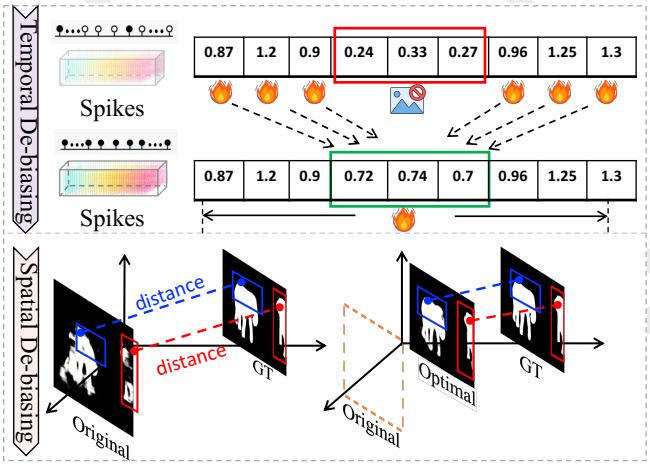


Figure 2: **Temporal and Spatial Debias.** Temporal debias captures subtle changes by exploring deep feature connections, while spatial debias constructs an OT map to minimize the distance between the spike saliency distribution and the real image distribution.

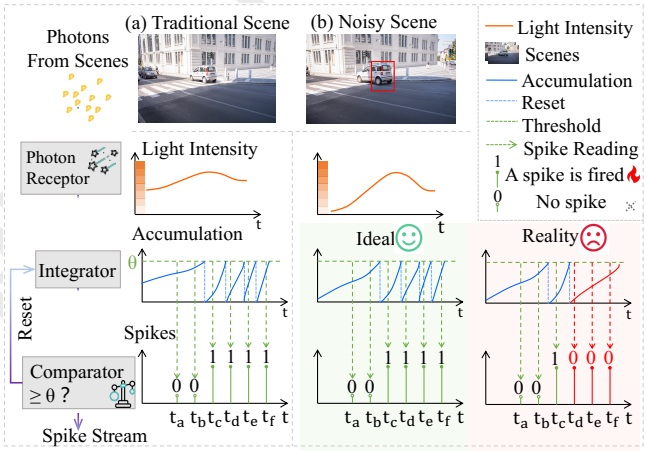


Figure 3: **Composite Noise Challenges in Spike Cameras.** (a) Traditional imaging forms the foundation of spike cameras. (b) Ideally, continuous spike streams and photons ensure smooth object motion. However, real-world light variations and background interference often result in missing photons, causing biased pixel representations.

leveraging spatiotemporal motion consistency to mitigate composite noise biases. SOTA propagates temporal information between samples to preserve the continuity of high-quality spatial regions, as shown in Fig. 2. It focuses on two key aspects:

1) Temporal Debias. To conquer the inconsistency in information among the temporal dimension, we use spiking neural networks (SNNs) to develop a multi-scale strategy for saliency map extraction, which captures deep feature connections across time steps and identifies subtle changes. SNNs generate saliency regions via a threshold mechanism, emitting spikes when the input strength surpasses a predefined threshold, creating a binary sequence. These saliency regions vary across time steps, each representing different confidence levels. We incorporate depthwise separable convolution (DWConv) into the framework to enhance local feature modeling while maintaining efficiency. By capturing deep feature dependencies and minor temporal variations, we improve interactions within the saliency map and mitigate confidence bias.

2) Spatial Debias. To facilitate the spatial alignment among visual representations, we address spatial saliency correction using an optimal transport (OT) strategy. The core idea is to find the OT map that minimizes the distance between the spike saliency distribution and the real-image distribution. Adversarial learning further refines this mapping to preserve the structural integrity of spike-extracted features. Building on SNN-based temporal bias correction, we extend the spatial subtle changes into temporal-spatial global debias, balancing local details and the global distribution of the spatiotemporal saliency map.

In summary, we introduce SOTA, a comprehensive framework for visual saliency detection. SOTA extends input features to multi-temporal and multi-scale representations, incorporating a cross-time-step attention mechanism with depthwise convolution for spike-based micro-debias (SM). We formulate saliency map detection as a Kantorovich problem and use adversarial learning to optimize spatial distribution via OT, mitigating spike-based global biases (SG).

Our main contributions are threefold:

- **Novel Modeling.** We bridge traditional imaging and downstream tasks by transforming composite noise in spike streams into domain gap analysis. We propose SOTA, which corrects domain bias across both local and global spatiotemporal dimensions, offering new insights for related tasks.
- **Innovative Method.** We use OT to rectify saliency distributions with global contextual information across long-term sequences. Additionally, we introduce contextual temporal debias, dynamically inferring positive micro-semantic associations across relevant time steps.
- **Generalization Validation.** We conduct extensive evaluations on real-world and synthetic datasets, thoroughly validating the feasibility and generalizability of SOTA. Experimental results in both single-step and multi-step settings demonstrate its effectiveness and confirm the validity of the identified problem.

2 Preliminaries and Motivation

Preliminaries. Unlike event cameras focus on dynamic changes only [Liu *et al.*, 2024a], spike cameras simulate the primate retina, with each pixel independently generating spikes in response to variations in light intensity [Huang *et al.*, 2022]. In traditional spike imaging, the CMOS photon receptor converts photons into voltage signals V, as shown in Fig. 3(a). Spikes are generated when the integrated signal reaches a threshold Θ, mapping physical signals to information. The spike generation process is expressed as:

$$\int_{t_s}^{t_e} I(t)dt \geq \Theta, \tag{1}$$

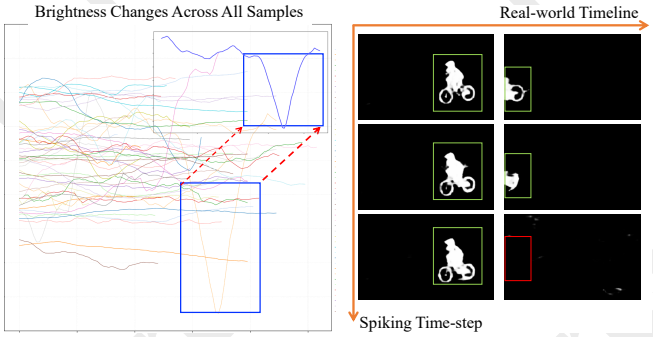


Figure 4: **Motivation of SOTA.** The left side shows variations in light conditions across samples (the x-axis denotes the frame index of each sample, and the y-axis represents average brightness), with different colored lines representing individual samples. Green boxes indicate correct predictions, while red boxes highlight errors.

where the voltage signal $V(t)$ at time t is captured from the light intensity $I(t)$ over the temporal interval $\mathcal{T} = [t_s, t_e]$. Spike signals are registered at rates of up to 40 kHz, forming a binary spike stream represented by 0s and 1s.

However, real-world photon accumulation is often imperfect. As shown in Fig. 3(b), a moving car is divided into two regions, with one half obscured by shadow noise in low-light conditions. Pixels in the shaded region may fail to reach the threshold Θ, preventing spike activation. This leads to reduced region consistency of the target object.

Motivation. We analyze the temporal variation of light intensity across samples in SPIKE-DAVIS, as shown in Fig. 4. Many samples exhibit significant fluctuations over time. A sample with noticeable "bumps" is selected for preliminary experiments, and the results show that RST predictions become inconsistent over time, with performance degrading as brightness decreases. Additionally, multi-step training reveals that information transmitted at different time steps is not always beneficial and may negatively impact predictions.

To mitigate this issue, we propose correcting biases across both spatial and temporal dimensions. First, the model should emphasize details from adjacent time steps in the temporal domain. Depthwise convolution, which processes channel and spatial information separately without increasing computational complexity, is key to introducing local debias.

The optimal transport (OT) problem transforms one distribution into another with minimal cost, either by finding the OT map (Monge problem) [Monge, 1781] or the OT plan (Kantorovich problem) [Kantorovich, 2006]. Unlike the Monge problem, the Kantorovich problem considers probabilistic transport, making it a well-defined convex problem with a unique optimal solution.

To strengthen our hypothesis, we reformulate composite bias correction as an OT problem between two probability distributions. The spatial scope of the saliency map is defined by the domains of the source distribution S and target distribution T, where x and y are sample points representing pixel-wise saliency scores. Our goal is to refine the initial spike saliency distribution S ($S \in P(X)$) into the target distribution T ($T \in P(Y)$), improving the saliency map's quality, as shown:

$$T^*_{\#}S = T, \tag{2}$$

where the mapping $T^* : Y \rightarrow X$ pushes the initial S to T, ensuring the transformed T^* aligns with the target.

3 Related Works

Spike Camera-Based Visual Tasks. Image reconstruction is fundamental to spike camera-based visual tasks [Chen *et al.*, 2022; Zhang *et al.*, 2023; Zhu *et al.*, 2019; Zheng *et al.*, 2023]. SpikeGS [Zhang *et al.*, 2024] combines 3D Gaussian Splatting with spike cameras for high-speed, high-quality 3D reconstruction, addressing challenges like motion blur and time-consuming rendering. SpikeNeRF [Zhu *et al.*, 2024b] further advances 3D reconstruction by using self-supervision and a spike-specific rendering loss to handle diverse illumination conditions. Spike streams have also been applied to depth estimation tasks, where spatiotemporal embeddings from spike camera streams improve accuracy [Zhang *et al.*, 2022]. Additionally, [Zhao *et al.*, 2024a] provides the first theoretical analysis of ultra-high-speed object recognition with spike cameras, proposing a robust representation. Recently, [Zhu *et al.*, 2024a] conducted the first study on visual saliency detection in continuous spike streams, inspiring future spike camera applications in saliency detection.

Sequential Saliency Detection. The main challenge in sequential saliency detection is maintaining temporal consistency while capturing smooth target regions in videos. Early methods relied on image-based saliency detection, which failed to capture temporal dynamics [Murray *et al.*, 2011]. Later research incorporated temporal information, with notable methods using motion extraction and fusion strategies based on optical flow [Chen *et al.*, 2017], along with approaches that jointly model spatial and temporal saliency [Guo *et al.*, 2024]. [Zhong *et al.*, 2025; Liu *et al.*, 2023] focus on action regions under multi-view conditions. These advancements significantly improved saliency detection in continuous streams, evolving from static image processing to fully integrated spatiotemporal modeling.

In this paper, we propose the novel SOTA framework, which effectively captures local changes and spatial information in temporal streams. Our method optimizes distribution alignment to mitigate domain gaps caused by temporal variations across time intervals.

4 Proposed Method

Our goal is to capture and refine smooth salient regions in motion areas reconstructed from the binary spike stream $\text{Spike}(x, y) = \{s(x, y, t)\}$, where $s(x, y, t) = 1$ when a spike is fired. Motivated by the success of Wasserstein GAN [Arjovsky *et al.*, 2017], we frame this process within an adversarial learning framework.

SOTA takes two inputs: the initial distribution S (sequentially reconstructed from $\text{Spike}(x, y)$) and the target distribution T. These are modeled by the Micro-debias network (T-Net) and the Global-debias network (F-Net), respectively, as shown in Fig. 5.

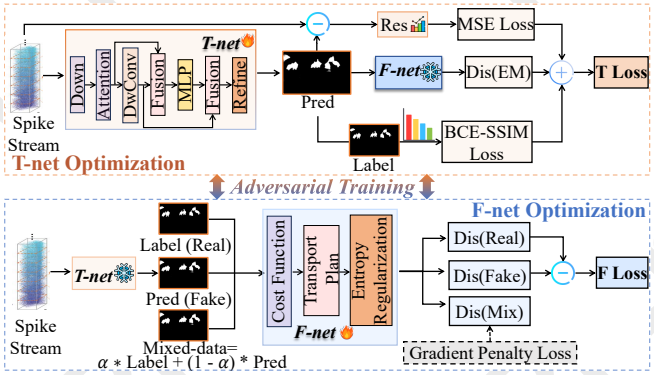


Figure 5: **Overview of the Proposed SOTA.** The two networks are optimized iteratively: T-Net generates temporal saliency maps with micro-detail debias, while F-Net refines global spatial debias. Together, they enhance motion associations in the spatiotemporal saliency map and model long-term dependencies.

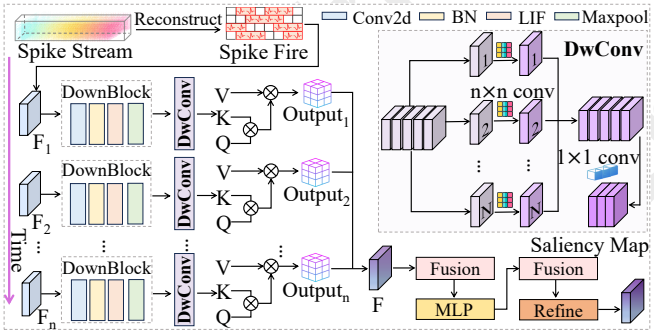


Figure 6: **Spike-based Micro-debias.** An adjacent-step attention mechanism aggregates temporal features, while DwConv captures fine-grained micro-dynamics.

4.1 Spike-Based Micro-Debias

Fig. 6 depicts the spike-based micro-debias (SM) module. Our goal is to enhance positive transfer across time steps and strengthen micro-temporal semantic associations.

Spiking Neuron Firing. Unlike continuous signals in traditional artificial neural networks (ANNs), spiking neural networks (SNNs) mimic biological neurons by using discrete spike signals [Xu *et al.*, 2023b]. The most widely used model is the leaky integrate-and-fire (LIF) [Gerstner *et al.*, 2014], where the input signal influences the neuron's membrane potential U. A spike is generated when U exceeds the firing threshold θ, expressed as:

$$\text{Spike}(x,y) = \begin{cases} 1, & \text{if } U > \theta, \\ 0, & \text{otherwise}, \end{cases} \tag{3}$$

The LIF neuron's event cycle includes a gradual increase in membrane potential due to the input current C_t, emission of a spike once the threshold is reached, and a subsequent reset. This process is described as:

$$U = U[t-1] + C_t, \tag{4}$$
$$S = \tau(U - U_\theta), \tag{5}$$
$$V = U(1-S) + V_{\text{reset}}S, \tag{6}$$

where τ is the Heaviside step function, and U_θ is the firing threshold.

Multi-Scale Spike Feature Modeling. To extract multi-scale features effectively, we construct a feature pyramid using successive downsampling:

$$F_i = \text{DownBlock}(F_{i-1}), \tag{7}$$

where $i = 1, 2, 3, 4$. Each downsampling step reduces spatial resolution while increasing channel depth, yielding a spatial dimension of $\frac{W}{2^i} \times \frac{H}{2^i}$ and a channel size of $C \times 2^i$. The initial feature F is extracted from $\text{Spike}(x, y)$ using a CBS block [Zhu *et al.*, 2024a], which includes a Conv2d layer, batch normalization, a LIF neuron, and a max-pooling layer.

Learning Multi-Temporal Micro-Interactions. Temporal inconsistencies in spike streams [Zhong *et al.*, 2024] can lead to significant performance degradation in SNNs. To mitigate this, we introduce a cross-step operation that reduces bias and enhances interactions across temporal contexts.

We use the final downsampled feature map F_4 as input, dividing it into segments F'_t corresponding to time steps t. A micro-interactive attention mechanism based on SSA [Zhu *et al.*, 2024a; Zhou *et al.*, 2023] is then constructed:

$$q = \text{DwConv}(F'_t, W_Q), \quad \text{for Query}, \tag{8}$$
$$k = \text{DwConv}(F'_{t+1}, W_K), \quad \text{for Key}, \tag{9}$$
$$v = \text{DwConv}(F'_{t+1}, W_V), \quad \text{for Value}, \tag{10}$$

where depthwise convolution DwConv enables SOTA to dynamically adjust spike responses, adapting to subtle sequential variations.

The features are processed using multi-head attention across adjacent time steps [Xu *et al.*, 2023a]. The attention result $\text{Att}(F')$ is combined with a projected version of F'_t, generating a refined salient feature. The updated feature is further processed through an MLP module, and the outputs are concatenated across time steps for refinement. We adopt the Refine module [Zhu *et al.*, 2024b] for this final step.

4.2 Spike-Based Global Debias

Kantorovich Modeling. The Kantorovich problem (KP) transforms one probability distribution ($\mathbb{S}$, modeled by T-Net) into another (the target distribution $\mathbb{T}$) while minimizing the "transportation cost". This cost is defined by a function $c(x, y)$, which quantifies the effort required to move mass from one location to another. The problem is formally defined as:

$$\text{Cost}(\mathbb{S}, \mathbb{T}) \overset{\text{def}}{=} \inf_{\pi \in \Pi(\mathbb{S},\mathbb{T})} \int_{X \times Y} c(x,y) d\pi(x,y), \tag{11}$$

where the minimum "transportation cost" is computed over all transport plans π, whose marginals correspond to $\mathbb{S}$ and $\mathbb{T}$. The optimal plan $\pi^* \in \Pi(\mathbb{S}, \mathbb{T})$ is the optimal transport plan.

Spike-Based Optimal Transport. We consider the dual form of KP and adapt it to spike-based optimal transport (OT). The corresponding formulation is:

$$\text{DP-Cost}(\mathbb{S}, \mathbb{T}) = \sup_{\varphi} \int_Y \varphi^c(y) d\mathbb{S}(y) + \int_X \varphi(x) d\mathbb{T}(x), \tag{12}$$

where $\varphi^c(y) = \inf_{x \in X}[c(x,y) - \varphi(x)]$. The goal is to find the supremum of $\varphi(x)$ and the infimum of $T(y)$. During training, we maximize φ and minimize T, leading to a max-min adversarial optimization process.

To model this, we use the initial spike distribution $\mathbb{S}$ to approximate $T(y)$ and the target distribution $\mathbb{T}$ using a simple CNN to estimate $\varphi(x)$. Our goal is to optimize Eq. (12), where the cost function $c(x,y)$ is a distance metric. The T-Net minimizes the discrepancy between $\mathbb{S}$ and $\mathbb{T}$, expressed as:

$$\int_Y [\tilde{c}(y, T(y)) - \varphi(T(y))]\, d\mathbb{P}(y). \tag{13}$$

The F-Net minimizes the transport cost between the target distribution and itself while maximizing the transport cost between the spike distribution $\mathbb{S}$ (from T-Net) and the target distribution $\mathbb{T}$, as:

$$\int_X \varphi(x) d\mathbb{Q}(x) + \int_Y [-\varphi(T(y))]\,. \tag{14}$$

4.3 Joint Optimization

For training, SOTA is optimized using two key components. The F-Net is updated using the Earth Mover's (EM) distance and a penalty loss inspired by Wasserstein GAN. The loss function for T-Net is defined as:

$$\mathcal{L}_T = \alpha \mathcal{L}_{\text{original}} + \mathcal{L}_{\text{MSE}} - \mathcal{D}_{\text{EM}}, \tag{15}$$

where $\mathcal{L}_{\text{original}}$ follows the RST framework [Zhu *et al.*, 2024a], incorporating binary cross-entropy, IoU loss, and SSIM loss. $\mathcal{D}_{\text{EM}}$ is the EM distance between the T-Net output and the target distribution. During inference, the OT processing is not applied.

5 Experimental Analysis

5.1 Datasets and Implementation Details

Synthetic Dataset. To evaluate the effectiveness of our proposed SOTA framework, we construct a bio-inspired video saliency detection dataset, SPIKE-DAVIS [Perazzi *et al.*, 2016], following prior works on related tasks [Zhang *et al.*, 2022]. We use XVFI [Sim *et al.*, 2021] to interpolate seven images between each frame pair. The dataset includes various actions with challenging attributes such as fast motion, occlusions, and motion blur.

Real-World Dataset. SVS dataset, captured using a spike camera with spatial resolution 250×400 and temporal resolution of 20,000 Hz [Zhu *et al.*, 2024a], consists of 130 sequences. Of these, 100 are used for training (24 high-condition, 76 low-condition) and 30 for validation (8 high-condition, 22 low-condition).

Implementation Details. All experiments are conducted on the PyTorch platform with an NVIDIA RTX 4090 24GB GPU. We optimize SOTA using the Adam optimizer [Kingma and Ba, 2015] with an initial learning rate of $2e{-}4$ and a weight decay of $2e{-}5$. The batch size is 2, and input images are resized to 256×256. For fair comparisons, we apply temporal spike representation and reconstruction on SVS as $M/\Delta t_{x,y}$, following [Zhu *et al.*, 2024a], where M is the maximum grayscale value and $\Delta t_{x,y}$ is the spike firing interval at pixel (x,y). We do not fine-tune parameters on SPIKE-DAVIS and use the same settings as for SVS.

Evaluation Metrics. We evaluate performance using the mean absolute error (MAE), mean F-measure score mF_β with $\beta^2 = 0.3$, maximum F-measure F_β^m, and Structure-measure S_m. MAE quantifies overall accuracy, while mF_β and F_β^m assess precision and recall. The S_m metric integrates object-level and region-level saliency performance. Additionally, we report peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) to evaluate the impact of image quality.

5.2 Comparisons with Existing Methods

Table 1 compares SOTA with state-of-the-art SNN-based methods. Quantitative evaluations are performed under both training settings. Spiking DeepLab and Spiking FCN [Kim *et al.*, 2022], relying on ANN-SNN conversion, suffer from conversion errors and domain gaps. Other networks struggle to capture dynamic details at each time step, leading to loss of subtle variations in saliency map extraction. Our SOTA framework outperforms methods like Spikformer [Zhou *et al.*, 2023], EVSNN [Zhu *et al.*, 2023], and RST [Zhu *et al.*, 2024a], achieving an MAE of 0.0478 in multi-step settings.

Spikformer-ADD and Spikformer-OR show improved performance in multi-step conditions, suggesting that they benefit from sequential processing. Additionally, SOTA achieves the highest $F_\beta^{\max}$ and mF_β scores, demonstrating its strong ability to balance precision and recall. Notably, SOTA's performance in the single-step setting is slightly lower, highlighting its effectiveness in modeling long-term temporal dependencies.

5.3 Ablation Studies

Influence of Different Components. We evaluate SOTA's generalization on SPIKE-DAVIS and SVS, as shown in Table 2. On SPIKE-DAVIS, the best MAE performance varies between single-step and multi-step approaches. In the single-step setting, SOTA achieves the highest S_m score of 0.6091 when the global component is included. This may be due to complex backgrounds in single-step predictions, where feature interactions are insufficient for capturing object-level details. However, introducing S_m leads to a slight MAE degradation by 0.0105.

In multi-step predictions, particularly with strong global dependencies, S_m combined with DwConv may struggle to capture long-range dependencies, leading to incomplete scene representations. In contrast, on SVS, S_m shows more consistent performance, emphasizing its effectiveness in focusing on localized receptive fields.

Influence of Micro Debias. Table 3 presents the performance of different fusion strategies within SM in the single-step training setting. The OR operation prioritizes overlapping regions between features, achieving an S_m score of 0.6968, but lacks continuity and cumulative effects. In saliency detection, this can lead to inadequate feature fusion, causing loss of critical information or poorly defined salient region boundaries.

Method	Venue	Single Step				Multi Step			
		MAE ↓	$F_\beta^{\max}$ ↑	mF_β ↑	S_m ↑	MAE ↓	$F_\beta^{\max}$ ↑	mF_β ↑	S_m ↑
Spiking Deeplab [Kim *et al.*, 2022]	NCE	0.1026	0.5310	0.5151	0.6599	0.0726	0.6175	0.6051	0.7125
Spiking FCN [Kim *et al.*, 2022]	NCE	0.1210	0.4779	0.4370	0.6070	0.0860	0.5970	0.5799	0.6911
EVSNN [Zhu *et al.*, 2023]	TPAMI	0.1059	0.5221	0.4988	0.6583	0.0945	0.6267	0.5850	0.7023
Spikformer-OR [Zhou *et al.*, 2023]	ICLR	0.1389	0.4527	0.4408	0.6068	0.0738	0.6526	0.6323	0.7161
Spikformer-ADD [Zhou *et al.*, 2023]	ICLR	0.1185	0.4638	0.4415	0.6119	0.0717	0.6890	0.6731	0.7563
RST [Zhu *et al.*, 2024a]	AAAI	0.0784	0.6313	0.6171	0.6970	0.0554	0.6981	0.6882	0.7591
RST [Zhu *et al.*, 2024a] ∗	AAAI	0.0776	0.6291	0.6152	0.6968	0.0612	0.6619	0.6515	0.7409
SOTA (Ours)		**0.0597**	**0.6978**	**0.6880**	**0.7569**	**0.0478**	**0.7402**	**0.7296**	**0.7912**

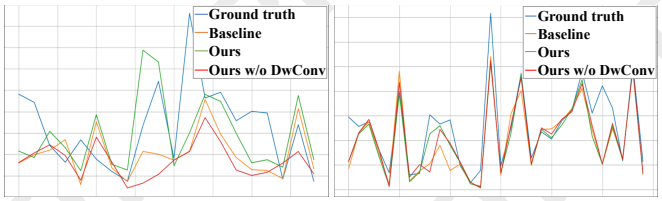
Table 1: **Performance Comparison on SVS.** ∗ denotes reproduced results under identical experimental settings. The number of time steps is set to 5.

Dataset	Base line	SM	SG	Single Step				Multi Step			
				MAE ↓	$F_\beta^{\max}$ ↑	mF_β ↑	S_m ↑	MAE ↓	$F_\beta^{\max}$ ↑	mF_β ↑	S_m ↑
SPIKE-DAVIS	✓			0.1119	0.3979	0.3646	0.5969	0.0827	0.4154	0.3974	0.6019
	✓		✓	0.1053	0.4191	0.3893	**0.6091**	**0.0785**	0.4179	0.3910	0.6027
	✓	✓	✓	**0.0861**	**0.4224**	**0.4132**	0.6076	0.0890	**0.4281**	**0.4172**	**0.6177**
SVS	✓			0.0776	0.6291	0.6152	0.6968	0.0612	0.6619	0.6515	0.7409
	✓		✓	0.0649	0.6945	0.6843	0.7539	0.0478	0.7355	0.7240	0.7908
	✓	✓	✓	**0.0597**	**0.6978**	**0.6880**	**0.7569**	**0.0478**	**0.7402**	**0.7296**	**0.7912**

Table 2: **Ablation Study of Components on SVS and SPIKE-DAVIS.** SM/SG stands for Spike-Based Micro/Global-Debias, respectively.

Fusion	MAE ↓	$F_\beta^{\max}$ ↑	mF_β ↑	S_m ↑
OR	0.0776	0.6291	0.6152	0.6968
ADD *w/* DwConv	0.0914	0.5978	0.5819	0.6895
ADD *w/o*	0.0758	0.6488	0.6355	0.7171
SOTA *w/o* DwConv	0.0649	0.6945	0.6843	0.7539
SOTA	**0.0597**	**0.6978**	**0.6880**	**0.7569**

Table 3: **Fusion Performance Comparison on SVS.** Results are for single-step training.



(a) SPIKE-DAVIS (b) SVS

Figure 7: **Saliency Pixel Ratio on SPIKE-DAVIS and SVS.** Colored lines represent different method variants. The x-axis denotes the number of categories, and the y-axis represents the ratio of foreground pixels.

ADD fusion integrates multi-modal information stably without excessive noise or complex interactions, but it captures high-level patterns, while DwConv is used for detailed local feature extraction. Combining ADD with DwConv weakens some important patterns, leading to a slight performance decline compared to ADD alone.

Fig. 7 compares foreground pixel ratios between ground truth (GT), baseline [Zhu *et al.*, 2024a], and SOTA. In most composite-bias scenarios, SOTA outperforms the baseline, reducing background dependency and improving pixel localization. Notably, removing DwConv significantly degrades performance, highlighting the importance of micro-debias. Furthermore, SOTA and the baseline show consistent performance in certain classes, suggesting the benefit of exploring adjacent temporal relations for network training optimization.

Influence of Global Debias. SOTA optimizes the optimal transport (OT) strategy by measuring the distance between the spike distribution and the target distribution. The choice of distance metric is crucial. We conduct an ablation study on different similarity metrics, with results shown in Table 4.

Kullback-Leibler (KL) and Jensen-Shannon (JS) divergence are effective for evaluating global distribution differences but struggle with local structural information and instability when handling zero probabilities or noise. In contrast, the Earth Mover's (EM) distance performs better in capturing spatial and sequential structures, which is why we use it for distribution alignment.

5.4 Visualization

Fig. 8 presents qualitative comparisons across multiple steps on SVS. The bar chart compares image quality metrics, with multi-step predictions outperforming single-step due to enhanced information transfer. SOTA shows substantial improvements in prediction accuracy.

To verify SOTA's effectiveness, we compare it with CNN-based DCFNet [Zhang *et al.*, 2021] on SPIKE-DAVIS in Fig. 9. The DCFNet under our Spike-R setting performs unstably among adjacent frames, while ours clearly identifies saliency regions without losing temporal continuity. How-

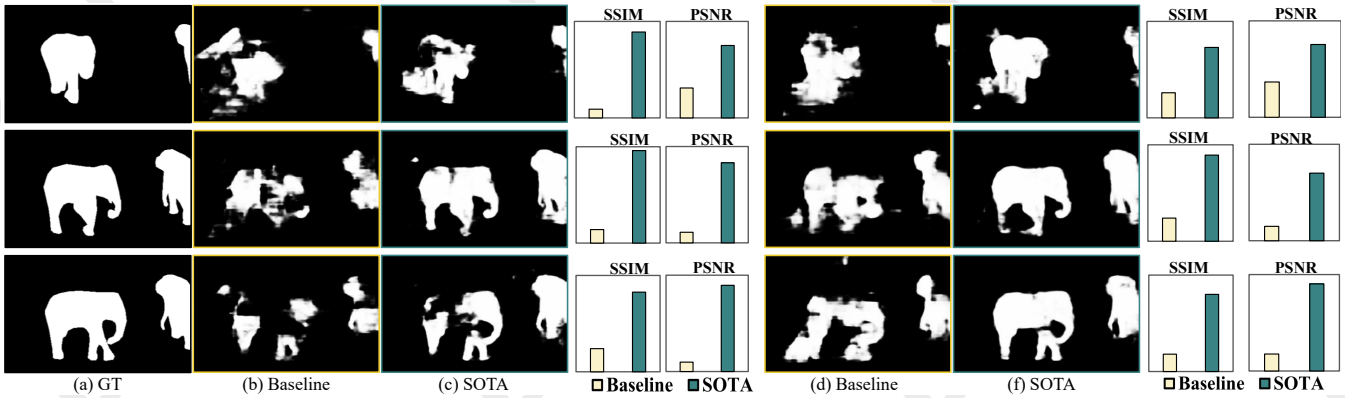


Figure 8: **Saliency Visualization Results on SVS.** Comparison of final image quality between RST, SOTA, and ground truths (GTs). (b) and (c) correspond to single-step training, while (d) and (f) represent multi-step training. Notably, the SSIM and PSNR values for multi-step training are lower than those for single-step training.

Distance	MAE ↓	$F_\beta^{\max}$ ↑	mF_β ↑	S_m ↑
ED	0.0933	0.5299	0.5120	0.6408
JS	0.1057	0.4988	0.4682	0.6078
KL	0.0849	0.5703	0.5494	0.6532
SOTA *w/* EM	**0.0597**	**0.6978**	**0.6880**	**0.7569**

Table 4: **Ablation Study of Distance Measures on SVS.** Each row presents performance using Euclidean Distance (ED), KL, and JS Divergence, with results for single-step training.

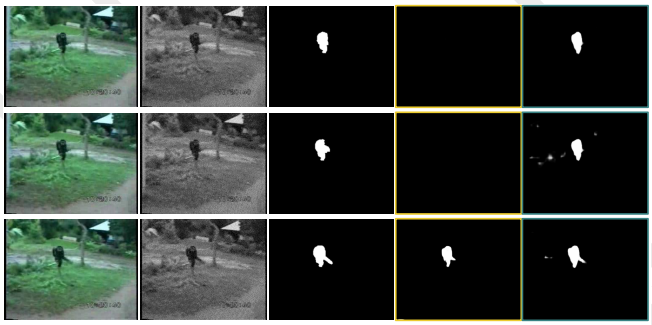


Figure 9: **Saliency Visualization Results on SPIKE-DAVIS.** Comparison of final image quality between GTs, CNN-based DCFNet, and SOTA. Notably, results are obtained under a multi-step training setting. Spike-R refers to our reconstruction frame.

ever, SOTA introduces slight noise artifacts, which inspired our future work.

5.5 Energy Efficiency Analysis

We evaluate the theoretical energy consumption for predicting saliency maps, summarized in Table 5. Following [Shi *et al.*, 2024], we analyze two variations: SOTA with and without DwConv on a GPU to assess energy efficiency. Unlike traditional ANNs, SNNs offer lower power consumption due to their binary nature and reliance on accumulate operations. Our SOTA model demonstrates superior energy efficiency compared to traditional methods.

Method	Single Step	Multi Step
RST [Zhu *et al.*, 2024a] ∗	1.627 mJ	14.394 mJ
SOTA *w/o* DwConv	1.660 mJ	**13.794** mJ
SOTA *w/* DwConv	**1.589** mJ	13.872 mJ

Table 5: **Efficiency Analysis on SPIKE-DAVIS.** ∗ denotes reproduced results under identical experimental settings.

6 Conclusion

This paper introduces SOTA, a novel method leveraging spike cameras to address the temporal domain gap caused by composite noise in visual saliency detection. The proposed Spike-based Micro-debias (SM) module refines temporal transfer across adjacent time steps, ensuring consistent saliency localization and continuous motion capture at the micro level. Meanwhile, the Spike-based Global-debias (SG) module enhances the stability of the macro distribution, improving overall visual map accuracy. In summary, we integrate adversarial training between these two components to effectively capture local variations and spatial information in sequential data. Additionally, optimal transport is used to align the overall feature distribution, mitigating domain gaps caused by temporal variations between frames. Our method is extensively evaluated on both real-world and synthetic benchmarks, including SVS and SPIKE-DAVIS, demonstrating superior effectiveness and generalization. In cases of complex scenes with fine-grained categories of objects, the model may struggle to capture crucial information accurately. It motivates us to consider implementing multiple image augmentation processing methods to better handle complex scenes.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants 62271361, 62422601, and U24B20140, the Hubei Provincial Key Research and Development Program under Grant 2024BAB039, and the Beijing Nova Program under Grants 20230484362 and 20240484703.

References

[Arjovsky *et al.*, 2017] Martín Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein GAN. *arxiv preprint arxiv:1701.07875*, 2017.

[Chen *et al.*, 2017] Chenglizhao Chen, Shuai Li, Yongguang Wang, Hong Qin, and Aimin Hao. Video saliency detection via spatial-temporal fusion and low-rank coherency diffusion. *IEEE Trans. Image Process.*, 26(7):3156–3170, 2017.

[Chen *et al.*, 2022] Shiyan Chen, Chaoteng Duan, Zhaofei Yu, Ruiqin Xiong, and Tiejun Huang. Self-supervised mutual learning for dynamic scene reconstruction of spiking camera. In *Proc. Int. Joint Conf. Artif. Intell.*, pages 2859–2866, 2022.

[Dong *et al.*, 2017] Siwei Dong, Tiejun Huang, and Yonghong Tian. Spike camera and its coding methods. In *Proc. IEEE Data Compress. Conf.*, page 437, 2017.

[Gerstner *et al.*, 2014] Wulfram Gerstner, Werner M Kistler, Richard Naud, and Liam Paninski. *Neuronal dynamics: From single neurons to networks and models of cognition*. Cambridge University Press, 2014.

[Guo *et al.*, 2024] Ruohao Guo, Dantong Niu, Liao Qu, Yanyu Qi, Ji Shi, Wenzhen Yue, Bowei Xing, Taiyan Chen, and Xianghua Ying. Instance-level panoramic audio-visual saliency detection and ranking. In *Proc. ACM Int. Conf. Multimedia*, pages 9426–9434, 2024.

[Hu *et al.*, 2022] Liwen Hu, Rui Zhao, Ziluo Ding, Lei Ma, Boxin Shi, Ruiqin Xiong, and Tiejun Huang. Optical flow estimation for spiking camera. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pages 17823–17832, 2022.

[Huang *et al.*, 2022] Tiejun Huang, Yajing Zheng, Zhaofei Yu, Rui Chen, Yuan Li, Ruiqin Xiong, Lei Ma, Junwei Zhao, Siwei Dong, Lin Zhu, Jianing Li, Shanshan Jia, Yihua Fu, Boxin Shi, Si Wu, and Yonghong Tian. 1000x faster camera and machine vision with ordinary devices. *arxiv preprint arxiv:2201.09302*, 2022.

[Kantorovich, 2006] Leonid V Kantorovich. On the translocation of masses. *J. Math. Sci.*, 133(4), 2006.

[Kim *et al.*, 2022] Youngeun Kim, Joshua Chough, and Priyadarshini Panda. Beyond classification: Directly training spiking neural networks for semantic segmentation. *Neuromorph. Comput. Eng.*, 2(4):44015, 2022.

[Kingma and Ba, 2015] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proc. Int. Conf. Learn. Represent.*, 2015.

[Liu *et al.*, 2022] Wenxuan Liu, Xian Zhong, Xuemei Jia, Kui Jiang, and Chia-Wen Lin. Actor-aware alignment network for action recognition. *IEEE Signal Process. Lett.*, 29:2597–2601, 2022.

[Liu *et al.*, 2023] Wenxuan Liu, Xian Zhong, Zhuo Zhou, Kui Jiang, Zheng Wang, and Chia-Wen Lin. Dual-recommendation disentanglement network for view fuzz in action recognition. *IEEE Trans. Image Process.*, 32:2719–2733, 2023.

[Liu *et al.*, 2024a] Zibin Liu, Banglei Guan, Yang Shang, Shunkun Liang, Zhenbao Yu, and Qifeng Yu. Optical flow-guided 6dof object pose tracking with an event camera. In *Proc. ACM Int. Conf. Multimedia*, pages 6501–6509, 2024.

[Liu *et al.*, 2024b] Zibin Liu, Banglei Guan, Yang Shang, Qifeng Yu, and Laurent Kneip. Line-based 6-dof object pose estimation and tracking with an event camera. *IEEE Trans. Image Process.*, 33:4765–4780, 2024.

[Monge, 1781] Gaspard Monge. Mémoire sur la théorie des déblais et des remblais. *Mem. Math. Phys. Acad. Royale Sci.*, pages 666–704, 1781.

[Murray *et al.*, 2011] Naila Murray, María Vanrell, Xavier Otazu, and C. Alejandro Párraga. Saliency estimation using a non-parametric low-level vision model. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pages 433–440, 2011.

[Perazzi *et al.*, 2016] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus H. Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pages 724–732, 2016.

[Shi *et al.*, 2024] Xinyu Shi, Zecheng Hao, and Zhaofei Yu. SpikingResformer: Bridging resnet and vision transformer in spiking neural networks. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pages 5610–5619, 2024.

[Sim *et al.*, 2021] Hyeonjun Sim, Jihyong Oh, and Munchurl Kim. XVFI: extreme video frame interpolation. In *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, pages 14469–14478, 2021.

[Sinha *et al.*, 2017] Raunak Sinha, Mrinalini Hoon, Jacob Baudin, Haruhisa Okawa, Rachel OL Wong, and Fred Rieke. Cellular and circuit mechanisms shaping the perceptual properties of the primate fovea. *Cell*, 168(3):413–426, 2017.

[Xu *et al.*, 2023a] Qi Xu, Yuyuan Gao, Jiangrong Shen, Yaxin Li, Xuming Ran, Huajin Tang, and Gang Pan. Enhancing adaptive history reserving by spiking convolutional block attention module in recurrent neural networks. In *Adv. Neural Inf. Process. Syst.*, 2023.

[Xu *et al.*, 2023b] Qi Xu, Yaxin Li, Jiangrong Shen, Jian K. Liu, Huajin Tang, and Gang Pan. Constructing deep spiking neural networks from artificial neural networks with knowledge distillation. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pages 7886–7895, 2023.

[Zhang *et al.*, 2021] Miao Zhang, Jie Liu, Yifei Wang, Yongri Piao, Shunyu Yao, Wei Ji, Jingjing Li, Huchuan Lu, and Zhongxuan Luo. Dynamic context-sensitive filtering network for video salient object detection. In *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, pages 1533–1543, 2021.

[Zhang *et al.*, 2022] Jiyuan Zhang, Lulu Tang, Zhaofei Yu, Jiwen Lu, and Tiejun Huang. Spike transformer: Monocular depth estimation for spiking camera. In *Proc. Eur. Conf. Comput. Vis.*, pages 34–52, 2022.

[Zhang *et al.*, 2023] Jiyuan Zhang, Shanshan Jia, Zhaofei Yu, and Tiejun Huang. Learning temporal-ordered representation for spike streams based on discrete wavelet transforms. In *Proc. AAAI Conf. Artif. Intell.*, pages 137–147, 2023.

[Zhang *et al.*, 2024] Jiyuan Zhang, Kang Chen, Shiyan Chen, Yajing Zheng, Tiejun Huang, and Zhaofei Yu. SpikeGS: 3D gaussian splatting from spike streams with high-speed camera motion. In *Proc. ACM Int. Conf. Multimedia*, pages 9194–9203, 2024.

[Zhao *et al.*, 2024a] Junwei Zhao, Shiliang Zhang, Zhaofei Yu, and Tiejun Huang. Recognizing ultra-high-speed moving objects with bio-inspired spike camera. In *Proc. AAAI Conf. Artif. Intell.*, pages 7478–7486, 2024.

[Zhao *et al.*, 2024b] Rui Zhao, Ruiqin Xiong, Jian Zhang, Xinfeng Zhang, Zhaofei Yu, and Tiejun Huang. Optical flow for spike camera with hierarchical spatial-temporal spike fusion. In *Proc. AAAI Conf. Artif. Intell.*, pages 7496–7504, 2024.

[Zhao *et al.*, 2024c] Xing Zhao, Haoran Liang, Peipei Li, Guodao Sun, Dongdong Zhao, Ronghua Liang, and Xiaofei He. Motion-aware memory network for fast video salient object detection. *IEEE Trans. Image Process.*, 33:709–721, 2024.

[Zheng *et al.*, 2023] 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 Trans. Pattern Anal. Mach. Intell.*, 45(7):8127–8142, 2023.

[Zhong *et al.*, 2024] Xian Zhong, Shengwang Hu, Wenxuan Liu, Wenxin Huang, Jianhao Ding, Zhaofei Yu, and Tiejun Huang. Towards low-latency event-based visual recognition with hybrid step-wise distillation spiking neural networks. In *Proc. ACM Int. Conf. Multimedia*, pages 9828–9836, 2024.

[Zhong *et al.*, 2025] Xian Zhong, Liang Chen, Wenxuan Liu, Shu Ye, Kui Jiang, Zheng Wang, and Chia-Wen Lin. Temporal-spatial semantics-driven progressive multi-view action de-biasing. *Comput Eng.*, 51(1):1–10, 2025.

[Zhou *et al.*, 2023] Zhaokun Zhou, Yuesheng Zhu, Chao He, Yaowei Wang, Shuicheng Yan, Yonghong Tian, and Li Yuan. Spikformer: When spiking neural network meets transformer. In *Proc. Int. Conf. Learn. Represent.*, 2023.

[Zhou *et al.*, 2024] Kangrui Zhou, Taihang Lei, Banglei Guan, and Qifeng Yu. Event-based depth estimation with dense occlusion. *Opt. Lett.*, 49:3376–3379, 2024.

[Zhu *et al.*, 2019] Lin Zhu, Siwei Dong, Tiejun Huang, and Yonghong Tian. A retina-inspired sampling method for visual texture reconstruction. In *Proc. IEEE Int. Conf. Multimedia Expo*, pages 1432–1437, 2019.

[Zhu *et al.*, 2023] 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 Trans. Pattern Anal. Mach. Intell.*, 45(1):1233–1249, 2023.

[Zhu *et al.*, 2024a] Lin Zhu, Xianzhang Chen, Xiao Wang, and Hua Huang. Finding visual saliency in continuous spike stream. In *Proc. AAAI Conf. Artif. Intell.*, pages 7757–7765, 2024.

[Zhu *et al.*, 2024b] Lin Zhu, Kangmin Jia, Yifan Zhao, Yunshan Qi, Lizhi Wang, and Hua Huang. SpikeNeRF: Learning neural radiance fields from continuous spike stream. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pages 6285–6295, 2024.