Transient Glimpses: Unveiling Occluded Backgrounds through the Spike Camera

Jiyuan Zhang[1, 2], **Shiyan Chen**[1, 2], **Yajing Zheng**[1, 2*], **Zhaofei Yu**[1, 2, 3*], **Tiejun Huang**[1, 2, 3]

[1]School of Computer Science, Peking University
[2]National Key Laboratory for Multimedia Information Processing, Peking University
[3]Institute for Artificial Intelligence, Peking University
{jyzhang,2301112005}@stu.pku.edu.cn, {yj.zheng,yuzf12,tjhuang}@pku.edu.cn

Abstract

The de-occlusion problem, involving extracting clear background images by removing foreground occlusions, holds significant practical importance but poses considerable challenges. Most current research predominantly focuses on generating discrete images from calibrated camera arrays, but this approach often struggles with dense occlusions and fast motions due to limited perspectives and motion blur. To overcome these limitations, an effective solution requires the integration of multi-view visual information. The spike camera, as an innovative neuromorphic sensor, shows promise with its ultra-high temporal resolution and dynamic range. In this study, we propose a novel approach that utilizes a single spike camera for continuous multi-view imaging to address occlusion removal. By rapidly moving the spike camera, we capture a dense stream of spikes from occluded scenes. Our model, SpkOccNet, processes these spikes by integrating multi-view spatial-temporal information via long-short-window feature extractor (LSW) and employs a novel cross-view mutual attention-based module (CVA) for effective fusion and refinement. Additionally, to facilitate research in occlusion removal, we introduce the S-OCC dataset, which consists of real-world spike-based data. Experimental results demonstrate the efficiency and generalization capabilities of our model in effectively removing dense occlusions across diverse scenes. Public project page: https://github.com/Leozhangjiyuan/SpikeDeOcclusion.

Introduction

The presence of dense occlusions poses challenges to visual algorithms. Recently, frame-based algorithms have been proposed to see the background scenes through occlusions assisted by leveraging multi-view image (Zhang et al. 2017; Wang et al. 2020; Zhang, Shen, and Lin 2021; Li et al. 2021a; Zhang et al. 2022b; Hur et al. 2023). The task is named ***Synthetic Aperture Imaging (SAI)***. However, these algorithms often rely on discrete multi-view exposures, which may not provide sufficient background information. Moreover, obtaining sharp frames in high-speed scenarios poses further challenges. In applications such as autonomous driving, effectively removing foreground occlusions (e.g., fences) is crucial for enhancing environment perception, particularly at high driving speeds. Consequently, the acquisition of *sharp* and *continuous-view* information in high-speed scenarios remains a significant and ongoing challenge.

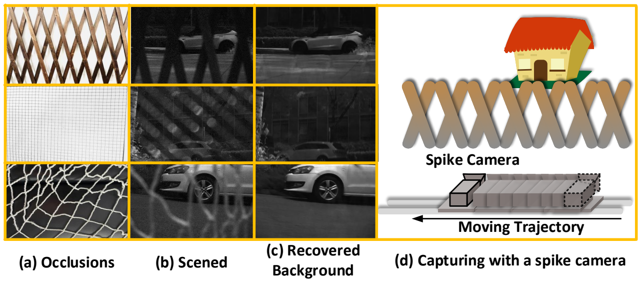

Figure 1: Outline of occlusion removal with a single spike camera. (a) Selected challenging occlusions, (b) Real-world Scenes with occlusions, (c) Recovered backgrounds with the proposed SpkOccNet, (d) We capture dataset with a fast-moving spike camera.

Recently, neuromorphic sensors (Gallego et al. 2020; Brandli et al. 2014; Huang et al. 2022) show remarkable performance in visual tasks. These sensors generate continuous signals asynchronously, enabling high temporal-resolution sampling. Two commonly used types of neuromorphic sensors are event cameras and spike cameras. Event cameras (Brandli et al. 2014; Lichtsteiner, Posch, and Delbruck 2008) asynchronously fire events in a differential manner when the light change surpasses a threshold, thus capturing rich motion information. Some studies have utilized event cameras for occlusion removal tasks (Zhang et al. 2021; Yu et al. 2022; Liao et al. 2022). However, event-based algorithms often require refocusing events to align them and provide accurate information for background reconstruction. This reliance on camera intrinsic parameters and distance information between objects and the camera restricts its applicability. Spike cameras (Huang et al. 2022) mimic the sampling mechanism of the fovea in the retina (Masland 2012; Wässle 2004), with each pixel capturing photons and asynchronously firing spikes when the accumulated intensity surpasses a threshold. The integration mechanism of spikes enables the recording of absolute light

[*]Corresponding authors

intensity (Huang et al. 2022; Zheng et al. 2021; Zhu et al. 2019), providing more texture information for reconstructing occluded regions. Compared to frame-based cameras, spike cameras offer more continuous and dense motion cues for occlusion removal (as elaborated in Sec.). In this paper, we propose, for the first time, to utilize spike cameras for foreground occlusion removal tasks, demonstrating their potential in effectively removing occlusions and reconstructing sharp backgrounds with camera motion.

How do we define the spike-based SAI task? Conventional cameras often suffer from motion blur when capturing scenes in motion, which hinders the acquisition of multiple perspective views with a single camera. Frame-based algorithms rely on camera arrays to compensate for the limited viewpoints, restricting their applicability in real-world scenarios. Our goal is to achieve foreground removal using only one fast-moving spike camera without complex equipment or calibration. Therefore, the spike-based SAI we defined possesses the following advantages: a) The high temporal resolution of the spike camera allows us to overcome the constraints imposed by the motion speed of the scene; b) a single spike camera is sufficient to capture continuous views, eliminating the need for multiple cameras.

How do we design the model? To deal with the spike-based SAI task, we build an end-to-end model named **SpkOccNet**. Specifically, to exploit the rich spatial-temporal information in spikes, we propose the Long-Short-Window (LSW) module to excavate and ensemble spatial-temporal features with different representations of long/short time windows from various views. To be specific, we segment spikes into three segments: one central and two end parts. We then utilize dense windows to transform spikes within three segments into dense representations while preserving their temporal characteristics, and simultaneously adopt a longer window to transform the central segment into a blurred image-like representation. Due to the motion displacement of foreground occlusions relative to the background being more significant and various as the view changes, spikes from different parts are complementary. Features from various views are fused using the cross-view mutual attention-based (Li et al. 2021b) module (CVA).

To enhance the generalization of our method in real-world scenarios, we construct the first real spike-based occlusion removal dataset **S-OCC**. As depicted in Fig.1(d), we mount a spike camera on a slider and moved it rapidly to capture diverse outdoor scenes featuring different occlusions. Fig.1(a) illustrates the various occlusion types included in the dataset, while Fig.1(b) showcases the occluded scenes contained in S-OCC. Furthermore, Fig.1(c) presents the deocclusion results obtained using the proposed **SpkOccNet**.

Our contributions are summarized as follows:

- We explore the spike-based SAI for the first time, utilizing sharp and continuous-view information from spike streams. Our approach incorporates information from dense viewpoints with long/short representations, leveraging mutual attention to fuse features.
- We contribute the first real-world spike-based dataset for occlusion removal, verifying the algorithm's generalization in real-world scenes.
- Experiments demonstrate the effectiveness of our method in occlusion removal, relying solely on a single camera with fast motion.

Related Works

Synthetic Aperture Imaging

Image-Based SAI For frame-based cameras, an earlier work (Vaish et al. 2004) utilized a camera array to align the information from multiple viewpoints to a reference viewpoint using coordinate relationships. However, its planar camera array requires stringent hardware calibrations. Vaish et al. (Vaish et al. 2006) take medians and entropy into consideration and proposed a more robust cost function. Zhao et al. (Pei et al. 2013) formulate an energy minimization problem to recognize each pixel from various views whether it belongs to the occlusion. Zhang et al. (Zhang et al. 2017) utilize a moving camera with its IMU data as the clue. Later method (Yang et al. 2014) is capable of predicting all-in-focus images. DeOccNet (Wang et al. 2020) includes a residual atrous spatial pyramid pooling module to enlarge receptive fields. Zhang et al. (Zhang, Shen, and Lin 2021) use shifted micro-lens images with a dynamic filter to explore information in the light field. Recent works mainly utilize stronger CNNs to remove occlusions (Li et al. 2021a; Zhang et al. 2022b; Hur et al. 2023).

Event-Based SAI Discrete images captured with traditional cameras fail to provide sufficient information in scenarios with extremely dense occlusions due to limited viewpoints. Event cameras show their potential to see through dense occlusions (Zhang et al. 2021; Yu et al. 2022; Liao et al. 2022) due to high temporal resolution. Zhang et al. (Zhang et al. 2021) propose to use SNNs as the encoder and CNNs as the decoder. Later work (Liao et al. 2022) combines events and images. However, the event-based SAI approaches require the camera intrinsic. The translation matrices of the camera and the target depth prior are complicated settings.

Spike-based Image Reconstruction

Spike cameras possess several advantages, including high temporal resolution (Zheng et al. 2023b), high dynamic range, and rich preservation of spatial texture. These advantages have led to wide applications in various downstream tasks, such as optical flow estimation (Hu et al. 2022; Zhao et al. 2022; Chen, Yu, and Huang 2023), object tracking (Zheng et al. 2023a), and depth estimation (Zhang et al. 2022a; Wang et al. 2022). Among these tasks, the reconstruction task (Zhang et al. 2023) serves as the fundamental basis. In the early stages, Zhu et al. (Zhu et al. 2019) propose to approximate the light intensity by statistically analyzing the spike stream. Zhu et al. (Zhu et al. 2021) and Zheng et al. (Zheng et al. 2021, 2023c) also develop biologically inspired reconstruction algorithms. Recently, Chen et al. explore self-supervised reconstruction methods (Chen et al. 2022) and spike-guided image deblurring (Chen et al. 2023). Existing methods have demonstrated the advantages of spike cameras in recovering textures from high-speed scenes.

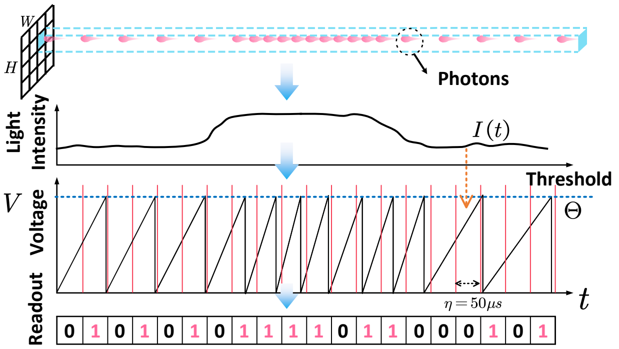

Figure 2: Illustration of the spike camera outputs spikes. The voltage always increases and resets with the light change, and spikes are read out with very short intervals.

Spike-based SAI

Theoretical Analysis

We focus on the physical imaging process to explain why spike cameras exhibit superior potential than frame-based cameras in addressing occlusion removal tasks. The photosensitive units of a spike camera consist of an array of $H \times W$ pixels, with each pixel independently capturing photons continuously. The photoelectric conversion unit transforms the captured photons into electrical current $I_{x,y}(t)$ and accumulates voltage $V_{x,y}$. When the voltage $V_{x,y}$ exceeds a dispatch threshold Θ, the pixel fire a spike, and subsequently, the voltage $V_{x,y}$ is reset to zero, shown as in Fig. 2. The entire process can be formulated as:

$$V^{+}_{x,y}(t) = \begin{cases} V^{-}_{x,y}(t) + I_{x,y}(t), & \text{if } V^{-}_{x,y}(t) < \Theta, \\ 0, & \text{otherwise,} \end{cases} \quad (1)$$

$$\mathbf{S}_{x,y}(k) = \begin{cases} 1, & \text{if } \exists t \in ((k-1)\eta, k\eta], V_{x,y}(t) = 0, \\ 0, & \text{if } \forall t \in ((k-1)\eta, k\eta], V_{x,y}(t) > 0, \end{cases} \quad (2)$$

where $V^{-}_{x,y}(t)$ and $V^{+}_{x,y}(t)$ denotes the voltage before and after receiving the electric current $I_{x,y}(t)$, $k \in \mathbb{R}$. The voltage is read out with the very short interval $\eta = 50\mu s$ and outputs a spike stream $\mathbf{S}$ with the size of $H \times W \times K$ after K times readout during T μs ($K = \frac{T}{\eta}$).

While capturing frame-based videos, the actual time interval between successive frames is T_{shutter}. To reduce motion blur, the exposure time T_{expo} per frame is kept shorter than T_{shutter}. Therefore, the continuous changes in light dynamics taking place during the time interval $T_{\text{void}} = T_{\text{shutter}} - T_{\text{expo}}$ are not captured.

Not considering the dynamic range, the image frame $\mathbf{B}_i$ in i-th capture can be formulated as:

$$\mathbf{B}_i = \frac{1}{T_{\text{expo}}} \int_{T_i}^{T_i + T_{\text{expo}}} \mathbf{L}_t dt \simeq \sum_{n=0}^{T_{\text{expo}}/\eta} \mathbf{S}(n), \quad (3)$$

where T_i is the timestamp exposure begins, and the L_t is the hidden sharp image at any exact moment t. With inappropriate T_{expo}, $\mathbf{B}_i$ would be blurry under fast motion from Eq. 3. As depicted in Fig. 3(c), a moving camera during the exposure time T_{expo} results in motion blur, thereby causing an enlargement of the foreground occlusion area. Due to the integrative sampling of the spike camera, the sum of spikes directly reflects the light intensity, thus $\mathbf{B}_i$ can be also approximately written in terms of $\mathbf{S}$.

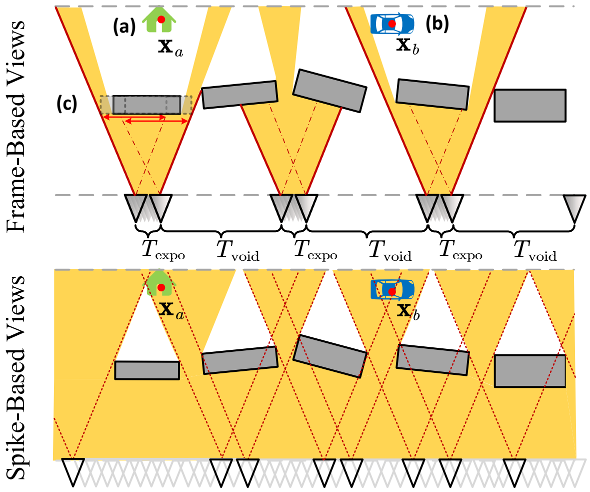

Figure 3: Illustration of the advantages of spike cameras over traditional cameras in seeing through backgrounds from the perspective of the imaging process. Orange regions represent the visible area of a frame or spike camera.

Imaging process comparison In T_{shutter}, we denoted the quantity of information captured by frames and spikes as Ω_{image} and Ω_{spike}, which can be formulated as:

$$\begin{aligned} \Omega_{\text{image}} &= \mathbf{B}_i + \emptyset , \\ \Omega_{\text{spike}} &= \sum_{k=0}^{T_{\text{expo}}/\eta} \mathbf{S}(k) + \sum_{k=T_{\text{expo}}/\eta}^{(T_{\text{expo}}+T_{\text{void}})/\eta} \mathbf{S}(k) , \end{aligned} \quad (4)$$

where $\emptyset$ denotes empty information. From Eq. 4, the **first term** indicates that within T_{expo}, the image records an $\mathbf{B}$, while the spike gets $\mathbf{S}$. However, $\mathbf{B}$ loses the temporal dimension, whereas $\mathbf{S}$ retains dense temporal information; the **second term** indicates that within T_{void}, the image captures nothing while the spike camera records continuously.

For SAI, in Fig. 3, the orange region represents the visible area during camera motion. The spike camera, due to its dense information as in Eq. 4, offers more clues from continuous viewpoints, allowing the observation of background objects $\mathbf{x}_a$ and $\mathbf{x}_b$, while the frame-based camera lose.

New Spike-based Dataset: S-OCC

Occlusion removal with spike cameras is a previously unexplored area, lacking relevant existing datasets to this work. Thus, we are dedicated to constructing the first dataset based on spike cameras with various occlusions and ground truths. We name the new dataset as **S-OCC**.

How We Set the Camera. In contrast to event-based and image-based approaches, this work strives for independence from camera intrinsic and extrinsic parameters, and scene prior knowledge, relying on a single spike camera. To record the scene, we place a spike camera on a slider and move it

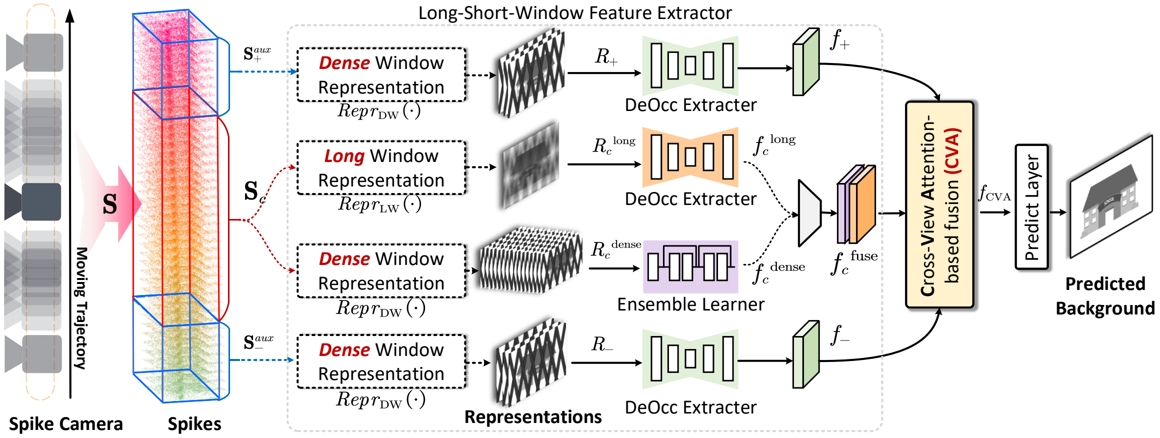

Figure 4: Architecture of the SpkOccNet. With the spike camera moving, continuous spikes are fed into the network, processed by the Long-Short-Window feature extractor (LSW) first then Cross-View Attention-based fusion (CVA).

quickly during each shot, taking about 0.1 seconds for each movement.

What the Occlusions are. We set five intricate occlusions to both enhance model robustness and unveil the potential of spike cameras. They encompassed **(1)** *a square iron mesh*, **(2)** *a hexagonal iron mesh*, **(3)** *a dense iron frame*, **(4)** *a fence*, and **(5)** *an irregular fabric net*. As Fig. 1(a) visually depicted, these occlusions characterize diverse levels of sparsity and density, thereby introducing complexity to this work. Notably, occluding objects are allowed their own motion during the camera capturing process.

How We Construct the Dataset. We record various outdoor scenes utilizing the aforementioned camera motion and occlusion setups, yielding a total of 128 sequences. Among these, 108 sequences are randomly picked for training, while the remaining 20 sequences are for testing. For static scenes without occlusion, We fix the spike camera to capture the background scenes and obtained grayscale images by calculating the spike firing rate (Zhu et al. 2019), which served as the ground truth for the dataset. As a result, each sample in this dataset comprises a spike stream alongside a clear background image with no occlusions. Besides, we claim that our dataset is captured with no sensitive, private information or societal implications involved.

Overall Architecture

We aim to predict the background image I through continuous spikes. We build the model called **SpkOccNet**, as shown in Fig. 4. Each input sample is a spike stream denoted as $\mathbf{S}$, generated by the camera through rapid sliding motion. $\mathbf{S}$ is in size of $H \times W \times T$, where $H \times W = 250 \times 400$ represents the spatial resolution and T is the number of time steps ($T = 0.1s/50\mu s = 2000$).

With camera motion, $\mathbf{S}$ records the dense and continuous changes in viewpoints. Upon the analysis from the previous section, we assert $\mathbf{S}$ contains all information for reconstructing the background. In $\mathbf{S}$, spikes at the two ends correspond to the largest viewpoint change, where the motion displacement of foreground occlusions relative to the background is more significant. Thus, the background texture captured by the spikes at the two ends is likely to complement the information recorded by the spikes in the middle. Motivated by this, we partition S into three segments for processing:

$$\mathbf{S} = \mathbf{S}_-^{aux} + \mathbf{S}_c + \mathbf{S}_+^{aux}, \tag{5}$$

where $\mathbf{S}_-^{aux} = H \times W \times T_-$, $\mathbf{S}_c = H \times W \times T_c$, $\mathbf{S}_+^{aux} = H \times W \times T_+$, and $T_- \in [0, W_{aux}]$, $T_c \in [W_{aux}, T - W_{aux}]$, $T_+ \in [T - W_{aux}, T]$, as shown in Fig.4. W_{aux} is the parameter that controls the window length. We set W_{aux} to 300, which is much shorter than T.

The SpkOccNet includes two stages, **long-short-window feature extractor (LSW)** and **cross-view attention-based fusion (CVA)**. In the first stage, each segment of spikes is first pre-processed with image-like representations, then extracted features with various modules. $\mathbf{S}_+^{aux}$ and $\mathbf{S}_-^{aux}$ is processed with *dense window reprensentaion* then *DeOcc* extracter, and output features f_+, f_-. $\mathbf{S}_c$ is processed with two branches, one is *long window reprensentaion* with *DeOcc* extractor, the other is *dense window reprensentaion* with *Ensemble Learner*, then output features f_c^{long}, f_c^{dense}.

In the second stage, the **CVA** module mainly aims to complement the texture of the center moment (f_c^{long}, f_c^{dense}) from the texture of two ends (f_+, f_-). Firstly, f_c^{long}, f_c^{dense} are fused with a channel-attention layer to get f_c^{fuse}. Then, the **CVA** input with f_c^{fuse}, f_+, f_- for feature fusion with the proposed attention mechanism in which features are enhanced with attention from others along both spatial and channel dimensions. Finally, the prediction layer consisted of Conv layers outputs the final reconstructed background image $\hat{I}$.

Long-Short-Window Feature Extractor

As in Sec. , we split the input $\mathbf{S}$ into three windows, $\mathbf{S}_-^{aux}$, $\mathbf{S}_c$, $\mathbf{S}_+^{aux}$, each of which contains continuous views. The timestamp for reconstruction is the center of $\mathbf{S}_c$. Thus, $\mathbf{S}_c$ provides the spatial structure for reference while $\mathbf{S}_+^{aux}$, $\mathbf{S}_-^{aux}$ provides complementary information for texture in $\mathbf{S}_c$. Firstly, $\mathbf{S}_+^{aux}$, $\mathbf{S}_-^{aux}$ are transformed with *dense window representation* ($Repr_{\mathrm{DW}}$). To be specific, we split spikes into

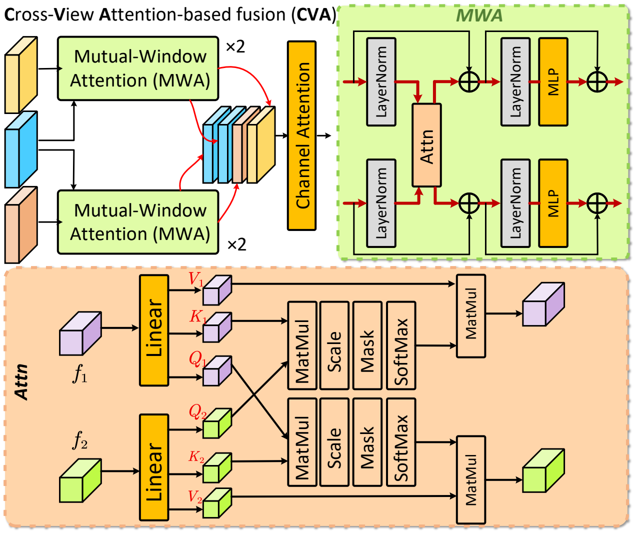

Figure 5: The structure of the proposed Cross-View Attention-based fusion (CVA), consisting of two mutual window attention (MWA) and a channel attention module.

dense groups by non-overlapping a sliding window whose length is $W_{dense} = 100$. In each group, spikes are accumulated in $\mathbf{S}_c$ along the time axis and transformed into W_{dense}/W_{aux} image-like representations with one channel. After $Repr_{\text{DW}}$, $\mathbf{S}^{aux}_{+}$, $\mathbf{S}^{aux}_{-}$ are transformed to R_+, R_-, then processed by the *DeOcc Extracter*, a U-shape network comprised of Conv encoders and decoders, respectively. We denote them as $M^{enc-dec}_{+}$, $M^{enc-dec}_{-}$. After that, features f_+, f_- are obtained. The above process is formulated as:

$$\begin{aligned} R_{+/-} &= Repr_{\text{DW}}(\mathbf{S}^{aux}_{+/-}), \\ f_{+/-} &= f_{M^{enc-dec}_{+/-}}(R_{+/-} : \theta_{M^{enc-dec}_{+/-}}). \end{aligned} \tag{6}$$

For $\mathbf{S}_c$, as it contains spikes from longer time intervals around the central point, we employ two different representations for its transformation.

(A) We utilize the *long window representation*($Repr_{\text{LW}}$) to transform spikes. The depth difference ensures that the foreground occlusions will exhibit larger displacements on the imaging plane than the background. For spikes at different moments, the visible regions in the background will change accordingly due to the variations in occlusions. Therefore, we accumulate spikes in $\mathbf{S}_c$ along the time axis with the full-length $W_{long} = T - 2W_{aux}$, getting the representation R^{long}_c. This transforms the foreground occlusions into a blurred effect similar to a long-exposure image, allowing the occluded background texture to have the "partially-see-through effect". The same U-shape feature extractor $M^{enc-dec}_c$ receives R_c and output feature f^{long}_c. Despite the presence of some motion blur in f^{long}_c, it still offers essential structural information for reconstructing the background.

(B) S_c is also processed with the $Repr_{\text{DW}}$ to get representation R^{dense}_c with continuous view changes. For R^{dense}_c, we employ residual Conv layers M^{res}_c to ensemble effective features f^{dense}_c from dense viewpoints. Finally, we use a channel attention (Zamir et al. 2022) to fuse the two features (f^{long}_c, f^{dense}_c) and get f^{fuse}_c. The above process can be formulated as:

$$\begin{aligned} R^{\text{long}}_c, R^{\text{dense}}_c &= Repr_{\text{LW}}(\mathbf{S}_c), Repr_{\text{DW}}(\mathbf{S}_c), \\ f^{\text{long}}_c &= f_{M^{res}_c}(R^{\text{long}}_c : \theta_{M^{res}_c}), \\ f^{\text{dense}}_c &= f_{M^{enc-dec}_c}(R^{\text{dense}}_c : \theta_{M^{enc-dec}_c}), \\ f^{\text{fuse}}_c &= \mathbf{Fuse}([f^{\text{long}}_c, f^{\text{dense}}_c]), \end{aligned} \tag{7}$$

Cross-View Attention-Based Fusion

In the second stage, we propose a novel Cross-View Attention-Based Fusion (**CVA**) module to fuse and refine the features from the first stage.

The CVA takes three features as input, $f^{\text{fuse}}_c, f_+, f_-$. As shown in Fig. 5, the CVA includes two modules with different attention mechanisms: one is the channel attention (CA) M_{CA} that is adapted from the multi-dconv head Transposed self-attention (MDTA) in Restormer (Zamir et al. 2022), the other is our proposed mutual window attention (MWA) M_{MWA} block.

Among $[f^{\text{fuse}}_c, f_+, f_-]$, f^{fuse}_c represents the center long-interval textures of $\mathbf{S}_c$ while f_- and f_+ offering textures from the shorter time windows on two ends of $\mathbf{S}$. We consider using the cross-view mutual attention mechanism for the following reasons:

(A) The occluded regions in f_-, f_+ and f^{fuse}_c differ due to the differences in viewpoints. Mutual attention can help to compensate for the occluded parts in one feature map by referring to the non-occluded parts in other feature maps.

(B) Due to the high-speed camera motion, there can still be some spatial displacement of the background scenes. The mutual attention mechanism can effectively align features to mitigate the impact of camera motion.

The MWA approximately comprises two Transformer blocks. Specifically, the standard Transformer block takes one input f, obtains its Q(*query*), K(*key*), and V(*value*) matrices and operates self-attention. For our MVA, it takes two inputs f_1 and f_2, obtains Q_1, K_1, V_1 and Q_2, K_2, V_2 matrices, and operates mutual attentions as followings:

$$\begin{aligned} f_{1,2}, f_{2,1} &= f_{M_{\text{MWA}}}(f_1, f_2 : \theta_{M_{\text{MWA}}}), \\ &= \mathbf{Attn}(Q_1, K_2, V_2), \mathbf{Attn}(Q_2, K_1, V_1). \end{aligned} \tag{8}$$

Considering the computing cost, we adopt window-based attention operation in the Swin Transformer blocks (Liu et al. 2021). In the MVA, we compute mutual attention twice, one between f^{fuse}_c and f_-, the other between f^{fuse}_c and f_+. The process can be formulated as:

$$\begin{aligned} f_{c,-}, f_{-,c} &= f_{M_{\text{MWA}}}(f^{\text{fuse}}_c, f_- : \theta_{M_{\text{MWA}}}), \\ f_{c,+}, f_{+,c} &= f_{M_{\text{MWA}}}(f^{\text{fuse}}_c, f_+ : \theta_{M_{\text{MWA}}}). \end{aligned} \tag{9}$$

Features are then concatenated together along the channel axis and processed by the channel attention layer which contains one MDTA layer for attention and one Conv layer for reducing output channels. It can be formulated as:

$$f_{\text{CVA}} = f_{M_{\text{CA}}}(\{f_{c,-}, f_{-,c}, f_{c,+}, f_{+,c}\} : \theta_{M_{\text{CA}}}). \tag{10}$$

Finally, the prediction layer consisted of Conv layers, input with f_{CVA}, outputs the predicted background image $\hat{I}$.

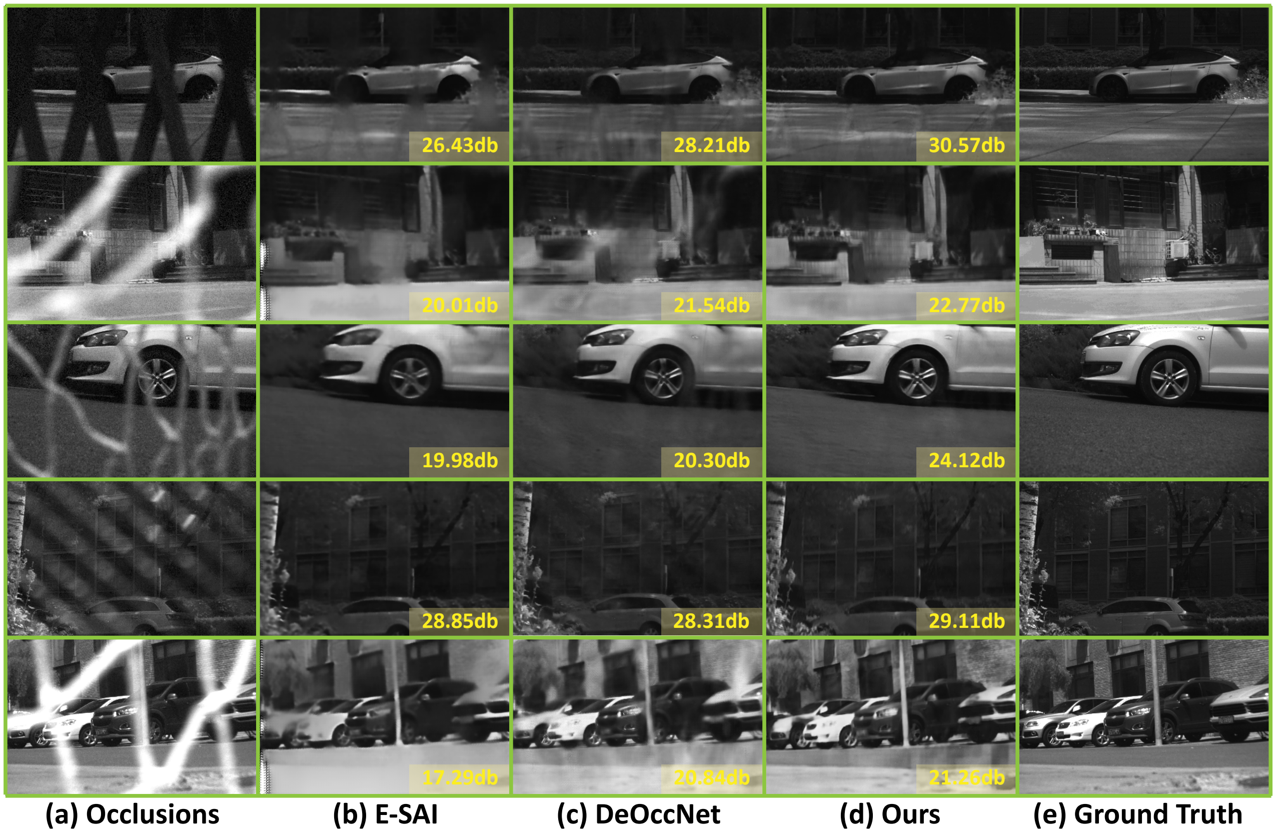

Figure 6: Results on S-OCC for our method compared with E-SAI (Zhang et al. 2021) and DeOccNet (Wang et al. 2020).

Experiments

We train all networks with PyTorch. L1 loss is used for optimization. Quantitative metrics are peak signal-to-noise ratio (PSNR) and structural similarity (SSIM). Detailed information about the S-OCC dataset, training and network implementation details, are included in the supplementary file.

Quantitative and Qualitative Results

We compare the proposed SpkOccNet with three other methods. Firstly, we use the U-shaped feature extractor used in our model as the baseline whose output channels are set to 1. Secondly, we include DeOccNet (Wang et al. 2020), a model that performs well in the image domain. For these two models, we simulated images as inputs. Specifically, we take the middle 300 spike frames (denoted as $\mathbf{S}_c^{mid}$) of $\mathbf{S}_c$ and process $\mathbf{S}_c^{mid}, \mathbf{S}_+, \mathbf{S}_-$ through $Repr_{\text{LW}}$ to obtain three accumulated grayscale images, which are similar to sequence captured by a camera with an exposure time of 15ms ($300\times50\mu$s) and the frame rate closed to 25 fps. Thirdly, we consider the model E-SAI (Zhang et al. 2021), which is based on event cameras and trained with a hybrid SNN-CNN model. We split the spike $\mathbf{S}$ into 30 dense windows as input according to the settings in E-SAI.

The quantitative results in terms of PSNR and SSIM are presented in Tab. 1. The table shows the performance of each model on the test set for five different occlusion scenarios, as well as the average performance. It can be observed that our model, SpkOccNet, achieves the best performance on the S-OCC dataset compared to the other three methods. SpkOccNet achieves a PSNR of 26.83 dB, which is approximately 1.76dB higher than DeOccNet and 1.32 dB higher than the E-SAI method. It is worth noting that SpkOccNet has a parameter size of only about 4.9M, while DeOccNet and E-SAI have parameter size of 39.04M and 18.59M, which are **8.0 and 3.8 times more than ours**. The performance and the parameter size demonstrate the stronger advantage of our proposed model for spike-based occlusion removal. Our model exhibits more pronounced advantages in dealing with ‘Fence’, ‘Hexagonal Mesh’, and ‘Fabric net’ occlusions. The ’Fence’ scenario involves extensive occlusions, whereas the ’Fabric Net’ exhibits densely-packed and irregular occlusions, both of which pose substantial challenges.

Fig. 6 presents visualized results. Our method achieves higher-quality image reconstructions. Specifically, our method is able to recover clearer background textures and overcome the issue of the change of illumination caused by occlusions. Although the E-SAI can reconstruct relatively smooth images, the reconstructed images appear blurry and the effect of under/overexposure exists in some regions. Besides, DeOccNet performs badly in removing severe occlusions. The results demonstrate that our method outperforms the other two image-domain and event-domain methods.

Occlusions	Methods	PSNR ↑	SSIM ↑
Fense	Baseline U-Net	25.89	0.774
	DeOccNet (2020)	26.06	0.782
	E-SAI (2021)	26.68	0.767
	SpkOccNet	**28.38**	**0.792**
Raster	Baseline U-Net	22.71	0.690
	DeOccNet (2020)	23.60	0.686
	E-SAI (2021)	25.13	0.745
	SpkOccNet	**26.28**	**0.746**
Square Mesh	Baseline U-Net	26.63	0.766
	DeOccNet (2020)	25.60	0.762
	E-SAI (2021)	28.26	0.772
	SpkOccNet	27.35	0.745
Hexagon Mesh	Baseline U-Net	28.42	0.713
	DeOccNet (2020)	28.17	0.759
	E-SAI (2021)	27.48	0.776
	SpkOccNet	**29.67**	**0.846**
Fabric Net	Baseline U-Net	19.50	0.640
	DeOccNet (2020)	20.90	0.660
	E-SAI (2021)	19.69	0.626
	SpkOccNet	**22.33**	**0.683**
Total	Baseline U-Net	24.60	0.761
	DeOccNet (2020)	24.77	0.771
	E-SAI (2021)	25.51	0.765
	SpkOccNet	**26.83**	**0.805**

Table 1: Comparison of various de-occlusion methods on S-OCC. The table shows results on five occlusion types.

$M^{enc-dec}_{c,+,-}$	M^{res}_{c}	M_{MVA}	PSNR ↑	SSIM ↑
✓			25.53	0.773
	✓		23.72	0.737
✓	✓		26.11	0.789
✓	✓	✓	**26.83**	**0.805**

Table 2: Ablation study of the proposed modules.

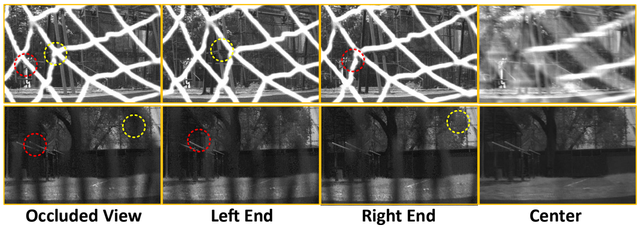

Figure 7: Visualized dense window representations of two-end spikes and long window representation of center spikes.

Input Length	45ms	75ms	105ms	Image
PSNR ↑	26.42	26.77	26.83	25.90
SSIM ↑	0.795	0.807	0.805	0.794

Table 3: Ablation study of input length of spikes and comparison with images as input.

Ablation Studies

Ablation on Modules. To validate the effectiveness of the proposed modules in SpkOccNet, we first conduct ablation experiments on modules, and the results are shown in Tab.2.

The effectiveness of the **LSW** is validated through **Row 1**∼**3**. Specifically, in the **Row 1**, only $R_c^{\text{long}}, R_+, R_-$ are used as inputs which are processed through their $M^{enc-dec}_{c,+,-}$ and the output features are simply concatenated for fusion. In the **Row 2**, only the dense representation R_c^{dense} of $\mathbf{S}_c$ is used as input and processed by the residual feature extractor M^{res}_c to obtain features. **Row 3** combines the approaches from the previous two rows. **Row 1** and **Row 3** demonstrate that the dense representation of spikes preserves temporally dense viewpoint information, thereby enhancing performance. **Row 2** and **Row 3** verify the effectiveness of the representations of the viewpoints at both ends and the longer time window representation in the middle.

Row 3 and **Row 4** validate the effectiveness of the **CVA** in the second stage. In **Row 4**, when performing feature fusion, the proposed M_{MVA} in CVA is used, while in **Row 3**, both simple concatenation and the Conv layer are employed for fusion. These two rows illustrate the effectiveness of **CVA** in facilitating feature complementarity across different viewpoints. Fig. 7 illustrates the dense window representations of $\mathbf{S}_+^{aux}$ and $\mathbf{S}_-^{aux}$, along with the long window representation R_c of $\mathbf{S}_c$. It is evident that $\mathbf{S}_+^{aux}$ and $\mathbf{S}_-^{aux}$ encompass complementary information from occluded viewpoints (red and yellow circles). R_c simulates the overall blurred texture similar to a "long exposure" effect.

Analysis on Input. We conducted experiments on the input spike length which represents the extent of viewpoint changes during camera motion. Additionally, to contrast the advantages of spike cameras against frame-based cameras, we simulated images as inputs of SpkOccNet, as described in Sec. . As reported in Tab. 3. In general, even with spikes only recording 45 ms (equivalent to only the time of one captured by a frame-based camera), the performance remains superior to that of image inputs. The result underscores the significance of viewpoint continuity in background reconstruction. As the input length increases, the PSNR improves, indicating that the magnitude of viewpoint changes also influences background reconstruction.

Conclusion

We explore the first spike-based SAI, utilizing spikes for recovering backgrounds from dense occlusions. Our model **SpkOccNet** integrates information from different viewpoints and window lengths, employing mutual attention for effective fusion and refinement. We contribute the first real-world spike-based dataset S-OCC for occlusion removal. Remarkably, our algorithm achieves impressive occlusion removal results using a single camera with fast motion.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (62176003, 62088102, 6230070113) and the China Postdoctoral Science Foundation (2022M720238, 2023T160015).

References

Brandli, C.; Berner, R.; Yang, M.; Liu, S.-C.; and Delbruck, T. 2014. A 240× 180 130 db 3 μs latency global shutter spatiotemporal vision sensor. *IEEE Journal of Solid-State Circuits*, 49(10): 2333–2341.

Chen, S.; Duan, C.; Yu, Z.; Xiong, R.; and Huang, T. 2022. Self-Supervised Mutual Learning for Dynamic Scene Reconstruction of Spiking Camera. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI)*, 2859–2866.

Chen, S.; Yu, Z.; and Huang, T. 2023. Self-Supervised Joint Dynamic Scene Reconstruction and Optical Flow Estimation for Spiking Camera. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(1): 350–358.

Chen, S.; Zhang, J.; Zheng, Y.; Huang, T.; and Yu, Z. 2023. Enhancing Motion Deblurring in High-Speed Scenes with Spike Streams. In *Thirty-seventh Conference on Neural Information Processing Systems*.

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.

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 (CVPR)*, 17844–17853.

Huang, T.; Zheng, Y.; Yu, Z.; Chen, R.; Li, Y.; Xiong, R.; Ma, L.; Zhao, J.; Dong, S.; Zhu, L.; Li, J.; Jia, S.; Fu, Y.; Shi, B.; Wu, S.; and Tian, Y. 2022. 1000× Faster Camera and Machine Vision with Ordinary Devices. *Engineering*.

Hur, J.; Lee, J. Y.; Choi, J.; and Kim, J. 2023. I See-Through You: A Framework for Removing Foreground Occlusion in Both Sparse and Dense Light Field Images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 229–238.

Li, Y.; Yang, W.; Xu, Z.; Chen, Z.; Shi, Z.; Zhang, Y.; and Huang, L. 2021a. Mask4D: 4D convolution network for light field occlusion removal. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2480–2484. IEEE.

Li, Z.; Sun, Y.; Zhang, L.; and Tang, J. 2021b. CTNet: Context-based tandem network for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12): 9904–9917.

Liao, W.; Zhang, X.; Yu, L.; Lin, S.; Yang, W.; and Qiao, N. 2022. Synthetic aperture imaging with events and frames. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 17735–17744.

Lichtsteiner, P.; Posch, C.; and Delbruck, T. 2008. A 128×128 120 dB 15μs Latency Asynchronous Temporal Contrast Vision Sensor. *IEEE Journal of Solid-State Circuits*, 43(2): 566–576.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 10012–10022.

Masland, R. H. 2012. The Neuronal Organization of the Retina. *Neuron*, 76(2): 266–280.

Pei, Z.; Zhang, Y.; Chen, X.; and Yang, Y.-H. 2013. Synthetic aperture imaging using pixel labeling via energy minimization. *Pattern Recognition*, 46(1): 174–187.

Vaish, V.; Levoy, M.; Szeliski, R.; Zitnick, C. L.; and Kang, S. B. 2006. Reconstructing occluded surfaces using synthetic apertures: Stereo, focus and robust measures. In *Proceedings of the 2004 IEEE 2006 Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, 2331–2338. IEEE.

Vaish, V.; Wilburn, B.; Joshi, N.; and Levoy, M. 2004. Using plane+ parallax for calibrating dense camera arrays. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, I–I. IEEE.

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

Wang, Y.; Wu, T.; Yang, J.; Wang, L.; An, W.; and Guo, Y. 2020. DeOccNet: Learning to see through foreground occlusions in light fields. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 118–127.

Wässle, H. 2004. Parallel processing in the mammalian retina. *Nature Reviews Neuroscience*, 5(10): 747–757.

Yang, T.; Zhang, Y.; Yu, J.; Li, J.; Ma, W.; Tong, X.; Yu, R.; and Ran, L. 2014. All-in-focus synthetic aperture imaging. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13*, 1–15. Springer.

Yu, L.; Zhang, X.; Liao, W.; Yang, W.; and Xia, G.-S. 2022. Learning to See Through with Events. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zamir, S. W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F. S.; and Yang, M.-H. 2022. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5728–5739.

Zhang, J.; Jia, S.; Yu, Z.; and Huang, T. 2023. Learning Temporal-Ordered Representation for Spike Streams Based on Discrete Wavelet Transforms. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(1): 137–147.

Zhang, J.; Tang, L.; Yu, Z.; Lu, J.; and Huang, T. 2022a. Spike Transformer: Monocular Depth Estimation for Spiking Camera. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII*, 34–52. Springer.

Zhang, S.; Chen, Y.; An, P.; Huang, X.; and Yang, C. 2022b. Light field occlusion removal network via foreground location and background recovery. *Signal Processing: Image Communication*, 109: 116853.

Zhang, S.; Shen, Z.; and Lin, Y. 2021. Removing Foreground Occlusions in Light Field using Micro-lens Dynamic Filter. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI)*, 1302–1308.

Zhang, X.; Liao, W.; Yu, L.; Yang, W.; and Xia, G.-S. 2021. Event-based synthetic aperture imaging with a hybrid network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14235–14244.

Zhang, X.; Zhang, Y.; Yang, T.; and Yang, Y.-H. 2017. Synthetic aperture photography using a moving camera-IMU system. *Pattern Recognition*, 62: 175–188.

Zhao, R.; Xiong, R.; Zhao, J.; Yu, Z.; Fan, X.; and Huang, T. 2022. Learning optical flow from continuous spike streams. *Advances in Neural Information Processing Systems (NeurIPS)*, 35: 7905–7920.

Zheng, Y.; Yu, Z.; Wang, S.; and Huang, T. 2023a. Spike-Based Motion Estimation for Object Tracking Through Bio-Inspired Unsupervised Learning. *IEEE Transactions on Image Processing*, 32: 335–349.

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*, 45(7): 8127–8142.

Zheng, Y.; Zheng, L.; Yu, Z.; Shi, B.; Tian, Y.; and Huang, T. 2021. High-speed image reconstruction through short-term plasticity for spiking cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6358–6367.

Zhu, L.; Dong, S.; Huang, T.; and Tian, Y. 2019. A retina-inspired sampling method for visual texture reconstruction. In *IEEE International Conference on Multimedia and Expo (ICME)*, 1432–1437.

Zhu, L.; Li, J.; Wang, X.; Huang, T.; and Tian, Y. 2021. NeuSpike-Net: High Speed Video Reconstruction via Bio-inspired Neuromorphic Cameras. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2400–2409.