This WACV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Spk2ImgMamba: Spiking Camera Image Reconstruction with Multi-scale State Space Models

Jiaoyang Yin[1,2♯] Bin Fan[3♯] Chao Xu[3] Tiejun Huang[1,2] Boxin Shi[1,2] ✉
[1]State Key Lab of Multimedia Info. Processing, School of Computer Science, Peking University
[2]Nat'l Eng. Research Ctr. of Visual Technology, School of Computer Science, Peking University
[3]Nat'l Key Lab of General AI, School of Intelligence Science and Technology, Peking University

Abstract

As a bio-inspired vision sensor, the spiking camera has showcased remarkable capability in high-speed imaging with a sampling rate of 40,000 Hz. Reconstructing clear images from continuous spike streams, which is obtained by each photosensor continuously detecting photons and firing them asynchronously, has garnered significant attention. Despite promising results, existing spike-to-image reconstruction methods face challenges in balancing global receptive fields and efficient computation due to the inherent limitations of their backbones. Recently, due to powerful long-range modeling and linear complexity, the state space model (SSM) has emerged as a competitive alternative to CNNs and Transformers. In this paper, we propose a lightweight spike-to-image reconstruction network that harnesses Mamba as the backbone. Our approach sequentially executes three core modules: temporal information integration, spatial feature enhancement, and progressive image reconstruction. The former accumulates cues across diverse temporal windows to explore both long-term and short-term contexts. Subsequently, to model global dependencies while heightening local detail perception, we develop a multi-scale SSM block characterized by multi-scale multi-direction scanning, which effectively boosts spatial feature representations. Finally, intensity images are decoded progressively from the enhanced light-intensity features. Extensive experiments on both synthetic and real-captured data demonstrate that our approach achieves state-of-the-art performance, with only 10% of the network parameters and nearly two orders of magnitude less computational effort. The code will be available at https://github.com/interstellarH/Spk2ImgMamba.

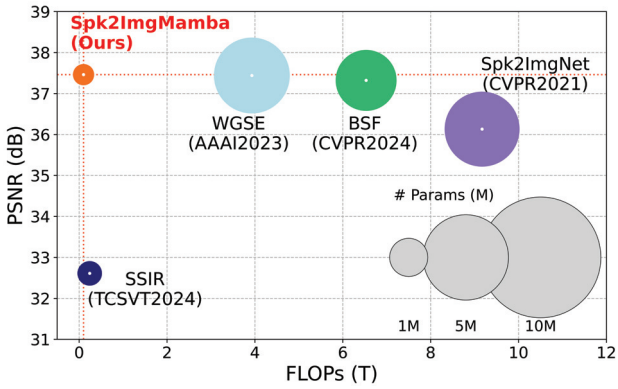


Figure 1. Model comparison of PSNR, FLOPs, and model size on the SREDS dataset [55]. Our Mamba-based architecture achieves excellent results in terms of reconstruction accuracy and computational efficiency. SSIR [55] is SNN-based, while Spk2ImgNet [53], BSF [58], and WGSE [51] are CNN-based methods.

1. Introduction

Inspired by the foveal sampling mechanism of the retina, a new type of neuromorphic visual sensor, known as the spiking camera [6, 28], has emerged in recent years. The spiking camera has been developed with a sampling rate of 40,000 Hz, demonstrating vast potential in high-speed imaging applications [2, 7, 46, 50, 56, 60, 66] such as drones and chemical particle observation. Unlike event cameras [1, 31, 34, 39], which record relative changes of light intensity (*i.e.*, differential sampling), each pixel of the spiking camera continuously accumulates photons and fires spikes asynchronously. This integration-based sampling allows the spiking camera to retain the absolute light intensity information of the scene, offering the possibility of reconstructing fine textures. However, binary spike streams are not intuitive for human visual perception and pose challenges for off-the-shelf computer vision algorithms to adapt di-

♯ Equal contribution.
✉ Corresponding author (shiboxin@pku.edu.cn).

rectly. Consequently, spike-to-image reconstruction, *i.e.*, reconstructing latent image frames from continuous spike streams, has become a research topic of great interest.

To reconstruct dynamic scenes from spike data, intrinsic temporal statistical characteristics [64] are often utilized by calculating the firing frequency or firing interval of each pixel, known as texture from playback (TFP) and texture from inter-spike-intervals (TFI), respectively. Also, other efforts focus on refining firing frequency estimates via retina-like visual imaging [65] and short-term plasticity [62,63]. Although these non-learning-based methods offer good interpretability, they often struggle to balance noise and motion blur due to limitations in the choice of temporal windows, resulting in less-than-ideal reconstruction results. In recent years, learning-based methods have gained popularity. Convolutional neural networks (CNNs) are particularly used to mine meaningful spatio-temporal information from spike inputs. However, to achieve sufficient receptive fields, existing CNN-based methods [4, 51, 53, 58] typically require stacking numerous convolutional blocks, which inevitably increases model complexity (*cf.*, Fig. 1). In contrast, while the Transformer-based method [29] can model long-range dependencies with the powerful expressive capability of Vision Transformers [8, 33], it is limited by the quadratic complexity concerning the input size. Overall, *current spike-to-image reconstruction backbones still face the dilemma of balancing global receptive fields with efficient computation*, hindering their practical application.

Recently, State Space Model (SSM)-based methods [18, 21, 42], especially the improved Mamba [20], have garnered attention and shown great potential in the fields of image restoration [22, 41], image enhancement [19, 59], object classification [3, 27], and target tracking [26, 48]. This is because Mamba is able to combine the best of both worlds: global receptive fields and linear complexity. Thus, Mamba presents a potentially effective approach for tackling the above dilemma in spike-to-image reconstruction tasks. However, directly adapting the standard Mamba, designed for 1D sequential data in NLP, to spike-to-image reconstruction remains challenging. To deal with the structured 2D image, a common approach is to divide the image into patches, and flatten it into a 1D sequence along a scanning path for recursive processing. To accommodate the non-causality of 2D images, ViM [67] devised bidirectional scanning to enhance the temporal relationships of sequence elements, and horizontal and vertical scanning were further introduced in VMamba [32]. Nevertheless, this single-scale, multi-direction scanning strategy, upon 1D unfolding, may render spatially adjacent pixels distant in the flattened sequence, *e.g.*, ears (red boxes) and eyes (blue boxes) in Fig. 2, disrupting the natural correlation between local pixels and impacting the model's accuracy in capturing local structures. We term this the *local correlation attenuation problem*. Additionally, as the arrival of photons follows a Poisson process, each asynchronously fired spike does not directly represent the cue corresponding to its space-time position [57]. The *fluctuation and sparsity of spike data* make it difficult to extract reliable light-intensity features.

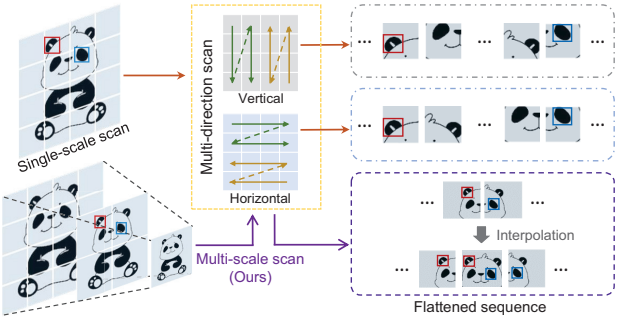


Figure 2. A toy example of different scanning methods on 2D images. *(top)* Single-scale scanning method [32] traverses the entire row or column axis, resulting in large distances between dependencies of spatially adjacent pixels (which have natural local correlations) in the flattened sequence. *(bottom)* We introduce a multi-scale scanning method that splits the features into a multi-scale structure, facilitating multi-direction scanning across scales, which enhances the ability to capture local dependencies.

In this paper, we propose Spk2ImgMamba, the *first* work to leverage Mamba as a powerful backbone for spike-to-image reconstruction. To enhance the model's perception of local details while maintaining the benefit of long-range dependency modeling, we design a tractable multi-scale SSM. It first utilizes channel splitting to package features into a multi-scale structure, then performs multi-direction scanning across scales, and finally fuses contextual features. This *multi-scale, multi-direction scanning* facilitates the adaptation of Mamba to non-sequential visual data. Moreover, we introduce a compact temporal information integration module to explore long-term and short-term contexts. Combined with the spatial feature enhancement module based on multi-scale SSM, the light-intensity features can be gradually extracted from input spike data for efficient image reconstruction. Extensive experiments show that our approach outperforms state-of-the-art (SOTA) CNN-based methods on synthetic and real-captured data, using *only 10% of the network parameters and nearly two orders of magnitude less computational effort*, as illustrated in Fig. 1. To summarize, this paper makes contributions by

- proposing Spk2ImgMamba, the first use of Mamba as a backbone for spike-to-image reconstruction;
- designing a multi-scale SSM, with a global receptive field and a keen sensitivity to local correlations; and
- achieving excellent reconstruction accuracy while offering significant advantages in model size and FLOPs.

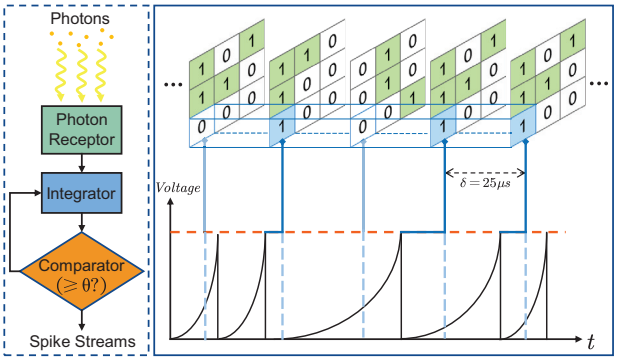


Figure 3. Illustration of the working mechanism of the spiking camera. Each pixel in a spike camera contains a photon receptor, integrator, and comparator. The receptor converts photons into photoelectrons, which the integrator accumulates. When the accumulated photoelectrons, represented as voltage values, surpass a certain threshold, a spike is generated, resetting the integrator.

2. Related Work

Spike-to-image Reconstruction. To better satisfy human visual perception and facilitate direct processing by machines, reconstructing the corresponding image frame from spikes has been widely studied. The temporal statistical properties of spiking cameras are usually leveraged. Zhu *et al*. [64] explored the spike generation principle and proposed two basic methods, *a.k.a*., TFP and TFI. Zhao *et al*. [54] hierarchically merged short- and long-term filtering. Another line of work aims to mimic human physiological mechanisms [62–65]. With the advent of deep learning, Spk2ImgNet [53] presented the first CNN-based architecture and achieved impressive results. The wavelet transform was combined with CNN-based learnable modules in [51]. Recently, an energy-efficient scheme was developed [55] based on the spiking neural network (SNN), but its performance is far below the desired level. Furthermore, self-supervised CNN-based methods [4, 5] were proposed to reduce data dependency. HDR and HFR videos were generated in [2] by introducing the rolling readout mechanism that is shown to be beneficial for HFR video reconstruction [9–16]. Yet, since CNNs fall short in encapsulating global context, designing complex network models using CNNs inevitably leads to high model complexity, increasing the marginal cost on resource-constrained devices.

State Space Models. State Space Models (SSMs), rooted in classical control theory, have recently gained popularity as a deep learning backbone in place of Transformers and CNNs. In modeling long-range relationships, their unique capacity to scale linearly with sequence length has garnered significant interest from the scholarly community. S4 [21] proposed normalizing the parameter matrices into a diagonal form to conserve computing power. Subsequent work [42] further introduced parallel scan techniques and multiple input multiple outputs. More recently, Mamba [20] set itself apart by incorporating a selective scan algorithm to acquire content awareness and a hardware-optimized algorithm to achieve remarkable computational efficiency. The primary challenge in adapting the original 1D sequence-based Mamba to vision tasks is serializing non-causal 2D vision data. VMamba [32] and Vim [67] pioneered the introduction of Mamba into vision backbone design via cross-scan and bidirectional SSM, achieving competitive results against ViTs and CNNs. Inspired by them, a series of works also explored various scanning strategies, such as atrous scanning [38], omnidirectional scanning [41], and continuous scanning [49]. However, merely altering the scanning direction still struggles to capture the subtle correlations between local pixels, thereby affecting detail recovery. Our proposed multi-scale scanning mechanism performs multi-direction scanning across scales, balancing both global and local receptive fields.

3. Preliminaries

3.1. Working Mechanism of Spiking Cameras

Spiking cameras employ an "integrate-and-fire" mechanism. As illustrated in Fig. 3, each pixel independently and continuously receives photons from the scene and converts them into photoelectrons, which are then accumulated via an integrator. When the accumulated photoelectrons exceed the predetermined threshold, the spiking camera asynchronously fires a spike, while clearing the pixel's photoelectrons to start a new accumulation cycle. The working mechanism can be formulated as [28, 54, 61]:

$$\mathcal{A}_t(\mathbf{x}) = \int_0^t \alpha \mathbf{I}_\tau(\mathbf{x}) \mathrm{d}\tau \mod \theta, \tag{1}$$

where $\mathcal{A}_t(\mathbf{x})$ reflects the number of photoelectrons accumulated at pixel $\mathbf{x} = (x, y)$ in the integrator. $\mathbf{I}_\tau(\mathbf{x})$ represents the light intensity at pixel $\mathbf{x}$ at time τ. The photoelectric conversion rate is denoted as α and the firing threshold is set to θ. Assume that δ (in microsecond level) is used to quantify a spike accumulation cycle, the spiking camera can output a dense binary spike plane with a spatial resolution of $H \times W$ at time $n\delta, n \in \mathbb{N}$. As a result, during an "integrate-and-fire" process with a temporal length of T, the spatio-temporal resolution of the spike stream $\mathcal{S}_t$ will reach $H \times W \times T$, where t denotes the central time of $\mathcal{S}_t$.

3.2. State Space Models

Formulation. SSMs are a class of linear time-invariant systems designed to map a 1-D continuous input signal $x(t) \in \mathbb{R}$ to a corresponding output signal $y(t) \in \mathbb{R}$ through an N-D latent state $h(t) \in \mathbb{R}^N$. Specifically, the

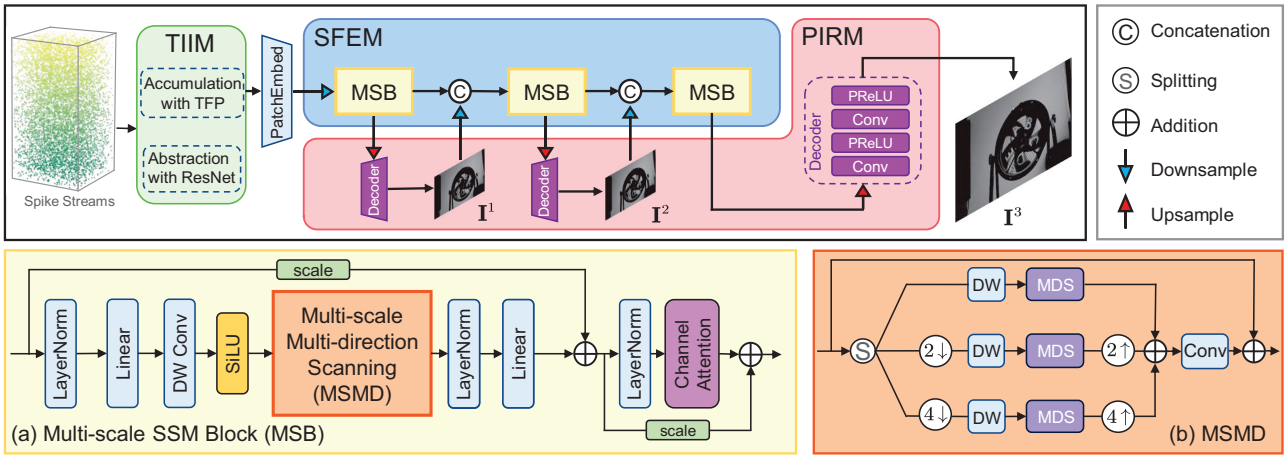


Figure 4. Overall architecture of Spk2ImgMamba. It mainly consists of three modules: Temporal Information Integration (TIIM), Spatial Feature Enhancement (SFEM), and Progressive Image Reconstruction (PIPM). First, TIIM is designed to explore long-term and short-term contexts in the temporal domain. Then, we propose a Multi-scale SSM block (MSB) to balance global and local dependencies, with multiple MSBs cascaded to form SFEM. In each MSB, features are packed into L scales by channel splitting, followed by Multi-Direction Scanning (MDS), and then cross-scale fusion is performed for feature enhancement. Finally, the intensity image is progressively decoded from the continuously enhanced light-intensity features.

continuous-time representation of SSMs can be formulated as a linear ordinary differential equation:

$$\begin{aligned} h'(t) &= \mathbf{A}h(t) + \mathbf{B}x(t), \\ y(t) &= \mathbf{C}h(t), \end{aligned} \tag{2}$$

where $\mathbf{A} \in \mathbb{R}^{N\times N}$, $\mathbf{B} \in \mathbb{R}^{N\times 1}$, and $\mathbf{C} \in \mathbb{R}^{1\times N}$ are learnable parameters.

Discretization. To adapt the continuous-time SSM to discrete input sequence $\mathbf{x} = (x_0, x_1, \dots) \in \mathbb{R}^M$ with a length of M, the S4 model discretizes these parameters using a step size Δ, which corresponds to the resolution of the continuous input $x(t)$. Specifically, the continuous parameters $\mathbf{A}$, $\mathbf{B}$ are transformed into their discrete counterparts $\overline{\mathbf{A}}$, $\overline{\mathbf{B}}$ through the zero-order hold technique, which is defined as:

$$\begin{aligned} \overline{\mathbf{A}} &= \exp(\Delta\mathbf{A}), \\ \overline{\mathbf{B}} &= (\Delta\mathbf{A})^{-1}(\exp(\Delta\mathbf{A}) - \mathbf{I}) \cdot \Delta\mathbf{B}. \end{aligned} \tag{3}$$

The discretized version of Eq. (2) with a step size Δ can be rewritten in the following RNN form, enabling *fast autoregressive inference*:

$$\begin{aligned} h_k &= \overline{\mathbf{A}}h_{k-1} + \overline{\mathbf{B}}x_k, \\ y_k &= \mathbf{C}h_k. \end{aligned} \tag{4}$$

Here, we use k instead of t to denote discretized timesteps. Furthermore, Eq. (2) can also be mathematically transformed into the following equivalent CNN form, enabling *efficient parallelizable training*:

$$\begin{aligned} \mathbf{y} &= \mathbf{x} * \overline{\mathbf{K}}, \\ \overline{\mathbf{K}} &= (\mathbf{C}\overline{\mathbf{B}}, \mathbf{C}\overline{\mathbf{A}\mathbf{B}}, \mathbf{C}\overline{\mathbf{A}}^{M-1}\overline{\mathbf{B}}), \end{aligned} \tag{5}$$

where $*$ denotes the convolution operation, and $\overline{\mathbf{K}} \in \mathbb{R}^M$ is a structured convolution kernel.

Selection. To solve the problem brought by linear time-invariant property when it comes to context-based reasoning, Mamba [20] introduces an input-dependent selection mechanism by making the parameters $\Delta, \mathbf{B}, \mathbf{C}$ functions of the input sequence $\mathbf{x}$. Thus, the time-varying SSM inherently possesses the capacity for context-aware reasoning while maintaining linear time complexity and memory efficiency. In the meantime, parallel scan algorithms and hardware-aware techniques are employed to facilitate efficient training.

4. Approach

Our approach aims to reconstruct the intermediate intensity image $\mathbf{I}_t$ corresponding to time t from spike streams $\mathcal{S}_t$ with a resolution of $H \times W \times T$. As depicted in Fig. 4, our approach consists of three core modules: temporal information integration (TIIM), spatial feature enhancement (SFEM), and progressive image reconstruction (PIRM). These modules are executed sequentially to gradually extract representative light-intensity features from input spike streams in a spatio-temporal aggregation manner. The TIIM focuses on long-term and short-term contexts in the temporal domain, incorporating both explicit spike accumulation and implicit feature abstraction capabilities. In the SFEM, we integrate a multi-scale SSM block, which models long-range dependencies and amplifies local correlations through *multi-scale, multi-direction scanning*. Notably, multi-direction scanning enables global interaction

between pixels, establishing long-range spatial dependencies; multi-scale scanning facilitates interactions between each pixel and its neighbors, thereby enhancing the capture of local spatial context, as depicted in Fig. 2. Finally, the intensity image is reconstructed progressively from the enhanced light-intensity features in the PIRM. We elaborate on these modules below.

4.1. Temporal Information Integration

The arrival of photons during spike firing follows a Poisson process and is influenced by thermal noise and quantization noise. Even with constant light intensity, the integration period of each spike may vary over time [58]. This brings fluctuation and sparsity in the spike data, making it challenging to directly extract spatio-temporal information from continuous spike streams. To address this, inspired by [17], we introduce a compact TIIM that fully exploits the time-series cues within the spike stream from both explicit spike accumulation and implicit feature abstraction perspectives.

Explicit Spike Accumulation. Preliminary spike-to-image reconstruction can be achieved similarly to TFP [64] by accumulating the spike planes within a defined temporal window ω ($0 < \omega \leq T$):

$$\mathbf{I}_t^{\omega}(\mathbf{x}) = \frac{\zeta}{\omega} \cdot \sum_{\tau=t-\omega/2}^{t+\omega/2} \mathcal{S}_{\tau}(\mathbf{x}), \tag{6}$$

where ζ is the maximum dynamic range. Different images can be viewed as summarizing spikes or extracting temporal correlations among spike planes with different ω. Notably, short-term ω provides better details but introduces noise, while long-term ω can suppress noise but tends to generate blurriness. Instead of balancing them, by setting diverse ω, we directly apply Eq. (6) as a baseline to explore both long-term and short-term contexts, aiming to generate a series of rough intensity images $\{\mathbf{I}_t^{\omega_i}\}_{i=1}^{N}$ corresponding to time t. Below, the subscript t is omitted for brevity.

Implicit Feature Abstraction. Inspired by [47], we leverage ResNet [23] to directly capture long-term contexts from the input spike stream and abstract feature representations $\mathbf{F}_{\text{IFA}} \in \mathbb{R}^{H \times W \times C_{\text{IFA}}}$. This can complement the insufficiency of feature modeling capabilities of explicit spike accumulation.

Next, we concatenate multiple sub-optimal images $\{\mathbf{I}_t^{\omega_i}\}_{i=1}^{N}$ with implicit feature representations $\mathbf{F}_{\text{IFA}}$ along the channel dimension, yielding light-intensity features $\mathbf{F}_{\text{T}}$ in the temporal domain, which effectively combines the advantages of explicit and implicit modeling to characterize time-series information in spike streams properly.

4.2. Spatial Feature Enhancement

After temporal information integration, the raw spike streams have been transformed into more tractable features $\mathbf{F}_{\text{T}}$, and the rich contextual information along the temporal axis is aggregated. However, in the spatial domain, the interaction between pixels is not exactly modeled. Considering that the single-scale scanning approach used in the vanilla SSM often leads to the problem of local correlation attenuation, we propose a multi-scale SSM block (MSB) equipped with multi-scale, multi-direction scanning (MSMD), which facilitates the full consideration of local pixel interactions while modeling globally. We cascade P MSBs for progressively refining the light-intensity features $\{\mathbf{F}_{\text{S}}^i\}_{i=1}^{P}$ in the spatial domain through P periods, which make up our SFEM. Note that each $\mathbf{F}_{\text{S}}^i$ can be used to reconstruct an intensity image $\mathbf{I}^P$, which is then re-concatenated into $\mathbf{F}_{\text{S}}^i$ and fed into the next MSB, serving as a reference for the following refinement. In particular, to save computational resources and improve computing efficiency, we perform downsampling before SFEM via overlapping patch embedding [45], using a 2D convolution with a kernel size of 5×5 and a stride of 2.

Multi-scale SSM Block (MSB). Inspired by VMamba [32] that consists of a single-scale, multi-direction scanning for global modeling as well as the versatile multi-scale structure [13, 25, 43, 44] for granular exploration, we design a MSB to enhance the modeling ability of local details while ensuring long-range dependencies. It is formed by cascading two residual structures.

In the first residual structure, the input feature is denoted as $\mathbf{X}^i \in \mathbb{R}^{C \times \frac{H}{2} \times \frac{W}{2}}$, where C represents the number of feature channels passing through the MSB. Note that from the second MSB onwards, $\mathbf{X}^i$ is the channel-wise concatenation of $\mathbf{F}_{\text{S}}^i$ and $\mathbf{I}^i$. As shown in Fig. 4 (a), $\mathbf{X}^i$ is firstly normalized by LayerNorm (LN), followed by a linear layer and depthwise separable convolution (DW). Then, nonlinearity is leveraged using the SiLU activation function [24], yielding $\mathbf{Y}^i \in \mathbb{R}^{C \times \frac{H}{2} \times \frac{W}{2}}$. At the core of the MSB lies the MSMD, which performs multi-scale and multi-direction scanning on $\mathbf{Y}^i$ to produce $\widetilde{\mathbf{Y}}^i \in \mathbb{R}^{C \times \frac{H}{2} \times \frac{W}{2}}$, which will be introduced in detail later. Finally, $\widetilde{\mathbf{Y}}^i$ is processed by a LayerNorm and Linear layer, and then is added to the initial input through the skip connection to obtain $\mathbf{Z}^i$. A learnable parameter, designated as s_1, regulates the flow of information in this process. The workflow outlined above can be depicted as follows:

$$\begin{aligned} \mathbf{Y}^i &= \text{SiLU}(\text{DW}(\text{Linear}(\text{LN}(\mathbf{X}^i)))), \\ \widetilde{\mathbf{Y}}^i &= \text{MSMD}(\mathbf{Y}^i), \\ \mathbf{Z}^i &= \text{Linear}(\text{LN}(\widetilde{\mathbf{Y}}^i)) + s_1 \cdot \mathbf{X}^i. \end{aligned} \tag{7}$$

In the second residual structure, similar to [22], channel attention (CA) is introduced to reduce channel redundancy by prioritizing critical channels. Here, a learnable parameter s_2 is employed to adjust the proportion of input features contributing to the final result. This process can be

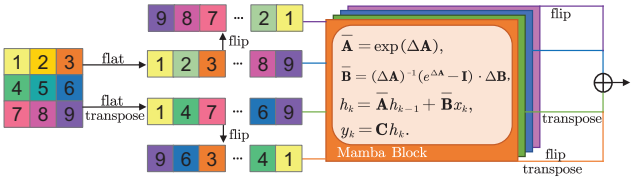


Figure 5. A specific illustration of multi-direction scanning (MDS). This corresponds to Fig. 4 (b). The feature map at each scale is divided into chunks and then flattened into sequences, where we use flip and transpose operations to create different scanning directions. Afterward, each sequence goes through a separate Mamba block to capture spatial contextual information. All sequences are eventually reshaped to the original resolution and summed to form the output feature.

described as:

$$\mathbf{X}^{i+1} = \mathrm{CA}(\mathrm{LN}(\mathbf{Z}^i)) + s_2 \cdot \mathbf{Z}^i. \tag{8}$$

Note that $\mathbf{X}^{i+1}$ can be denoted as $\mathbf{F}_{\mathrm{S}}^{i+1}$. The above steps constitute $(i+1)$-th MSB. By iterating MSBs for P periods, a series of spatially enhanced light-intensity features $\{\mathbf{F}_{\mathrm{S}}^i\}_{i=1}^P$ are obtained for progressive image reconstruction.

Multi-scale Multi-direction Scanning (MSMD). To better address the problem of local correlation attenuation, MSMD is organized in a multi-scale architecture, *i.e.*, multi-scale, multi-direction scanning is embedded to enable each pixel to fully interact with its spatially close neighboring pixels (*cf.*, Fig. 2). As shown in Fig. 4 (b), $\mathbf{Y}^i$ is first splitted along the channel dimension into L chunks $\{\mathbf{Y}_j^i\}_{j=1}^L$. Then, each chunk is downsampled to a specific resolution via average pooling, and processed by depthwise separable convolution and multi-direction scanning (MDS) to obtain L separate features $\{\widetilde{\mathbf{Y}}_j^i\}_{j=1}^L$. These features are then upsampled back to the original resolution and fused by addition and convolution, yielding $\widetilde{\mathbf{Y}}^i$. In this way, Mamba can better utilize the natural correlations between local pixels while modeling the global receptive field, improving the model's ability to accurately capture local structures.

In our implementation, each MDS follows the standard design of VMamba [32]. Specifically, as shown in Fig. 5, the 2D feature map is flattened into a 1D sequence using four distinct traversal pathways, with each sequence passing through a Mamba block. The sequences are then reshaped back to their original resolution and summed. Note that since the 1D sequences have different lengths at different scales, we additionally adopt *linear interpolation* to transform them to the original resolution. Through scanning in four interleaved directions, MDS effectively integrates spatial contextual information from neighboring pixels in various directions, contributing to extracting more reliable light-intensity features.

The aforementioned process can be articulated as:

$$\begin{aligned}
&[\mathbf{Y}_1^i, \mathbf{Y}_2^i, ..., \mathbf{Y}_L^i] = S(\mathbf{Y}^i), \\
&\widetilde{\mathbf{Y}}_j^i = \mathrm{MDS}(\mathrm{DW}(\mathcal{D}_{\frac{1}{2^{j-1}}}(\mathbf{Y}_j^i))), j = 1, 2, ..., L, \\
&\widetilde{\mathbf{Y}}^i = \mathbf{Y}^i + \mathrm{Conv}(\mathcal{U}(\widetilde{\mathbf{Y}}_1^i) ... + \mathcal{U}(\widetilde{\mathbf{Y}}_L^i)),
\end{aligned} \tag{9}$$

where S denotes channel splitting, $\mathcal{D}_{\frac{1}{2^{j-1}}}$ refers to the average pooling operation to downsample the features to the size of $\frac{1}{2^{j-1}}$, and $\mathcal{U}$ signifies linear interpolation that upsamples 1D features to the original resolution.

In summary, by employing multi-direction scanning, global interactions between any two pixels are established, aiding in capturing spatial long-range dependencies. Multi-scale scanning, on the other hand, enhances interactions between each pixel and its surrounding pixels, thereby better capturing spatial local context relationships. Compared to embedding the single-scale, multi-direction scanning in VMamba into an Unet-like architecture for multi-scale perception, our MSMD has superior reconstruction ability due to its interactive consideration of local and global correlations in SSM, as evidenced by the ablations in Table 2.

4.3. Progressive Image Reconstruction

Through each MSB period in SFEM, we obtain a light-intensity feature that is continuously optimized in the spatial domain. Consequently, we can sequentially derive P enhanced light-intensity features $\{\mathbf{F}_{\mathrm{S}}^i\}_{i=1}^P$, which can be firstly upsampled by PixelShuffle [40] and then fed into a lightweight image decoder to progressively recover the intensity images $\{\mathbf{I}^i\}_{i=1}^P$. As shown in Fig. 4, the image decoder consists of two simple 2D convolutions, each of which is accompanied by a PReLU activation, with the final layer outputting an intensity image of size $H \times W \times 1$. Note that $\mathbf{I}^P$ is the final reconstructed intensity image.

4.4. Loss Function

Similar to [55], we adopt the reconstruction loss $\mathcal{L}_{\mathrm{rec}}$ and the perceptual loss $\mathcal{L}_{\mathrm{per}}$ to train our network:

$$\mathcal{L} = \mathcal{L}_{\mathrm{rec}} + \lambda_{\mathrm{per}} \mathcal{L}_{\mathrm{per}}, \tag{10}$$

where λ_{per} is empirically set to 0.2. To supervise the intermediate images $\{\mathbf{I}^i\}_{i=1}^{P-1}$ and the final image $\mathbf{I}^P$ reconstructed by PIRM, we calculate the ℓ_1 distance between the estimate and the ground-truth:

$$\mathcal{L}_{\mathrm{rec}} = \frac{1}{HWP} \sum_{i=1}^{P} \frac{1}{2^{i-1}} \left\| \mathbf{I}^i - \mathbf{I}^{GT} \right\|_1. \tag{11}$$

We also leverage $\mathcal{L}_{\mathrm{per}}$ to mitigate blur artifacts and enhance the visual perception of the final reconstructed image $\mathbf{I}^P$:

$$\mathcal{L}_{\mathrm{per}} = \frac{1}{HW} \| \phi_{\mathrm{vgg}}(\mathbf{I}^P) - \phi_{\mathrm{vgg}}(\mathbf{I}^{GT}) \|_2, \tag{12}$$

where ϕ_{vgg} represents the feature extractor of the pre-trained VGG network.

Table 1. Quantitative comparisons against SOTA methods on the synthetic SREDS dataset [55] and real-captured dataset [65]. The numbers in **red** and blue represent the best and second-best performance. We report the number of parameters and computational complexity (FLOPs). SNN-based methods excel in computational efficiency, and CNN-based methods excel in reconstruction performance. Our approach achieves an excellent balance across these critical aspects.

Method	Backbone	Params (M)	FLOPs (T)	SREDS Dataset					Real-captured Dataset	
				PSNR↑	SSIM↑	LPIPS↓	NIQE↓	BRISQUE↓	NIQE↓	BRISQUE↓
TFP [64]	–	–	–	25.35	0.690	0.2547	5.970	43.074	9.342	45.202
TFI [64]	–	–	–	18.50	0.638	0.2590	4.518	44.933	10.09	58.309
TFSTP [63]	–	–	–	20.68	0.618	0.2761	5.348	51.697	10.92	64.566
SSIR [55]	SNN	0.382	0.238	32.61	0.919	0.0500	3.467	15.664	5.750	25.341
Spk2ImgNet [53]	CNN	3.761	9.169	36.13	0.950	0.0294	3.084	15.348	5.662	16.518
BSF [58]	CNN	2.477	6.532	37.32	0.956	0.0263	2.998	15.018	5.716	15.949
WGSE [51]	CNN	3.852	3.930	37.44	0.958	0.0241	3.032	15.555	5.620	16.154
Spk2ImgMamba (Ours)	Mamba	**0.282**	**0.099**	**37.46**	**0.958**	**0.0239**	**2.846**	**14.671**	**5.307**	**15.823**

5. Experiments

5.1. Implementation Details

Datasets. In current spike-to-image reconstruction methods, the common evaluation approach is to train on synthetic spike data and then perform generalization tests on real-world datasets. Similar to this paradigm, we adopt the recently released SREDS dataset [55], which is synthesized based on the REDS dataset [37], to train our network. It comprises 240 training scenes and 30 testing scenes, each containing 24 consecutive frames. Each frame has a spatial resolution of 1280×720 and corresponds to spike streams with a temporal length of 20. During training, each scene is cropped non-overlappingly to 96×96, yielding a total of 21,840 image patches and 10,483,200 spike planes. For generalization testing, we utilize real-captured spike data, including "recVidarReal2019" [65] and "momVidarReal2021" [61] datasets, which are also employed for model evaluation in [51, 62]. They have a spatial resolution of 400×250, and contain high-speed camera/object motion scenarios under complex indoor and outdoor conditions, which is sufficient to demonstrate the generalization performance of our approach.

Training Details. Our model is trained using the Adam optimizer [30] for 250 epochs with a batch size of 8 and 1000 iterations per epoch. The initial learning rate is 0.0005 and decays by a factor of 0.7 every 25 epochs. The temporal length of the input spike stream is $T = 61$. In TIIM, ω in Eq. (6) is set to 15 equidistant values, *i.e.*, $\{5, 9, 13, ..., 61\}$, for long-term and short-term accumulation, and C_{IFA} is set to 17, yielding feature channels of 32 for $\mathbf{F}_{\text{T}}$. With patch embedding, we increase the number of feature channels to $C = 48$ for subsequent processing. In SFEM, the number of stacked MSBs and the number of multi-scale scanning layers are both set to 3, *i.e.*, $P = L = 3$, striking a good balance of accuracy and efficiency. To augment the training data, spikes and ground-truth images are randomly flipped vertically as well as rotated 90°, 180°, or 270°. Our model is trained and tested on a single NVIDIA RTX 4090 GPU.

Evaluation Metrics. We adopt the widely-used PSNR and SSIM metrics and learned perceptual metric LPIPS [52] to evaluate the reconstruction quality. Besides, we also employ two non-reference image quality assessment metrics NIQE [36] and BRISQUE [35]. A higher PSNR/SSIM (↑) or lower LPIPS/NIQE/BRISQUE (↓) score indicates better performance.

5.2. Comparison with SOTA Methods

We perform comparisons against the following baselines. 1) **Non-learning-based methods**: TFP [64], TFI [64], and TFSTP [63]. 2) **Learning-based methods**: SSIR [55], designed for energy-efficient spike-to-image reconstruction with the help of spiking neural networks. Spk2ImgNet [53], BSF [58], and WGSE [51], which are popular spike-to-image reconstruction methods based on the CNN backbone. Although the recently proposed SwinSF [29] is a Transformer-based method, it requires the testing and training datasets to have the same spatial resolution, making it unsuitable for comparison with our approach.

Synthetic Data. Table 1 reports the quantitative results. Limited by modeling capabilities, non-learning-based methods struggle to achieve satisfactory accuracy. Although computationally efficient, SNN-based methods have difficulty in restoring high-quality images. While CNN-based methods show promising performance, they suffer from high computational costs. In contrast, thanks to Mamba's powerful modeling capability, our method achieves comparable accuracy using only 10% of the parameters of CNN-based methods, and FLOPs are reduced by nearly two orders of magnitude, which facilitates practical cost-effective applications.

Real-captured Data. Fig. 6 illustrates the qualitative results. Our method can reconstruct high-quality images in high-speed scenarios with fine textures and sharp edges, *e.g.*, a fast-moving car (100 km/h) and a high-speed rotating

For more visual comparisons on other datasets, please refer to the supplementary material.

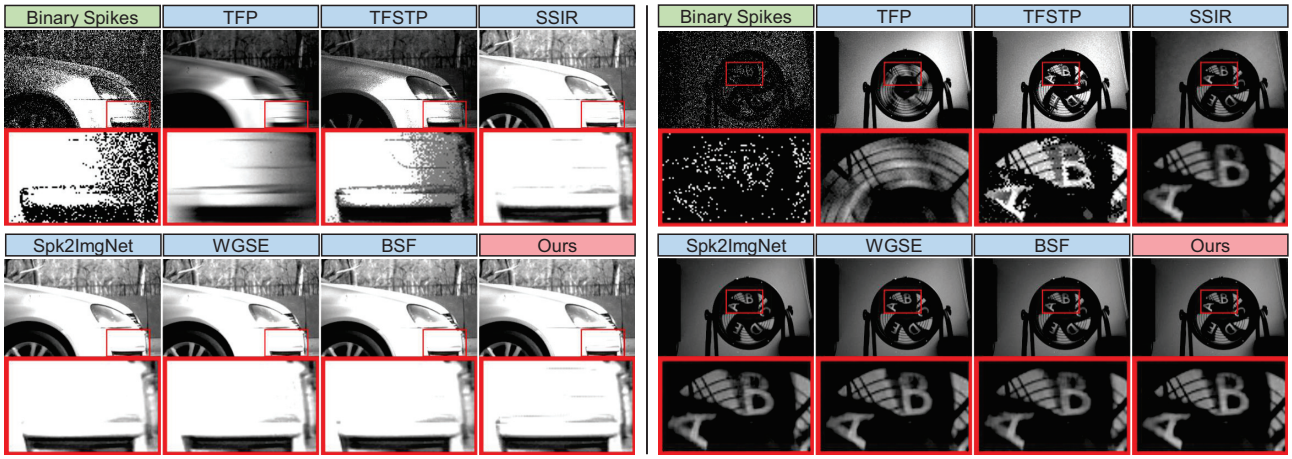


Figure 6. Qualitative comparison on the real-captured "recVidarReal2019" [65] dataset. The image reconstructed by our approach exhibits clearer boundaries and richer details. Please zoom in for more details.

Table 2. Ablation studies on the network architecture on the SREDS dataset [55].

Setting	PSNR↑ (dB)	SSIM↑	Params (M)
Model Design			
Replace TIIM with Conv.	36.83	0.956	0.296
Single-scale + UNet	36.55	0.953	0.696
Our full model	**37.46**	**0.958**	0.282
Numbers of Stacked MSBs			
$P = 1$	36.93	0.956	0.159
$P = 2$	37.18	0.957	0.221
$P = 3$ (full model)	**37.46**	**0.958**	0.282
Layers for Multi-scale Scanning			
$L = 1$	36.54	0.953	0.270
$L = 2$	37.02	0.956	0.276
$L = 3$ (full model)	**37.46**	**0.958**	0.282

fan (2600 rpm). Thus, our method has excellent generalization ability in real-world scenarios, extending the spiking camera's capability for high-speed imaging applications, especially in resource-constrained devices.

5.3. Ablation Studies

We analyze the role of each component in the proposed network architecture. The ablation results are presented in Table 2. It can be seen that replacing TIIM with a 2D convolution leads to sub-optimal results, demonstrating the importance of exploring time-series cues in spike streams. In addition, we embed the single-scale scanning in VMamba [32] into an Unet-like architecture for multi-scale modeling, *i.e.*, "Single-scale + UNet". Our multi-scale multi-direction scanning exhibits more accurate reconstruction performance due to its collaborative perception of local and global correlations.

We also investigate the efficacy of the number of stacked MSBs and the number of multi-scale scanning layers. Integrating more MSBs in the SFEM is conducive to purifying more representative light-intensity features from binary spike streams, thereby facilitating the restoration of intermediate images. Using more scanning scales for SSM enables the model to better capture local details, which is crucial for reconstructing high-quality images. Empirically, using three scales is sufficient to strike a good balance between performance and efficiency.

6. Conclusion

The spiking camera is a novel neuromorphic sensor that outputs spike streams to record high-speed dynamic scenes with extremely high temporal resolution. In this paper, we propose a Mamba-based spike-to-image reconstruction method called Spk2ImgMamba, which leverages the merits of global receptive fields and linear complexity. To address the challenge of local correlation attenuation, we introduce a multi-scale scanning mechanism, enhancing the model's intrinsic ability to capture local structures useful for image reconstruction. Experiments on various datasets demonstrate its promising performance in reconstructing dynamic scenes and highlight its significant advantages in terms of model size and computational complexity. In the future, we will extend our model to generate a coherent video sequence from spike streams to explore correlations in continuous space-time.

Acknowledgments. This work was supported by National Science and Technology Major Project (2021ZD0109803), Beijing Natural Science Foundation (L233024), Beijing Municipal Science & Technology Commission, Administrative Commission of Zhongguancun Science Park (Z241100003524012), and National Natural Science Foundation of China (62088102, 62136001, 62276007, 62401021). Bin Fan was also supported by China National Postdoctoral Program for Innovative Talents (BX20230013) and China Postdoctoral Science Foundation (2024M750101).

References

[1] Christian Brandli, Raphael Berner, Minhao Yang, Shih-Chii Liu, and Tobi Delbruck. A 240× 180 130 db 3 μs latency global shutter spatiotemporal vision sensor. *IEEE Journal of Solid-State Circuits (JSSC)*, 49(10):2333–2341, 2014. 1

[2] Yakun Chang, Yeliduosi Xiaokaiti, Yujia Liu, Bin Fan, Zhaojun Huang, Tiejun Huang, and Boxin Shi. Towards HDR and HFR video from rolling-mixed-bit spikings. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 25117–25127, 2024. 1, 3

[3] Keyan Chen, Bowen Chen, Chenyang Liu, Wenyuan Li, Zhengxia Zou, and Zhenwei Shi. RSMamba: Remote sensing image classification with state space model. *IEEE Geoscience and Remote Sensing Letters*, 21:1–5, 2024. 2

[4] Shiyan Chen, Chaoteng Duan, Zhaofei Yu, Ruiqin Xiong, and Tiejun Huang. Self-supervised mutual learning for dynamic scene reconstruction of spiking camera. In *Proceedings of the International Joint Conferences on Artificial Intelligence (IJCAI)*, page 2859–2866, 2022. 2, 3

[5] Shiyan Chen, Zhaofei Yu, and Tiejun Huang. Self-supervised joint dynamic scene reconstruction and optical flow estimation for spiking camera. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 350–358, 2023. 3

[6] Siwei Dong, Tiejun Huang, and Yonghong Tian. Spike camera and its coding methods. In *Proceedings of the Data Compression Conference (DCC)*, pages 437–437, 2017. 1

[7] Yanchen Dong, Ruiqin Xiong, Jing Zhao, Jian Zhang, Xiaopeng Fan, Shuyuan Zhu, and Tiejun Huang. Joint demosaicing and denoising for spike camera. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 1582–1590, 2024. 1

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16×16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[9] Bin Fan and Yuchao Dai. Inverting a rolling shutter camera: Bring rolling shutter images to high framerate global shutter video. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 4228–4237, 2021. 3

[10] Bin Fan, Yuchao Dai, and Mingyi He. SUNet: Symmetric undistortion network for rolling shutter correction. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 4541–4550, 2021. 3

[11] Bin Fan, Yuchao Dai, and Mingyi He. Rolling shutter camera: Modeling, optimization and learning. *Machine Intelligence Research*, 20(6):783–798, 2023. 3

[12] Bin Fan, Yuchao Dai, and Hongdong Li. Rolling shutter inversion: Bring rolling shutter images to high framerate global shutter video. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 45(5):6214–6230, 2022. 3

[13] Bin Fan, Yuchao Dai, and Hongdong Li. Learning bilateral cost volume for rolling shutter temporal super-resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 46(5):3862–3879, 2024. 3, 5

[14] Bin Fan, Yuchao Dai, Zhiyuan Zhang, Qi Liu, and Mingyi He. Context-aware video reconstruction for rolling shutter cameras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17572–17582, 2022. 3

[15] Bin Fan, Ying Guo, Yuchao Dai, Chao Xu, and Boxin Shi. Self-supervised learning for rolling shutter temporal super-resolution. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 3

[16] Bin Fan, Yuxin Mao, Yuchao Dai, Zhexiong Wan, and Qi Liu. Joint appearance and motion learning for efficient rolling shutter correction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5671–5681, 2023. 3

[17] Bin Fan, Jiaoyang Yin, Yuchao Dai, Chao Xu, Tiejun Huang, and Boxin Shi. Spatio-temporal interactive learning for efficient image reconstruction of spiking cameras. *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 5

[18] Daniel Y Fu, Tri Dao, Khaled K Saab, Armin W Thomas, Atri Rudra, and Christopher Ré. Hungry hungry hippos: Towards language modeling with state space models. *arXiv preprint arXiv:2212.14052*, 2022. 2

[19] Hu Gao, Bowen Ma, Ying Zhang, Jingfan Yang, Jing Yang, and Depeng Dang. Learning enriched features via selective state spaces model for efficient image deblurring. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, 2024. 2

[20] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023. 2, 3, 4

[21] Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*, 2021. 2, 3

[22] Hang Guo, Jinmin Li, Tao Dai, Zhihao Ouyang, Xudong Ren, and Shu-Tao Xia. MambaIR: A simple baseline for image restoration with state-space model. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2024. 2, 5

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 5

[24] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 5

[25] Xiaotao Hu, Zhewei Huang, Ailin Huang, Jun Xu, and Shuchang Zhou. A dynamic multi-scale voxel flow network for video prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6121–6131, 2023. 5

[26] Ju Huang, Shiao Wang, Shuai Wang, Zhe Wu, Xiao Wang, and Bo Jiang. Mamba-FETrack: Frame-event tracking via state space model. *arXiv preprint arXiv:2404.18174*, 2024. 2

[27] Tao Huang, Xiaohuan Pei, Shan You, Fei Wang, Chen Qian, and Chang Xu. LocalMamba: Visual state space model with

windowed selective scan. *arXiv preprint arXiv:2403.09338*, 2024. 2

[28] 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. 1000× faster camera and machine vision with ordinary devices. *Engineering*, 25:110–119, 2022. 1, 3

[29] Liangyan Jiang, Chuang Zhu, and Yanxu Chen. SwinSF: Image reconstruction from spatial-temporal spike streams. *arXiv preprint arXiv:2407.15708*, 2024. 2, 7

[30] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015. 7

[31] Patrick Lichtsteiner, Christoph Posch, and Tobi Delbruck. A 128×128 120 db 15μs latency asynchronous temporal contrast vision sensor. *IEEE Journal of Solid-State Circuits (JSSC)*, 43(2):566–576, 2008. 1

[32] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, and Yunfan Liu. VMamba: Visual state space model. *arXiv preprint arXiv:2401.10166*, 2024. 2, 3, 5, 6, 8

[33] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 10012–10022, 2021. 2

[34] Hanyue Lou, Jinxiu Liang, Minggui Teng, Bin Fan, Yong Xu, and Boxin Shi. Zero-shot event-intensity asymmetric stereo via visual prompting from image domain. *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 1

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

[36] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters (SPL)*, 20(3):209–212, 2012. 7

[37] Seungjun Nah, Sungyong Baik, Seokil Hong, Gyeongsik Moon, Sanghyun Son, Radu Timofte, and Kyoung Mu Lee. NTIRE 2019 challenge on video deblurring and super-resolution: Dataset and study. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1996–2005, 2019. 7

[38] Xiaohuan Pei, Tao Huang, and Chang Xu. EfficientVMamba: Atrous selective scan for light weight visual mamba. *arXiv preprint arXiv:2403.09977*, 2024. 3

[39] Christoph Posch, Daniel Matolin, and Rainer Wohlgenannt. A QVGA 143 db dynamic range frame-free PWM image sensor with lossless pixel-level video compression and time-domain CDS. *IEEE Journal of Solid-State Circuits (JSSC)*, 46(1):259–275, 2010. 1

[40] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1874–1883, 2016. 6

[41] Yuan Shi, Bin Xia, Xiaoyu Jin, Xing Wang, Tianyu Zhao, Xin Xia, Xuefeng Xiao, and Wenming Yang. VmambaIR: Visual state space model for image restoration. *arXiv preprint arXiv:2403.11423*, 2024. 2, 3

[42] Jimmy TH Smith, Andrew Warrington, and Scott W Linderman. Simplified state space layers for sequence modeling. *arXiv preprint arXiv:2208.04933*, 2022. 2, 3

[43] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8934–8943, 2018. 5

[44] Zachary Teed and Jia Deng. RAFT: Recurrent all-pairs field transforms for optical flow. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 402–419. Springer, 2020. 5

[45] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvtv2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 8(3):1–10, 2022. 5

[46] Lujie Xia, Ziluo Ding, Rui Zhao, Jiyuan Zhang, Lei Ma, Zhaofei Yu, Tiejun Huang, and Ruiqin Xiong. Unsupervised optical flow estimation with dynamic timing representation for spike camera. *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 36, 2023. 1

[47] Lujie Xia, Jing Zhao, Ruiqin Xiong, and Tiejun Huang. SVFI: Spiking-based video frame interpolation for high-speed motion. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 2910–2918, 2023. 5

[48] Changcheng Xiao, Qiong Cao, Zhigang Luo, and Long Lan. MambaTrack: A simple baseline for multiple object tracking with state space model. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, 2024. 2

[49] Chenhongyi Yang, Zehui Chen, Miguel Espinosa, Linus Ericsson, Zhenyu Wang, Jiaming Liu, and Elliot J Crowley. PlainMamba: Improving non-hierarchical mamba in visual recognition. *arXiv preprint arXiv:2403.17695*, 2024. 3

[50] Siqi Yang, Zhaojun Huang, Yakun Chang, Bin Fan, Zhaofei Yu, and Boxin Shi. Real-data-driven 2000 FPS color video from mosaicked chromatic spikes. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2024. 1

[51] Jiyuan Zhang, Shanshan Jia, Zhaofei Yu, and Tiejun Huang. Learning temporal-ordered representation for spike streams based on discrete wavelet transforms. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 137–147, 2023. 1, 2, 3, 7

[52] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 586–595, 2018. 7

[53] Jing Zhao, Ruiqin Xiong, Hangfan Liu, Jian Zhang, and Tiejun Huang. Spk2ImgNet: Learning to reconstruct dynamic scene from continuous spike stream. In *Proceedings*

of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 11996–12005, 2021. 1, 2, 3, 7

[54] Jing Zhao, Ruiqin Xiong, Jiyu Xie, Boxin Shi, Zhaofei Yu, Wen Gao, and Tiejun Huang. Reconstructing clear image for high-speed motion scene with a retina-inspired spike camera. *IEEE Transactions on Computational Imaging (TCI)*, 8:12–27, 2021. 3

[55] Rui Zhao, Ruiqin Xiong, Jian Zhang, Zhaofei Yu, Shuyuan Zhu, Lei Ma, and Tiejun Huang. Spike camera image reconstruction using deep spiking neural networks. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 34(6):5207–5212, 2024. 1, 3, 6, 7, 8

[56] 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 *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 7496–7504, 2024. 1

[57] Rui Zhao, Ruiqin Xiong, Jing Zhao, Zhaofei Yu, Xiaopeng Fan, and Tiejun Huang. Learning optical flow from continuous spike streams. *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 35:7905–7920, 2022. 2

[58] Rui Zhao, Ruiqin Xiong, Jing Zhao, Jian Zhang, Xiaopeng Fan, Zhaofei Yu, and Tiejun Huang. Boosting spike camera image reconstruction from a perspective of dealing with spike fluctuations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24955–24965, 2024. 1, 2, 5, 7

[59] Zou Zhen, Yu Hu, and Zhao Feng. FreqMamba: Viewing mamba from a frequency perspective for image deraining. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, 2024. 2

[60] Yajing Zheng, Zhaofei Yu, Song Wang, and Tiejun Huang. Spike-based motion estimation for object tracking through bio-inspired unsupervised learning. *IEEE Transactions on Image Processing (TIP)*, 32:335–349, 2022. 1

[61] Yajing Zheng, Jiyuan Zhang, Rui Zhao, Jianhao Ding, Shiyan Chen, Ruiqin Xiong, Zhaofei Yu, and Tiejun Huang. SpikeCV: Open a continuous computer vision era. *arXiv preprint arXiv:2303.11684*, 2023. 3, 7

[62] 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 Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 45(7):8127–8142, 2023. 2, 3, 7

[63] Yajing Zheng, Lingxiao Zheng, Zhaofei Yu, Boxin Shi, Yonghong Tian, and Tiejun Huang. High-speed image reconstruction through short-term plasticity for spiking cameras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6358–6367, 2021. 2, 3, 7

[64] Lin Zhu, Siwei Dong, Tiejun Huang, and Yonghong Tian. A retina-inspired sampling method for visual texture reconstruction. In *Proceedings of the IEEE International Conference on Multimedia & Expo (ICME)*, pages 1432–1437, 2019. 2, 3, 5, 7

[65] Lin Zhu, Siwei Dong, Jianing Li, Tiejun Huang, and Yonghong Tian. Retina-like visual image reconstruction via spiking neural model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1438–1446, 2020. 2, 3, 7, 8

[66] Lin Zhu, Kangmin Jia, Yifan Zhao, Yunshan Qi, Lizhi Wang, and Hua Huang. SpikeNeRF: Learning neural radiance fields from continuous spike stream. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6285–6295, 2024. 1

[67] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision Mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*, 2024. 2, 3