EventMamba: Enhancing Spatio-Temporal Locality with State Space Models for Event-Based Video Reconstruction

Chengjie Ge, Xueyang Fu, Peng He, Kunyu Wang, Chengzhi Cao, Zheng-Jun Zha*

University of Science and Technology of China, China
cjge@mail.ustc.edu.cn, xyfu@ustc.edu.cn, {hp0618,kunyuwang,chengzhicao}@mail.ustc.edu.cn, zhazj@ustc.edu.cn

Abstract

Leveraging its robust linear global modeling capability, Mamba has notably excelled in computer vision. Despite its success, existing Mamba-based vision models have overlooked the nuances of event-driven tasks, especially in video reconstruction. Event-based video reconstruction (EBVR) demands spatial translation invariance and close attention to local event relationships in the spatio-temporal domain. Unfortunately, conventional Mamba algorithms apply static window partitions and standard reshape scanning methods, leading to significant losses in local connectivity. To overcome these limitations, we introduce EventMamba—a specialized model designed for EBVR tasks. EventMamba innovates by incorporating random window offset (RWO) in the spatial domain, moving away from the restrictive fixed partitioning. Additionally, it features a new consistent traversal serialization approach in the spatio-temporal domain, which maintains the proximity of adjacent events both spatially and temporally. These enhancements enable EventMamba to retain Mamba's robust modeling capabilities while significantly preserving the spatio-temporal locality of event data. Comprehensive testing on multiple datasets shows that EventMamba markedly enhances video reconstruction, drastically improving computation speed while delivering superior visual quality compared to Transformer-based methods.

Introduction

Event cameras, also known as neuromorphic cameras, draw inspiration from biological systems and offer substantial advancements over traditional visual sensors. They provide exceptional temporal resolution (1 μs), superior dynamic range (140 dB), and ultra-low power consumption (5 mW) (Gallego et al. 2020; Delbrück et al. 2010; Benosman et al. 2013; Fu et al. 2024). Unlike conventional cameras, event cameras capture data asynchronously and sparsely, which complicates direct interpretation and integration with standard computer vision techniques. To address this, converting event data into more conventional intensity images is essential for bridging the technological divide in computer vision applications.

Over the past decades, deep learning has achieved remarkable progress in computer vision (Ge, Fu, and Zha 2022; Ge et al. 2024; Zhang, Yang, and Wang 2023; Zhang et al. 2024; Zhang, Yang, and Hu 2023; Li et al. 2023; Shi et al. 2022; Peng et al. 2024; Wang et al. 2024), especially in event-based video reconstruction (EBVR) (Zhu et al. 2022; Rebecq et al. 2019; Cadena et al. 2023; Scheerlinck et al. 2020; Gallego et al. 2020). Despite these strides, EBVR algorithms still have room for improvement. Current methods typically utilize Convolutional Neural Networks (CNNs) or Transformers to reconstruct frames from event data. While CNNs focus on local details, often at the expense of global context, this can lead to increased susceptibility to noise and blurring, resulting in unclear visual outputs (Jang, McCormack, and Tong 2021). Conversely, Transformers excel in capturing extensive non-local information through their self-attention mechanisms (Weng, Zhang, and Xiong 2021; Xu et al. 2024). However, this approach scales quadratically with the input size, noted as $O(n^2)$. This scaling is problematic for high-resolution data, such as the 1280×720 output from Prophesee EVK4 cameras, which demands significant computational resources and challenges deployment on resource-limited devices.

Recently, the Mamba module, particularly within the State Space Model (SSM) framework, has introduced a groundbreaking approach to address previous challenges (Gu, Goel, and Ré 2021). As a subset of SSMs, the advanced Mamba modules (Gu and Dao 2023) have shown substantial advancements by employing a sophisticated selection mechanism and hardware optimizations (Zhu et al. 2024; Liu et al. 2024b; Xing et al. 2024; Liu et al. 2024a). However, these vision Mamba modules are not directly suitable for the EBVR task due to two main reasons. Firstly, EBVR tasks require translation invariance in the spatial domain, necessitating a location-independent approach to map events to video frames. Traditional vision Mamba networks, such as VisionMamba and VmambaIR, use fixed non-overlapping windows that constrain SSM operations to these local windows, thereby imposing spatial priors inappropriate for EBVR tasks, leading to a loss of translation invariance and incomplete local relationship capture. Secondly, event data in the spatio-temporal domain is critical, and the common practice of flattening temporal features into a one-dimensional sequence processed recursively disrupts the natural spatio-temporal event relationships, resulting in a loss of local information.

*Corresponding author.

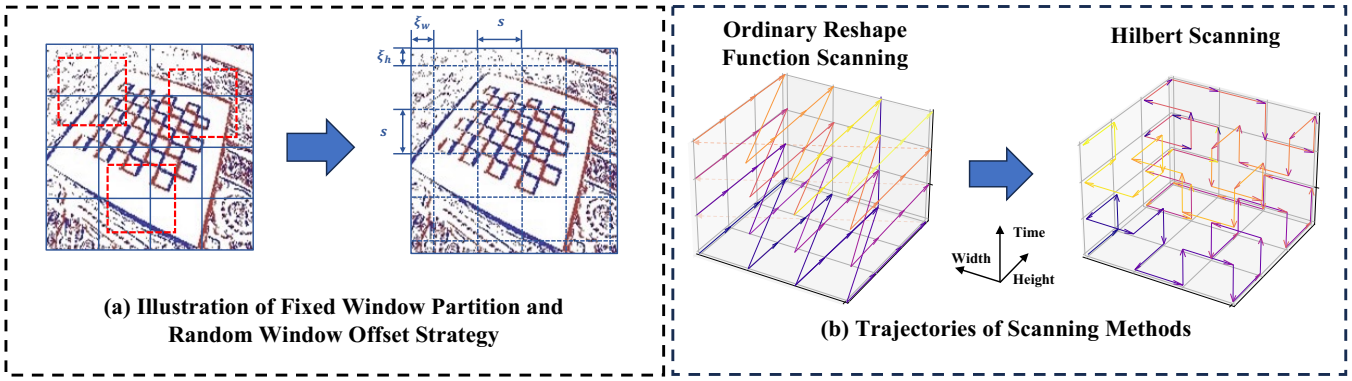

Figure 1: (a) Example illustrating the loss of locality in the fixed window strategy (spatial locality loss in the red box), and our proposed Random Window Offset solution. (b) Demonstration of the loss of spatio-temporal locality in conventional space-filling curves contrasted with our introduced Hilbert space-filling curve technique.

To overcome these challenges, we introduce EventMamba, a specialized SSM network designed for high-speed EBVR tasks. In the spatial domain, EventMamba employs a novel random window offset (RWO) strategy, which uses randomly offset windows to encompass the entire feature map, rather than restricting it to fixed partitions (see Figure 1(a)). This RWO strategy ensures preservation of translation invariance and more comprehensive local relationship mapping. In the spatio-temporal domain, EventMamba employs a unique Hilbert Space Filling Curve (HSFC) scanning mechanism. Compared to other space-filling curves, the Hilbert curve exhibits superior locality preserving properties and a lower space-to-linear ratio (Chen et al. 2022; Bauman 2006; Wu et al. 2024). This implies that the Hilbert curve is more effective at retaining the local characteristics of the original data and offers higher efficiency when mapping multi-dimensional space to a one-dimensional linear sequence. EventMamba leverages these inherent advantages of Hilbert and trans-Hilbert curves to convert spatio-temporal pixels into a one-dimensional sequence along the curve trajectory, enabling a fine-grained and locality-preserving recovery of event data (see Figure 1(b)). Through these innovations, EventMamba maintains the spatial and temporal locality of event data while leveraging the powerful linear global modeling capabilities of Mamba.

In summary, our contributions are as follows:

- We conduct a critical analysis of the limitations of existing Transformer and CNNs-based methods and introduce EventMamba, a pioneering model that integrates State Space Models for event-based video reconstruction.
- We elaborately design a random window offset strategy for reconstruction tasks to compensate for the loss of translation invariance caused by previous vision Mamba models when using fixed partitioned windows, thereby better modeling local information in the spatial domain.
- We design a Hilbert Space Filling Curve mechanism tailored for EBVR tasks to address the disruption of spatio-temporal locality in previous vision Mamba models, significantly enhancing the model's ability to capture spatio-temporal relationships.

Our comprehensive experimental results show that EventMamba markedly outperforms existing models, enhancing both subjective and objective performance metrics. On the IJRR dataset (Mueggler et al. 2017), EventMamba increases the SSIM value by 2.9% compared to previous state-of-the-art approaches.

Related Works

Event-based Video Reconstruction

Early studies on EVBR tasks were primarily based on physical priors of the event stream, which were significantly limited by specific conditions of the photographic scenes (Kim et al. 2008; Lagorce et al. 2016; Munda, Reinbacher, and Pock 2018; Chen et al. 2020; Gehrig et al. 2021; Schaefer, Gehrig, and Scaramuzza 2022; Shiba, Aoki, and Gallego 2022; Tulyakov et al. 2022; Freeman, Singh, and Mayer-Patel 2023). Kim *et al.* developed a method based on the Kalman filter to reconstruct gradient video frames from a rotating event camera, and used Poisson integration techniques to recover luminance frames with temporal dimensions (Kim et al. 2008). Bardow *et al.* proposed a variational energy minimization framework that allows simultaneous recovery of video frames and dense optical flow from the sliding window of an event camera (Barua, Miyatani, and Veeraraghavan 2016). Zhang *et al.* formulated the event-based video frame reconstruction task as an optical flow-based linear inverse problem, demonstrating that this approach could generate luminance video frames of quality comparable to those trained with deep neural networks, without the need for training deep networks (Zhang, Yezzi, and Gallego 2021).

In recent years, with the development of deep learning, data-driven neural network algorithms have made significant breakthroughs in the field of EBVR. Rebecq *et al.* developed E2VID, a model for event-based video frame reconstruction that combines the advantages of CNNs and Recurrent Neural Network (RNNs), improving the quality of video frame reconstruction through controlled incremental updates of event sequences (Rebecq et al. 2019). To address the vanishing gradient problem in long sequence

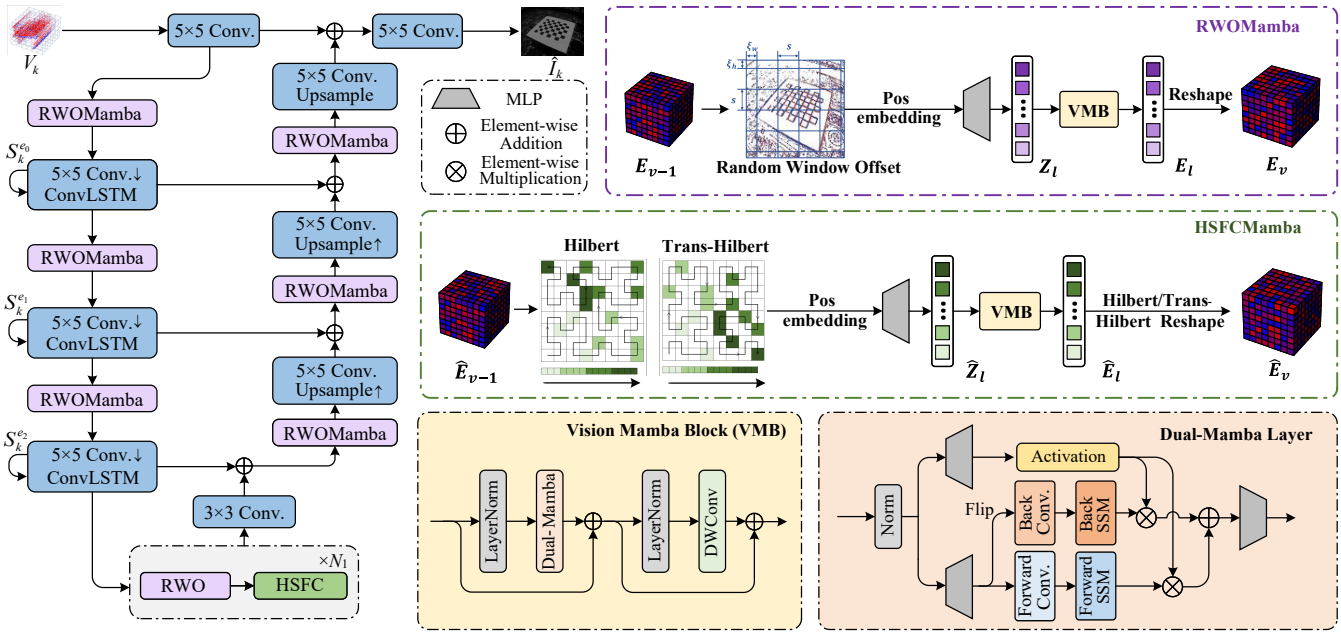

Figure 2: The EventMamba architecture is U-Net-like, processing event voxels (V_k) to predict intensity images. It incorporates two key components: RWOMamba and HSFCMamba, which are designed to maintain the translation invariance and spatio-temporal locality of event features, respectively. The number of N_1 is set to 2 in our EventMamba architecture.

data processing, Rebecq *et al.* later introduced E2VID+, which incorporates multi-layer ConvLSTM units to stabilize gradients during backpropagation (Stoffregen et al. 2020). Cadena *et al.* used spatially-adaptive denormalization (SPADE) layers in the E2VID framework and proposed that the SPADE module enhanced the quality of reconstructed frames in the video (Scheerlinck et al. 2020). Wen *et al.* were among the first to apply the Transformer architecture to the field of event-based video reconstruction, combining the local feature extraction capabilities of CNNs with the global information processing advantages of Transformer to further enhance the quality of event-based video reconstruction (Weng, Zhang, and Xiong 2021). Zhu *et al.* proposed a novel EBVR network based on spiking neural networks, their meet the comparable performance with ANN methods while saves the energy consumption (Zhu et al. 2022). Cadena *et al.* considered the sparsity of the event stream in event-based video reconstruction and were the first to employ stacked sparse convolutional modules in the reconstruction network, effectively reducing the network's complexity (Cadena et al. 2023).

State Space Model

State Space Models (SSMs), first introduced in the S4 model (Gu, Goel, and Ré 2021), model global information more efficiently than CNNs or Transformers. S5 (Smith, Warrington, and Linderman 2022) reduced the complexity to linear levels using MIMO SSMs and parallel scanning. H3 (Mehta et al. 2022) added gating units to enhance expressiveness, enabling it to compete with Transformers in language modeling. Mamba (Gu and Dao 2023) introduced an input-adaptive mechanism, outperforming similarly-scaled Transformers in inference speed, throughput, and overall performance.

Motivated by the success of Mamba in language modeling, various Mamba-based models have been proposed for vision tasks (Zhu et al. 2024; Liu et al. 2024b; Xing et al. 2024; Shi et al. 2024; Li et al. 2024; Wu et al. 2024). However, these models often directly apply fixed partitioned windows to vision tasks without fully considering the unique characteristics of event-based tasks. EBVR tasks require higher translation invariance and spatio-temporal locality compared to classification and segmentation tasks, as previously discussed. This is because EBVR tasks necessitate feature integration across both temporal and spatial dimensions for pixel-level regression. To effectively address EBVR tasks, it is crucial to develop a mechanism that overcomes the limitations of current vision Mamba models in terms of translation invariance and spatio-temporal locality.

Methodology

In this section, we introduce the fundamental concepts of SSMs. We then provide a detailed explanation of how we integrate SSMs with the EBVR task. This includes a description of the model framework, modular design, training strategies, and the loss functions employed.

Preliminaries

State Space Sequence Models and the Mamba framework are based on linear dynamics principles, transforming a one-dimensional sequence $x(t) \in \mathbf{R}$ through a hidden state

space $h(t) \in \mathbf{R}^N$ to produce an output $y(t)$. The transformation involves matrices $\mathbf{A} \in \mathbf{R}^{N\times N}$ (state evolution), $\mathbf{B} \in \mathbf{R}^{N\times 1}$ (input-to-state), and $\mathbf{C} \in \mathbf{R}^{1\times N}$ (state-to-output). The system dynamics are described by:

$$h'(t) = \mathbf{A}h(t) + \mathbf{B}x(t), \tag{1}$$

$$y(t) = \mathbf{C}h'(t). \tag{2}$$

The S4 and Mamba modules adapt these models to discrete time, using a time scale Δ to convert $\mathbf{A}$ and $\mathbf{B}$ to their discrete counterparts $\bar{\mathbf{A}}$ and $\bar{\mathbf{B}}$, employing the Zero-Order Hold (ZOH) method (Karafyllis and Krstic 2011) :

$$\bar{\mathbf{A}} = \exp(\mathbf{\Delta A}), \tag{3}$$

$$\bar{\mathbf{B}} = (\mathbf{\Delta A})^{-1}(\exp(\mathbf{\Delta A} - \mathbf{I})) \cdot \mathbf{\Delta B}. \tag{4}$$

Finally, the output is obtained through global convolution:

$$h_t = \bar{\mathbf{A}}h_{t-1} + \bar{\mathbf{B}}x_t, \tag{5}$$

$$y_t = \mathbf{C}h_t. \tag{6}$$

The output is generated by a structured convolution:

$$\bar{\mathbf{K}} = (\mathbf{C}\bar{\mathbf{B}}, \mathbf{C}\bar{\mathbf{A}}\bar{\mathbf{B}}, \ldots, \mathbf{C}\bar{\mathbf{A}}^{\mathbf{M-1}}\bar{\mathbf{B}}), \tag{7}$$

$$\mathbf{y} = x * \bar{\mathbf{K}}, \tag{8}$$

wherein $\mathbf{M}$ represents the length of the sequence x, and $\bar{\mathbf{K}} \in \mathbf{R}^M$ denotes a structured convolutional kernel.

Event Representation

We consider an event stream e_i containing N_E events over a duration of T seconds, where each event $e_i = (x_i, y_i, t_i, p_i)$ encodes the position (x_i, y_i), timestamp t_i, and polarity p_i of the i-th brightness change detected by the sensor. The goal is to generate a stream of video frames $\hat{I}_k$ from the same T-second interval, where each video frame $\hat{I}_k \in [0,1]^{W\times H}$ represents a 2D grayscale representation of the scene's absolute brightness. H and W represent the height and width. The proposed method limits each generated video frame to rely solely on past events.

We sort events into groups corresponding to the timestamps of the video frames. Given N_I frames, each identified by a timestamp s_k, we define the k-th event group as:

$$G_k \doteq \{e_i \mid s_{k-1} \le t_i < s_k\}, \quad \text{for } k = 1, \ldots, N_I. \tag{9}$$

To input these organized events into CNNs, the events are amassed into a voxel grid $V_k \in \mathbf{R}^{W\times H\times B}$, where W and H correspond to the grid's dimensions, and B is the number of distinct bins. The timestamp t_i of each event is normalized to the range $[0, B-1]$, yielding the normalized timestamp t_i^*:

$$t_i^* = \frac{(B-1)(t_i - T_k)}{\Delta T}. \tag{10}$$

We then use an interpolation method to distribute the polarity contribution of each event across the nearest two voxels in the temporal dimension:

$$V_k(x,y,t) = \sum p_i \max(0, 1 - |t - t_i^*|)\delta(x - x_i, y - y_i). \tag{11}$$

In the experiments, $B = 5$ is selected as the count of discrete intervals for the time dimension, resulting in each event cluster being represented by a voxel grid with dimensions $W \times H \times 5$, which can then be fed into DNNs for further processing and analysis.

Network Structures

Random Window Offset Mamba. In previous iterations of the vision Mamba model, image features were divided into non-overlapping windows and processed through the Mamba module for experimental purposes. However, for reconstruction tasks, all window partitions contain equally important information. Therefore, the use of fixed window partitions results in a loss of translation invariance. To address this issue, we propose the Random Window Offset (RWO) strategy to endow the EventMamba model with translation invariance and to fully utilize the local relationships in the spatial domain as shown in Figure 2. The vision mamba block (VMB) is a residual structure that combines a dual-mamba layer, inspired by the VisionMamba (Zhu et al. 2024), with a depth-wise convolution (DWConv) (Chollet 2017) layer. The computation is defined as follows:

$$\begin{aligned} &z_l = \text{MLP}(\text{RWO}(E_{v-1}; s, \xi_h^l, \xi_w^l)),\ (\xi_h^l, \xi_w^l) \sim \mathbb{U}(\mathfrak{R}_s), \\ &E_l = \text{VMB}(z_l), \end{aligned} \tag{12}$$

where E_{v-1} represents the input event feature from the previous stage, s is the spatial size of local window, z_l denotes the output sequence after the MLP layer, and E_l denotes the output sequence after the VMB layer. $\mathfrak{R}_s$ includes all possible offsets within the uniform distribution $\mathbb{U}(\mathfrak{R}_s)$. The representation of $\mathfrak{R}_s$ can be simplified as:

$$\mathfrak{R}_s := [0, \ldots, s-1] \times [0, \ldots, s-1]. \tag{13}$$

In the training process, the symbol $\times$ denotes the Cartesian product. The parameters (ξ_h^l, ξ_w^l) are considered as independently and identically distributed random variables sampled from the uniform distribution $\mathbb{U}(\mathfrak{R}_s)$. Assuming the total number of Mamba layers is N, the random displacements $\{(\xi_h^l, \xi_w^l)\}_{l=0}^{N-1}$ are intentionally designed to be independent to ensure the maintenance of faithful locality and translation invariance at the layer level. Through this approach, although each individual layer considers all possible displacements, only one set of sampled displacements $\{(\xi_h^l, \xi_w^l)\}_{l=0}^{N-1}$ is required for each forward propagation, thus the training time remains consistent with that of fixed window partitioning. Following the RWO strategy, the event features are input into the VMB module and reshape to its original shape E_v.

Regarding the testing process, the conventional strategy is to use layer-wise expectation inspired by Dropout (Srivastava et al. 2014) to approximate the overall output of the model, which is expressed as:

$$EM^{test}(x) = \mathop{\mathbb{E}}_{s}[EM(x; s, \xi_h^l, \xi_w^l)], \tag{14}$$

where EM stands for our EventMamba network for simplicity. However, for our task, using layer-wise expectation requires traversing all combinations of random windows, this computation method will greatly increase the computational burden, which contradicts our original intention of introducing the Mamba model. Therefore, we use another Monte Carlo approximation to estimate the output of the model,

which is expressed as:

$$EM^{test}(x) \approx \frac{1}{M}\sum_{i=1}^{M}[EM(x; s_i, \xi_h^l, \xi_w^l)]. \quad (15)$$

It may seem that the testing time would scale linearly with M, the number of averaged forward passes. However, modern accelerators can perform multiple forward passes concurrently, significantly reducing the testing time. This acceleration is achieved by transferring an input to the GPU(s) and creating a mini-batch that consists of the same input repeated multiple times. EventMamba performs RWO operations independently along the batch dimension, allowing for parallel processing. After a single forward pass through EventMamba, the Monte Carlo estimate is obtained by averaging over the mini-batch. For readability, we omit the standard deviations in the testing results presented in the main text. In our experiments, we set M equal to 8. Detailed testing results, including standard deviations, will be provided in the supplementary materials [1].

Hilbert Space-Filling Curve Mamba. In terms of temporal modeling, although some methods have attempted to facilitate channel-wise modeling by applying the same operations as in the spatial domain to the channel dimension of feature maps or using the reshape function in PyTorch to serialize event features, these methods inevitably lead to the loss of spatio-temporal correlation among adjacent pixels. To overcome this challenge, drawing inspiration from (Wu et al. 2024), we introduce the Hilbert Space-Filling Curve Mamba (HSFCMamba), which utilizes space-filling curves to convert unstructured event features into regular sequences as shown in Figure 2. Specifically, we select two representative space-filling curves, the Hilbert curve and its transposed variant Trans-Hilbert curve, to scan event features. Compared to ordinary reshape function scanning curves, space-filling curves like the Hilbert curve exhibits superior locality preserving properties and a lower space-to-linear ratio (Chen et al. 2022), meaning that adjacent keypoints in the scanned one-dimensional sequence typically have geometrically close positions in the three-dimensional space. We believe this property can largely retain the spatial relationships among points, which is essential for accurate feature representation and analysis in event data. As a complement, the Trans-Hilbert curve exhibits similar characteristics but scans from a different perspective, providing a diversified view of spatial locality. Concretely, we use the following expression for serialization:

$$\begin{aligned} z_h, \bar{z}_h &= \text{MLP}(\text{HSFC}(\hat{E}_{v-1}) + \text{pos\_embedding}), \\ \hat{E}_l &= \text{VMB}(\text{concat}(z_h, \bar{z}_h)), \end{aligned} \quad (16)$$

where HSFC represents the Hilbert scan and Trans-Hilbert scan methods, z_h and $\bar{z}_h$ is the serialized output respectively. Subsequently, we concatenate the sequences z_h and $\bar{z}_h$ obtained from the two different scanning methods, each with a shape of $(B, \frac{H}{8} \times \frac{W}{8} \times 8C)$, into a total sequence $\hat{z}_l$ with a shape of $(B, \frac{H}{8} \times \frac{W}{8} \times 8C \times 2)$. The sequence $\hat{z}_l$ is then fed into the VMB module and reshaped through Hilbert/Trans-Hilbert index to obtain its original shape, constructing the enhanced spatio-temporal event features $\hat{E}_l$. Through our HSFCMamba, our network better preserves spatio-temporal correlations among event data.

[1] Supplementary materials are available at https://github.com/ndwsfkba/EventMamba

Experiments

Settings

Implementation Details. The network is trained for 400 epochs on an NVIDIA 3090 GPU using the AdamW optimizer (Kingma and Ba 2014) with a batch size of 4, a patch size of 128×128, an initial learning rate of 1e-4, and an exponential decay strategy with a gamma of 0.99.

Loss Functions. Our loss function over t-times can be written as:

$$\mathcal{L} = \sum_{t=0}^{N_0} \mathcal{L}_{LPIPS}^t + \lambda_{TC} \sum_{t=L_0}^{N_0} \mathcal{L}_{temp}^t, \quad (17)$$

where $\mathcal{L}_{LPIPS}$ represents the LPIPS loss(Zhang et al. 2018), and $\mathcal{L}_{temp}$ represents the temporal consistency loss (Li et al. 2021; Zhang et al. 2021). A comprehensive description of these loss functions will be provided in the supplementary materials. The hyperparameters N_0, L_0, and λ_{TC} are set to 20, 2, and 0.5, respectively.

Training Datasets. We utilize the ESIM (Rebecq, Gehrig, and Scaramuzza 2018) multi-object 2D renderer option to generate a synthetic training dataset. This renderer captured multiple moving objects within the 2D motion range of the camera. The dataset consists of 280 sequences, each with a duration of 10 seconds. The contrast threshold for event generation ranged from 0.1 to 1.5. Each sequence includes generated event streams, ground truth video frames, and optical flow maps, with an average frequency of 51 Hz. The resolution of both the event camera and frame camera is 256×256. These sequences encompass up to 30 foreground objects with varying velocities and trajectories, randomly selected from the MS-COCO dataset (Lin et al. 2014).

Testing Datasets. We compare the video quality of event stream reconstruction on three publicly available training datasets: HQF (Stoffregen et al. 2020), IJRR (Mueggler et al. 2017), and MVSEC (Zhu et al. 2018).

Evaluation Metrics. In order to quantitatively evaluate the structural quality of video reconstruction, we follow the approach of E2VID and use the following commonly used metrics: Mean Squared Error (MSE), Structural Similarity Index (SSIM) (Hore and Ziou 2010), and Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al. 2018).

Evaluation Methods. In our comparative analysis, we evaluate the performance of EventMamba against multiple cutting-edge techniques in the field, namely FireNet+ (Scheerlinck et al. 2020), E2VID+ (Stoffregen et al. 2020), ET-Net (Weng, Zhang, and Xiong 2021), SPADE-E2VID (Cadena et al. 2021), EVSNN (Zhu et al. 2022), and HyperE2VID (Ercan et al. 2024).

Quantitative Comparison

Table 1 showcases the quantitative comparison results between our proposed EventMamba network and previous

Method	HQF			IJRR			MVSEC		
	MSE↓	SSIM↑	LPIPS↓	MSE↓	SSIM↑	LPIPS↓	MSE↓	SSIM↑	LPIPS↓
E2VID+	0.036	0.533	**0.252**	0.055	0.518	<u>0.261</u>	0.132	0.264	0.514
FireNet+	0.041	0.471	0.316	0.062	0.464	0.318	0.218	0.212	0.569
SPADE-E2VID	0.077	0.400	0.486	0.079	0.462	0.422	0.138	0.266	0.589
EVSNN	0.065	0.424	0.502	0.093	0.413	0.531	0.149	0.253	0.576
ET-Net	0.035	0.558	0.274	0.053	0.552	0.296	0.115	0.315	0.493
HyperE2VID	<u>0.033</u>	<u>0.563</u>	0.272	<u>0.047</u>	<u>0.569</u>	0.278	<u>0.096</u>	<u>0.319</u>	<u>0.489</u>
EventMamba	**0.031**	**0.575**	<u>0.261</u>	**0.039**	**0.586**	**0.254**	**0.073**	**0.328**	**0.475**

Table 1: Comparison on HQF, IJRR and MVSEC Datasets. Best and second best indexs are marked in bold and underline.

Config.	(a)	(b)	(c)	(d)	(e)	Ours
RWO	×	✓	✓	✓	×	✓
HSFC	✓	Hilbert	Trans	×	×	✓
MSE ↓	0.048	0.041	0.041	0.045	0.049	0.039
SSIM↑	0.562	0.582	0.581	0.574	0.559	0.586

Table 2: Ablation study on network structures on IJRR.

methods. In terms of MSE and SSIM values, our Mamba network surpasses all existing event-based video reconstruction networks. Regarding the LPIPS metric, our proposed method also outperforms the majority of existing approaches. However, there is a slight decrease compared to the E2VID+ method in the HQF dataset. Our EventMamba demonstrates superior performance in MSE, SSIM, and LPIPS metrics, achieving state-of-the-art results across multiple datasets. Additionally, our EventMamba model shows greater computational efficiency than Transformer based method ET-Net, with detailed findings available in the **Ablation Studies and Analysis**.

Qualitative Comparison

Figure 3 demonstrates the qualitative reconstruction results of our EventMamba and all baseline methods on images from video clips of the HQF, IJRR, and MVSEC datasets. Ground Truth (GT) video frames are also presented for comparison. It is observed that the frames reconstructed by FireNet+ and E2VID+ lack accuracy in brightness, leading to an inferior overall visual quality of the images. The reconstruction outcomes of ETNet and HyperE2VID are visually more appealing than those of FireNet+ and E2VID+, rendering a more lifelike scene. In contrast, our EventMamba further enriches the final reconstruction with more intricate details, whilst mitigating common artifacts seen in E2VID+. Furthermore, the image contrast of our reconstructed frames closely matches that of the GT images. These qualitative results corroborate the data presented in Table 1. To further support our comparative analysis, we capture a series of high-definition video sequences using the Prophesee EVK4 camera. Figure 4 showcases a series of comparison images. EventMamba delivers superior visual outcomes, exhibiting fewer artifacts than other methods.

Ablation Studies and Analysis

Our ablation experiments are conducted on the IJRR dataset unless otherwise specified. More ablation studies and analysis are presented in the supplementary materials.

	Params (M)	Times (ms)
FireNet+ (L)	0.04	11.8
FireNet+ (H)	0.04	47.73
E2VID+ (L)	10.71	14.55
E2VID+ (H)	10.71	81.82
SPADE (L)	11.46	24.09
SPADE (H)	11.46	210.45
ET-Net (L)	22.18	36.36
ET-Net (H)	22.18	1.85(s)
HyperE2VID (L)	10.15	16.54
HyperE2VID (H)	10.15	126.93
Ours (L)	11.21	18.82
Ours (H)	11.21	136.14

Table 3: Computational complexity of different networks.

Investigation of RWOMamba and HSFCMamba. To explore the impact of RWOMamba and HSFCMamba on experimental outcomes, we design several experimental configurations to validate the efficacy of the proposed methods, as shown in Table 2. We investigate different configurations including changing the window partition strategy in RWOMamba to fixed window partition (Config. (a)), using the Hilbert curve scan alone (Config. (b)), using the trans-Hilbert curve scan alone (Config. (c)), and the Pytorch reshape function scanning (Config. (d)). Config. (e) involves using fixed window partition along with the Pytorch reshape function scanning. The experimental results demonstrate that both RWOMamba and HSFCMamba contribute to significant performance improvements compared to the baseline configurations. These findings validate the effectiveness of our proposed methods in capturing and preserving spatio-temporal dependencies in event features, ultimately leading to enhanced outcomes in the EBVR task.

Comparison of Computational Complexity. To verify the efficiency of our network, we analyze its computational complexity compared to other approaches, focusing on the total number of parameters and inference time on a 3090 GPU. To make our comparison more meaningful, we select resolutions of two common types of event camera sensors currently in use: low resolution 346×240 (L) and high resolution 1280×720 (H). Our experimental results are presented in Table 3. In the table, the model's parameter count is measured in millions (M), and inference time in milliseconds (ms). From Table 3, it is clear that our method strikes a good balance among computational complexity and perfor-

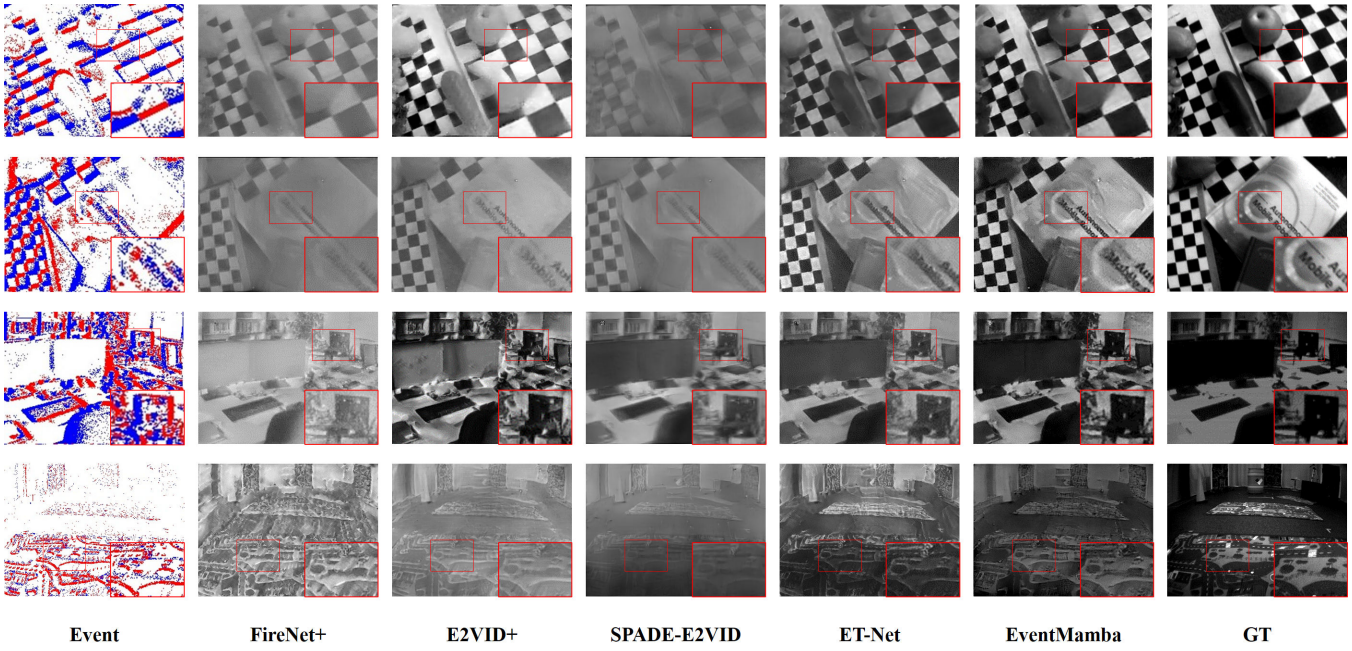

Figure 3: Qualitative comparisons on three benchmarks from HQF (row1-2), IJRR (row3), and MVSEC (row4).

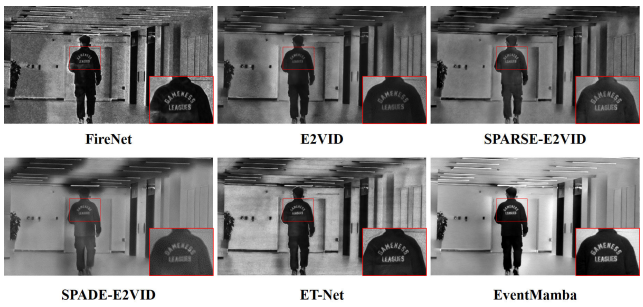

Figure 4: Qualitative comparisons on sequences captured by the Prophesee EVK4 camera.

Network	Params (M)	MSE↓	SSIM↑
FireNet+	0.04	0.062	0.464
E2VID+	10.71	0.055	0.518
SPADE-E2VID	11.46	0.079	0.462
ET-Net	22.18	0.053	0.553
HyperE2VID	10.15	0.047	0.569
Ours (8)	0.70	0.063	0.493
Ours (16)	2.61	0.057	0.532
Ours (24)	5.86	0.047	0.568
Ours (32)	11.21	**0.039**	<u>0.586</u>
Ours (48)	22.06	**0.039**	**0.587**

Table 4: Ablation study on the number of base channels in terms of MSE/SSIM.

mance. On one hand, compared to Transformer-based methods such as ET-Net, our method not only has half the parameters but also significantly reduces inference time when processing large-size inputs of 1280×720. On the other hand, compared to CNNs-based methods, our approach achieves substantial performance improvements with only a slight increase in computational costs, demonstrating the superiority of our proposed method.

Investigation of Base Channel. We further analyze the effect of the base channel C within our network on the final outcomes. Quantitative comparisons with other methods are displayed in Table 4. The results demonstrate that setting the network's base number C to 24 yields the highest performance compared to all prior methods. While increasing the base number beyond 32 offers some metric improvements, the gains begin to plateau. Consequently, we opt for a base channel of 32 to balance performance and efficiency.

Conclusion

In this paper, we critically analyze the shortcomings of existing vision Mamba models in addressing Event-Based Video Reconstruction (EBVR) tasks. To better adapt vision Mamba models to the specific characteristics of EBVR tasks, we propose random window offset (RWO) and Hilbert space filling curve (HSFC) strategies in the spatial and temporal domains, respectively. Specifically, in the spatial domain, we replace fixed window partitioning during training with randomly offset window partitioning to ensure translation invariance. In the temporal domain, we employ Hilbert/trans-Hilbert scanning strategies for serialization to maintain the spatio-temporal locality of events. Based on these strategies, our proposed EventMamba model demonstrates outstanding performance on multiple datasets and achieves significant improvements in computational speed compared to Transformer-based models.

Acknowledgments

This work was supported by National Key R&D Program of China under Grant 2020AAA0105702, National Natural Science Foundation of China (NSFC) under Grants 62225207, 62436008, 62422609 and 62276243.

References

Barua, S.; Miyatani, Y.; and Veeraraghavan, A. 2016. Direct face detection and video reconstruction from event cameras. In *WACV*, 1–9. IEEE.

Bauman, K. E. 2006. The dilation factor of the Peano-Hilbert curve. *Mathematical Notes*, 80: 609–620.

Benosman, R.; Clercq, C.; Lagorce, X.; Ieng, S.-H.; and Bartolozzi, C. 2013. Event-based visual flow. *TNNLS*, 25(2): 407–417.

Cadena, P. R. G.; Qian, Y.; Wang, C.; and Yang, M. 2021. Spade-e2vid: Spatially-adaptive denormalization for event-based video reconstruction. *TIP*, 30: 2488–2500.

Cadena, P. R. G.; Qian, Y.; Wang, C.; and Yang, M. 2023. Sparse-E2VID: A Sparse Convolutional Model for Event-Based Video Reconstruction Trained With Real Event Noise. In *CVPR*, 4149–4157.

Chen, G.; Cao, H.; Conradt, J.; Tang, H.; Rohrbein, F.; and Knoll, A. 2020. Event-based neuromorphic vision for autonomous driving: A paradigm shift for bio-inspired visual sensing and perception. *IEEE Signal Processing Magazine*, 37(4): 34–49.

Chen, W.; Zhu, X.; Chen, G.; and Yu, B. 2022. Efficient point cloud analysis using hilbert curve. In *European Conference on Computer Vision*, 730–747. Springer.

Chollet, F. 2017. Xception: Deep learning with depthwise separable convolutions. In *CVPR*, 1251–1258.

Delbrück, T.; Linares-Barranco, B.; Culurciello, E.; and Posch, C. 2010. Activity-driven, event-based vision sensors. In *ISCAS*, 2426–2429. IEEE.

Ercan, B.; Eker, O.; Saglam, C.; Erdem, A.; and Erdem, E. 2024. Hypere2vid: Improving event-based video reconstruction via hypernetworks. *IEEE Transactions on Image Processing*.

Freeman, A. C.; Singh, M.; and Mayer-Patel, K. 2023. An Asynchronous Intensity Representation for Framed and Event Video Sources. In *Proceedings of the 14th Conference on ACM Multimedia Systems*, 74–85.

Fu, X.; Cao, C.; Xu, S.; Zhang, F.; Wang, K.; and Zha, Z.-J. 2024. Event-Driven Heterogeneous Network for Video Deraining. *International Journal of Computer Vision*, 1–21.

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. *TPAMI*, 44(1): 154–180.

Ge, C.; Fu, X.; He, P.; Wang, K.; Cao, C.; and Zha, Z.-J. 2024. Neuromorphic Event Signal-Driven Network for Video De-raining. In *AAAI*, volume 38, 1878–1886.

Ge, C.; Fu, X.; and Zha, Z.-J. 2022. Learning Dual Convolutional Dictionaries for Image De-raining. In *ACM MM*, 6636–6644.

Gehrig, M.; Aarents, W.; Gehrig, D.; and Scaramuzza, D. 2021. Dsec: A stereo event camera dataset for driving scenarios. *RA-L*, 6(3): 4947–4954.

Gu, A.; and Dao, T. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv:2312.00752*.

Gu, A.; Goel, K.; and Ré, C. 2021. Efficiently modeling long sequences with structured state spaces. *arXiv:2111.00396*.

Hore, A.; and Ziou, D. 2010. Image quality metrics: PSNR vs. SSIM. In *ICPR*, 2366–2369. IEEE.

Jang, H.; McCormack, D.; and Tong, F. 2021. Noise-trained deep neural networks effectively predict human vision and its neural responses to challenging images. *PLoS biology*, 19(12): e3001418.

Karafyllis, I.; and Krstic, M. 2011. Nonlinear stabilization under sampled and delayed measurements, and with inputs subject to delay and zero-order hold. *IEEE Transactions on Automatic Control*, 57(5): 1141–1154.

Kim, H.; Handa, A.; Benosman, R.; Ieng, S.-H.; and Davison, A. J. 2008. Simultaneous mosaicing and tracking with an event camera. *J. Solid State Circ*, 43: 566–576.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv:1412.6980*.

Lagorce, X.; Orchard, G.; Galluppi, F.; Shi, B. E.; and Benosman, R. B. 2016. Hots: a hierarchy of event-based time-surfaces for pattern recognition. *TPAMI*, 39(7): 1346–1359.

Li, D.; Zhu, J.; Wang, M.; Liu, J.; Fu, X.; and Zha, Z.-J. 2023. Edge-aware regional message passing controller for image forgery localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8222–8232.

Li, K.; Li, X.; Wang, Y.; He, Y.; Wang, Y.; Wang, L.; and Qiao, Y. 2024. Videomamba: State space model for efficient video understanding. *arXiv preprint arXiv:2403.06977*.

Li, S.; Luo, Y.; Zhu, Y.; Zhao, X.; Li, Y.; and Shan, Y. 2021. Enforcing temporal consistency in video depth estimation. In *ICCV*, 1145–1154.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *ECCV*, 740–755. Springer.

Liu, J.; Yang, H.; Zhou, H.-Y.; Xi, Y.; Yu, L.; Yu, Y.; Liang, Y.; Shi, G.; Zhang, S.; Zheng, H.; et al. 2024a. Swin-umamba: Mamba-based unet with imagenet-based pretraining. *arXiv:2402.03302*.

Liu, Y.; Tian, Y.; Zhao, Y.; Yu, H.; Xie, L.; Wang, Y.; Ye, Q.; and Liu, Y. 2024b. Vmamba: Visual state space model. *arXiv:2401.10166*.

Mehta, H.; Gupta, A.; Cutkosky, A.; and Neyshabur, B. 2022. Long range language modeling via gated state spaces. *arxiv:2206.13947*.

Mueggler, E.; Rebecq, H.; Gallego, G.; Delbruck, T.; and Scaramuzza, D. 2017. The event-camera dataset and simulator: Event-based data for pose estimation, visual odometry, and SLAM. *IJRR*, 36(2): 142–149.

Munda, G.; Reinbacher, C.; and Pock, T. 2018. Real-time intensity-image reconstruction for event cameras using manifold regularisation. *IJCV*, 126: 1381–1393.

Peng, L.; Cao, Y.; Sun, Y.; and Wang, Y. 2024. Lightweight Adaptive Feature De-drifting for Compressed Image Classification. *IEEE Transactions on Multimedia*.

Rebecq, H.; Gehrig, D.; and Scaramuzza, D. 2018. ESIM: an open event camera simulator. In *Conference on robot learning*, 969–982.

Rebecq, H.; Ranftl, R.; Koltun, V.; and Scaramuzza, D. 2019. Events-to-video: Bringing modern computer vision to event cameras. In *CVPR*, 3857–3866.

Schaefer, S.; Gehrig, D.; and Scaramuzza, D. 2022. Aegnn: Asynchronous event-based graph neural networks. In *CVPR*, 12371–12381.

Scheerlinck, C.; Rebecq, H.; Gehrig, D.; Barnes, N.; Mahony, R.; and Scaramuzza, D. 2020. Fast image reconstruction with an event camera. In *WACV*, 156–163.

Shi, Y.; Xia, B.; Jin, X.; Wang, X.; Zhao, T.; Xia, X.; Xiao, X.; and Yang, W. 2024. Vmambair: Visual state space model for image restoration. *arXiv preprint arXiv:2403.11423*.

Shi, Z.; Chen, Z.; Xu, Z.; Yang, W.; and Huang, L. 2022. AtHom: Two divergent attentions stimulated by homomorphic training in text-to-image synthesis. In *Proceedings of the 30th ACM International Conference on Multimedia*, 2211–2219.

Shiba, S.; Aoki, Y.; and Gallego, G. 2022. Secrets of event-based optical flow. In *ECCV*, 628–645. Springer.

Smith, J. T.; Warrington, A.; and Linderman, S. W. 2022. Simplified state space layers for sequence modeling. *arXiv:2208.04933*.

Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1): 1929–1958.

Stoffregen, T.; Scheerlinck, C.; Scaramuzza, D.; Drummond, T.; Barnes, N.; Kleeman, L.; and Mahony, R. 2020. Reducing the sim-to-real gap for event cameras. In *ECCV*, 534–549. Springer International Publishing.

Tulyakov, S.; Bochicchio, A.; Gehrig, D.; Georgoulis, S.; Li, Y.; and Scaramuzza, D. 2022. Time lens++: Event-based frame interpolation with parametric non-linear flow and multi-scale fusion. In *CVPR*, 17755–17764.

Wang, K.; Fu, X.; Ge, C.; Cao, C.; and Zha, Z.-J. 2024. Towards Generalized UAV Object Detection: A Novel Perspective from Frequency Domain Disentanglement. *International Journal of Computer Vision*, 1–29.

Weng, W.; Zhang, Y.; and Xiong, Z. 2021. Event-based video reconstruction using transformer. In *ICCV*, 2563–2572.

Wu, H.; Yang, Y.; Xu, H.; Wang, W.; Zhou, J.; and Zhu, L. 2024. RainMamba: Enhanced Locality Learning with State Space Models for Video Deraining. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 7881–7890.

Xing, Z.; Ye, T.; Yang, Y.; Liu, G.; and Zhu, L. 2024. Segmamba: Long-range sequential modeling mamba for 3d medical image segmentation. *arXiv:2401.13560*.

Xu, S.; Sun, Z.; Zhu, J.; Zhu, Y.; Fu, X.; and Zha, Z.-J. 2024. DemosaicFormer: Coarse-to-Fine Demosaicing Network for HybridEVS Camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 1126–1135.

Zhang, F.; Li, Y.; You, S.; and Fu, Y. 2021. Learning temporal consistency for low light video enhancement from single images. In *CVPR*, 4967–4976.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 586–595.

Zhang, Y.; Shi, Z.; Yang, W.; Wang, S.; Wang, S.; and Xue, Y. 2024. GenSeg: On Generating Unified Adversary for Segmentation. In *IJCAI*.

Zhang, Y.; Yang, W.; and Hu, R. 2023. BAProto: Boundary-Aware Prototype for High-quality Instance Segmentation. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, 2333–2338. IEEE.

Zhang, Y.; Yang, W.; and Wang, S. 2023. FGNet: Towards Filling the Intra-class and Inter-class Gaps for Few-shot Segmentation. In *IJCAI*, 1749–1758.

Zhang, Z.; Yezzi, A.; and Gallego, G. 2021. Formulating event-based image reconstruction as a linear inverse problem with deep regularization using optical flow. *arXiv:2112.06242*.

Zhu, A. Z.; Thakur, D.; Özaslan, T.; Pfrommer, B.; Kumar, V.; and Daniilidis, K. 2018. The multivehicle stereo event camera dataset: An event camera dataset for 3D perception. *Robotics and Automation Letters*, 3(3): 2032–2039.

Zhu, L.; Liao, B.; Zhang, Q.; Wang, X.; Liu, W.; and Wang, X. 2024. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv:2401.09417*.

Zhu, L.; Wang, X.; Chang, Y.; Li, J.; Huang, T.; and Tian, Y. 2022. Event-based video reconstruction via potential-assisted spiking neural network. In *CVPR*, 3594–3604.