E-mamba: Using state-space-models for direct event processing in space situational awareness

Alejandro Hernández Díaz[1], Rebecca Davidson[2], Steve Eckersley[2], Christopher Bridges[1], Simon Hadfield[1]

[1]*University of Surrey, Guildford, United Kingdom*
[2]*Surrey Satellite Technology Limited, Guildford, United Kingdom*

The planning and execution of modern space missions rely on traditional SSA methods for detecting and tracking orbiting hazards. This often leads to sub-optimal responses due to remote sensing inaccuracies and transmission delays. On the other hand, deploying and maintaining space-based sensors is expensive and technically challenging due to the inadequacy of current vision technologies. In this paper, we propose a novel perception framework to enhance in-orbit autonomy and address the shortcomings of traditional SSA methods. We leverage the advances of neuromorphic cameras for a vastly superior sensing performance under space conditions. Additionally, we maximize the advantageous characteristics of the sensor by harnessing the modelling power and efficient design of selective State Space Models. Specifically, we introduce two novel event-based backbones, E-Mamba and E-Vim, for real-time on-board inference with linear scaling in complexity w.r.t. input length. Extensive evaluation across multiple neuromorphic datasets demonstrate the superior parameter efficiency or our approaches (<1.3M params), while yielding comparable performance to the state of the art in both detection and dense-prediction tasks. This opens the door to a new era of highly-efficient intelligent solutions to improve the capabilities and safety of future space missions.

1 Introduction

Space Situational Awareness (SSA) stands as a cornerstone of the global space exploration initiative. Its essence lies on the identification, analysis and tracking of near-earth objects' orbits. These efforts allow for the integration of said trajectories into new mission designs and facilitate the strategic rerouting of existing systems when necessary. However, most of the work carried out in this field relies on the high-fidelity detection and mathematical modeling of orbits from earth. This significantly prolongs the response time of current satellites to any unforeseen incoming object. In addition, low-latency autonomous approaches to SSA are often discouraged, as traditional visual sensors are ill-suited for space-related applications due to their sub-optimal characteristics e.g. slow capture-rate, high power consumption and susceptibility to low light environments.

In terrestrial research, event cameras emerged as a neurologically inspired visual sensor which significantly deviates from the operational principles of traditional frame-based cameras. Unlike standard cameras, which capture full-frame luminance levels at predetermined intervals, event cameras document per-pixel intensity changes asynchronously in real-time. These new sensors offer several advantages over their RGB counterparts, particularly benefiting on-board SSA applications: Their asynchronous nature drastically diminishes their power consumption and data output volume, rendering them ideal for long-term in-orbit deployment. Moreover, their higher dynamic range enhances capture performance in the challenging lighting conditions of space, including direct sunlight and deep shadows. Additionally, their reduced latency enables fast reactivity, which is vital for collision avoidance and maneuver planning tasks. However, their asynchronous nature also causes a radical shift in output representation, which in this case is a stream of spatio-temporal events. Unfortunately, given that most traditional computer vision algorithms rely on dense and synchronous pixel measurements, adapting them to accommodate the stream generated by event cameras can pose significant challenges.

The current landscape of event-based processing architectures can be divided into two categories. **Point-based methods** process the generated events in their natural sequence form, by employing sparse computational paradigms. The architectural choices inside this line of research include Point Networks [1] [2], Graph Neural Networks (GNNs) [3] [4] [5], or Spiking Neural Networks (SNNs) [6] [7] [8]. However, despite the high compatibility of these approaches with the natural characteristics of event sequences, they tend to offer limited performance or need specialised hardware to function. To address the performance issues, several works have proposed a second processing paradigm which involves converting events into image-like representations, making them compatible with modern vision architectures i.e **Frame-based methods**. The preferred architectures inside this category are Convolutional Neural Networks (CNNs) which have been successfully applied to multiple event-based task such as Optical Flow estimation [9] [10], Depth prediction [11] [12], object detection [13] and object classification [14]. Following the landscape changes in synchronous vision, Transformers were also presented as viable alternatives to CNNs, obtaining state-of-the-art results across several benchmarks [15] [16] at the cost of longer training cycles due to the lack of convolutional inductive biases. In spite of the competitive performance shown by this category of models, accumulating the events reduces the advantageous reaction time of the sensor while also introducing redundant computation due to empty pixels/voxels. This makes them less well suited to space applications and their computational constraints.

In this paper we propose selective State Space Models (SSMs) [17] [18] [19] as a potential alternative for both of these regimes. Specifically, we build our architectures on top of the Mamba processing approach to address the shortcomings present in previous work. Mamba is a sequence modeling architecture which has recently emerged in the Natural Language Processing (NLP) domain. Initial experiments have demonstrated its notably reduced computational requirements, as unlike transformers, it scales linearly in complexity w.r.t. sequence length. Additionally, it has proven to be significantly more parameter efficient than existing state-of-the-art techniques. The joint characteristics of the proposed pipeline hold promise for enabling a lightweight space-ready computer vision system, characterised by high efficiency spanning from sensor input to computational output. In turn, this would allow for the introduction of novel intelligent assistive frameworks in future space missions, where the on-board perception components effectively communicate with human operators to alert of possible hazards or suggest re-routing maneuvers.

2 State Space Models (SSMs)

We define a State Space Model that describes a 1-D sequence-to-sequence mapping from $u(t) : \mathbb{R} \rightarrow \mathbb{R}$ to $y(t) : \mathbb{R} \rightarrow \mathbb{R}$ through an N-D hidden state $h(t) \in \mathbb{R}^N$. Such a model is parametrised by two projection matrices $\mathbf{B}$, $\mathbf{C}$ dependent on the input sequence (i.e. $\mathbf{B} = S_{\mathbf{B}}(u) : \mathbb{R} \rightarrow \mathbb{R}^N$, $\mathbf{C} = S_{\mathbf{C}}(u) : \mathbb{R} \rightarrow \mathbb{R}^N$), and an evolution matrix $\mathbf{A} \in \mathbb{R}^{N \times N}$. $S_{\mathbf{B}}$ and $S_{\mathbf{C}}$ represent linear projections to dimension N i.e. $\text{Linear}_N(\cdot)$.

This can be formulated as linear ordinary differential equations, where h' is the gradient of the state h.

$$\begin{aligned} h'(t) &= \mathbf{A}h(t) + \mathbf{B}u(t), \\ y(t) &= \mathbf{C}h(t). \end{aligned} \tag{1}$$

In order to integrate this SSM design into deep learning algorithms, we transform the continuous-time parameters $\mathbf{A}$, $\mathbf{B}$ into the discrete-time parameters $\overline{\mathbf{A}}$, $\overline{\mathbf{B}}$ using the Zero Order Hold (ZOH) discretisation method.

$$\overline{\mathbf{A}} = \exp(\Delta\mathbf{A}) \quad \overline{\mathbf{B}} = (\Delta\mathbf{A})^{-1}(\exp(\Delta\mathbf{A}) - \mathbf{I}) \cdot \Delta\mathbf{B}. \tag{2}$$

This reformulation adds a new step size parameter Δ representing the input's resolution, theoretically controlling how much to focus on or ignore each measurement. We ensure that the step size is also dependent on the input by $\Delta = \tau_\Delta(\Delta + S_\Delta(u))$ where $S_\Delta = \text{Broadcast}_D(\text{Linear}_1(\cdot))$ and $\tau_\Delta = \text{softplus}$.

We leverage the HiPPO theory of continuous-time memorization [20] as the initialization mechanism for our evolution matrix $\mathbf{A}$, allowing the state to integrate recent inputs with higher fidelity than those further in the past.

$$\mathbf{A}_{nk} = -\begin{cases} (2n+1)^{1/2}(2k+1)^{1/2} & \text{if } n > k \\ n+1 & \text{if } n = k \\ 0 & \text{if } n < k \end{cases} \tag{3}$$

3 Selective SSMs for Event Vision

Let $I(x,y,t)$ denote the log intensity at location (x,y) and time t. An event e is triggered at (x,y,t) whenever the change in log intensity surpasses a predefined contrast threshold C. Each event is characterised by its polarity p, indicating either a positive or negative variation in I. Hence, the output of an event camera is an asynchronous stream of N events

$$\mathcal{E} = \{e_k \mid I(x_k, y_k, t_k) - I(x_{k-1}, y_{k-1}, t_{k-1}) \geq C\}_{k=1}^N \tag{4}$$

In this paper, we propose two main approaches to efficient SSM-based event processing for space applications:

Point-based. Current state-of-the-art models i.e. Transformers are not able to perform point-based event processing due to their poor scalability with input length as well as their limited context windows. However, Mamba's proven ability to scale up to 1M-length sequences with no compromise in performance finally enables us to efficiently perform raw event-stream processing, preserving all the speed and memory benefits for on-board space applications. To explore this regime we propose **E-Mamba**, our event-sequence encoder comprised of a stack of unidirectional SSM blocks.

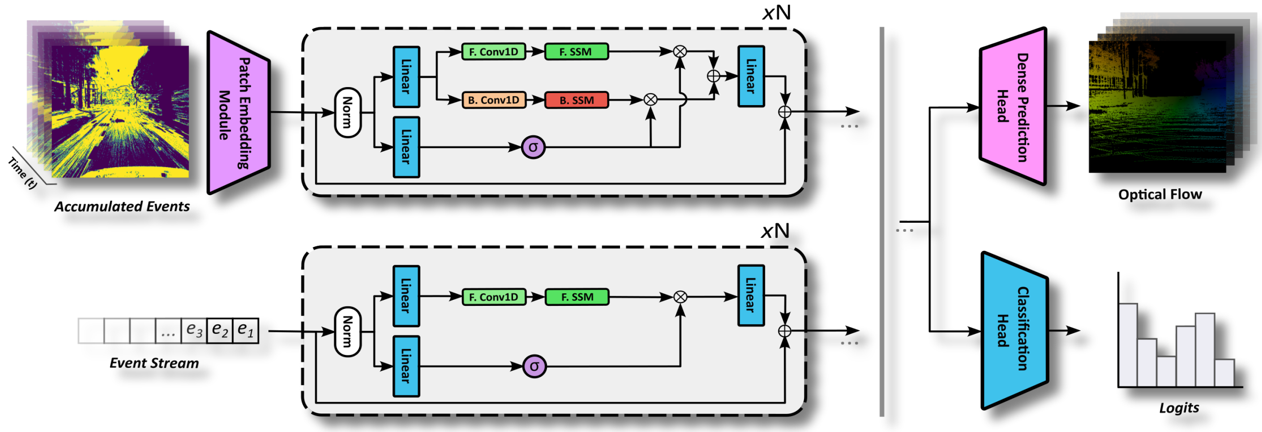

Figure 1: Overview of the **E-Vim** (top) and **E-Mamba** (bottom) architectures. The semantically rich features extracted by our models can be subsequently utilised for both classification or dense-prediction tasks.

Dataset	Model	Acc. (↑)	# params
[22]	EvT+ [23]	97.57	0.66M
	TORE [24]	96.2	5.94M
	S-former [25]	98.96	9.28M
	E-Mamba	60.02	75.5K
	E-Vim$_\text{S}$	84.03	383K
	E-Vim$_\text{L}$	80.09	1.2M
[26]	EvS [27]	68.0	N/A
	NDA [28]	81.7	132.8M
	S-former [25]	81.4	9.28M
	E-Mamba	32.4	75.5K
	E-Vim$_\text{S}$	59.7	1.2M
	E-Vim$_\text{L}$	59.1	2.9M

Table 1: Evaluation of Base **E-Mamba** and **E-Vim** variants on DVS128-Gesture [22] and CIFAR10-DVS [26].

In this case, each individual event in the stream is transformed into a 4D tensor $e_k = [x_k, y_k, (t_{k-1} - t_k), p_k]$ encoding its spatial location within the DVS's sensor array, its polarity and the difference in timestamp w.r.t. the previous event. These sensors produce streams in the order of millions of events per second, hence in order to ease the learning process we encode relative time information as timestamp differences.

Frame-based. To fuse the advantages of frame-base methods with the parameter and implementation efficiency of SSMs, we propose **E-Vim**, our novel frame-based event encoder. In this case, we extend our point-based encoder block with bi-directional processing of the input sequence using forward and backward SSMs, and a spatio-temporal patch embedding module. First, we transform the input streams into voxelised representations by dividing the sequence into t temporal windows of length l. Subsequently, we convert the events in each window into a frame representation $F \in \mathbb{R}^{2\times B\times H\times W}$. We focus on two representation strategies: traditional volumetric voxel-grids, following the pipeline described in [11], and event histograms, where each pixel location in F is represented by two histogram-like vectors of B bins, one per polarity $p \in \{0,1\}$.

We introduce a spatio-temporal patch embedding strategy to encode the input frames by using a 3D strided convolution with a kernel $k \in \mathbb{R}^{1\times P_H\times P_W}$. This in turn creates a sequence S of length $(B \times \frac{H}{P_H} \times \frac{W}{P_W})$ comprised of non-overlapping spatio-temporal patch features, subsequently flattened and projected to the dimensionality of the encoder. Following BERT's [21] conventions, the model adds a learnable [CLS] token $\text{cls} \in \mathbb{R}^D$ to the sequence plus learnable positional encodings $E_{\text{pos}} \in \mathbb{R}^{(L+1)\times D}$.

4 Sequence Classification Results

We begin by assessing our selected models on the two primary event-based sequence classification tasks, namely gesture recognition on the DVS128-Gesture dataset [22] and object detection using CIFAR10-DVS [26]. These are used as proxy tasks with relevance to space-situational awareness due to the nature of their acquisition methodologies, assessing the adaptability of our proposed systems to both static-camera dynamic-scene and dynamic-camera static-scene scenarios.

Point-based processing Our architecture is composed of 2 Mamba encoder blocks, each with an embedding size of 100. Following the encoding process, the embedding of the last event is linearly projected through a classification head to obtain the logits. We optimize the framework using Negative Log-likelihood loss and a base learning rate of 0.001.

Frame-based processing The reduced computational demands of frame-based processing allows us to explore two different model sizes for our architectures: **E-Vim$_\text{S}$** and **E-Vim$_\text{L}$** of encoder depths equal to 1 and 4 for DVS128-Gesture, and 4 and 10 for CIFAR10-DVS. We first train the two baselines on

Dataset	Model	$\mathrm{N_{bins}}$		lr		Repre.		Aug.		Time*
		15	2	$1e^{-3}$	$3e^{-4}$	Voxel	Hist.	NDA	Rand.	
DVS128 [22]	E-Mamba	-	-	-	-	-	-	90.97	-	$1.45ms$
	E-Vim$_\mathrm{S}$	84.03	79.51	84.03	80.9	84.03	91.01	86.46	83.4	$1.66ms$
CIFAR10-DVS [26]	E-Mamba	-	-	-	-	-	-	35.97	-	$1.45ms$
	E-Vim$_\mathrm{S}$	59.7	53.8	59.1	58.9	59.7	60.01	62.5	54.3	$4.01ms$

* Avg. inference time over 1000 runs.

Table 2: Additional evaluation of the best-performing **E-Vim** variant and **E-Mamba** (with neuromorphic augmentations) on DVS128-Gesture [22] and CIFAR10-DVS [26].

($96 \times 96 \times \mathrm{N_{bins}}$) spatio-temporal crops using the same experimental details described above. In this case, we pre-process the input sequences into voxel-grid representations by discretising the time dimension in $\mathrm{N_{bins}} = 15$ bins prior to feeding them to the models. The final embedding of the learned [CLS] token is linearly projected to obtain the predictions.

As shown in Table 1, the compact variants of our **E-Vim** model outperform their larger counterparts in both gesture and object recognition tasks despite having 32% and 55% fewer parameters respectively. This demonstrates the exceptional parameter efficiency of our State Space Models. As expected, our frame-based architectures outperformed the proposed **E-Mamba** in all datasets. We hypothesize that this is caused by the lack of convolutional inductive biases present in our point-based approach, making it harder for the network to encode spatial relationships between non-neighboring events in the sequence. However, our **E-Mamba** model is able to perform perform streaming inference on an event-by-event basis and with minimal computational requirements, by using the recurrent interpretation of its discretised SSMs. This is an extremely valuable property in on-board use cases where fast-reactivity is required.

We now select the best-performing frame-based variant in each dataset for further investigation. These additional experiments assess how various design choices impact task performance i.e. hyperparameter optimization, temporal granularity of the input volumes, event-representation and data augmentation. On this last front we both explore a naive pipeline of affine transforms (e.g. random flips, rotations, shears etc.), and the state-of-the-art Neuromorphic Data Augmentation (NDA) framework proposed in [29]. Additionally, we also investigate the effect of related point-based transforms (event-drop, rotations, translations and spatio-temporal jitter) on the generalization capabilities of our **E-Mamba** architectures.

According to Table 2, our **E-Vim$_\mathbf{S}$** model exhibits comparable performance to the state of the art in the DVS128-Gesture dataset with less than $400K$ parameters. This suggests significant potential for our model as a lightweight & powerful encoder.

Dataset	Model	L1 (↓)	# params
DSEC-flow [10]	E-Vim$_\mathrm{S}$	0.523	$1.2M$
	E-Vim$_\mathrm{L}$	0.549	$8.1M$

Table 3: Evaluation of Base **E-Vim** variants on our custom validation split of DSEC-flow [10].

Upon reducing the input's temporal granularity, we observed a substantial drop in performance in both benchmarks. This indicates that our architectures benefit from the rich temporal information in the events, instead of relying on appearance cues only.

Adequately applying relevant transformations to the input data also proved to be extremely valuable for our models. Naive augmentations hindered the performance of our SSMs, but using the event-specific NDA transforms improved their generalization capabilities w.r.t. the baseline case in both tasks.

5 Dense-prediction Tasks

Additionally, we evaluate our SSM variants on a dense pixel-wise prediction task: Optical flow estimation on the DSEC-flow dataset [10]. This is not only a highly relevant task for autonomous space applications, but also enables us to observe the behaviour of our SSMs in more complex dynamic- camera dynamic-scene scenarios. Nonetheless, the lengthier sequences in DSEC ($> 3M$) render it infeasible to effectively employ point-wise training on the data. Instead we focus training-efficient frame processing techniques.

Frame-based processing Here we also explore two different model sizes, **E-Vim$_\mathbf{S}$** and **E-Vim$_\mathbf{L}$** of encoder depths equal to 8 and 20. We use the same voxel-grid representation but increase our random crop size to ($480 \times 480 \times 15$) to account for the higher resolution. The final embedding of the learned [CLS] token is reshaped into a $2D$ tensor and fed into a decoder-like head with 4 [Upsample+Conv2D] blocks to obtain the flow predictions. We use an L1 loss and a base learning rate of $3e^{-4}$.

Table 3 presents the results obtained by **E-Vim** in the dense prediction task. Consistently with earlier ob-

Dataset	N_{bins}		lr		Repre.		# params
	30	40	$1e^{-3}$	$5e^{-5}$	Hist.	Voxel	
DSEC-flow [10]	0.5636	0.572	0.5384	0.59	0.502	0.523	$1.2M$

Table 4: Evaluation of the best-performing Base **E-Vim** variant on the additional DSEC-flow [10] experiments.

servations, the smaller model demonstrates superior performance compared to its larger counterpart, regardless of their $6.8M$ parameter difference. These results illustrate the high-adaptability of our approach to multiple task types and use-case scenarios, while maintaining competitive performance. We now select the best-performing variant and conduct a similar array of extra evaluations to those presented in the classification experiments.

Table 4 Illustrates the performance of our **E-Vim**$_{\mathbf{S}}$ across said extra evaluations. Motivated by the results obtained in the classification tasks, we further quantised the input volume into 30 and 40 bins for a finer temporal history of the sequence. Nevertheless, the redundant computation introduced by the additional empty pixels prevented any significant improvement.

We could also observe significant benefits across tasks as a consequence of using the simpler histogram frames. However, we show that our model is able to successfully extract information from multiple representations without hindering task results.

6 Discussion

In this paper we proposed neuromorphic cameras as an ideal sensor for on-board Space Situational Awareness scenarios, where fast-reactivity and computational efficiency are essential. Additionally, we introduced two SSM-based architectures designed to extract task-agnostic features from event streams, while fully leveraging the advantageous characteristics of the sensing device.

Experimental results in several event-based benchmarks have verified the modeling capabilities and parameter/computational efficiency of our architectures. The integration between the proposed components yields a lightweight, space-ready perception framework. This can be employed alongside current SSA methods to enhance both the detection accuracy of potential hazards and the reaction time to such threats. We hope that our work encourages new avenues of research and applications in on-board intelligent systems for the next generation of space missions.

7 Acknowledgement

This work was partially supported by Surrey Satellite Technology Ltd.

References

1. Wang, Q., Zhang, Y., Yuan, J. & Lu, Y. *Space-Time Event Clouds for Gesture Recognition: From RGB Cameras to Event Cameras* in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)* (IEEE, Waikoloa Village, HI, USA, Jan. 2019), 1826–1835. isbn: 978-1-72811-975-5. (2024).
2. Sekikawa, Y., Hara, K. & Saito, H. *EventNet: Asynchronous Recursive Event Processing* in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, Long Beach, CA, USA, June 2019), 3882–3891. isbn: 978-1-72813-293-8. (2024).
3. Bi, Y., Chadha, A., Abbas, A., Bourtsoulatze, E. & Andreopoulos, Y. *Graph-Based Object Classification for Neuromorphic Vision Sensing* in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (IEEE, Seoul, Korea (South), Oct. 2019), 491–501. isbn: 978-1-72814-803-8. (2024).
4. Bi, Y., Chadha, A., Abbas, A., Bourtsoulatze, E. & Andreopoulos, Y. Graph-Based Spatio-Temporal Feature Learning for Neuromorphic Vision Sensing. *IEEE Transactions on Image Processing* **29,** 9084–9098. issn: 1941-0042. (2024) (2020).
5. Deng, Y., Chen, H., Liu, H. & Li, Y. *A Voxel Graph CNN for Object Classification with Event Cameras* in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, New Orleans, LA, USA, June 2022), 1162–1171. isbn: 978-1-66546-946-3. (2024).
6. Orchard, G. *et al.* HFirst: A Temporal Approach to Object Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **37,** 2028–2040. issn: 0162-8828, 2160-9292, 1939-3539. (2024) (Oct. 2015).
7. Zhou, Z. *et al. Spikformer: When Spiking Neural Network Meets Transformer* in *The Eleventh International Conference on Learning Representations* (Sept. 2022). (2024).
8. Zhou, Z. *et al. Spikformer V2: Join the High Accuracy Club on ImageNet with an SNN Ticket* Jan. 2024. arXiv: 2401.02020 [cs]. (2024).
9. Zhu, A., Yuan, L., Chaney, K. & Daniilidis, K. *EV-FlowNet: Self-Supervised Optical Flow Estimation for Event-based Cameras* in *Robotics: Science and Systems XIV* (Robotics: Science and Systems Foundation, June 2018). isbn: 978-0-9923747-4-7. (2024).
10. Gehrig, M., Millhäusler, M., Gehrig, D. & Scaramuzza, D. *E-RAFT: Dense Optical Flow from Event Cameras* in *2021 International Conference on 3D Vision (3DV)* (Dec. 2021), 197–206. (2023).
11. Zhu, A. Z., Yuan, L., Chaney, K. & Daniilidis, K. *Unsupervised Event-Based Learning of Optical Flow, Depth, and Egomotion* in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), 989–997. (2023).
12. Ye, C., Mitrokhin, A., Fermüller, C., Yorke, J. A. & Aloimonos, Y. *Unsupervised Learning of Dense Optical Flow, Depth and Egomotion with Event-Based Sensors* in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (Oct. 2020), 5831–5838. (2023).
13. Perot, E., de Tournemire, P., Nitti, D., Masci, J. & Sironi, A. *Learning to Detect Objects with a 1 Megapixel Event Camera* in *Advances in Neural Information Processing Systems* **33** (Curran Associates, Inc., 2020), 16639–16652. (2024).

14. Deng, Y., Chen, H. & Li, Y. MVF-Net: A Multi-View Fusion Network for Event-Based Object Classification. *IEEE Transactions on Circuits and Systems for Video Technology* **32,** 8275–8284. ISSN: 1051-8215, 1558-2205. (2024) (Dec. 2022).

15. Tian, Y. Event Transformer FlowNet for Optical Flow Estimation (2022).

16. Li, Y. *et al. BlinkFlow: A Dataset to Push the Limits of Event-based Optical Flow Estimation* Mar. 2023. arXiv: 2303.07716 [cs]. (2023).

17. Gu, A. & Dao, T. *Mamba: Linear-Time Sequence Modeling with Selective State Spaces* Dec. 2023. arXiv: 2312.00752 [cs]. (2024).

18. Zhu, L. *et al. Vision Mamba: Efficient Visual Representation Learning with Bidirectional State Space Model* Jan. 2024. arXiv: 2401.09417 [cs]. (2024).

19. Guo, H. *et al. MambaIR: A Simple Baseline for Image Restoration with State-Space Model* Mar. 2024. arXiv: 2402.15648 [cs]. (2024).

20. Gu, A., Dao, T., Ermon, S., Rudra, A. & Ré, C. *HiPPO: Recurrent Memory with Optimal Polynomial Projections* in *Advances in Neural Information Processing Systems* **33** (Curran Associates, Inc., 2020), 1474–1487. (2024).

21. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (eds Burstein, J., Doran, C. & Solorio, T.) (Association for Computational Linguistics, Minneapolis, Minnesota, June 2019), 4171–4186. (2024).

22. Amir, A. *et al. A Low Power, Fully Event-Based Gesture Recognition System* in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, Honolulu, HI, July 2017), 7388–7397. ISBN: 978-1-5386-0457-1. (2024).

23. Sabater, A., Montesano, L. & Murillo, A. C. Event Transformer+. A Multi-Purpose Solution for Efficient Event Data Processing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–8. ISSN: 1939-3539. (2023) (2023).

24. Baldwin, R. W., Liu, R., Almatrafi, M., Asari, V. & Hirakawa, K. Time-Ordered Recent Event (TORE) Volumes for Event Cameras. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **45,** 2519–2532. ISSN: 1939-3539. (2024) (Feb. 2023).

25. Li, Y., Lei, Y. & Yang, X. *Spikeformer: A Novel Architecture for Training High-Performance Low-Latency Spiking Neural Network* Nov. 2022. arXiv: 2211.10686 [cs]. (2024).

26. Li, H., Liu, H., Ji, X., Li, G. & Shi, L. CIFAR10-DVS: An Event-Stream Dataset for Object Classification. *Frontiers in Neuroscience* **11.** ISSN: 1662-453X. (2024) (May 2017).

27. Li, Y. *et al. Graph-Based Asynchronous Event Processing for Rapid Object Recognition* in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (IEEE, Montreal, QC, Canada, Oct. 2021), 914–923. ISBN: 978-1-66542-812-5. (2024).

28. Li, Y., Kim, Y., Park, H., Geller, T. & Panda, P. *Neuromorphic Data Augmentation for Training Spiking Neural Networks* July 2022. arXiv: 2203.06145 [cs]. (2024).

29. Li, Y., Kim, Y., Park, H., Geller, T. & Panda, P. *Neuromorphic Data Augmentation for Training Spiking Neural Networks* ECCV 2022, 2022. (2024).