

Asynchronous Bioplausible Neuron for Spiking Neural Networks for Event-Based Vision

Sanket Kachole¹, Hussain Sajwani^{2,3}, Fariborz Baghaei Naeini^{1,4},
Dimitrios Makris¹, and Yahya Zweiri^{2,3}

¹ Department of Computer Science, Kingston University, London, UK

² Advanced Research and Innovation Center(ARIC), Khalifa University, UAE

³ Department of Aerospace Engineering, Khalifa University of Science and
Technology, Abu Dhabi, UAE

⁴ Ipsotek, an Eviden Company, London

⁵ {K1742163,f.baghaeinaeini,d.makris}@kingston.ac.uk¹
{hussain.sajwani,yahya.zweiri}@ku.ac.ae^{2,3}

1 Spiking Neural Network

Spiking Neural Networks represent a paradigm shift in neural network technology, drawing closer to the biological intricacies of the human brain. Unlike traditional neural networks, which process information continuously, SNNs transmit information through discrete events known as spikes. This not only makes them more biologically plausible but also potentially more efficient in terms of computational resources.

The concept of SNNs stems from the desire to mimic the brain's extraordinary ability to process information efficiently. Historically, the development of SNNs has been part of the broader field of neuromorphic computing, aiming to create computer architectures that mirror neural structures. The fundamental building blocks of SNNs are neurons and synapses, interconnected in ways that allow for the dynamic transmission of spike signals.

SNNs stand out due to their potential applications in areas where traditional networks may not be as efficient, particularly in tasks involving temporal data processing, like speech and gesture recognition. The unique architecture of SNNs, where the timing of a spike carries information, allows for a more nuanced and dynamic approach to learning patterns and making predictions.

1.1 Encoding Method in Spiking Neural Networks

The encoding method in SNNs is a critical aspect that determines how information is represented and processed. Unlike conventional neural networks that use continuous values, SNNs rely on spikes, where the timing and frequency of these spikes represent information. There are various methods of encoding in SNNs, each with its strengths and applications.

Rate encoding, one of the most common methods, encodes information in the frequency of the spikes. High-frequency spikes can represent a high value,

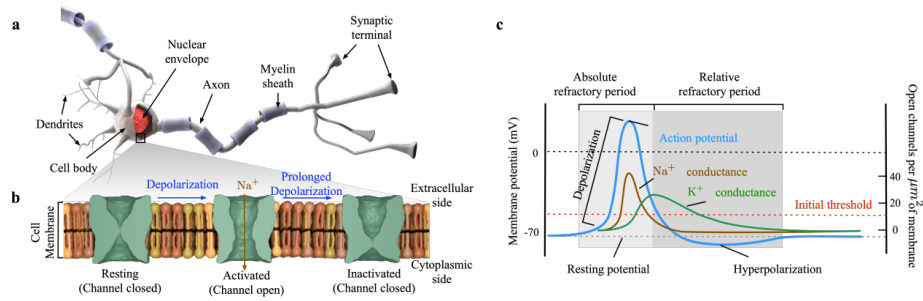


Fig. 1: Illustrative depiction of neuronal anatomy and electrophysiology. Panel a shows the neuron’s structure, highlighting the dendrites, cell body, axon, myelin sheath, and synaptic terminal. Panel b details the states of a sodium (Na^+) channel during and after action potential initiation: resting with the channel closed, activated with the channel open, and inactivated with the channel closed, awaiting potential normalization. Panel c presents the action potential mechanism as described by the Hodgkin-Huxley model, showcasing the dynamic changes in Na^+ and K^+ conductances that facilitate neuronal firing, alongside the periods of absolute and relative refractory phases following the spike [2]

and vice versa. However, this method can be less efficient in terms of spike usage and might not capture the temporal aspects of the information.

Temporal encoding, on the other hand, utilizes the precise timing of spikes to convey information. This method can be more efficient than rate encoding and is particularly useful in tasks that require processing of temporal patterns, such as speech recognition.

Population encoding involves using a group of neurons to represent a single value, providing a balance between rate and temporal encoding methods. It enhances the robustness of the representation and can handle a wider range of values more efficiently.

1.2 Biological Concepts in Spiking Neural Networks

The inspiration for SNNs comes directly from the way biological neural networks in the brain process information. The neurons in these networks communicate through electrical impulses or spikes, which are brief and discrete events. This biological realism in SNNs is not just limited to the use of spikes; it extends to the modeling of neurons and synapses.

Neuron models in SNNs, such as the Hodgkin-Huxley and integrate-and-fire models, aim to replicate the electrical characteristics of biological neurons. These models describe how neurons accumulate input signals and generate an output spike once a certain threshold is reached.

Synaptic transmission and plasticity are also crucial biological concepts replicated in SNNs. Synapses in the brain strengthen or weaken over time based on the activity of the neurons they connect, a concept known as synaptic plasticity. This mechanism is the basis for learning and memory in the brain and is

mirrored in SNNs through learning rules like Spike-Timing-Dependent Plasticity (STDP).

1.3 Mathematical Framework in Spiking Neural Networks

The mathematical framework of SNNs is essential for understanding their operation and designing efficient networks. This framework involves equations and algorithms that describe how neurons and synapses behave and interact.

Neuron dynamics in SNNs are typically described by differential equations that model the change in a neuron's membrane potential over time. These equations take into account the inputs received from other neurons and determine when a neuron should emit a spike.

Learning in SNNs is governed by mathematical rules that adjust the strength of synapses based on the timing of spikes. STDP, for instance, is a rule where the synaptic strength is increased if a presynaptic neuron's spike precedes a postsynaptic neuron's spike and decreased in the opposite scenario. This rule is critical for unsupervised learning in networks.

In summary, Spiking Neural Networks represent a sophisticated and biologically inspired approach to neural computation. Their unique properties, such as spike-based information processing and the incorporation of temporal dynamics, offer promising avenues for research and application in fields ranging from robotics to neuroscience.

2 Datasets

2.1 Neuromorphic-MNIST (N-MNIST) Dataset Overview

The N-MNIST dataset represents a significant advancement in the field of neuromorphic vision systems, offering a novel adaptation of the widely recognized MNIST dataset. This "spiking" version maintains the structural integrity of the original dataset, featuring 60,000 training and 10,000 testing samples, each with a visual scale of 28x28 pixels. N-MNIST was developed using the ATIS sensor on a motorized pan-tilt unit, enabling dynamic capture of MNIST examples from an LCD monitor. This method closely mimics the temporal dynamics of real-world visual perception, a crucial aspect for advanced neuromorphic computing. The dataset's development is thoroughly detailed in [6].

2.2 CIFAR-10 DVS Dataset Overview

The CIFAR-10 DVS dataset is an innovative adaptation of the CIFAR-10 dataset, tailored for neuromorphic vision research. It transforms frame-based images into event streams using a Dynamic Vision Sensor (DVS). This dataset comprises 10,000 converted images, with 1,000 from each of the 10 classes of the original CIFAR-10 dataset, which includes 60,000 32x32 color images. The conversion employs a repeated closed-loop smooth (RCLS) movement of images, generating more realistic and applicable event streams. These streams possess complex

spatio-temporal structures, positioning the CIFAR-10 DVS as a moderate-level dataset in terms of complexity. This dataset supports the development of event-driven algorithms for object classification tasks, employing methodologies such as spike-based forward networks and support vector machines with bag of events (BOE) features [5].

2.3 DVS128 Gesture Recognition Dataset Overview

The DVS128 Gesture Dataset, crucial for real-time gesture recognition systems, features data recorded using a DVS128. It includes 11 hand gestures from 29 subjects under three illumination conditions. Each trial has two files: a data file (.aedat) containing DVS128 events and an annotation file (.csv) with the start and stop times of each gesture. The DVS data is stored in the AEDAT 3.1 file format as Polarity Events, where each event includes x, y coordinates and polarity.

2.4 ESD-1 and ESD-2 Dataset Overview

The ESD dataset, elucidated in [3], emerges as an expansive dataset for investigating robotic grasping scenarios. Captured using a DAVIS346 sensor mounted on a robotic arm, it integrates both conventional RGB frames and asynchronous events. The dataset is distinguished by its detailed, instance-specific annotations, categorized into six classes: bottle, box, pouch, book, mouse, and platform, comprising 17,186 annotated images and 177 event streams. Divided into ESD-1 and ESD-2, ESD-1 is tailored for known object segmentation challenges, containing 10 objects across 13,984 training images and 3,202 testing images. ESD-2, dedicated to unknown object segmentation, is solely for testing, featuring five unique objects not included in ESD-1. Both subsets exhibit variations in camera motion direction, arm speed, lighting conditions, and object clutter. These include linear, rotational, and partial-rotational motions, arm speeds of 0.15 m/s, 0.3 m/s, and 1 m/s, alongside normal and low lighting conditions. Object clutter varies from 2 to 10. To align the event data with the 40 Hz RGB frames, a temporal window of 25 ms is used, ensuring synchronization between the frame modalities.

3 Metrics

3.1 Classification Accuracy for Image classification

Classification Accuracy is a performance metric for classification tasks that measures the percentage of test data points for which the class with the highest predicted probability matches the true class label.

Mathematically, Classification Accuracy accuracy can be defined as:

$$\text{Classification Accuracy} = \frac{\text{Number of Correct Top-1 Predictions}}{\text{Total Number of Samples}} \times 100 \quad (1)$$

Here, “Number of Correct Top-1 Predictions” refers to the count of samples where the class with the highest predicted probability is the same as the true class, and “Total Number of Samples” is the size of the dataset being evaluated.

3.2 Event-wise Segmentation Accuracy

Event-wise accuracy is used to evaluate the performance of asynchronous event-based object segmentation which essentially predicts class labels for each event to its true label. This approach provides a way to evaluate object detection models based on event data and accounts for the sparsity of events:

$$Acc(d, d') = \frac{1}{N} \sum_1^N \frac{d_i}{d'_i} \times 100 \quad (2)$$

where d , d' , and N represent the ground truth event set, the predicted event set, and the total number of events respectively [4].

3.3 Homeostasis

To understand the capacity of neurons to self-regulate and maintain a consistent level of activity over time we also measure the homeostasis of the host SNNs. Specifically, we employ three statistical indicators: FR_m , FR_{m_std} , and FR_{s_std} , all of which are based on the neuron firing rate. FR_m signifies the average neuron firing rate of an SNN over all P trials. FR_{m_std} represents the mean of P standard deviations, where each deviation corresponds to the neuron firing rates of an SNN during an individual trial. Conversely, FR_{s_std} illustrates the variability of the P standard deviations. The firing rate is calculated as follows:

$$f_i^{l,p} = \frac{1}{T_p} \sum_{t=1}^{T_p} s_i^l(t_p), \quad (3)$$

where $f_i^{l,p}$ denotes the firing rate of the i -th neuron in the l -th layer during the p -th trial and T_p represents the duration of the p -th trial. FR_m^p symbolizes the average firing rate of all neurons in an SNN for the p -th trial, while FR_{std}^p corresponds to the standard deviation of the firing rates of all neurons within an SNN throughout the p -th trial. [2]

3.4 Energy Consumption

To evaluate the energy efficiency of Spiking Neural Networks (SNNs), we focus on calculating both multiply-accumulate (MAC) and accumulate (AC) operations. While MAC operations consider multiplicative interactions, AC operations in SNNs focus on accumulative interactions, especially given the binary nature of spikes. In addition to the MAC and AC operations, the energy consumption is further influenced by the average spiking activity, ζ_l . This activity represents the

ratio of the total number of spikes in layer l over all events to the total number of neurons in that layer. When dealing with asynchronous events, the focus is on the temporal sequence of events. The adapted energy formula for SNNs, incorporating both MAC and AC operations in the context of asynchronous events, is:

$$E_{\text{SNN}} = (1.69 \text{ pJ} \times \text{MAC} + 0.38 \text{ pJ} \times \text{AC}) \times \zeta_l \quad (4)$$

where MAC and AC signify the count of multiply-accumulate and accumulate operations, respectively. As per the study in [1], the coefficient 1.69pJ is energy consumed per MAC operation, while 0.38pJ is the energy for each AC operation. The term ζ_l captures the average spiking activity for layer l , shedding light on the spiking dynamics of the network in the context of asynchronous events. This term quantifies the energy expenditure associated with the sequence of spikes, especially pertinent when addressing asynchronous event-driven inputs.

4 Ablation Study

4.1 Weighting Factor Selection

Table 1: Performance Analysis for different values of k_1 , k_2 , and k_3 , on CIFAR and DVS128 datasets.

k_1	k_2	k_3	CIFAR		DVS128	
			Acc.	FR_m	Acc.	FR_m
0.05	0.15	0.05	67.39	0.841	69.34	0.872
0.15	0.25	0.15	86.71	0.625	91.02	0.713
0.25	0.5	0.25	94.74	0.352	97.83	0.421
0.5	0.75	0.5	61.82	0.936	76.48	0.947
0.75	0.85	0.75	52.01	1.62	57.01	1.707

This study aims to experimentally select the weighting factors k_1 , k_2 , and k_3 , which correspond to the three elements of the ABN. The experiment utilized the CIFAR-DVS and DVS128 datasets for gesture recognition to determine the optimal combination of k_1 , k_2 , and k_3 . This was achieved by evaluating the combinations through image classification accuracy and firing rate. It was observed that as the combination of k_1 , k_2 , and k_3 started from 0.05, 0.15, and 0.05 respectively and incrementally increased, the accuracy of image classification improved. Peak performance and the lowest firing rates were achieved at the combination of $k_1 = 0.5$, $k_2 = 0.25$, and $k_3 = 0.25$. Beyond this point, further increases in the k values led to a decrease in accuracy and a surge in firing rates. Therefore, having identified the optimal combination, it becomes crucial to examine the most influential element of the ABN, which is the focus of the subsequent subsection.

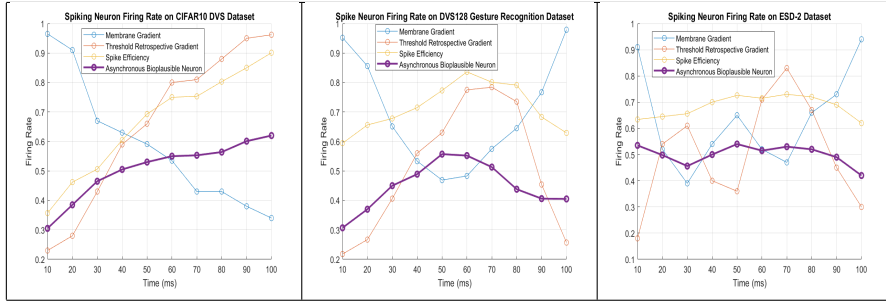


Fig. 2: Ablation Study Spike Activity Analysis of each element of the ABN

4.2 Spike Activity Analysis

The ablation study of the proposed ABN model, focused on the distinct behaviors of its constituent elements MG, TRG, and SE, observed across the CIFAR-10 DVS, ESD-2, and DVS-128 Gesture Recognition datasets, as depicted in Fig. 2. The MG demonstrated a robust increase in firing rates over time, while the TRG exhibited an inverse trend, decreasing over the same intervals. These opposing dynamics suggest that while each element responds distinctly to temporal stimuli, they collectively contribute to a regulatory effect, mitigating abrupt fluctuations and promoting threshold stability in the ABN model. The SE component further modulates this balance, optimizing the firing rates for computational efficiency. In addition, the variability in the firing rate patterns across different datasets reflects the distinct event frequencies encountered within each dataset. The CIFAR-10 DVS dataset, with its more uniform event timing, facilitates a steady increase in the MG’s influence, whereas the ESD-2’s irregular events lead to more erratic firing rates, challenging threshold stabilization. The DVS-128 Gesture Recognition dataset, with its complex event sequences, showcases the ABN model’s adaptability to rapid temporal changes. This highlights the ABN model’s capacity to modulate neural firing in response to the diverse temporal structures of sensory input.

4.3 ABN Analysis

This ablation study, as presented in Table 2, examines the individual and combined contributions of each functional element (MG, TRG, SE) to the overall performance on the CIFAR and DVS128 datasets. The results clearly indicate that all three factors are crucial, as each one independently enhances performance. Specifically, the integration of TRG demonstrates the most significant impact on performance, followed by MG and SE, in terms of their contribution to the final outcomes. Using optimal weights k_1 , k_2 , and k_3 , the model attains peak performance, with accuracies of 94.74% on CIFAR and 97.83% on DVS128. It showcases an inverse relationship between firing rate and accuracy, particularly in the full ABN model, achieving the highest accuracy at minimal firing rates of

Table 2: Ablation Study of the ABN

Feat.	Params			CIFAR		DVS128	
	k_1	k_2	k_3	Acc.	FR_m	Acc.	FR_m
MG	0.25	0	0	74.48	0.534	79.41	0.682
TRG	0	0.5	0	83.03	0.516	87.34	0.603
SE	0	0	0.25	56.35	0.832	61.4	0.854
MG+TRG	0.25	0.5	0	78.52	0.523	82.01	0.671
TRG+SE	0	0.5	0	81.3	0.481	84.01	0.601
MG+SE	0.25	0	0.25	59.26	0.704	63.28	0.853
All	0.25	0.5	0.25	94.74	0.352	97.83	0.421

0.352 on CIFAR and 0.421 on DVS128, indicating efficient spike utilization for enhanced accuracy.

5 Qualitative Evaluation

The qualitative results illustrated in the attached images demonstrate the efficacy of the proposed method in semantic segmentation and image classification tasks using specialized datasets. For semantic segmentation, the method was evaluated on the ESD1 and ESD2 datasets, showcasing its ability to distinguish and classify multiple objects within a scene. As seen in the images, the method’s performance was tested across varying complexities—from scenes with two objects to those cluttered with ten objects—as well as under challenging conditions such as movement at 1 m/s speed, rotational changes, varied distances (like 83cm), and low light environments. The ground truth images display a rich tapestry of colors, each representing a distinct class, against which the predictions are compared. The predicted segmentation closely mirrors the ground truth, with precise color-coded object delineation, even as the number of objects increases or the conditions become more adverse.

In the image classification task using the N-MNIST dataset, the method’s robustness was further validated. The dataset, a neuromorphic rendition of the traditional MNIST, presents digits as spike events, requiring the algorithm to interpret spatiotemporal data. The results are compelling, with the predicted classifications forming recognizable representations of the digits zero through nine, depicted in a scatter plot-like format that aligns well with the original dot patterns. This indicates a high degree of accuracy in the network’s ability to classify temporal patterns, a task critical for neuromorphic computing systems that process data in a manner akin to biological neural networks.

References

1. Aydin, A., Gehrig, M., Gehrig, D., Scaramuzza, D.: A Hybrid ANN-SNN Architecture for Low-Power and Low-Latency Visual Perception (3 2023)
2. Ding, J., Dong, B., Heide, F., Ding, Y., Zhou, Y., Yin, B., Yang, X.: Biologically Inspired Dynamic Thresholds for Spiking Neural Networks. *Neural Information Processing Systems* (2022)

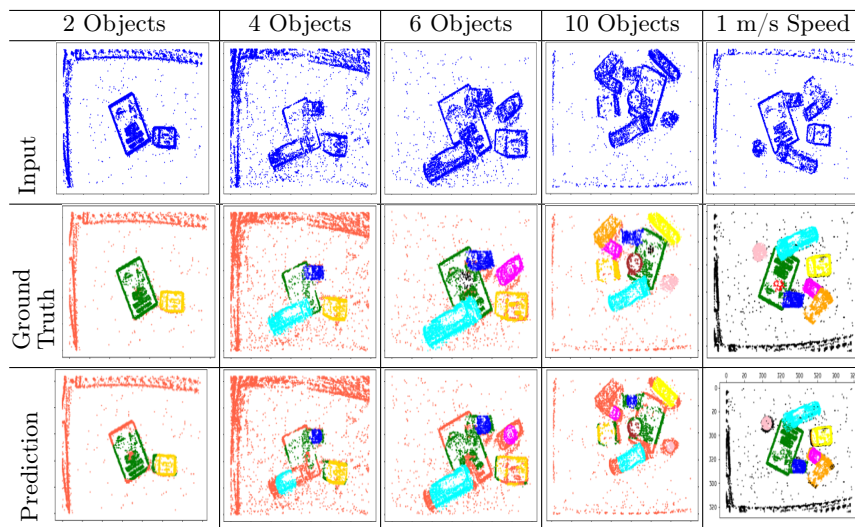


Fig. 3: Qualitative Results - Semantic Segmentation on ESD1 dataset performance of the proposed neural network across scenarios with different object counts and motion speeds. The top row shows the input point clouds, the middle row the ground truth labels, and the bottom row the network’s predictions. The semantic segmentation capabilities of the model are evident, handling from 2 to 10 objects and motion at 1 m/s with high fidelity to the ground truth.

- Huang, X., Sanket, K., Ayyad, A., Naeini, F.B., Makris, D., Zweiri, Y.: A Neuromorphic Dataset for Object Segmentation in Indoor Cluttered Environment (2 2023), <http://arxiv.org/abs/2302.06301>
- Kachole, S., Alkendi, Y., Baghaei Naeini, F., Makris, D., Zweiri, Y.: Asynchronous Events-based Panoptic Segmentation using Graph Mixer Neural Network. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 4083–4092 (2023). <https://doi.org/10.1109/CVPRW59228.2023.00429>
- Li, H., Liu, H., Ji, X., Li, G., Shi, L.: CIFAR10-DVS: An Event-Stream Dataset for Object Classification. *Frontiers in Neuroscience* **11** (5 2017). <https://doi.org/10.3389/fnins.2017.00309>
- Orchard, G., Jayawant, A., Cohen, G.K., Thakor, N.: Converting Static Image Datasets to Spiking Neuromorphic Datasets Using Saccades. *Frontiers in Neuroscience* **9** (2015). <https://doi.org/10.3389/fnins.2015.00437>, <https://www.frontiersin.org/articles/10.3389/fnins.2015.00437>

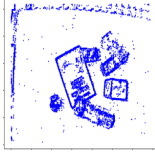
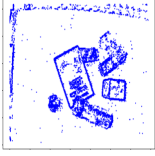
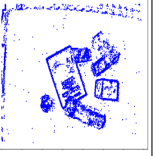
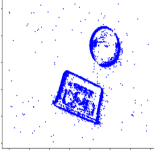
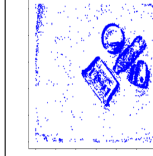
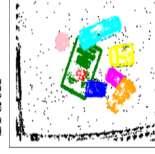

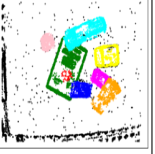
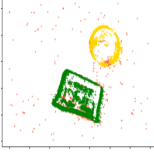
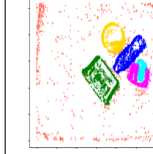
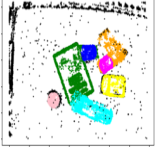
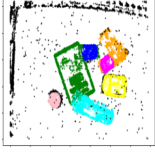
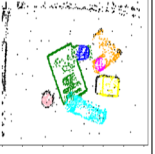
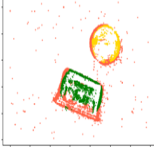
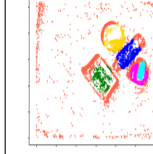
	Rotational	83cm Distance	Low Light	Unknown objects	Unknown Objects
Input					
Ground Truth					
Prediction					

Fig. 4: Qualitative Results - Robust Semantic Segmentation under Varied Conditions - Illustrated here are the segmentation results of the proposed model on the ESD1 and ESD2 datasets under diverse environmental challenges. The model’s predictions demonstrate resilience and adaptability to rotations, distance variations (83cm shown), dim lighting, and the presence of unknown objects, maintaining consistency with the ground truth across all test cases.

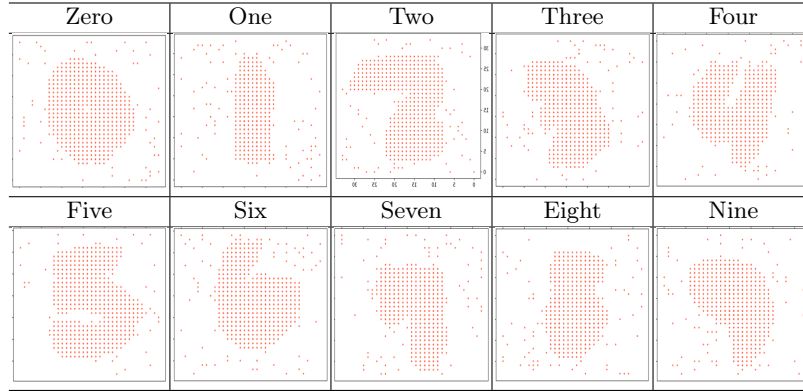


Fig. 5: Qualitative Results - Image Classification Results on N-MNIST - This figure showcases the classification performance of the proposed method on the N-MNIST dataset, where each subplot represents the spatiotemporal spike patterns associated with the digits zero through nine. The method’s ability to discern and classify the neuromorphic data into accurate digit representations is highlighted, reflecting its potential for applications in dynamic, event-driven vision systems.