

# Efficient Event-Based Object Detection: A Hybrid Neural Network with Spatial and Temporal Attention

Soikat Hasan Ahmed\*, Jan Finkbeiner†, Emre Neftci  
 Forschungszentrum Jülich, RWTH Aachen University  
 {s.ahmed, j.finkbeiner, e.neftci}@fz-juelich.de

## Abstract

*Event cameras offer high temporal resolution and dynamic range with minimal motion blur, making them promising for robust object detection. While Spiking Neural Networks (SNNs) on neuromorphic hardware are often considered for energy efficient and low latency event-based data processing, they often fall short of Artificial Neural Networks (ANNs) in accuracy and flexibility. Here, we introduce Attention-based Hybrid SNN-ANN backbones for event-based object detection to leverage the strengths of both SNN and ANN architectures. A novel Attention-based SNN-ANN bridge module captures sparse spatial and temporal relations from the SNN layer and converts them into dense feature maps for the ANN part of the backbone. Additionally, we present a variant that integrates DWConvLSTMs to the ANN blocks to capture slower dynamics. This multi-timescale network combines fast SNN processing for short timesteps with long-term dense RNN processing, effectively capturing both fast and slow dynamics. Experimental results demonstrate that our proposed method surpasses SNN-based approaches by significant margins, with results comparable to existing ANN and RNN-based methods. Unlike ANN-only networks, the hybrid setup allows us to implement the SNN blocks on digital neuromorphic hardware to investigate the feasibility of our approach. Extensive ablation studies and implementation on neuromorphic hardware confirm the effectiveness of our proposed modules and architectural choices. Our hybrid SNN-ANN architectures pave the way for ANN-like performance at a drastically reduced parameter, latency, and power budget.*

## 1. Introduction

Over the past decade, deep learning has made significant advances in object detection. State-of-the-art approaches predominantly rely on frame-based cameras which cap-

ture frames at a fixed rate. Frame cameras provide dense intensity data but have limitations in dynamic range and frame rates, leading to motion blur. Dynamic Vision Sensors (DVS), or event cameras, offer an alternative by asynchronously capturing pixel-level illumination changes, achieving low latency ( $\sim 10\mu\text{s}$ ), higher temporal resolution, and an extended dynamic range (140 dB vs. 60 dB) [12]. These characteristics make them well-suited for low-light and fast-motion scenarios. However, due to the sparse, high-temporal-resolution data they generate, effectively processing event data for object detection remains a challenging and emerging research area. Early adopters of event-based object detection ANN models often naively repurpose architectures originally designed for frame-based cameras [5, 18, 30, 35, 44]. ANN models generally achieve good accuracy but tend to be large in terms of parameter count and MAC operations, making them less suitable for deployment on power-efficient edge or neuromorphic devices. Furthermore, the high sparsity and temporal resolution is often discarded in favor of dense representations to leverage GPUs' dense, vector-based representations. In contrast, SNNs implemented in neuromorphic hardware are ideally suited to leverage the sparsity of event-based inputs, offering significant reductions in computational cost, power consumption, and latency [4, 8, 34]. However, SNNs tend to be less accurate at the task level compared to their ANN counterparts.

In this work, we create a hybrid SNN-ANN-based backbone architecture to combine the efficient, event-driven processing of SNNs on neuromorphic hardware with the efficient learning and representation capabilities of ANNs. The SNN extracts low-level features with high temporal resolution from the event-based sensors and converts them into intermediate features, which then change to slower timescales before being processed by the ANN with dense activations. Additionally, we feature a variant that adds DWConvLSTMs [14, 36] to the ANN block. This multi-timescale RNN variant combines the sparse SNN processing of short timesteps with long time horizon processing via the dense RNN with the extracted long timesteps to efficiently cap-

\*Conceptual design, algorithm development and experimentation.

†Conceptual design, hardware analysis and deployment.

ture both fast and slow dynamics of the data. For hybrid models, the SNN can be efficiently deployed at the edge in a power efficient manner, as we demonstrate with an implementation on a digital neuromorphic chip. ANN processing can occur either at the edge or, with reduced data rates, in a cloud setting. By training the network jointly, the SNN component can leverage backpropagated errors for efficient training via the surrogate gradient approach [31].

Information in SNNs is communicated by spike events. In our hybrid models, these must be efficiently converted into dense representations without discarding valuable spatiotemporal features. In our model, this is achieved by an attention-based SNN-ANN bridge module, which bridges SNN representations to ANN representations.

The attention module contains two attention modules named Event-Rate Spatial (ERS) attention and Spatial Aware Temporal (SAT) attention. The SAT attention module addresses the challenge of sparse event inputs by enhancing the model’s understanding of irregular structures and temporal attention to discern temporal relationships within the data. On the other hand, the ERS attention module focuses on highlighting spatial areas by leveraging the activity of events. Moreover, we implement the SNN blocks in digital neuromorphic hardware to demonstrate the feasibility of our approach. We report the performance of our model on large-scale event-based object detection benchmarks. The contributions of our work can be summarized as follows:

- A novel hybrid backbone-based event object detection model. To the best of our knowledge, this is the first work to propose a hybrid object detection approach for benchmark object detection task. Evaluation on the Gen1 and Gen4 Automotive Detection datasets [9, 33] shows that the proposed method outperforms SNN-based methods and achieves comparable results to ANN and RNN-based approaches.
- An attention-based SNN-ANN bridge module ( $\beta_{asab}$ ) to convert spatiotemporal spikes into a dense representation, enabling the ANN part of the network to process it effectively while preserving valuable spatial and temporal features through the Event-Rate Spatial (ERS) and Spatial-Aware Temporal (SAT) attention mechanisms.
- A multi-timescale RNN variant that includes both the high-temporal resolution SNN block followed by a slower, long time horizon DWConvLSTMs in the ANN block, operating on larger timesteps extracted via the  $\beta_{asab}$  module.
- Implementation of the SNN blocks on digital neuromorphic hardware to validate its performance and efficiency.

## 2. Related Work

Recent studies demonstrated the potential of event cameras in object detection tasks. In the earlier stages of adopting

event cameras, the focus primarily revolved around adapting existing frame-based feature extractors and a detection head for object detection using event data [5, 18]. In [18], researchers integrated event-based data into off-the-shelf frame-based object detection networks. They employed an InceptionNet-based backbone for feature extraction and a single-shot detector (SSD) for detection [29, 40]. Similarly, [33] utilized a frame-based object detection model called RetinaNet, which incorporates a spatial pooling-based feature extractor [27] along with a detection head, applied to event data.

Additionally, methods such as [14, 25, 33] have incorporated recurrent neural networks (RNNs) as feature extractors for event data. [48] uses SSM to improve training time, and [42] proposes a training schema with efficient ground truth label utilization. [30] introduced sparse convolution as a method for event feature extraction. To address the challenges of efficiently extracting spatiotemporal features, [35] investigates the usability of a graph neural network-based approach as a feature extractor.

Recently, SNN-based methods have become popular for event data processing due to their spike-based working principle, similar to event cameras, which enables efficient processing. Research conducted by [6] and [39] showcases the effective utilization of SNNs in object detection tasks. Specifically, [6] and [39] delve into assessing the performance of converting widely-used ANN-based backbone architectures such as SqueezeNet [19], VGG [37], MobileNet [16], DenseNet [17], and ResNet [15] into SNN architecture for event-based object detection. Nonetheless, optimizing intermediate and high-level features for detection with SNNs results in a significant drop in accuracy.

Recognizing the distinct advantages offered by both SNNs and ANNs, researchers have explored merging these networks into hybrid architectures [23]. For instance, [47] presents a framework that leverages hierarchical information abstraction for meta-continual learning with interpretable multimodal reasoning. Building on this idea, [43] introduces DashNet, which integrates SNNs with ANN-based feature extraction for high-speed object tracking. Similarly, [28] improves SNN performance through a hybrid top-down attention mechanism, while [24] demonstrates that hybrid models can achieve energy-efficient optical flow estimation with enhanced robustness. Complementing these advances, [2, 23] develops an architecture that fuses SNN backbones with ANN heads for event-based vision tasks. By leveraging the complementary strengths of each, these hybrid networks show promise for simpler tasks. However, the bridge between SNNs and ANNs is still overlooked to harness the best of both worlds.

Moreover, the full extent of their capabilities remains largely unexplored, especially in tackling state-of-the-art benchmark vision tasks, such as object detection on popu-

lar datasets like Gen1 Automotive Detection dataset [9] and Gen4 Automotive Detection dataset [33].

### 3. Hybrid Object Detection Network

The overall hybrid network as shown in Figure 1 comprises two key parts: an attention-based hybrid backbone designed to extract spatio-temporal features, and detection heads tasked with identifying objects. In the following section, we will delve into the details of the core components of the network.

#### 3.1. Event Representation

An event is represented as  $e_n = (x_n, y_n, t_n, p_n)$ , where  $(x_n, y_n)$  is the pixel location,  $t_n$  is the time, and  $p_n$  is polarity which indicates the change in light intensity (i.e., positive or negative). The event data is pre-processed to convert it into a time-space tensor format. Following [14], we start by creating a 4D tensor  $Events[t_{k-1}, t_k] \in \mathbb{R}^{T \times 2 \times H \times W}$ , where  $T$  represents number of time discretization steps, 2 denotes polarity features which contain the count of positive and negative events in each discretized time step, and  $H$  and  $W$  signify the height and width of the event camera, respectively. Given the event set  $\mathcal{E} = \{e_1, e_2, \dots, e_N\}$ , the event tensor  $Events[t_{k-1}, t_k]$  is constructed from the discretized time variable  $t'_n = \left\lceil \frac{t_n - t_a}{t_b - t_a} \cdot T \right\rceil$  as follows:

$$Events[t_{k-1}, t_k](t, p, x, y) = \sum_{e_n \in \mathcal{E}} \delta(p - p_n) \delta(x - x_n) \delta(y - y_n) \delta(t - t'_n). \quad (1)$$

While training, event tensors are created. However, during inference, given an input sparsity of  $\sim 98\%$  (for Gen 1), this results in significant efficiency gains due to sparse processing in neuromorphic hardware compared to the dense processing in a GPU.

#### 3.2. Attention-based Hybrid Backbone

The proposed hybrid backbone architecture, as shown in Figure 1, consists of three fundamental components: a low-level spatio-temporal feature extractor  $f_{snn}$ , an ANN-based high-level spatial feature extractor  $f_{ann}$ , and a novel Attention-based SNN-ANN Bridge (ASAB) module  $\beta_{asab}$ .

The first module, denoted as  $f_{snn}$ , is an event-level feature extractor operating in the spatio-temporal domain and consists of multiple convolutional SNN blocks. Each block follows a structured sequence of operations: standard convolution, batch normalization [20], and Parametric Leaky Integration and Fire (PLIF) spiking neuron [10]. The neural dynamics of PLIF with trainable time constant  $\tau = \text{sigmoid}(w)^{-1}$  given input  $X[t]$  can be expressed as

follows:

$$V[t] = V[t-1] + \frac{1}{\tau}(X[t] - (V[t-1] - V_{reset})). \quad (2)$$

The  $f_{snn}$  module receives  $Events[t_{k-1}, t_k]$  as its input and generates events  $\mathbf{E}_{spike} = f_{snn}(Events[t_{k-1}, t_k]) \in \mathbb{R}^{T \times C \times H' \times W'}$ . As SNNs operate on a faster timescale and utilize sparse representations and ANNs operate on dense representations, efficiently translating valuable spatio-temporal information into dense representations is essential. To achieve this translation, the  $\mathbf{E}_{spike}$  is subsequently fed into a proposed  $\beta_{asab}$  module, which bridges the SNN and the ANN parts. The events  $\mathbf{E}_{spike}$  are converted into dense, non-binary features while preserving spatial and temporal information in the form of spatial feature maps. The output of  $\beta_{asab}$  is represented by  $\mathbf{F}_{out} = \beta_{asab}(\mathbf{E}_{spike})$ , with dimensions  $C \times H' \times W'$  which is compatible with traditional 2D convolution-based networks, allowing for smooth processing and integration of information across both spatial and temporal dimensions. The attention module is further described in Section 3.3.

The third component,  $f_{ann}$ , extracts high-level spatial features using multiple ANN blocks with standard ANN components. Each ANN block consists of standard convolution operations, normalization [3, 20], and ReLU activation functions, enabling the extraction of detailed high-level spatial features from the densely encoded  $\mathbf{F}_{out}$ .

In addition to the proposed model, we explore a variant that features an added RNN module, incorporating two Depth-Wise separable Convolutional LSTM (DWConvLSTM) units similar to those in [14], as illustrated in Figure 1.

The  $f_{snn}$  processes fast dynamics with small timesteps from the event-based camera, while the DWConvLSTM operates on larger timesteps extracted from the  $\beta_{asab}$ -module to capture slower dynamics. The resulting outputs from the ANN blocks are then fed to the detection head for the final object detection output.

#### 3.3. Attention-based SNN-ANN Bridge Module

The bridge module  $\beta_{asab}$  comprises two attention modules: i) Spatial-aware Temporal (SAT) attention and ii) Event-Rate Spatial (ERS) attention. The SAT attention module dynamically captures local spatial context within the irregular spatial spike-structure to uncover temporal relations. Meanwhile, the ERS attention submodule focuses on attending to spatial areas utilizing the spatial event activities. Below, we describe these two submodules.

##### 3.3.1 Spatial-aware Temporal (SAT) Attention

The SAT attention contains three crucial operations: i) Channel-wise Temporal Grouping to group relevant fea-

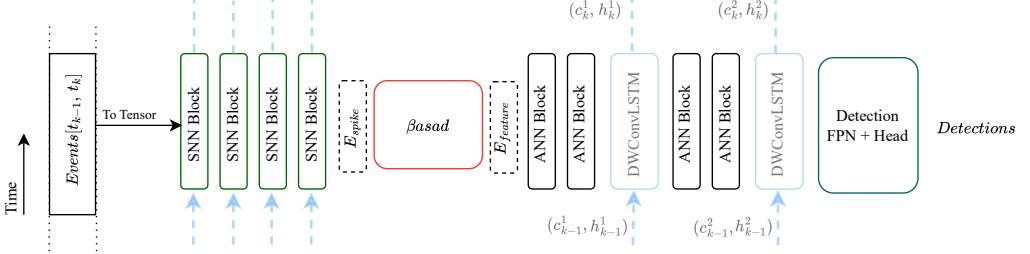


Figure 1. Architecture of the hybrid model with object detection head and SNN-ANN hybrid backbone, including the SNN part,  $\beta_{asab}$  bridge module and ANN part. The DWConvLSTM modules and dashed blue arrows are only part of the proposed hybrid + RNN variant.

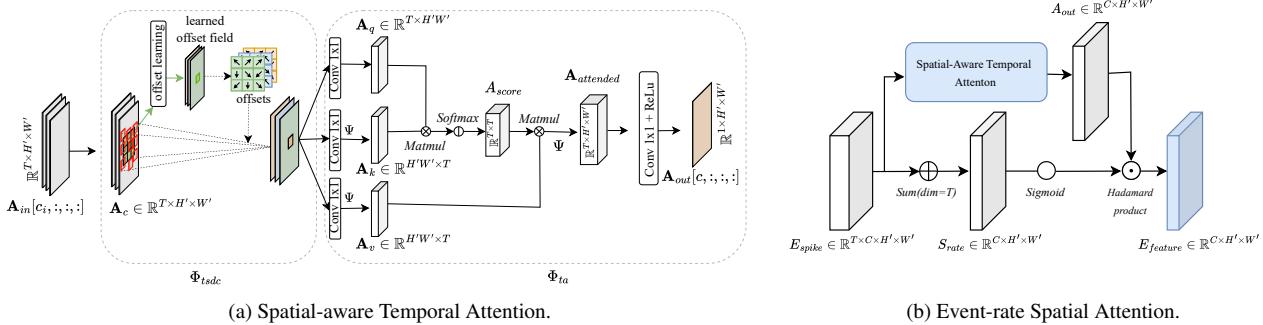


Figure 2. Visualization of the proposed attention module’s components. (a) Spatial-aware Temporal Attention: Highlights relevant temporal features in spatial regions to enhance temporal coherence in event-based data. (b) Event-rate Spatial Attention: Emphasizes spatial regions based on event rates, allowing for adaptive focus on areas with significant event activity. Together, these components improve feature extraction in spatiotemporal data processing.

tures from different time dimensions, ii) Time-wise Separable Deformable Convolution (TSDC) denoted as  $\Phi_{tsdc}$  for capturing channel-independent local spatial context from sparse spike features, and iii) Temporal attention module  $\Phi_{ta}$ , which uses local spatial context features to extract temporal relations to accumulate and translate temporal information into spatial information.

**Time-wise Separable Deformable Convolutions (TSDC):** At first, we apply the channel-wise temporal grouping operation to the input data so that each feature channel is processed separately while capturing spatial and temporal relations. This operation transforms the input spike tensor  $E_{spike} \in \mathbb{R}^{T \times C \times H' \times W'}$  into  $A_{in} \in \mathbb{R}^{C \times T \times H' \times W'}$ .

As shown in Figure 2a, the  $\Phi_{tsdc}$  operation operates on individual channel-wise temporal groups  $A_{in}[c_i, :, :, :]$ , denoted as  $A_c$ , where  $i$  is the channel index. This operation extracts the local spatial context of the sparse, irregularly shaped spike-based representations. We posit that this irregular representation is better extracted using a deformed kernel rather than a standard square grid kernel, as discussed in Section 4.3. We implemented the TSDC as a time-wise separable convolution to capture spatial details independently

of time-based changes, as motion occurs over time. Isolating spatial aspects enables a clearer understanding of the structure and layout of features, separate from their movement.

The time-wise separated spatial context is then passed to the  $\Phi_{ta}$  module for further processing to determine the temporal relation of different time dimensions.

For the implementation of TSDC, we utilize deformable convolution introduced by [7] which adjusts sampling points on the standard grid by dynamically predicting kernel offsets based on input features. During training, an additional convolutional layer called “offset learning” (refer to Figure 2a) is trained to predict these offsets. Moreover, to independently process each temporal dimension, we set the group of deformable convolution kernels equal to the number of time steps  $T$ . This is done to encourage the network to focus on the spatial context of the data while maintaining temporal relations intact for further processing.

**Temporal Attention:** To learn relationships between different time steps, we pass the local spatial context  $A_{sc} = \Phi_{tsdc}(A_c)$  through a temporal attention module. This module leverages the multi-head softmax self-attention mechanism introduced in [41]. In our case, we apply self-attention

along the temporal dimension to extract temporal relations.

Firstly, we calculate the keys, queries, and values for temporal self-attention by employing  $1 \times 1$  convolutions, followed by a reshape operation, which we denote as  $\mathbf{A}_k$ ,  $\mathbf{A}_q$ , and  $\mathbf{A}_v$ , respectively. These operations output tensors of shapes  $\mathbb{R}^{H'W' \times T}$ ,  $\mathbb{R}^{T \times H'W'}$ , and  $\mathbb{R}^{T \times H'W'}$ , respectively.

$$\mathbf{A}_k = \omega_k(A_{sc}), \quad \mathbf{A}_q = \Psi(\omega_q(A_{sc})), \quad \mathbf{A}_v = \Psi(\omega_v(A_{sc})) \quad (3)$$

Where  $\Psi$  denotes the reshape operation and  $\omega$  denotes the  $1 \times 1$  convolution operation. Next, the temporal attention scores denoted as  $\mathbf{A}_{score} \in \mathbb{R}^{T \times T}$  are computed by performing matrix multiplication between  $\mathbf{A}_q$  and  $\mathbf{A}_k$ , followed by applying a softmax operation:

$$\mathbf{A}_{score} = \text{softmax}(\mathbf{A}_q \mathbf{A}_k). \quad (4)$$

To obtain the attended temporal features,  $A_v$  is multiplied with  $A_{score}$ , followed by a reshape operation to output  $A_{attended} \in \mathbb{R}^{T \times H' \times W'}$ :

$$\mathbf{A}_{attended} = \Psi(\mathbf{A}_v \mathbf{A}_{score}). \quad (5)$$

Finally, a weighted-sum along the temporal dimension using a  $1 \times 1$  convolution produces the output  $\mathbf{A}_{out}[c, :, :] \in \mathbb{R}^{H' \times W'}$ . This operation effectively combines the attended temporal features to produce the final output.

### 3.3.2 Event-rate Spatial Attention

This attention module extracts spatial correlation as spatial weights, utilizing dynamic event activity from intermediate spikes generated by the  $f_{snn}$  module. To identify active regions, an Event-rate Spatial Attention mechanism takes the input  $\mathbf{E}_{spike}$ , and sums the time dimension to calculate the event rates  $\mathbf{S}_{rate}$ , resulting in a shape of  $\mathbb{R}^{C \times H' \times W'}$ :  $\mathbf{S}_{rate} = \sum_{t=1}^T E_{spike}(t, :, :, :)$ . The  $\mathbf{S}_{rate}$  is first normalized using a sigmoid function to provide a spatial attention score based on the event activity. This attention score is then utilized as a weight to adjust the output of the SAT module through a Hadamard product, as visualized in Figure 2b:

$$\mathbf{E}_{feature} = \text{sigmoid}(\mathbf{S}_{rate}) \odot \mathbf{A}_{out} \quad (6)$$

The resulting tensor  $F_{out}$  is then fed into ANN blocks, which are subsequently utilized to predict the object detection bounding box by a detection head [13].

## 4. Experiments

### 4.1. Setup

**Datasets:** To conduct the training and evaluation of our network, we utilized two event-based object detection datasets: Gen1 Automotive Detection dataset [9] and Gen4 Automotive Detection dataset [33]. The Gen1 and Gen4

datasets comprise 39 and 15 hours of event camera recordings at a resolution of  $304 \times 240$  and  $720 \times 1280$ , respectively, with bounding box annotations for car, pedestrian, and two-wheeler (Gen4 only) classes.

**Implementation Details:** The model is implemented in PyTorch [32] with the SpikingJelly library [11] and trained end-to-end for 50 epochs on the Gen 1 dataset and 10 epochs on the Gen 4 dataset. The ADAM optimizer [22] is used with a OneCycle learning rate schedule [38], which decays linearly from a set maximum learning rate. The kernel size for  $\Phi_{tsdc}$  is set to 5. The training pipeline incorporates data augmentation methods such as random horizontal flips, zoom, and crop, based on [14]. Event representations for the SNN are constructed from 5 ms bins. During training, object detections are generated every 50 ms, using the SNN’s output from the last 10 time bins, while inference allows higher temporal resolution, bounded by the SNN timestep. The YOLOX framework [13] is used for object detection, incorporating IOU loss, class loss, and regression loss. For the Gen 1 dataset, models are trained with a batch size of 24 and a learning rate of  $2 \times 10^{-4}$ , requiring approximately 8 hours on four 3090 GPUs. On the Gen 4 dataset, the batch size is 8 with a learning rate of  $3.5 \times 10^{-4}$ , taking around 1.5 days on four 3090 GPUs.

When using an RNN variant, we follow previous methods with a sequence length of 21 for fair comparison. This RNN-based network, trained for 400,000 steps with a batch size of 2, requires approximately 6 days to complete training.

### 4.2. Benchmark comparisons

**Comparison Design** To the best of our knowledge, this work presents the first hybrid object detection model implemented in large-scale benchmark datasets, rendering comparisons to other work challenging. Therefore, we design our comparison in three setups - (i) comparison with existing ANN-based methods (ii) comparison with SNN-based object detection methods, and (iii) comparison with RNN-based models.

**Evaluation Procedure:** Following the evaluation protocol established in prior studies [6, 14, 33], the mean average precision (mAP) [26] is used as the primary evaluation metric to compare the proposed methods’ effectiveness with existing approaches. Since most methods do not offer open-source code, the reported numbers from the corresponding papers were used.

**Comparison design with ANN-based methods:** The efficacy of the proposed method was evaluated against ANN-based models. The results presented in Table 1 provide

Table 1. Comparative analysis of various ANN-based models for event-based object detection on the Gen1 [9] and Gen4 [33] Automotive Detection datasets, where mAP denotes mAP(.5:.05:.95).  $A^*$  suggests that this information was not directly available and estimated based on the publication.

Models	Type	Params	Gen 1	Gen 4
			mAP	mAP
AEGNN [35]	GNN	20M	0.16	-
SparseConv [30]	ANN	133M	0.15	-
Inception + SSD [18]	ANN	$> 60M^*$	0.3	0.34
RRC-Events [5]	ANN	$> 100M^*$	0.31	0.34
Events-RetinaNet [33]	ANN	33M	0.34	0.18
E2Vid-RetinaNet [33]	ANN	44M	0.27	.25
RVT-B W/O LSTM [14]	Transformer	$16.2M^*$	0.32	-
<b>Proposed</b>	Hybrid	6.6M	0.35	.27

Table 2. Comparative analysis of various SNN-based models for event-based object detection on the Gen1 Automotive Detection dataset .

Models	Type	Params	mAP
VGG-11+SDD [6]	SNN	13M	0.17
MobileNet-64+SSD [6]	SNN	24M	0.15
DenseNet121-24+SSD [6]	SNN	8M	0.19
FP-DAGNet[45]	SNN	22M	0.22
EMS-RES10 [39]	SNN	6.20M	0.27
EMS-RES18 [39]	SNN	9.34M	0.29
EMS-RES34 [39]	SNN	14.4M	0.31
SpikeFPN [46]	SNN	22M	0.22
<b>Proposed</b>	Hybrid	6.6M	0.35

Table 3. Comprehensive evaluation of different RNN-based models for event-based object detection tasks on the Gen1 Automotive Detection dataset. Here 'TF' denotes Transformer.

Models	Type	Params	mAP
S4D-ViT-B [48]	TF + SSM	16.5M	0.46
S5-ViT-B [48]	TF + SSM	18.2M	0.48
S5-ViT-S [48]	TF + SSM	9.7M	0.47
RVT-B [14]	TF + RNN	19M	0.47
RVT-S [14]	TF + RNN	10M	0.46
RVT-T [14]	TF + RNN	4M	0.44
ASTMNet [25]	(T)CNN + RNN	100M	0.48
RED [33]	CNN + RNN	24M	0.40
<b>Proposed+RNN</b>	Hybrid + RNN	7.7M	0.43

a compelling comparison of various ANN-based networks

and performance on the event-based Gen 1 dataset. Notably, the Proposed hybrid model stands out with only 6.6M parameters, significantly smaller than other models such as SparseConv (133M) and RRC-Events (100M). Despite its compact size, our proposed model achieves an accuracy of 0.35, outperforming larger models like SparseConv, which achieves 0.15, and closely matching the performance of Events-RetinaNet (33M, 0.34). In contrast, for the Gen 4 dataset, the analysis also includes architectures like Events-RetinaNet and E2Vid-RetinaNet. Events-RetinaNet achieves a lower mean Average Precision (mAP) of 0.18, while E2Vid-RetinaNet performs slightly better with an mAP of 0.25. The AEGNN model, which utilizes a graph neural network approach, achieves an mAP of 0.16 with 20 million parameters; however, its performance is overshadowed by the proposed hybrid model, which achieves an mAP of 0.27 while maintaining a compact size of only 6.6 million parameters. We observe that larger models tend to achieve higher accuracy due to their increased parameter counts. However, their significant size makes them less suitable for deployment on hardware with limited resources, such as edge devices or neuromorphic systems.

**Comparison Design with SNN-based Methods:** The proposed hybrid method was compared against several SNN-based methods, specifically, VGG-11+SDD [6], MobileNet-64+SSD [6], DenseNet121-24+SSD [6], FP-DAGNet[45], EMS-RES10 [39], EMS-RES18 [39], EMS-RES34 [39] and SpikeFPN [46].

Table 2 displays the comparison results with various state-of-the-art SNN-based object detection methods. More specifically, VGG-11+SDD and MobileNet-64+SSD achieve mAP values of 0.17 and 0.15 with parameter counts of 13M and 24M, respectively. DenseNet121-24+SSD, with a smaller parameter size of 8M, slightly outperforms these models with an mAP of 0.19. FP-DAGNet

and SpikeFPN, both at 22M parameters, attain an mAP of 0.22. The EMS-RES series showcases incremental improvements, with EMS-RES10 (6.2M) achieving 0.27, EMS-RES18 (9.34M) at 0.29, and EMS-RES34 (14.4M) achieving the highest mAP among SNNs at 0.31. In contrast, the Proposed Hybrid Model surpasses all these SNN models with an mAP of 0.35 while maintaining an efficient parameter size of only 6.6M. This superiority can be attributed to our method’s incorporation of a hybrid feature extraction approach with both spatial and temporal attention modules, which are lacking in other methods.

#### Comparison design with RNN and SSM-based methods:

Although the performance of the RNN-based models generally outperforms models with spiking components, this comparison aims to investigate how the proposed hybrid model is comparable to the RNN. We provide comparisons with RNN-based methods such as RED [33], ASTMNet [25], RVT-B [14], RVT-S [14], RVT-T [14], S4D-ViT-B [48], S5-ViT-B [48], and S5-ViT-S [48]. Table 3 presents a comparison of results obtained with RNN-based models. It can be seen that two of the works, RED [33] and ASTM-Net [25], have substantially larger parameter counts and are therefore expected to perform better. RVT [14] demonstrates good accuracy at a parameter count comparable to the proposed hybrid network. Our fully recurrent backbone allows high-frequency detection without re-computing recurrent layers, unlike RVT’s non-causal attention which requires re-computation for every prediction. CNNs are more efficient than MLPs at small batch sizes due to higher arithmetic intensity; in our method, less than 5% of MACs are from attention and MLPs versus 67% in RVT which is harder to deploy on energy-efficient edge and neuromorphic hardware.

#### 4.3. Ablation Study

Table 4. Ablation study for ASAB module.

Models	mAP(.5)	mAP
Variant 1(w/o - $\Phi_{ta}$ )	0.57	0.33
Variant 2 (w/o deform)	0.59	0.34
Variant 3 (w/o - ESA)	0.59	0.34
Variant 4 (w/o - ASAB)	0.53	0.30
<b>Variant 5 (Proposed)</b>	<b>0.61</b>	<b>0.35</b>

**ASAB module:** The ablation study highlights the importance of each ASAB module component in enhancing model accuracy on the Gen1 Automotive Detection dataset. In Table 4, Variant 1 (excluding  $\Psi_{ta}$ ) achieves an mAP of

0.33, revealing reduced temporal capture. In variant 2, we replaced the deformable convolution with a standard convolution, which shows irregular sampling helps with sparse data. In variant 3, removing the ERS module shows some accuracy drop, indicating limited spatial flexibility and attention. Variant 4, replacing ASAB with a simple accumulation operation, results in the lowest mAP of 0.30. The complete model (Variant 5) reaches the highest mAP of 0.35, emphasizing the value of each component. Figure 3 illustrates how the bridge module enhances detection by reducing false predictions. Additionally, we performed an ablation study on various DWConvLSTM configurations by toggling layers. The proposed setting achieved a mAP of 0.43. Please refer to the supplementary (Table 8) for more details .

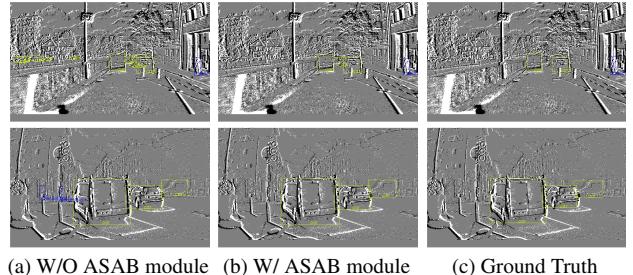


Figure 3. Visual comparison of object detection outputs between the baseline hybrid event object detection method (left) and the proposed method (right) for the Gen 4 dataset. From left to right: (a) object detection output without the ASAB module, (b) object detection output with the ASAB module, and (c) ground truth object boundaries. More samples are in the Supplementary document.

## 5. Hardware Implementation, Energy and Computational Efficiency Analysis

### 5.1. Hardware Implementation

Table 5. Power and time measurements of the SNN block on Loihi 2 for several input sizes and number of weight bits. The power is measured in Watts and the execution time per step is in milliseconds. The mean and standard deviation of the measurements averaged over 12 inputs for a total of 100k steps are reported.

Weight quant.	# chips	Power [W]	Time/Step
int8	6	$1.73 \pm 0.10$	2.06
int6	6	$1.71 \pm 0.11$	2.06
int4	6	$1.95 \pm 0.33$	1.16

In order to demonstrate the suitability of the chosen hybrid SNN-ANN approach for energy efficient inference on

the edge, we implemented the SNN backbone in hardware. In the proposed architecture the SNN block transforms sensor data into intermediate representations and therefore underlies the strictest latency requirements. Due to the clear separation between SNN and ANN parts in the model’s architecture, the SNN blocks can be implemented in specialized hardware. As hardware, we chose Intel’s Loihi 2 [8], a digital, programmable chip based on event-based communication and computation. Only minor adjustments are necessary for execution on Loihi 2: The kernel-weights of the convolutional layers are quantized to `int8` via a per-output-channel quantization scheme showing no resulting loss in accuracy (mAP: 0.348 (`float16`) vs 0.343 (`int8`)). The batchnorm (BN) operations and quantization scaling are fused into the LIF-neuron dynamics by scaling and shifting the inputs according to the following equations:

$$\text{scale} = \frac{q_{\text{scale}} \text{ weight}_{\text{BN}}}{\tau \sqrt{\text{Var}_{\text{BN}} + \varepsilon_{\text{BN}}}} \quad (7)$$

$$\text{shift} = (\text{bias}_{\text{conv}} - \text{mean}_{\text{BN}}) \frac{\text{weight}_{\text{BN}}}{\tau \sqrt{\text{Var}_{\text{BN}} + \varepsilon_{\text{BN}}}} + \frac{\text{bias}_{\text{BN}}}{\tau} \quad (8)$$

where  $q_{\text{scale}}$  is the scaling factor introduced by the quantization and  $\tau$  is the PLIF neurons time constant. Given this approach, spike times are almost exactly reproduced on Loihi 2 compared to the PyTorch `int8` implementation. For benchmarking purposes, the inputs to the network are simulated with an additional neuron population, due to the current IO limitations of the chip. With this approach the spiking statistics in the input and SNN layers are reproduced.

Table 5 reports power and time measurement results of the 4-layer SNN block running on Loihi 2 for inputs of size (2, 256, 160). The network runs at  $(1.7 \pm 0.1)$  W and  $(1.9 \pm 0.8)$  ms per step, which is faster than real-time in the currently chosen architecture (5 ms per step). These results compare favorably to commercially available chips for edge computing like the NVIDIA Jetson Nano (7 W – 15 W) [1] and demonstrate the suitability of an SNN backbone for event-based data processing.

## 5.2. Computational Analysis

We trained a variant of our model where the modified PLIF neuron acts like a ReLU with proposed attention module to investigate a comparison between a similar artificial neural network (ANN) and the hybrid network. This variant, *Baseline<sub>ann</sub>*, with a mean  $15.34 \times 10^9$  multiply-accumulate operations (MACs), makes it resource-intensive and unsuitable for energy-constrained hardware. In contrast, our Proposed hybrid model computes only  $1.63 \times$

This research was funded by the German Federal Ministry of Education and Research (BMBF) under the projects "GreenEdge-FuE" (16ME0521) and "Cluster4Future" (03ZU1106CB). Access to JUWELS [21] has been granted by GCS ([www.gauss-centre.eu](http://www.gauss-centre.eu)) under project neuroml. We thank Intel for access to Loihi 2.

Table 6. Comparison of different baselines complexities.

Models	mAP(.5)	MACs	ACs
<i>Baseline<sub>ann</sub></i>	0.61	$15.34e9$	0.0
<i>Baseline<sub>w/o β<sub>asab</sub></sub></i>	0.53	$1.18e9$	$0.97e9$
<b>Proposed<sub>w/β<sub>asab</sub></sub></b>	0.61	<b><math>1.63e9</math></b>	<b><math>0.97e9</math></b>
<i>Proposed<sub>snn+</sub></i>	0.58	$0.87e9$	$1.59e9$

$10^9$  MACs, making it more practical for edge devices. The  $\beta_{\text{asab}}$  module does incur additional operations but leads to significant accuracy improvement (compare to *Baseline<sub>w/o β<sub>asab</sub></sub>*). Additionally, the spiking neural network (SNN) variant, *Proposed<sub>snn+</sub>* (Increasing one SNN layer and reducing one ANN layer), reduces computational demands further to  $0.87 \times 10^9$  MACs and is highly efficient on neuromorphic hardware, running significantly faster and with less energy on Intel’s Loihi 2 compared to the dense-activation ANNs.

We analyze power consumption across methods following [2]. Among SNNs, DenseNet121+SSD uses 0.9 mJ (0.0 MACs,  $2.3 \times 10^9$  ACs), while VGG-11+SSD requires 4.2 mJ ( $11.1 \times 10^9$  ACs). In ANNs, Inception+SSD is the most demanding at 19.3 mJ ( $11.4 \times 10^9$  MACs). Events-RetinaNet consumes 5.4 mJ ( $3.2 \times 10^9$  MACs), and RVT-B W/O LSTM requires 3.9 mJ ( $2.3 \times 10^9$  MACs). Our method achieves  $1.6 \times 10^9$  MACs,  $1.0 \times 10^9$  ACs, and 3.1 mJ, significantly reducing energy and computational costs compared to most ANNs.

## 6. Conclusion

In this work, we introduced a hybrid attention-based SNN-ANN backbone for event-based visual object detection. A novel attention-based SNN-ANN bridge module is proposed to capture sparse spatial and temporal relations from the SNN layer and convert them into dense feature maps for the ANN part of the backbone. Additionally, we demonstrate the effectiveness of combining RNNs on multiple timescales: hardware-efficient SNNs for fast dynamics on short timescales with ConvLSTMs for longer timescales that operate on the extracted features of the bridge-module. Experimental results demonstrate that our proposed method surpasses baseline hybrid and SNN-based approaches by significant margins, with results comparable to existing ANN-based methods. The efficacy of our proposed modules and architectural choices is confirmed through extensive ablation studies. Additionally, we demonstrate the effectiveness of our architectural choice with separate SNN and ANN blocks by implementing the SNN blocks on digital neuromorphic hardware, Intel’s Loihi 2. The neuromorphic hardware implementation achieves sub-real-time processing and improved power consumption compared to commercially available edge computing hardware. The achieved accuracy and hardware

implementation results pave the way toward a hybrid SNN-ANN architecture that achieves ANN-like performance at a drastically reduced parameter and power budget.

## References

- [1] Nvidia jetson orin nano datasheet, revision 4. <https://openzeka.com/wp-content/uploads/2023/03/jetson-orin-nano-datasheet-r4-web.pdf>, 2023. Accessed: 2025-03-09. 8
- [2] Asude Aydin, Mathias Gehrig, Daniel Gehrig, and Davide Scaramuzza. A hybrid ann-snn architecture for low-power and low-latency visual perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2024. 2, 8
- [3] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 3
- [4] Francesco Barchi, Gianvito Urgese, Alessandro Siino, Santa Di Cataldo, Enrico Macii, and Andrea Acquaviva. Flexible on-line reconfiguration of multi-core neuromorphic platforms. *IEEE Transactions on Emerging Topics in Computing*, 9(2):915–927, 2019. 1
- [5] Nicholas FY Chen. Pseudo-labels for supervised learning on dynamic vision sensor data, applied to object detection under ego-motion. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 644–653, 2018. 1, 2, 6
- [6] Loïc Cordone, Benoît Miramond, and Philippe Thierion. Object detection with spiking neural networks on automotive event data. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2022. 2, 5, 6
- [7] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017. 4
- [8] Mike Davies, Narayan Srinivasa, Tsung-Han Lin, Gautham Chinya, Yongqiang Cao, Sri Harsha Choday, Georgios Dimou, Prasad Joshi, Nabil Imam, Shweta Jain, et al. Loihi: A neuromorphic manycore processor with on-chip learning. *Ieee Micro*, 38(1):82–99, 2018. 1, 8
- [9] Pierre De Tournemire, Davide Nitti, Etienne Perot, Davide Migliore, and Amos Sironi. A large scale event-based detection dataset for automotive. *arXiv preprint arXiv:2001.08499*, 2020. 2, 3, 5, 6
- [10] Wei Fang, Zhaofei Yu, Yanqi Chen, Timothée Masquelier, Tiejun Huang, and Yonghong Tian. Incorporating learnable membrane time constant to enhance learning of spiking neural networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2661–2671, 2021. 3
- [11] Wei Fang, Yanqi Chen, Jianhao Ding, Zhaofei Yu, Timothée Masquelier, Ding Chen, Liwei Huang, Huihui Zhou, Guoqi Li, and Yonghong Tian. Spikingjelly: An open-source machine learning infrastructure platform for spike-based intelligence. *Science Advances*, 9(40):eadi1480, 2023. 5
- [12] Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J Davison, Jörg Conradt, Kostas Daniilidis, et al. Event-based vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(1):154–180, 2020. 1
- [13] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021. 5, 2
- [14] Mathias Gehrig and Davide Scaramuzza. Recurrent vision transformers for object detection with event cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13884–13893, 2023. 1, 2, 3, 5, 6, 7
- [15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2
- [16] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 2
- [17] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 2
- [18] Massimiliano Iacono, Stefan Weber, Arren Glover, and Chiara Bartolozzi. Towards event-driven object detection with off-the-shelf deep learning. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1–9. IEEE, 2018. 1, 2, 6
- [19] Forrest N Iandola, Song Han, Matthew W Moskewicz, Khalid Ashraf, William J Dally, and Kurt Keutzer. Squeezeenet: Alexnet-level accuracy with 50x fewer parameters and 0.5 mb model size. *arXiv preprint arXiv:1602.07360*, 2016. 2
- [20] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. pmlr, 2015. 3
- [21] Jülich Supercomputing Centre. JUWELS Cluster and Booster: Exascale Pathfinder with Modular Supercomputing Architecture at Juelich Supercomputing Centre. *Journal of large-scale research facilities*, 7(A138), 2021. 8
- [22] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5
- [23] Alexander Kugele, Thomas Pfeil, Michael Pfeiffer, and Elisabetta Chicca. Hybrid snn-ann: Energy-efficient classification and object detection for event-based vision. In *DAGM German Conference on Pattern Recognition*, pages 297–312. Springer, 2021. 2
- [24] Chankyu Lee, Adarsh Kumar Kosta, Alex Zihao Zhu, Kenneth Chaney, Kostas Daniilidis, and Kaushik Roy. Spikeflownet: event-based optical flow estimation with energy-efficient hybrid neural networks. In *European Conference on Computer Vision*, pages 366–382. Springer, 2020. 2

- [25] Jianing Li, Jia Li, Lin Zhu, Xijie Xiang, Tiejun Huang, and Yonghong Tian. Asynchronous spatio-temporal memory network for continuous event-based object detection. *IEEE Transactions on Image Processing*, 31:2975–2987, 2022. [2](#), [6](#), [7](#)
- [26] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V I3*, pages 740–755. Springer, 2014. [5](#)
- [27] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. [2](#)
- [28] Faqiang Liu and Rong Zhao. Enhancing spiking neural networks with hybrid top-down attention. *Frontiers in Neuroscience*, 16:949142, 2022. [2](#)
- [29] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I I4*, pages 21–37. Springer, 2016. [2](#)
- [30] Nico Messikommer, Daniel Gehrig, Antonio Loquercio, and Davide Scaramuzza. Event-based asynchronous sparse convolutional networks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII I6*, pages 415–431. Springer, 2020. [1](#), [2](#), [6](#)
- [31] E. O. Neftci, H. Mostafa, and F. Zenke. Surrogate gradient learning in spiking neural networks: Bringing the power of gradientbased optimization to spiking neural networks. *IEEE Signal Processing Magazine*, 36(6):5163, 2019. [2](#)
- [32] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. [5](#)
- [33] Etienne Perot, Pierre De Tournemire, Davide Nitti, Jonathan Masci, and Amos Sironi. Learning to detect objects with a 1 megapixel event camera. *Advances in Neural Information Processing Systems*, 33:16639–16652, 2020. [2](#), [3](#), [5](#), [6](#), [7](#)
- [34] Jun Sawada, Philipp Akopyan, Andrew S Cassidy, Brian Taba, Michael V Debole, Pallab Datta, Rodrigo Alvarez-Icaza, Arnon Amir, John V Arthur, Alexander Andreopoulos, et al. Truenorth ecosystem for brain-inspired computing: scalable systems, software, and applications. In *SC’16: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 130–141. IEEE, 2016. [1](#)
- [35] Simon Schaefer, Daniel Gehrig, and Davide Scaramuzza. Aenn: Asynchronous event-based graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12371–12381, 2022. [1](#), [2](#), [6](#)
- [36] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28, 2015. [1](#)
- [37] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. [2](#)
- [38] Leslie N Smith and Nicholay Topin. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial intelligence and machine learning for multi-domain operations applications*, pages 369–386. SPIE, 2019. [5](#)
- [39] Qiaoyi Su, Yuhong Chou, Yifan Hu, Jianing Li, Shijie Mei, Ziyang Zhang, and Guoqi Li. Deep directly-trained spiking neural networks for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6555–6565, 2023. [2](#), [6](#)
- [40] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. [2](#)
- [41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. [4](#)
- [42] Ziyi Wu, Mathias Gehrig, Qing Lyu, Xudong Liu, and Igor Gilitschenski. Leod: Label-efficient object detection for event cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16933–16943, 2024. [2](#)
- [43] Zheyu Yang, Yujie Wu, Guanrui Wang, Yukuan Yang, Guoqi Li, Lei Deng, Jun Zhu, and Luping Shi. Dashnet: A hybrid artificial and spiking neural network for high-speed object tracking. *arXiv preprint arXiv:1909.12942*, 2019. [2](#)
- [44] Jason Yik, Soikat Hasan Ahmed, Zergham Ahmed, Brian Anderson, Andreas G Andreou, Chiara Bartolozzi, Arindam Basu, Douwe den Blanken, Petrut Bogdan, Sander Bohte, et al. Neurobench: Advancing neuromorphic computing through collaborative, fair and representative benchmarking. *arXiv preprint arXiv:2304.04640*, 2023. [1](#)
- [45] Hu Zhang, Liziwei Leng, Kaiwei Che, Qian Liu, Jie Cheng, Qinghai Guo, Jiangxing Liao, and Ran Cheng. Automotive object detection via learning sparse events by temporal dynamics of spiking neurons. *arXiv preprint arXiv:2307.12900*, 2023. [6](#)
- [46] Hu Zhang, Yanchen Li, Liziwei Leng, Kaiwei Che, Qian Liu, Qinghai Guo, Jianxing Liao, and Ran Cheng. Automotive object detection via learning sparse events by spiking neurons. *IEEE Transactions on Cognitive and Developmental Systems*, 2024. [6](#)
- [47] Rong Zhao, Zheyu Yang, Hao Zheng, Yujie Wu, Faqiang Liu, Zhenzhi Wu, Lukai Li, Feng Chen, Seng Song, Jun Zhu, et al. A framework for the general design and computation of hybrid neural networks. *Nature communications*, 13(1):3427, 2022. [2](#)

- [48] Nikola Zubic, Mathias Gehrig, and Davide Scaramuzza. State space models for event cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5819–5828, 2024. [2](#), [6](#), [7](#)

# Efficient Event-Based Object Detection: A Hybrid Neural Network with Spatial and Temporal Attention

## Supplementary Material

### Network Architecture

Table 9 displays the network architecture of the proposed hybrid backbone. Additionally, Figure 4 illustrates the basic SNN and ANN blocks. For better understanding, we have also included an additional figure (Figure 5) to clarify the channel-wise temporal grouping operation from Spatial-aware Temporal attention, where similar features from different time dimensions are grouped. In Figure 6, we show additional visual detection outputs from Gen 1 dataset.

### Hardware Implementation

This section provides more performance details on the hardware implementation of the spiking layers of the backbone on a digital neuromorphic chip, Intel’s Loihi 2 [8]. In Table 10, we present the Power and time measurements of the SNN block on Loihi 2 for various input sizes, each tested with different weight quantization settings. In Table 7, effects of accuracy due to the quantization of the weights for the SNN blocks, aimed at making them compatible with neuromorphic hardware.

### Channel-wise Temporal Grouping:

Consider a scenario where two objects are moving in different directions within a scene. Since event cameras detect changes in light intensity, most events captured by the event camera in this scenario will be triggered by the edges of these objects. Additionally, due to their movements over time, there will be spatial shifting, as illustrated in Figure 5a, denoted as  $S^{t1}$  and  $S^{t2}$ . To extract low-level features from these events, a feature extractor similar to  $f_{snn}$  processes the spatio-temporal spikes  $E_{toy}$ , learning various features such as edges and structures across multiple channels. For illustration, consider two moving features: one round

Table 7. This table shows the effects of accuracy due to the quantization of the weights for the SNN blocks, aimed at making them compatible with neuromorphic hardware.

Models	mAP(.5)	mAP(.5:.05:.95)
Variant 1 (float16)	0.613	0.348
Variant 2 (int8)	0.612	0.349
Variant 3 (int6)	0.612	0.348
Variant 4 (int4)	0.610	0.347
Variant 5 (int2)	0.432	0.224

Table 8. Ablation study for DWConvLSTM module in Gen 1 dataset.

Models	L5	L6	L7	L8	mAP
Variant 1	✓	✓	✓	✓	0.42
Variant 2	✗	✓	✓	✓	0.42
<b>Proposed+RNN</b>	✗	✓	✗	✓	<b>0.43</b>

and another lightning-shaped features, as shown in Figure 5. We would like to group together events that are produced by one object. This can be accomplished by transposing the  $C$  and  $T$  dimensions. We call this procedure channel-wise temporal grouping. Note that the input and the output of the feature extractor from the input are simplified in the figure for easier understanding.

### Effect of number of SNN blocks

In Table 11, we show additional experiments to analyze the effect of a number of SNN blocks. Three network variants were examined to assess the impact of different SNN and ANN layer numbers in the proposed architecture. The feature extractor comprised eight layers. Variant 1 decreased SNN layers and increased ANN layers in the 3–4 setup, resulting in a slight performance boost with mAP(0.75) rising from 0.34 to 0.35. Variant 3, increasing SNN layers in the 5–3 setup, led to reduced accuracy across all metrics due to fewer ANN layers to extract high-level features. Variant 2, utilizing the 4–4 setting, balanced between the two, achieving comparable accuracy to Variant 1 with reduced computational overhead from additional ANN blocks. Hence, this configuration was adopted for all subsequent experiments.

### Effect of DWConvLSTMs:

The ablation study examines configurations of the DWConvLSTM [14, 36] module in the backbone hybrid architecture. In Table 8, Variant 1, with all layers (L5-L8), sets a baseline mAP of 0.42. Variant 2, omitting L5 but keeping L6-L8, also achieves an mAP of 0.42, showing L5’s minimal impact. The Proposed+RNN variant, excluding L5 and L7 while adding an RNN with L6 and L8, reaches the highest mAP of 0.43 in the Gen 1 dataset. The RNN variant shows a similar tendency on the Gen 4 dataset (from 0.27 to 0.34 mAP).

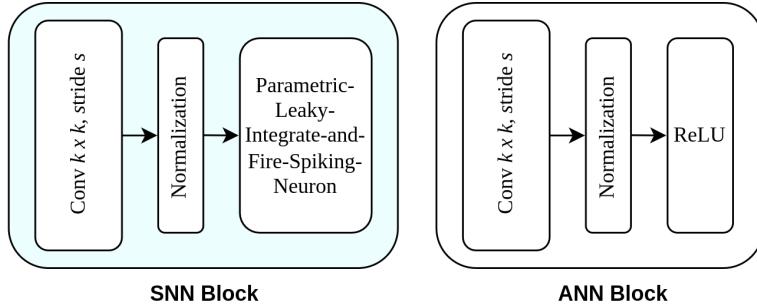


Figure 4. Basic SNN and ANN blocks.

Table 9. Hybrid + RNN architecture with DWConvLSTM.

Layer	Kernel	Output Dimensions	Layer Type
Input	-	$T \times 2 \times H \times W$	
1	64c3p1s2	$T \times 64 \times \frac{1}{2}H \times \frac{1}{2}W$	SNN Layers
2	128c3p1s2	$T \times 128 \times \frac{1}{4}H \times \frac{1}{4}W$	
3	256c3p1s2	$T \times 256 \times \frac{1}{8}H \times \frac{1}{8}W$	
4	256c3p1s1	$T \times 256 \times \frac{1}{8}H \times \frac{1}{8}W$	
5	-	$256 \times \frac{1}{8}H \times \frac{1}{8}W$	$\beta_{asab}$
6	256c3p1s1	$256 \times \frac{1}{8}H \times \frac{1}{8}W$	ANN Layers
7	256c3p1s2	$256 \times \frac{1}{16}H \times \frac{1}{16}W$	
8	-	$256 \times \frac{1}{16}H \times \frac{1}{168}W$	DWConvLSTM
9	256c3p1s1	$256 \times \frac{1}{16}H \times \frac{1}{16}W$	
10	256c3p1s2	$256 \times \frac{1}{32}H \times \frac{1}{32}W$	
11	-	$256 \times \frac{1}{8}H \times \frac{1}{8}W$	DWConvLSTM
Detection Head YoloX[13]			

Table 10. Power and time measurements of the SNN block on Loihi 2 for several input sizes and number of weight bits. The power is measured in Watts and the execution time per step in milliseconds. The mean and standard deviation of the measurements averaged over 12 inputs for a total of 100k steps are reported.

Input size (C,W,H)	Weight quantization	Number of chips	Total Power [W]	Execution Time Per Step [ms]
(2, 256, 160)	int8	6	$1.73 \pm 0.10$	$2.06 \pm 0.74$
	int6	6	$1.71 \pm 0.11$	$2.06 \pm 0.74$
	int4	6	$1.95 \pm 0.33$	$1.16 \pm 0.49$
(2, 128, 160)	int8	4	$1.51 \pm 0.56$	$0.80 \pm 0.32$
	int6	4	$1.81 \pm 0.48$	$0.45 \pm 0.18$
	int4	3	$1.39 \pm 0.44$	$0.49 \pm 0.18$

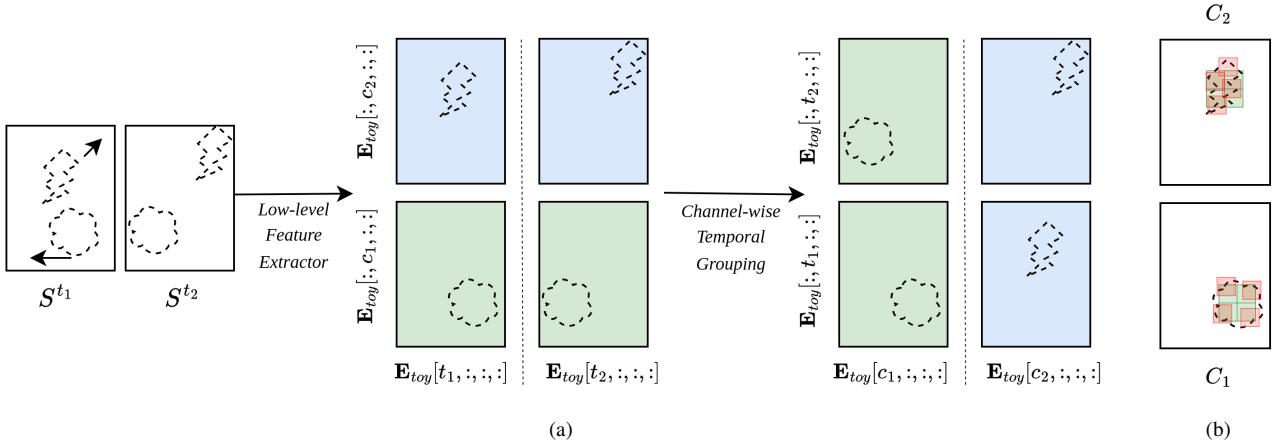


Figure 5. Toy examples: Figure 5a illustrates the channel-wise temporal grouping operation. In Figure 5b, deformed kernels highlighted in red are compared with a regular grid marked in green. For visualization, a  $2 \times 2$  kernel is shown as an example.

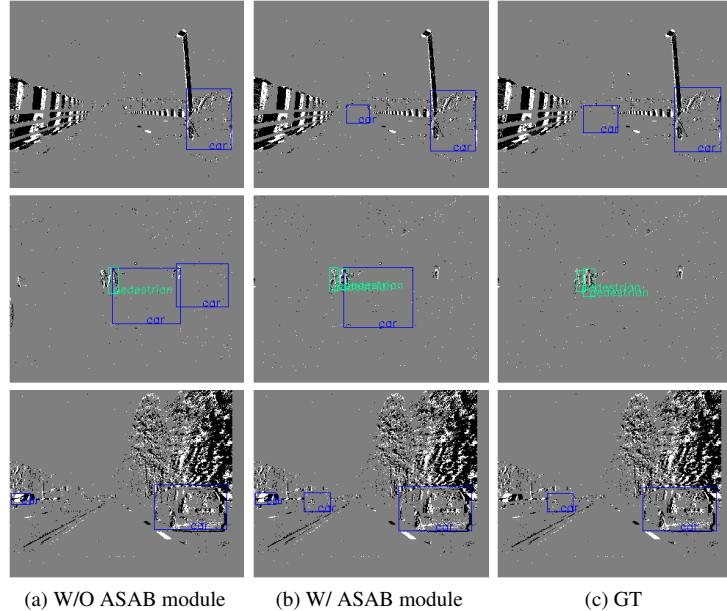


Figure 6. Visual comparison with the baseline hybrid event object detection method for the Gen 1. From left to right, (a) object detection output of the without proposed bridge module, (b) object detection output of the proposed method (with proposed bridge module), and (c) ground-truth (GT) object boundaries. These visualizations demonstrate that our proposed method significantly improves the detection of smaller objects and mitigates false predictions.

Table 11. Ablation study for different settings of the number of SNN and ANN blocks. SNN - ANN represents the number of SNN and ANN blocks in the network.

Models	SNN - ANN	mAP(.5)	mAP(.75)	mAP(.5:.05:.95)
Variant 1	3 - 5	0.61	0.35	0.35
<b>Variant 2(proposed)</b>	4 - 4	0.61	0.34	0.35
Variant 3	5 - 3	0.58	0.33	0.33