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

Supplementary Material

Network Architecture

Table 3 displays the network architecture of the proposed hybrid backbone. Additionally, Figure 1 illustrates the basic SNN and ANN blocks. For better understanding, we have also included an additional figure (Figure 2) to clarify the channel-wise temporal grouping operation from Spatial-aware Temporal attention, where similar features from different time dimensions are grouped. In Figure 3, 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 [**?**]. In Table 4, 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 1, 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 2a, 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 $\mathbf{E}_{toy}$, learning various features such as edges and structures across multiple channels. For illustration, consider two moving features: one round and another lightning-shaped features, as shown in Figure 2. 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.

Table 1. 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 2. Ablation study for DWConvLSTM module in Gen 1 dataset.

Models	L5	L6	L7	L8	mAP
Variant 1	✓	✓	✓	✓	0.42
Variant 2	×	✓	✓	✓	0.42
Proposed+RNN	×	✓	×	✓	**0.43**

Effect of number of SNN blocks

In Table 5, we shows 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 [**?** **?**] module in the backbone hybrid architecture. In Table 2, 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).

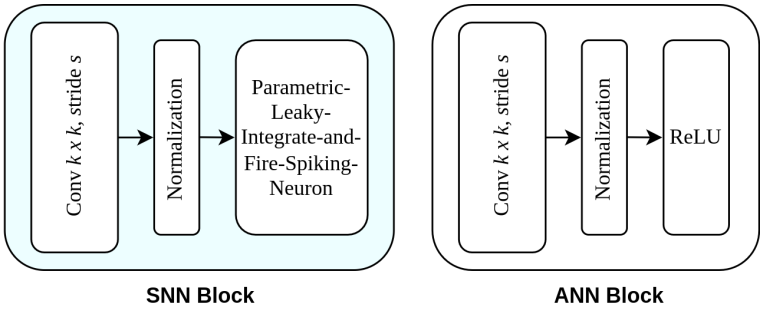


Figure 1. Basic SNN and ANN blocks.

Table 3. Hybrid + RNN architecture with DWConvLSTM.

Layer	Kernel	Output Dimensions	Layer Type
Input	-	$T \times 2 \times H \times W$	
1	$64c3p1s2$	$T \times 64 \times \frac{1}{2}H \times \frac{1}{2}W$	
2	$128c3p1s2$	$T \times 128 \times \frac{1}{4}H \times \frac{1}{4}W$	SNN Layers
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	$256c3p1s1$	$256 \times \frac{1}{8}H \times \frac{1}{8}W$	
7	$256c3p1s2$	$256 \times \frac{1}{16}H \times \frac{1}{16}W$	ANN Layers
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[**?**]			

Table 4. 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 qunatization	Number of chips	Total Power [W]	Execution Time Per Step [ms]
(2, 256, 160)	int8	6	1.73 ± 0.10	2.06 ± 0.74
	int6	6	1.71 ± 0.11	2.06 ± 0.74
	int4	6	1.95 ± 0.33	1.16 ± 0.49
(2, 128, 160)	int8	4	1.51 ± 0.56	0.80 ± 0.32
	int6	4	1.81 ± 0.48	0.45 ± 0.18
	int4	3	1.39 ± 0.44	0.49 ± 0.18

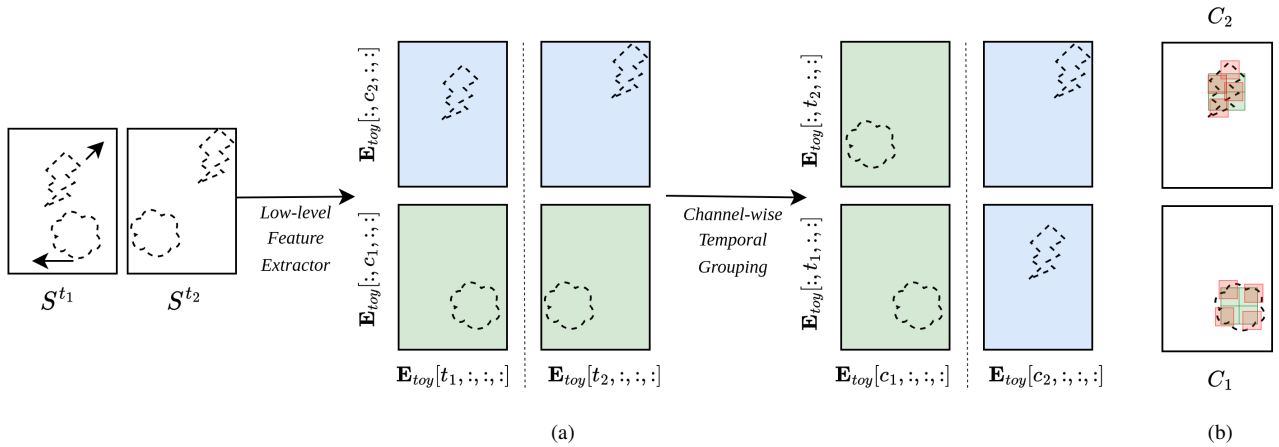
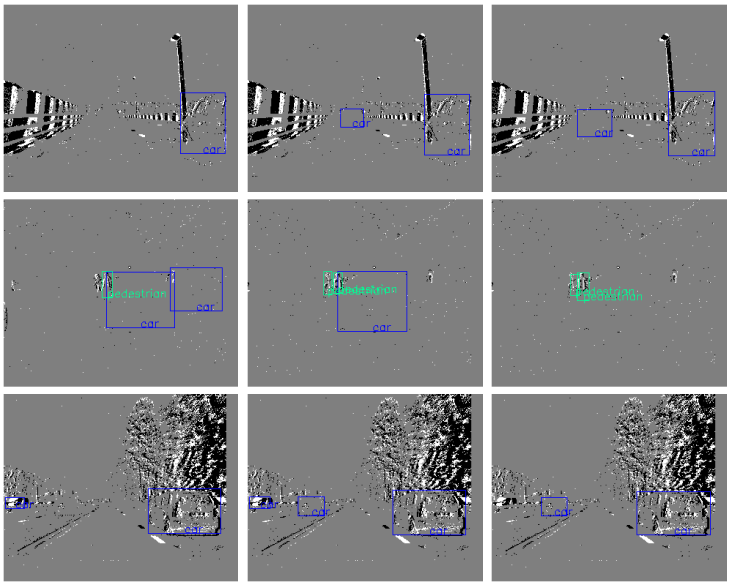


Figure 2. Toy examples: Figure 2a illustrates the channel-wise temporal grouping operation. In Figure 2b, deformed kernels highlighted in red are compared with a regular grid marked in green. For visualization, a 2×2 kernel is shown as an example.



(a) W/O ASAB module (b) W/ ASAB module (c) GT

Figure 3. 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 5. 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
Variant 2(proposed)	4 - 4	0.61	0.34	0.35
Variant 3	5 - 3	0.58	0.33	0.33