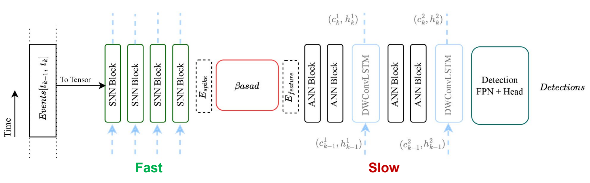
Efficient Event-Based Object Detection: A Hybrid Neural Network with Spatial and Temporal Attention

Soikat Hasan Ahmed*, Jan Finkbeiner†, Emre Neftci
Forschungszentrum Jülich, RWTH Aachen University

Introduction

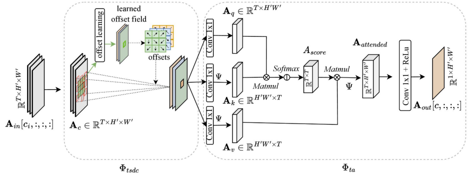
- Hybrid Event Object Detector: First hybrid SNN-ANN model for benchmark event-based object detection.
- β_{ASAB} Bridge Module: Attention-based module converting spikes to dense features via ERS and SAT.
- Multi-Timescale RNN: Combines fast SNN and slow DWConvLSTM for temporal feature learning.
- Neuromorphic Deployment: SNN blocks validated on digital neuromorphic hardware for efficiency.

Overall Network



- Architecture of the hybrid model featuring an object detection head and an SNN-ANN **hybrid backbone**, which includes the SNN block, the $\boldsymbol{\beta}_{\mathbf{ASAB}}$ bridge module, and the ANN block. The DWConvLSTM modules and dashed blue arrows are specific to the hybrid + RNN variant.

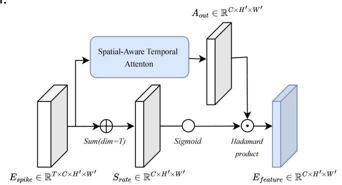
Spatial-aware Temporal Attention (SAT)



- Channel-wise Temporal Grouping to group together temporally relevant features for better temporal understanding
- Time-wise Separable Deformable Convolution for spatial context, and
- Temporal Attention to translate temporal cues into spatial features.

Event-rate Spatial Attention (ESA)

- Computes event rates by summing spikes over the time dimension.
- Normalizes event rates using a sigmoid function to create spatial attention scores.
- Uses these scores to weight the SAT module output via element-wise multiplication.



Datasets

- We train our hybrid network end-to-end using Prophesee Gen1 and Gen4 automotive event camera datasets.
- Gen1 (39 hrs, 304×240) and Gen4 (15 hrs, 720×1280) provide annotated events for cars, pedestrians, and two-wheelers (Gen4 only).

Quantitative Results

Models	Type	Params	Gen 1 mAP	Gen 4 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	-
Proposed	Hybrid	6.6M	0.35	.27

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
Proposed	Hybrid	6.6M	0.35

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
Proposed+RNN	Hybrid + RNN	7.7M	0.43

- The proposed hybrid model achieves higher accuracy than SNNs and matches ANN/RNN models with lower power and latency.

Neuromorphic Hardware Implementation

- The SNN-blocks of hybrid model was deployed on **Intel's Loihi 2 neuromorphic chip**, leveraging its event-based architecture for energy-efficient inference.
- Convolutional weights were quantized at different levels using a per-output-channel scheme, revealing negligible accuracy loss.
- Spike dynamics and BatchNorm were fused into LIF neuron behavior for efficient deployment, with q_{scale} as the quantization scaling factor and τ as the PLIF neuron time constant.

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

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

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

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

Models	Gen 1/Gen 4 mAP	MACs / ACs	Energy [mJ]
VGG-11+SDD	0.17 / -	0.0 / $11.1e9$	4.2
MobileNet-64+SSD	0.15 / -	0.0 / $4.3e9$	1.6
DenseNet121+SSD	0.19 / -	0.0 / $2.3e9$	0.9
Inception + SSD	0.3 / 0.34	$11.4e9^*$ / 0.0	19.3
Events-RetinaNet	0.34 / 0.18	$3.2e9^*$ / 0.0	5.4
E2Vid-RetinaNet	0.27 / .25	$> 3.2e9^*$ / 0.0	> 5.4
RVT-B W/O LSTM	0.32 / -	$2.3e9$ / 0.0	3.9
Proposed	0.35 / .27	1.6e9 / $1.0e9$	3.1

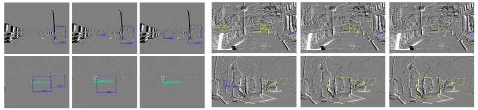
Ablation study

- An in-depth ablation study was conducted for each component of the proposed **ASAB** module, along with various configurations of the hybrid architecture.

Models	mAP(.5)	mAP
Variant 1(w/o - Φta)	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
Variant 5 (Proposed)	0.61	**0.35**

Models	mAP(.5)	MACs	ACs
$Baseline_{ann}$	0.61	$15.34e9$	0.0
$Baseline_{w/o\,\beta_{asab}}$	0.53	$1.18e9$	$0.97e9$
$\mathbf{Proposed_{w/\beta_{asab}}}$	0.61	**1.63e9**	**0.97e9**
$Proposed_{snn+}$	0.58	$0.87e9$	$1.59e9$

Visual Results



From left to right: Without ASAB, With ASAB, and Ground Truth. The first three columns correspond to the Prophesee GEN1 dataset, and the last three to the GEN4 dataset.

Scan for Details