TRIP: **T**rainable **R**egion-of-**I**nterest **P**rediction for Hardware-Efficient Neuromorphic Processing on Event-based Vision

Cina Arjmand, Yingfu Xu, Kevin Shidqi, Alexandra F Dobrita, Kanishkan Vadivel, Paul Detterer, Manolis Sifalakis, Amirreza Yousefzadeh, and Guangzhi Tang

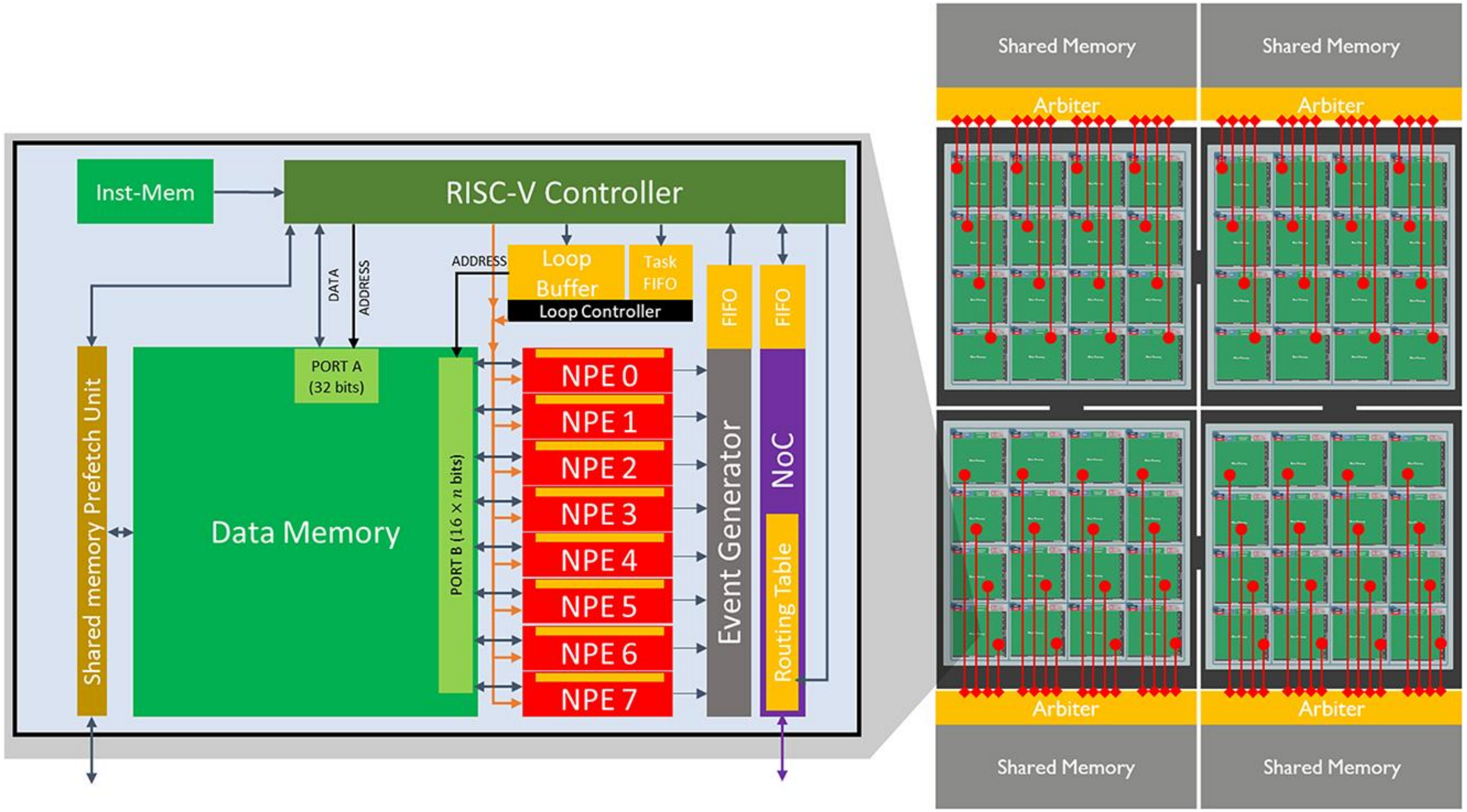
Guangzhi Tang

guangzhi.tang@maastrichtuniversity.nl

DACS, Maastricht University

International Conference on Neuromorphic Systems (ICONS 2024)

SENECA - Multicore Digital Neuromorphic Processor



Tang, et al., *SENECA: building a fully digital neuromorphic processor, design trade-offs and challenges,* Front. Neurosci., 2023

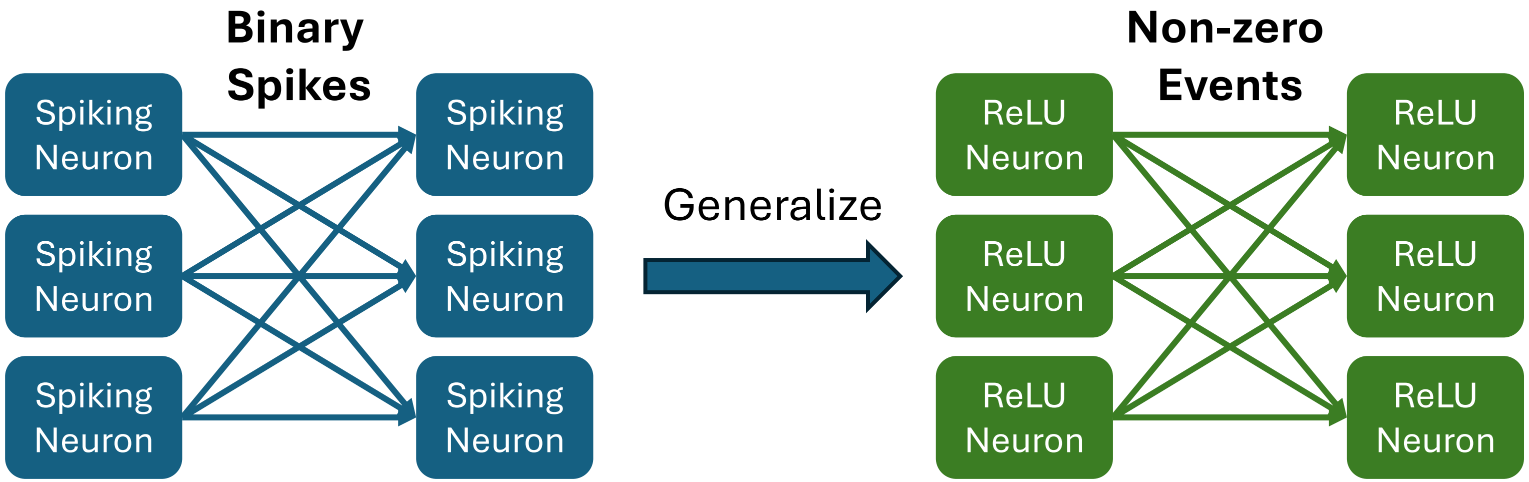
Generalize digital neuromorphic processing to **compete** with efficient deep learning

Flexible: RISC-V controller and fully programable neural processing

Scalable: Multicasting NoC and core-to-core asynchrony

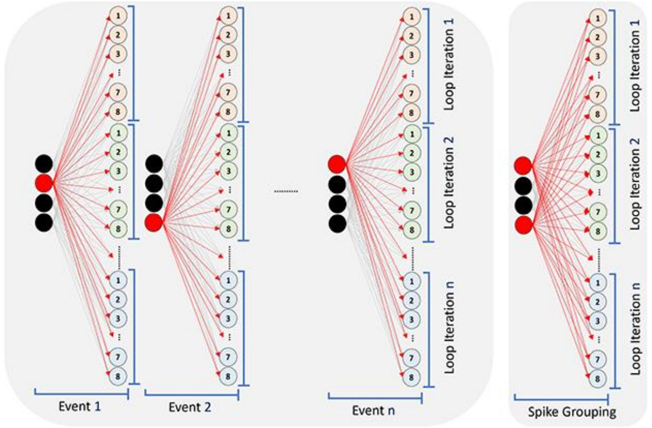
Efficient: Multiplexing and unified programable hierarchical memory

From Spiking Networks to **Event-based Neural Networks**

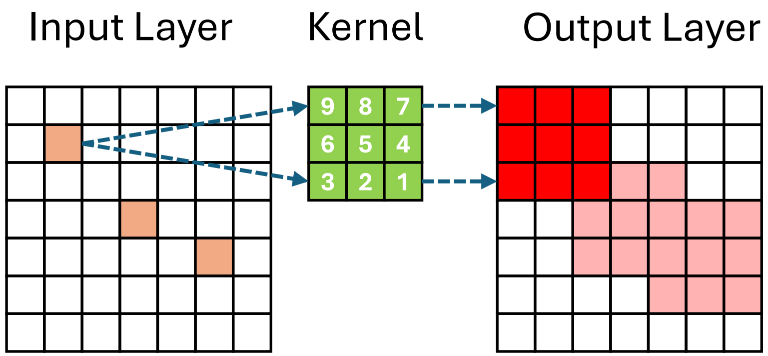


- **Efficiency**: Data movements dominate
- **Data Rate**: Address codes dominate
- **Improved Capability**: Higher accuracy

Spike Grouping for Data Reuse

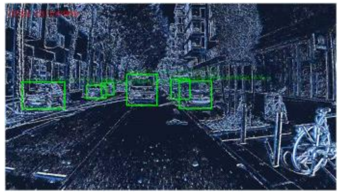


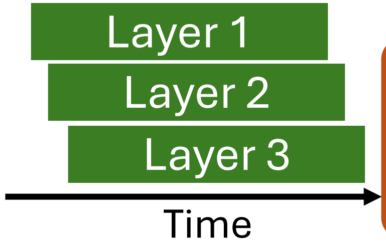
Sparse Event-driven CNN



Sparse Depth-first CNN

- Spike ASAP
- Low memory
- Low latency





Lower area, latency and energy than SNN

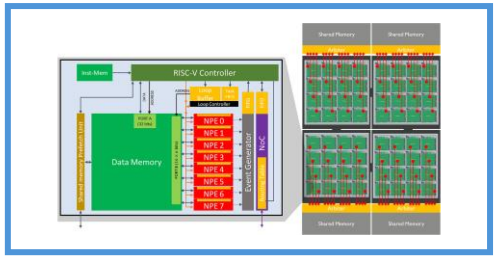
Xu, et al., *Optimizing event-based neural networks on digital neuromorphic architecture: a comprehensive design space exploration,* Front. Neurosci., 2024

Neuromorphic Processing on Event-based Vision

Event-based Camera

Sparse Events

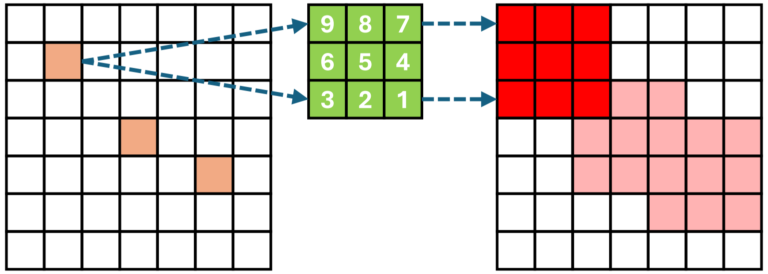
Neuromorphic Processor



- **Exploit input sparsity**
- **Low latency and energy efficient**

Event-driven Convolution

Input Layer Kernel Output Layer

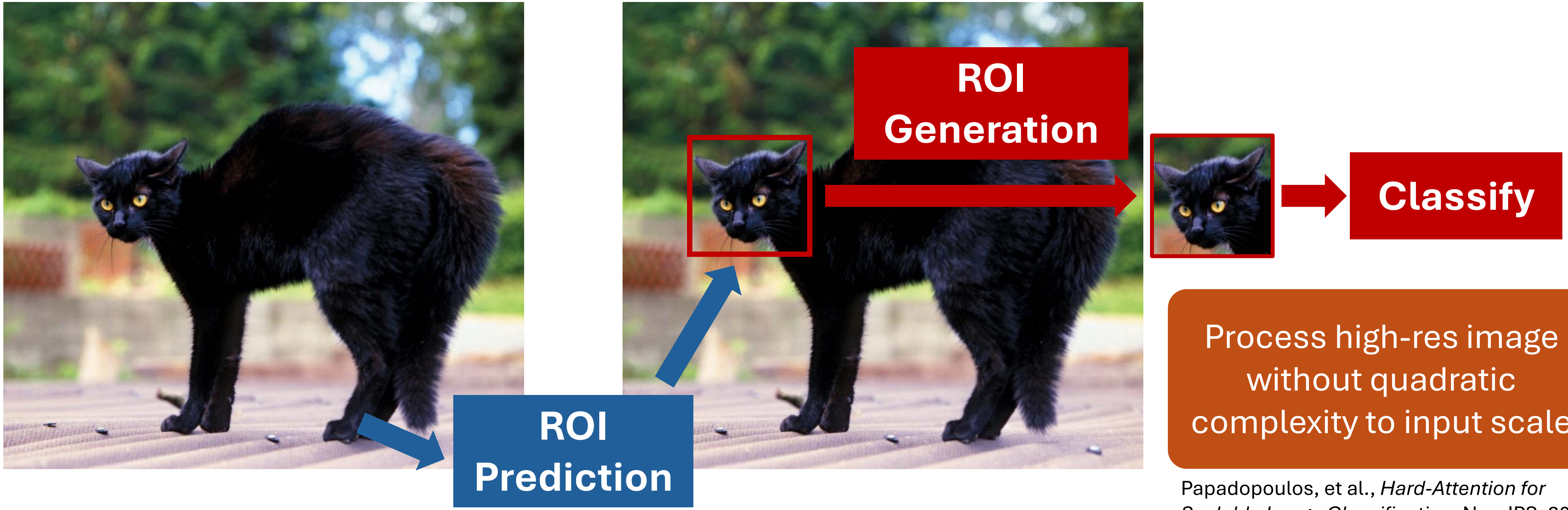


Challenge 1: Memory Cost
- Large neural state memory
- High area cost

Challenge 2: Compute Cost
- Large number of events
- High computation cost

Unbearable for High-Res Event-based Vision

Hard Attention for Efficient Image Classification



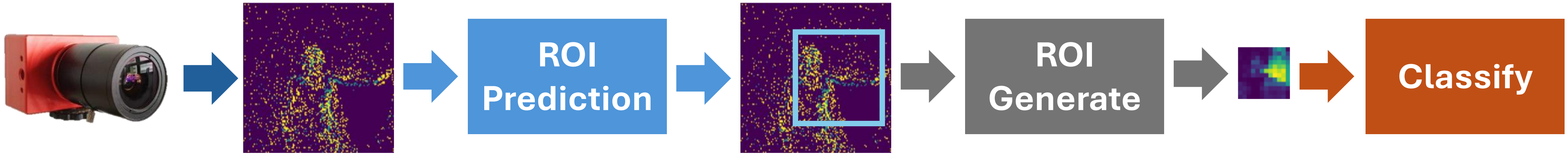
Process high-res image without quadratic complexity to input scale

Papadopoulos, et al., *Hard-Attention for Scalable Image Classification,* NeurIPS, 2021

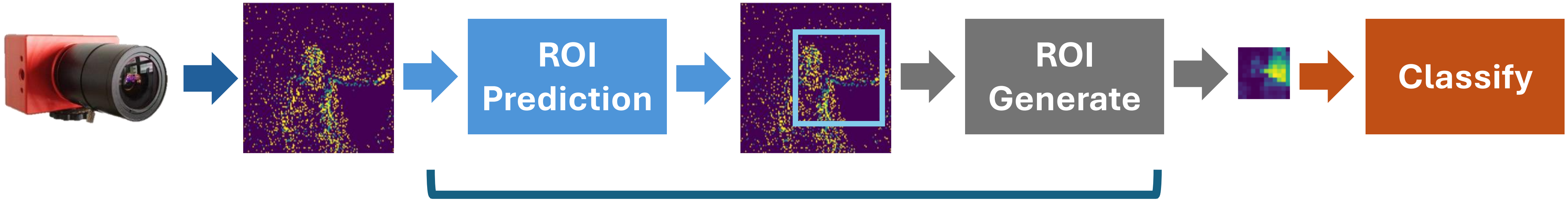
Challenge 1: Overhead for ROI
Costly ROI prediction and generation for complex scenes

Challenge 2: Training Complexity
Hard to perform end-to-end training with simple architecture

TRIP: Hard Attention for Event-based Vision



TRIP: Hard Attention for Event-based Vision

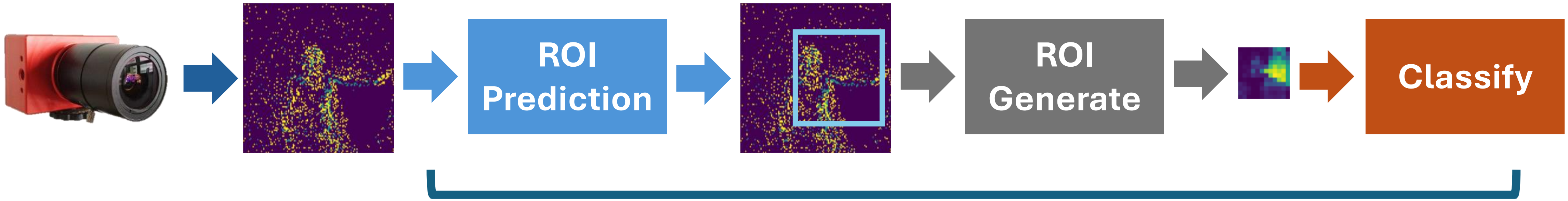


- Reduced inputs for ROI prediction
- Simple ROI prediction network
- Efficient ROI generation algorithm

Challenge 1: Overhead for ROI
Costly ROI prediction and generation for complex scenes

Challenge 2: Training Complexity
Hard to perform end-to-end training with simple architecture

TRIP: Hard Attention for Event-based Vision

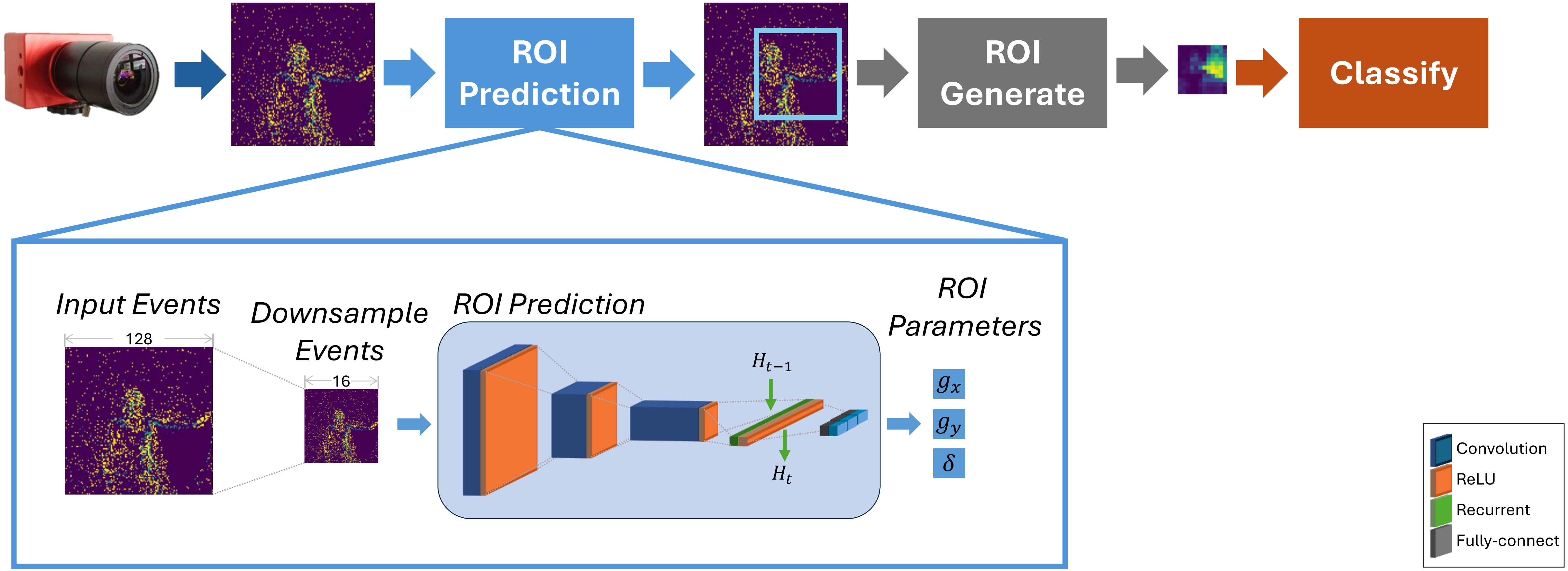


- End-to-end training
- Differentiable ROI generation
- Only require class label and classification loss

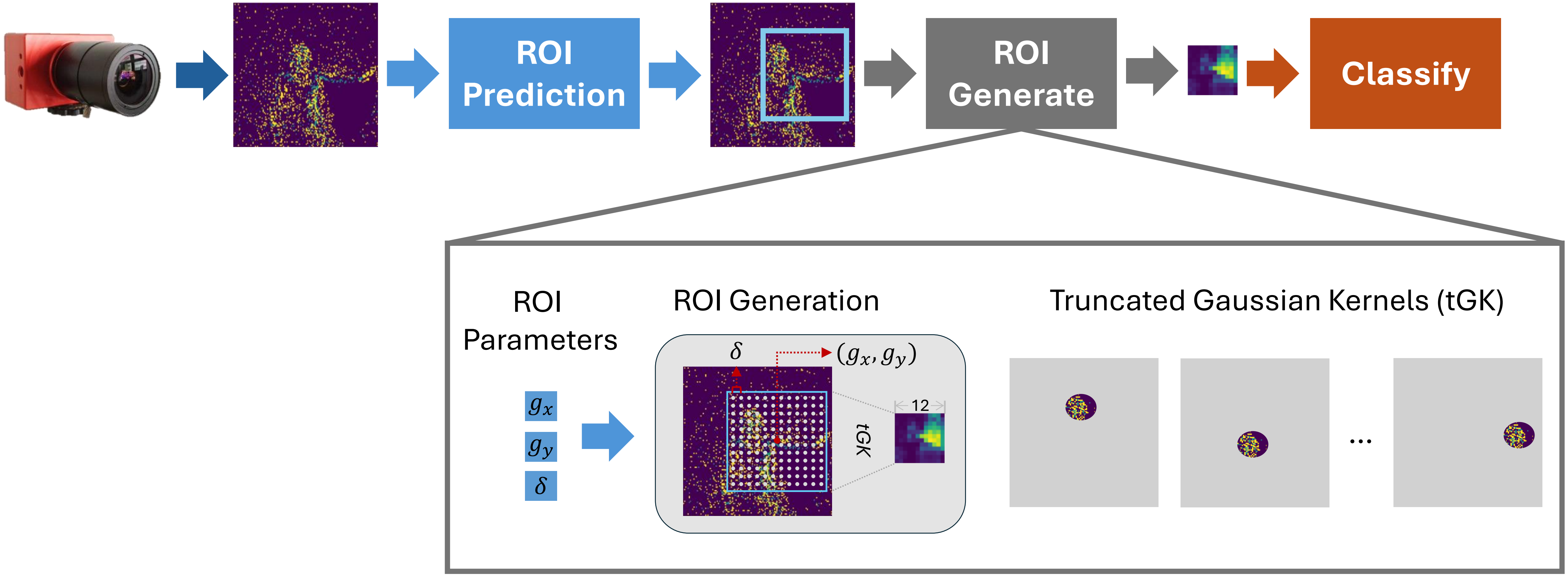
Challenge 1: Overhead for ROI
Costly ROI prediction and generation for complex scenes

Challenge 2: Training Complexity
Hard to perform end-to-end training with simple architecture

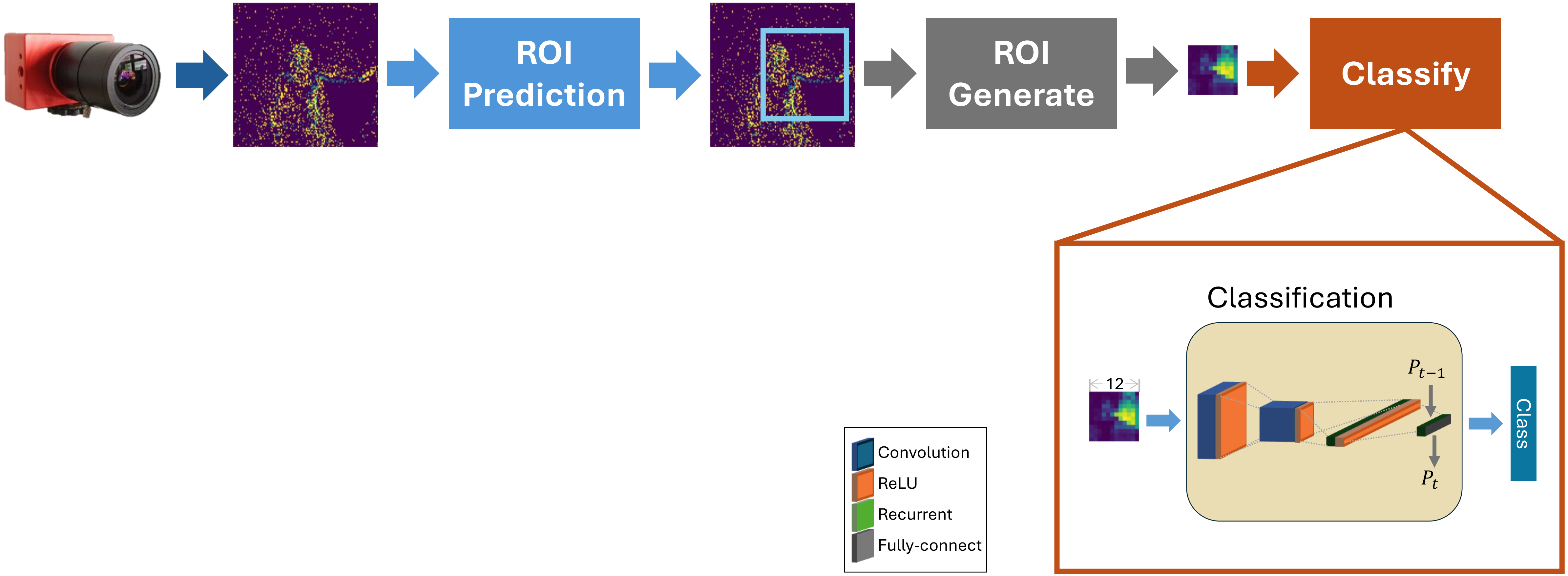
TRIP: Hard Attention for Event-based Vision



TRIP: Hard Attention for Event-based Vision



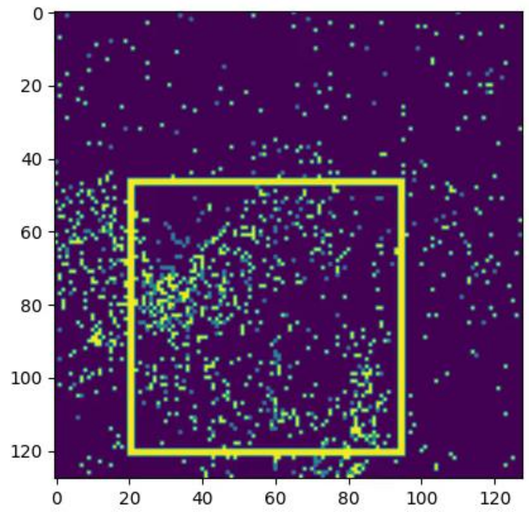
TRIP: Hard Attention for Event-based Vision



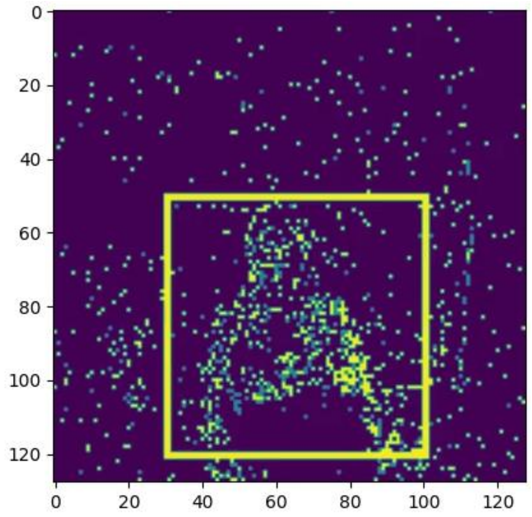
Performance on the **DvsGesture** Dataset

Architecture	Input Resolution	Param	Effective MACs (Single Timebin)	Accuracy [%] (mean $\pm$ std)	Accuracy [%] (Maximum)
LSTM [20]	32×32	7.4M	3.9M	–	86.8
AlexNet+LSTM [21]	128×128	8.3M	601.3M	–	97.7
CNN+EGRU [13]	128×128	4.8M	80.6M	97.3± 0.4	97.8
ConvLIAF [22]	32×32	**0.22M**	113.3M	–	97.6
TRIP (Ours)	16×16+12×12	0.46M	**1.75M**	**97.6 ± 0.5**	**98.6**

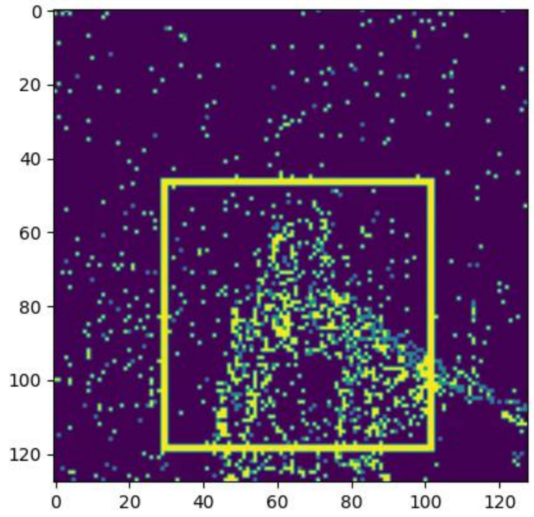
Right Hand Wave



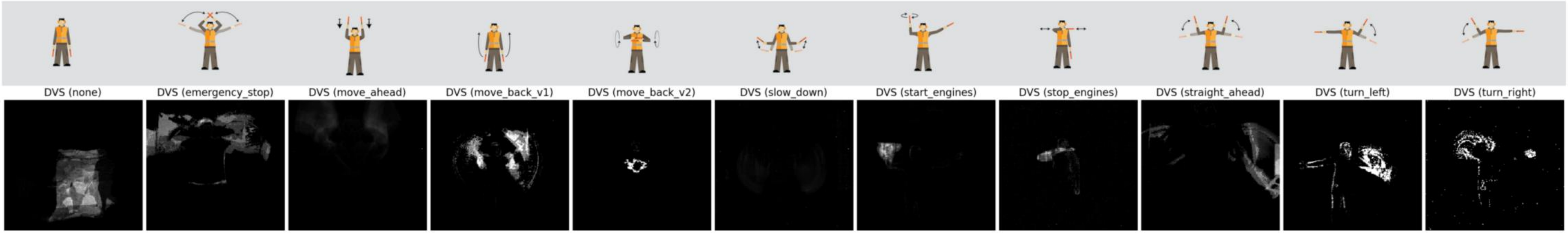
Right Hand Counter-Clockwise



Left Hand Clockwise



Performance on the Marshalling Signals Dataset



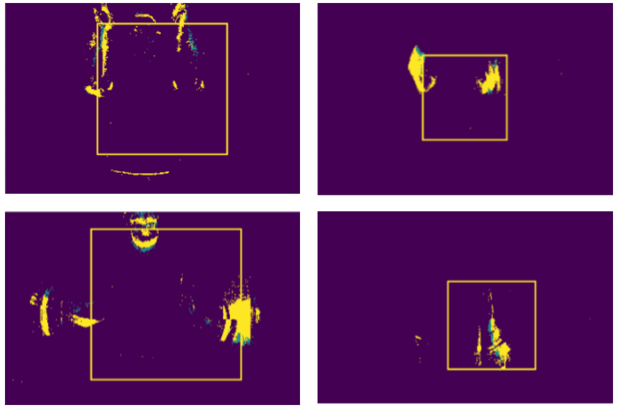
- 11 Gestures
- 8 Distances (1m-4.5m)
- DAVIS 346 (346x260)

Muller, et al., *Aircraft marshalling signals dataset of fmcw radar and event-based camera for sensor fusion,* RadarConf, 2023

Distance 1.5 m　　Distance 2.5 m

Move Back

Slow Down



Architecture	Param	FLOPs	Accuracy [%]
ResNet18 [12]	11.7M	1810M	74.6
EfficientNet-B1 [12]	7.794M	690M	82.6
TRIP (Ours)	**4.13M**	**37.0M**	**83.6**

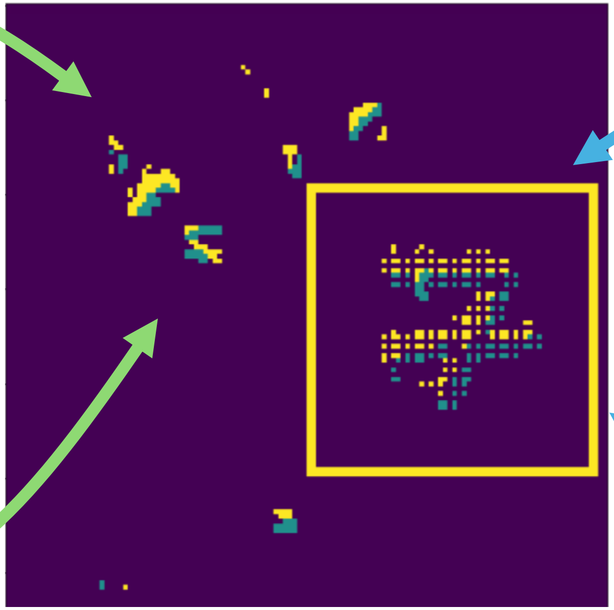
Experiments on our **Synthetic N-MNIST** Dataset

Add structured noises

Randomly select 8 other N-MNIST digits

Randomly crop 8x8 regions from each digit

Randomly place the cropped regions



Randomly scale 34x34 N-MNIST digits by 1-2 times

Randomly place scaled N-MNIST on 128x128 canvas

Same number of samples as the original N-MNIST (60K Training and 10K Testing)

Experiments on our **Synthetic N-MNIST** Dataset

TRIP

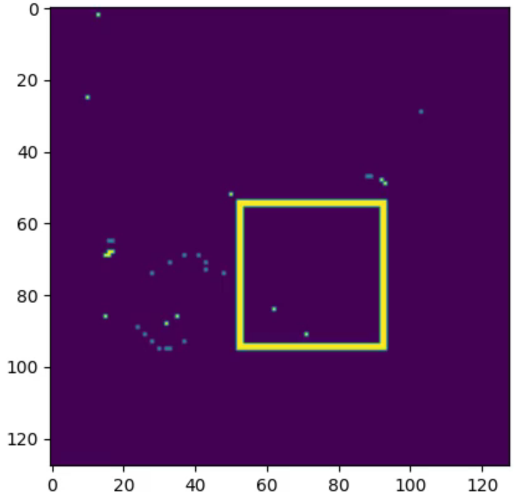
Input Dim	Param	FLOPs	Accuracy
16x16	0.30M	16.0M	95.4
32x32	0.65M	28.0M	96.1

TRIP achieves better performance than one level higher input-res

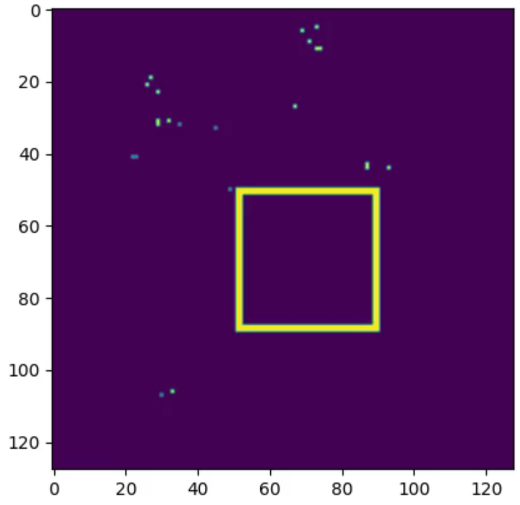
Single CNN with same number of layers

Input Dim	Param	FLOPs	Accuracy
16x16	0.31M	6.0M	71.8
32x32	0.67M	24.4M	93.0
64x64	0.67M	57.4M	96.2

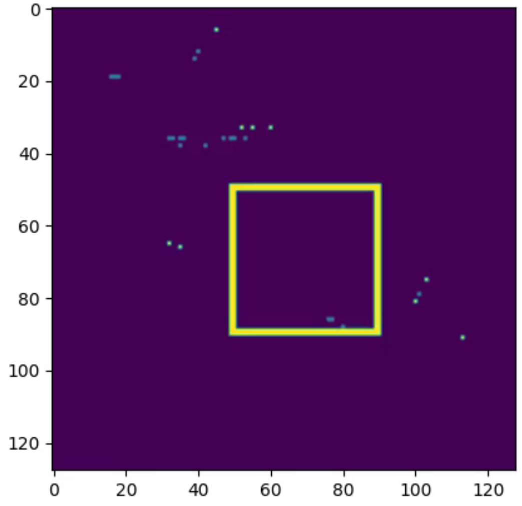
Digit 0



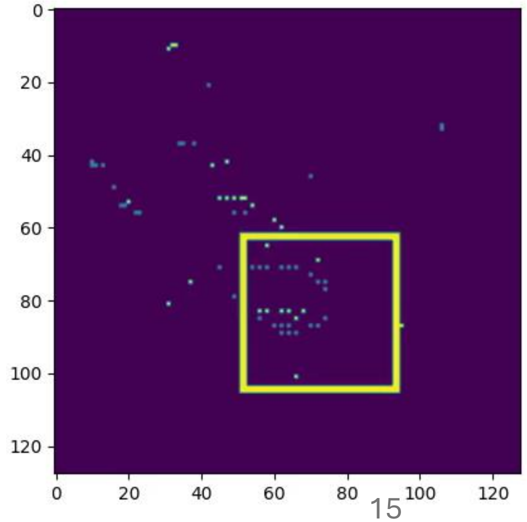
Digit 1



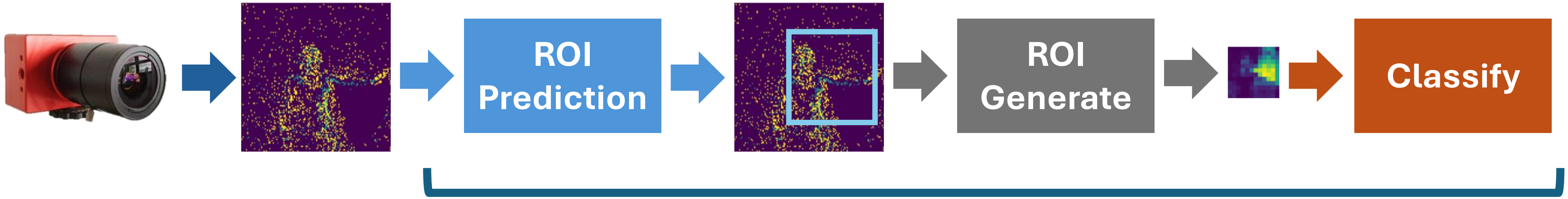
Digit 2



Digit 3



Deploy TRIP on **SENECA** Neuromorphic Processor



End-to-end deployment on the SENECA neuromorphic processor

HW-aware Optimization

- Activation Sparsity-aware Training
- 4-bit Quantization-aware Training

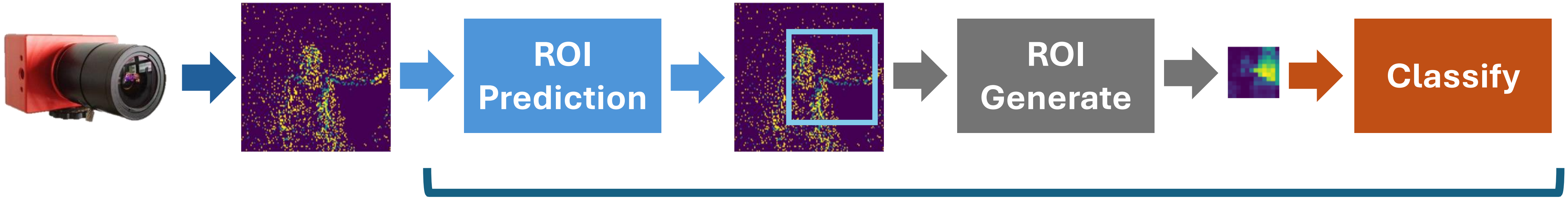
RISCV-efficient ROI Generation

Use Dynamic Average Pooling (DAP) to replace costly tGK

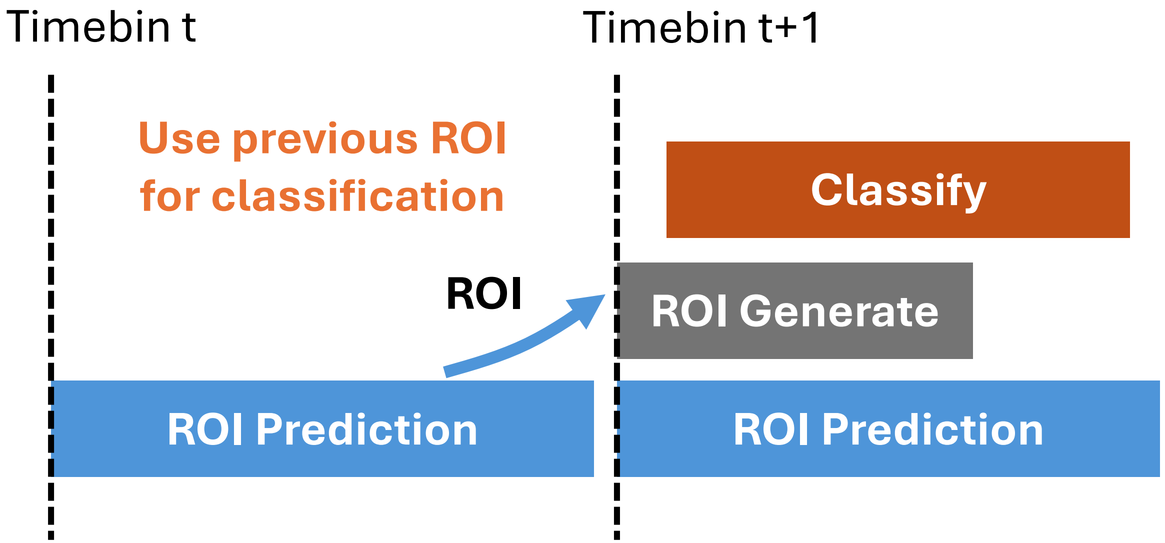
Pretrain with tGK → Finetune with DAP

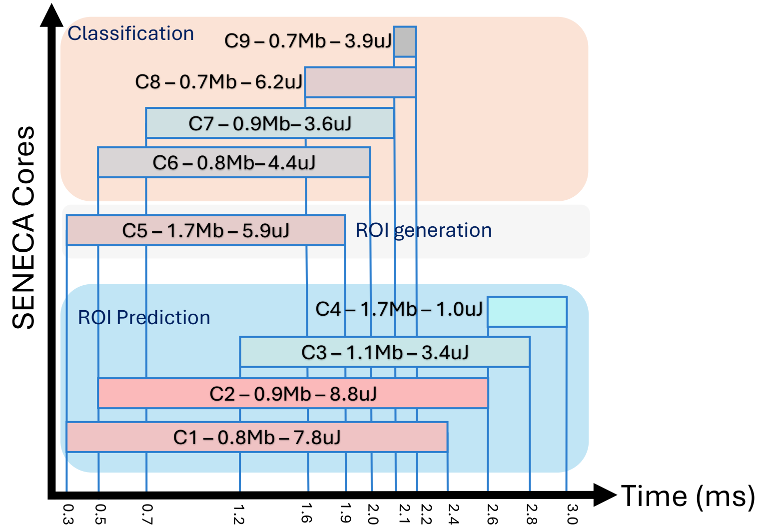
Xu, et al., *Optimizing event-based neural networks on digital neuromorphic architecture: a comprehensive design space exploration,* Front. Neurosci., 2024

Deploy TRIP on **SENECA** Neuromorphic Processor

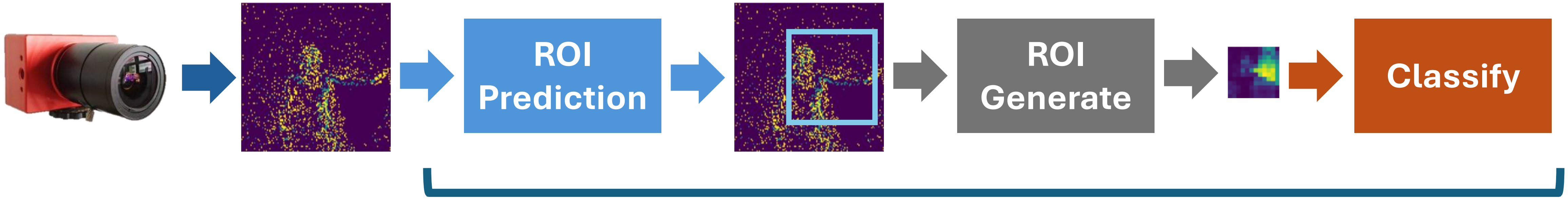


Causal processing increases latency of classification





Deploy TRIP on SENECA Neuromorphic Processor



Hardware	Solutions	Technology	Core [#]	Area [mm^2]	Single Timebin Latency [ms]	Single Timebin E_{inf} [uJ]	Single Timebin Accuracy [%]	Multiple Timebins Latency [ms]	Multiple Timebins E_{inf} [uJ]	Multiple Timebins Accuracy [%]
Loihi [26]	Spiking CNN [3]	Intel 14 nm	>20	>8.20	11	–	89.6	–	–	–
Loihi [26]	Spiking CNN [14]	Intel 14 nm	59	24.19	–	–	–	**22.0**	2731	96.2
TrueNorth [27]	Spiking CNN [11]	Samsung 28 nm	3838	383.8	–	–	**91.8**	104.6	18702	94.6
SENECA [10]	Event-based CNN	GF FDX 22 nm	7	**3.29**	–	–	–	78.9	1069.2	97.3
SENECA [10]	TRIP	GF FDX 22 nm	9	4.23	**2.7**	**35.86**	91.1	25.8	**430.32**	**98.3**

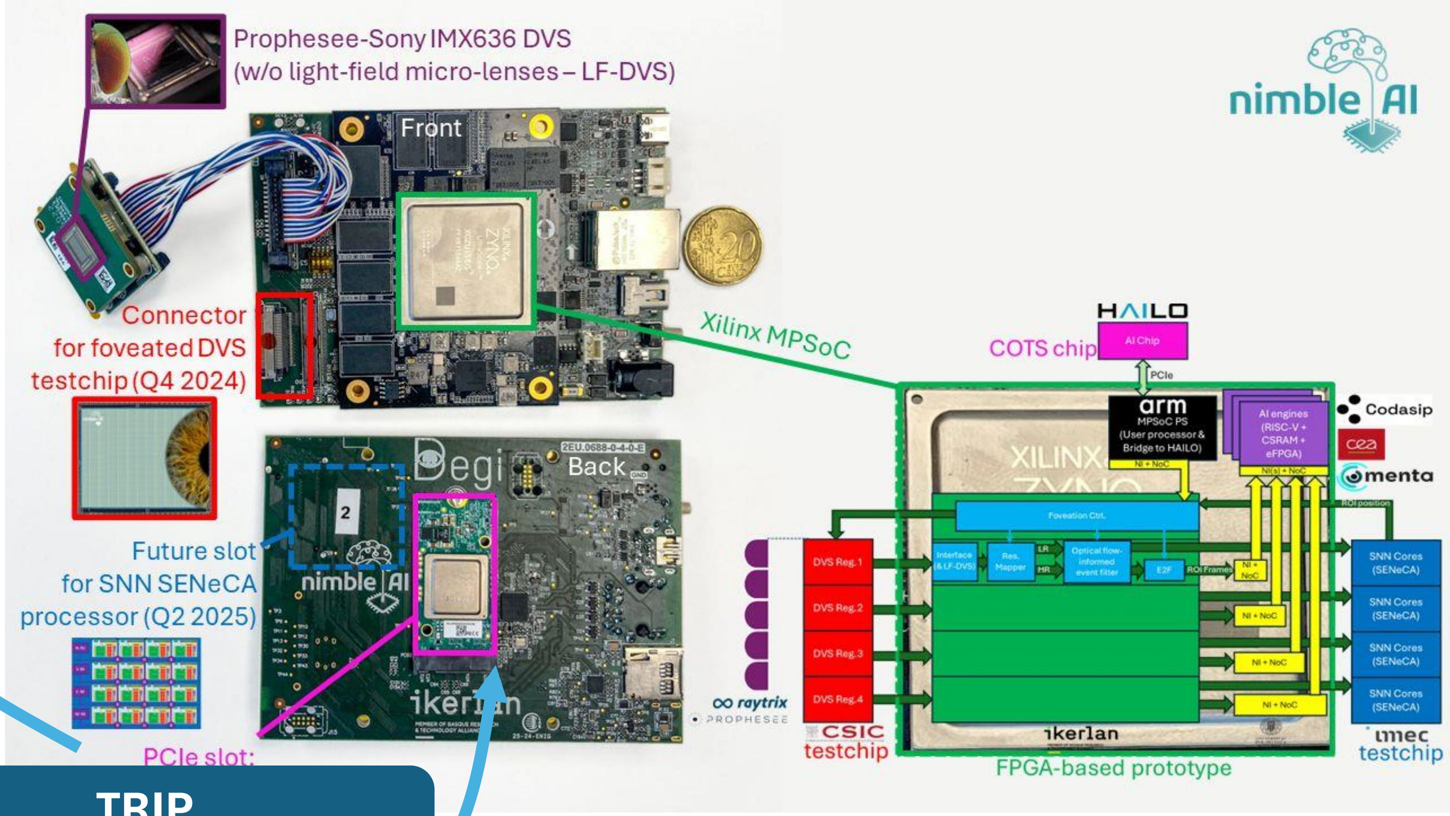
Future Integration of Sensing and Processing

Near-DVS for low-latency and low-power processing

SynSense Speck

iCatch iEVCam

Ongoing Nimble-AI (EU Horizon) Project for Near-DVS Processing



TRIP
Cost-efficient Solution

Acknowledgements


Cina Arjmand


Yingfu Xu


Kevin Shidqi

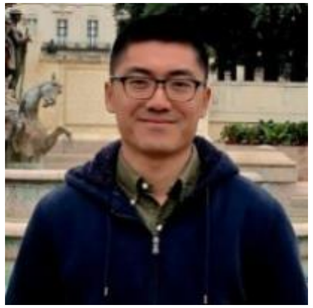

Alexandra Dobrita


Kanishkan Vadivel


Paul Detterer


Manolis Sifalakis


Amirreza Yousefzadeh


Guangzhi Tang

This work was partially funded by European research and innovation projects REBECCA (KDT JU under grant agreement No. 101097224), NeuroKIT2E (KDT JU under grant agreement No. 101112268), and NimbleAI (Horizon EU under grant agreement 101070679)

Q&A