Seminar Report: Efficient Processing of Event Data with Neural Networks

Andrii Chumak

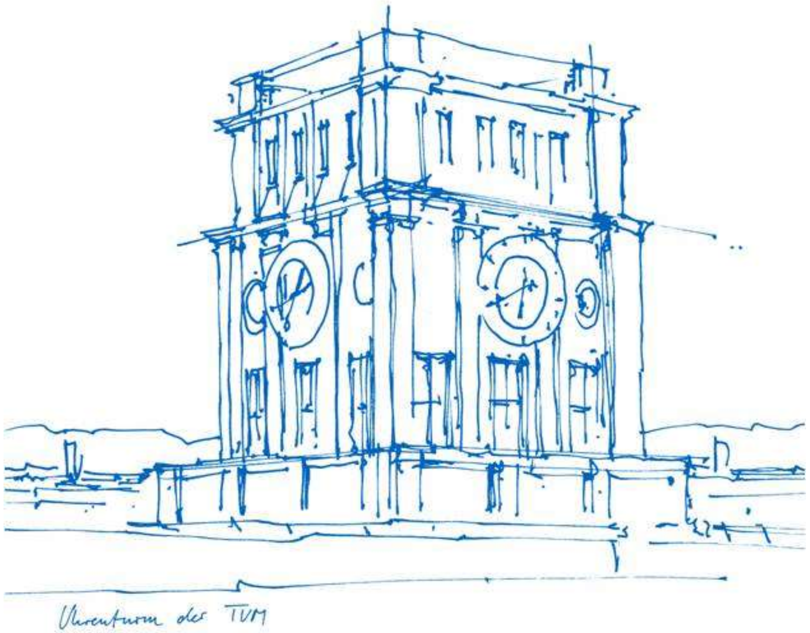
Supervisor: Yannick Burkhardt

Smart Robotic Lab

TUM School of Computation, Information, and Technology

Technical University of Munich

03.12.2024



Introduction: Event Cameras

Sub-millisecond latency: Multiple thousands fps time-resolution equivalent

=> Faster and more accurate object detection possible

High-dynamic range: Robustness in difficult lighting conditions

Data efficiency: Only the pixels sensing the changes generate events

However, most computer vision algorithms are designed for dense data

Outline

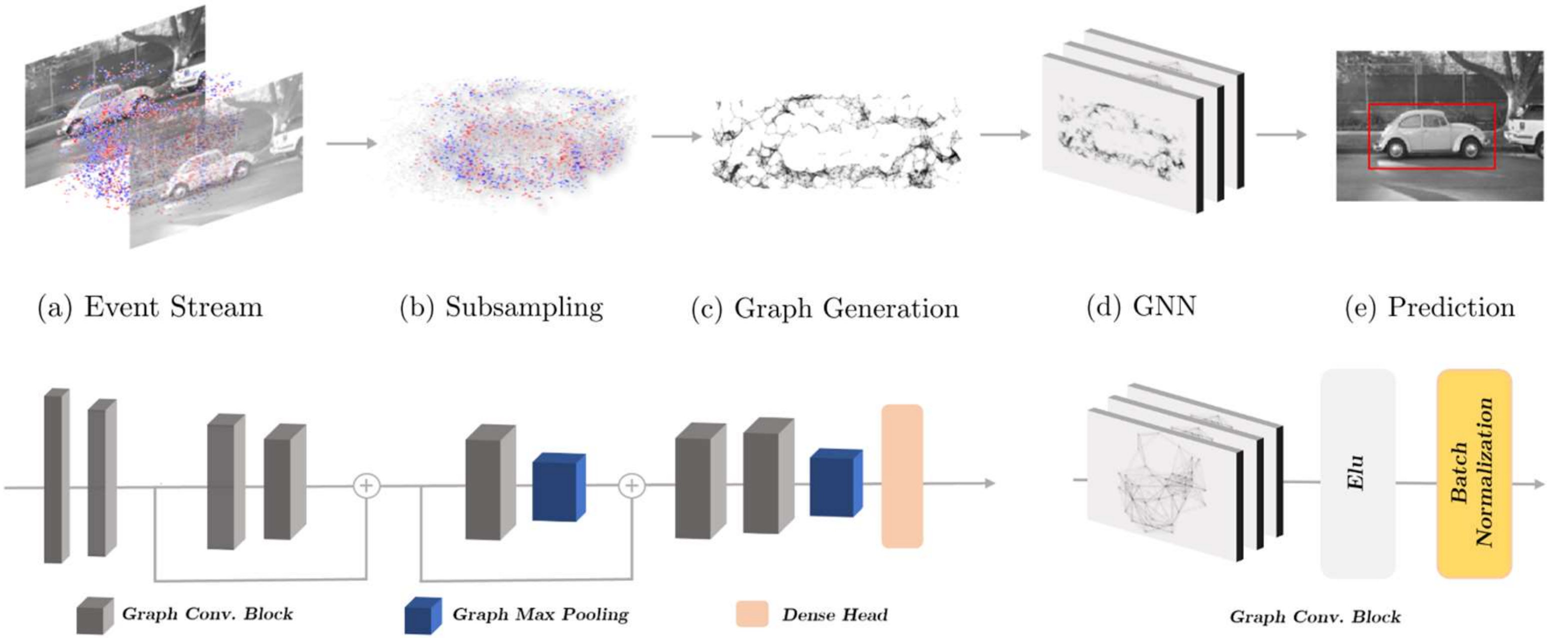
1. Related Work
2. Method Descriptions
3. Experiments and Results
4. Personal Comments
5. Future Work
6. Summary

Related Work

Research directions:

	✓	✗	
• Conventional dense networks, e.g., CNNs	Reuse existing architectures	Discards inherent sparsity and temporal resolution	
• Spiking Neural Networks	Model asynchronous data efficiently	Difficult to train	Accuracy to be improved
• Graph Neural Networks	Best computational efficiency		
• Transformers for spatio-temporal data	Good accuracy and inference time		

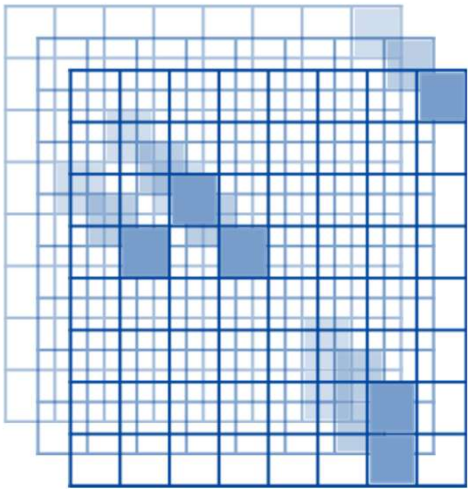
Method Descriptions: Asynchronous Event-based Graph Neural Networks (AEGNN)



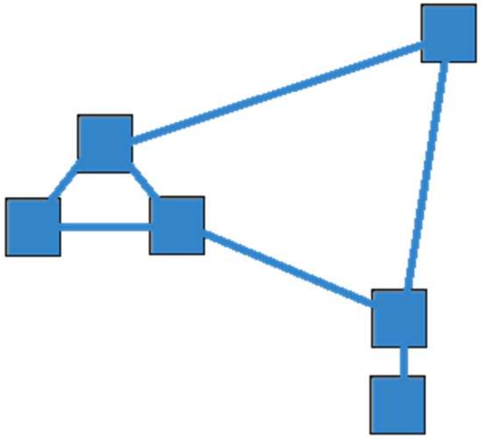
[1]

Method Descriptions: Asynchronous Event-based Graph Neural Networks (AEGNN)

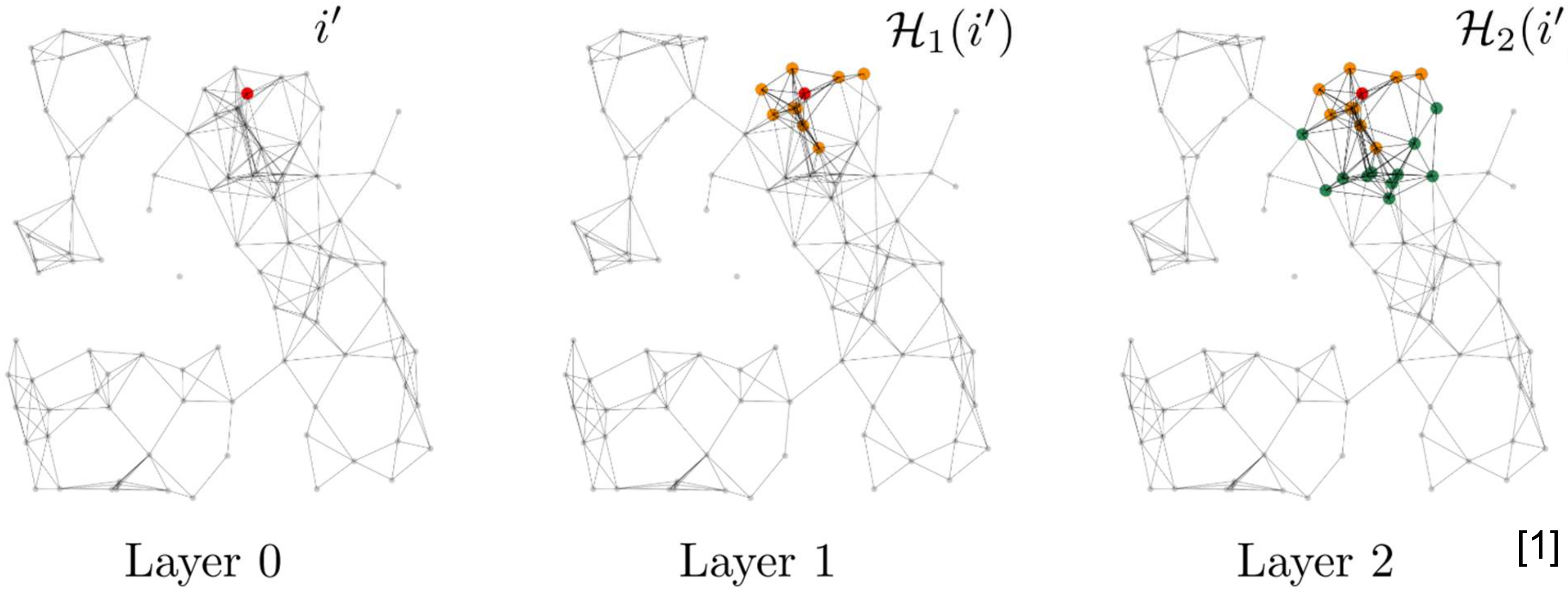
CNN



GNN

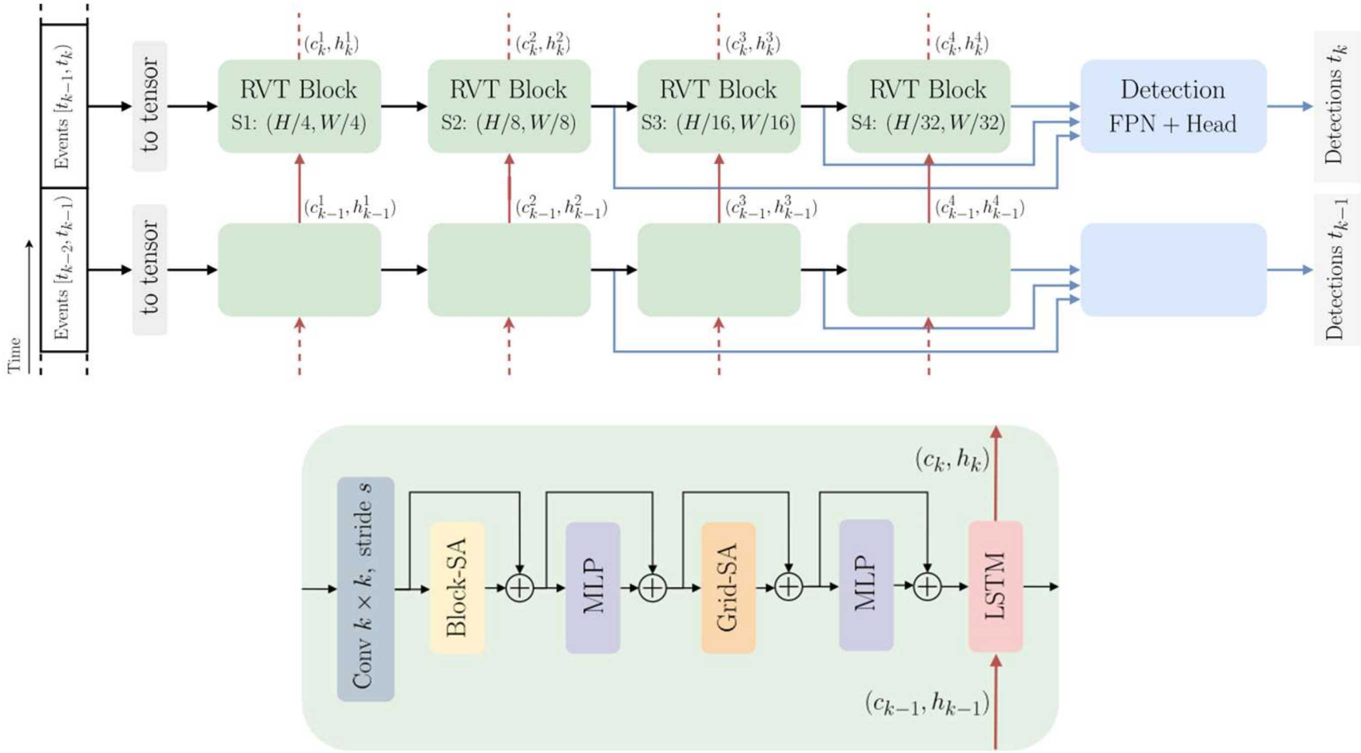


Method Descriptions: Asynchronous Event-based Graph Neural Networks (AEGNN)



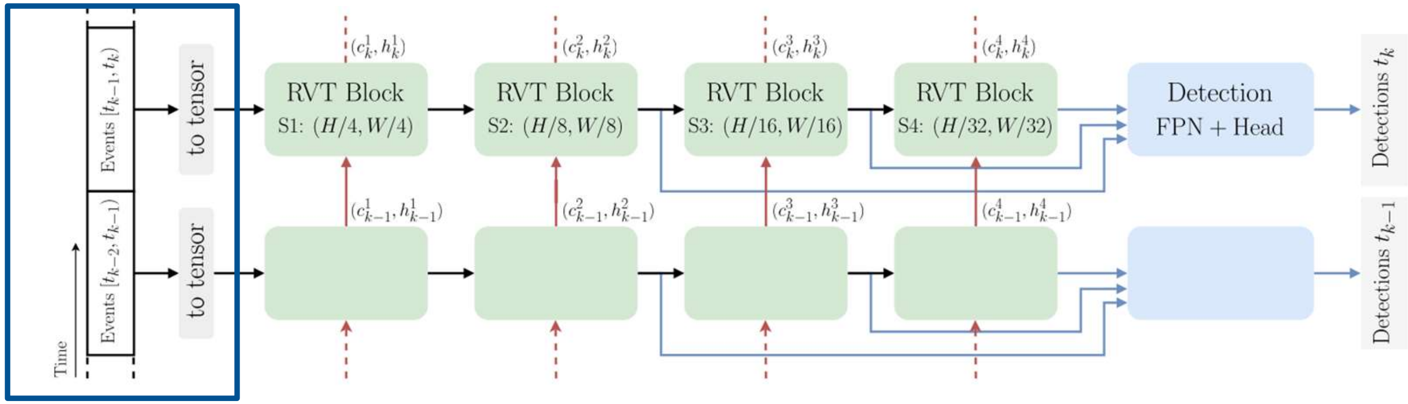
[1]

Method Descriptions: Recurrent Vision Transformers (RVT)

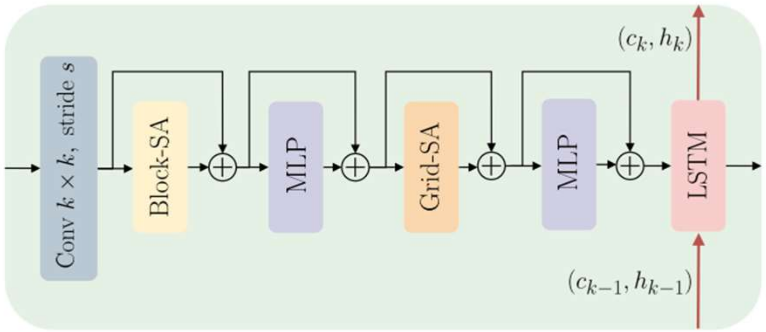


[2]

Method Descriptions: Recurrent Vision Transformers (RVT)

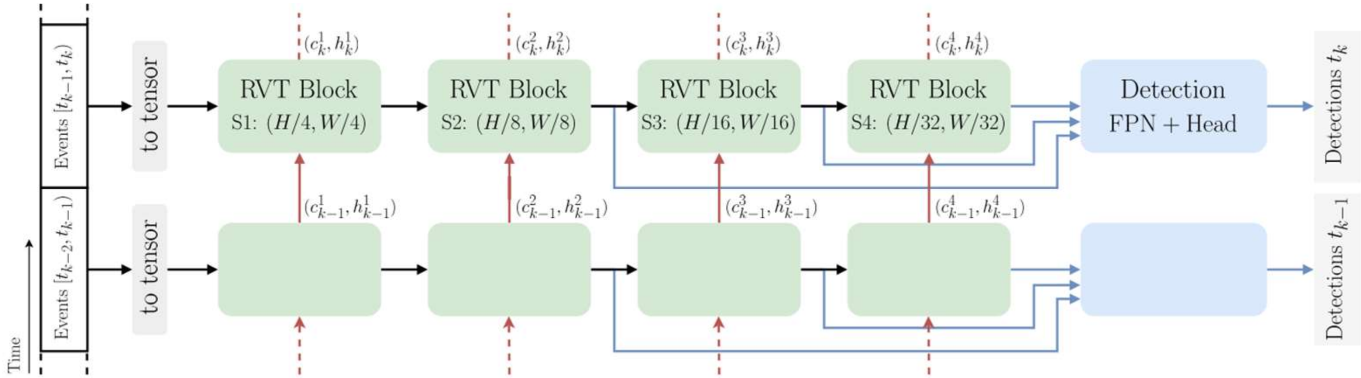


- Collect events for T discretized steps of time
- Convert data into tensor suitable for convolutions



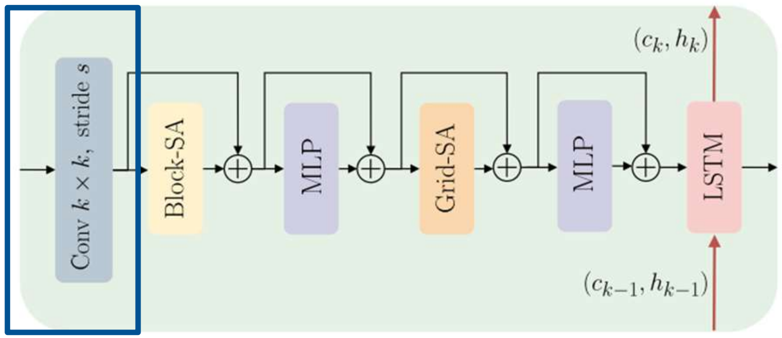
[2]

Method Descriptions: Recurrent Vision Transformers (RVT)



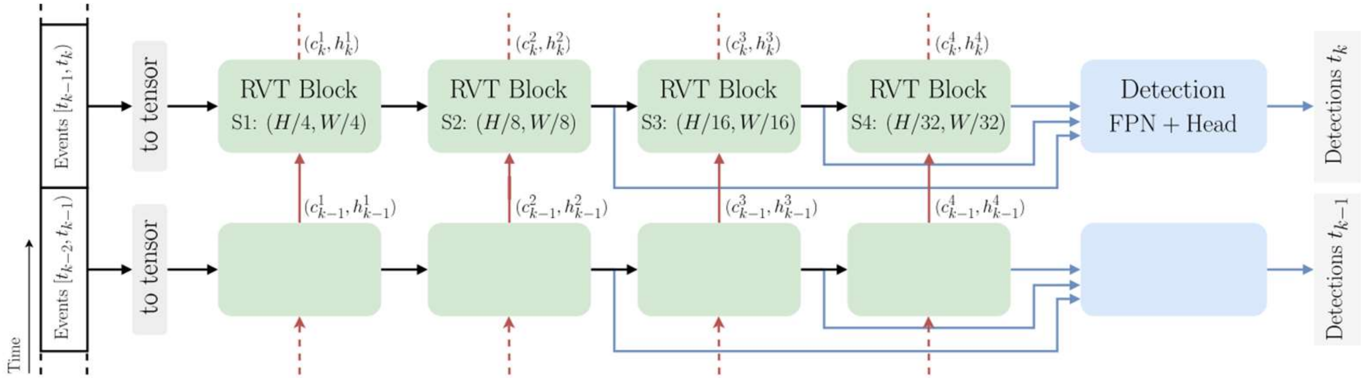
- Apply 2D convolutions to extract spatial features and downsample data

-> Positional embeddings

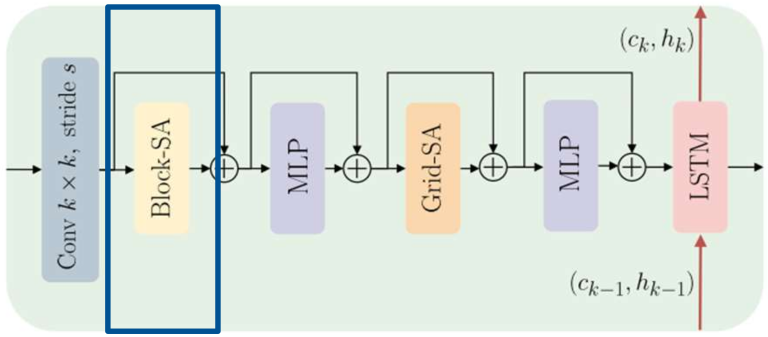


[2]

Method Descriptions: Recurrent Vision Transformers (RVT)

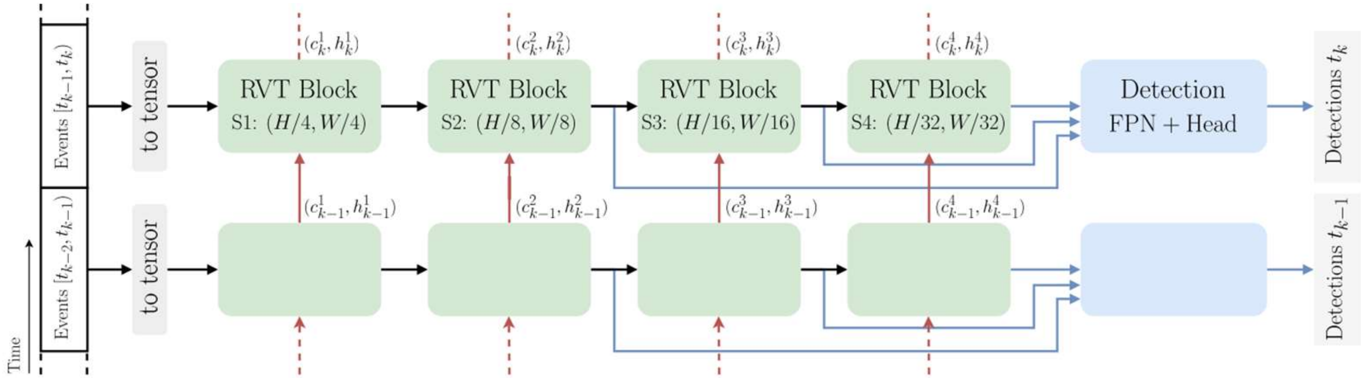


- Self Attention block to model local interactions

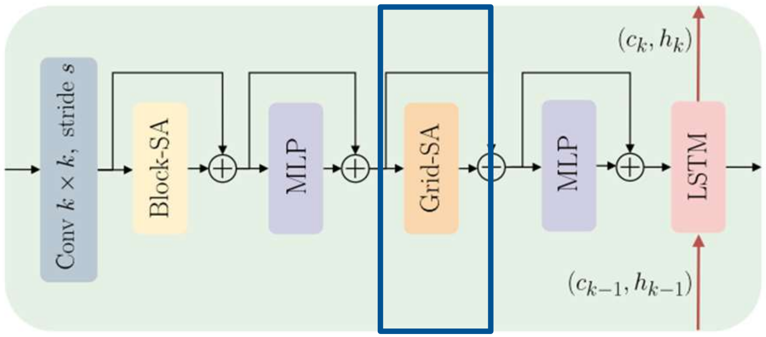


[2]

Method Descriptions: Recurrent Vision Transformers (RVT)

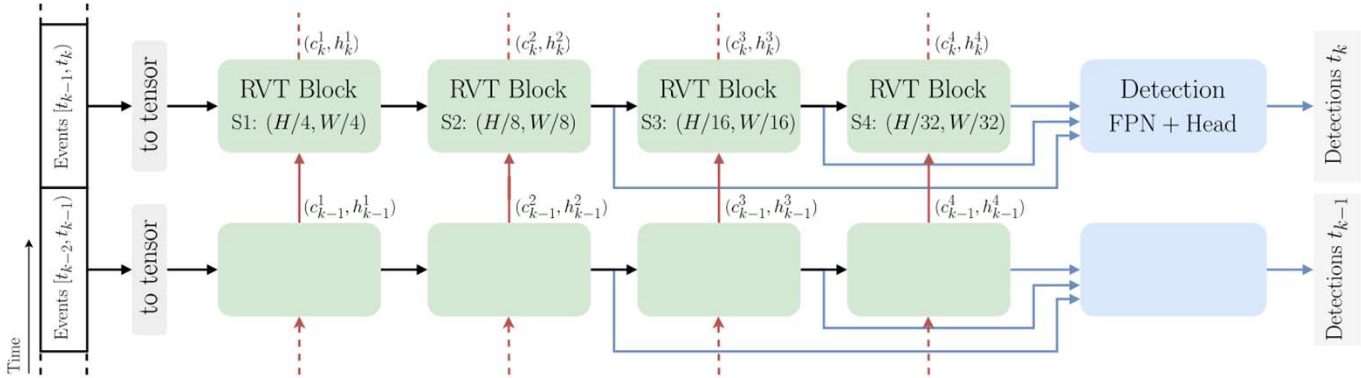


- Grid Self Attention block to model global features

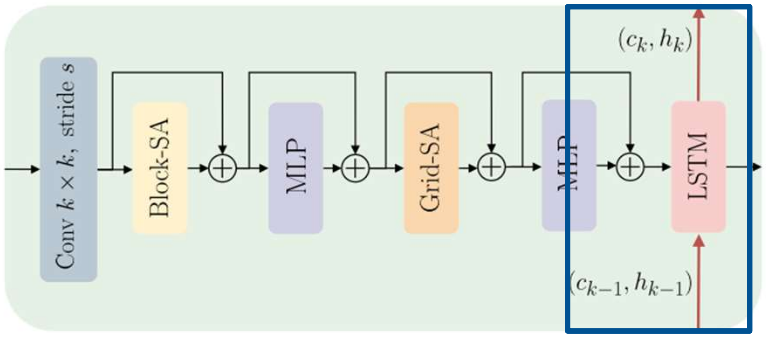


[2]

Method Descriptions: Recurrent Vision Transformers (RVT)

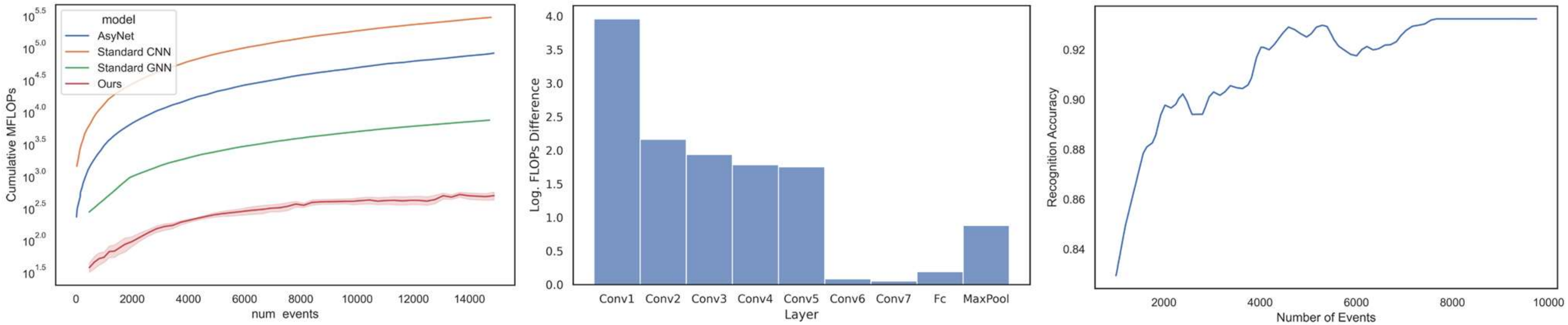


- LSTM models temporal feature aggregation



[2]

Experiments and Results: AEGNN



(a) MFLOPS over events

(b) MFLOP savings per layer

(c) Accuracy over events

[1]

Experiments and Results: AEGNN

			N-Caltech101		Gen1	
Methods	Representation	Async.	mAP ↑	MFLOP/ev ↓	mAP ↑	MFLOP/ev ↓
YOLE [7]	Event-Histogram	✓	0.398	3682	-	-
Asynet [36]	Event-Histogram	✓	**0.643**	200	0.129	205
RED [43]	Event-Volume	✗	-	-	**0.40**	4712
NVS-S [32]	Graph	✓	0.346*	7.8	0.086*	7.8
Ours	Graph	✓	0.595	**7.41**	0.163	**5.26**

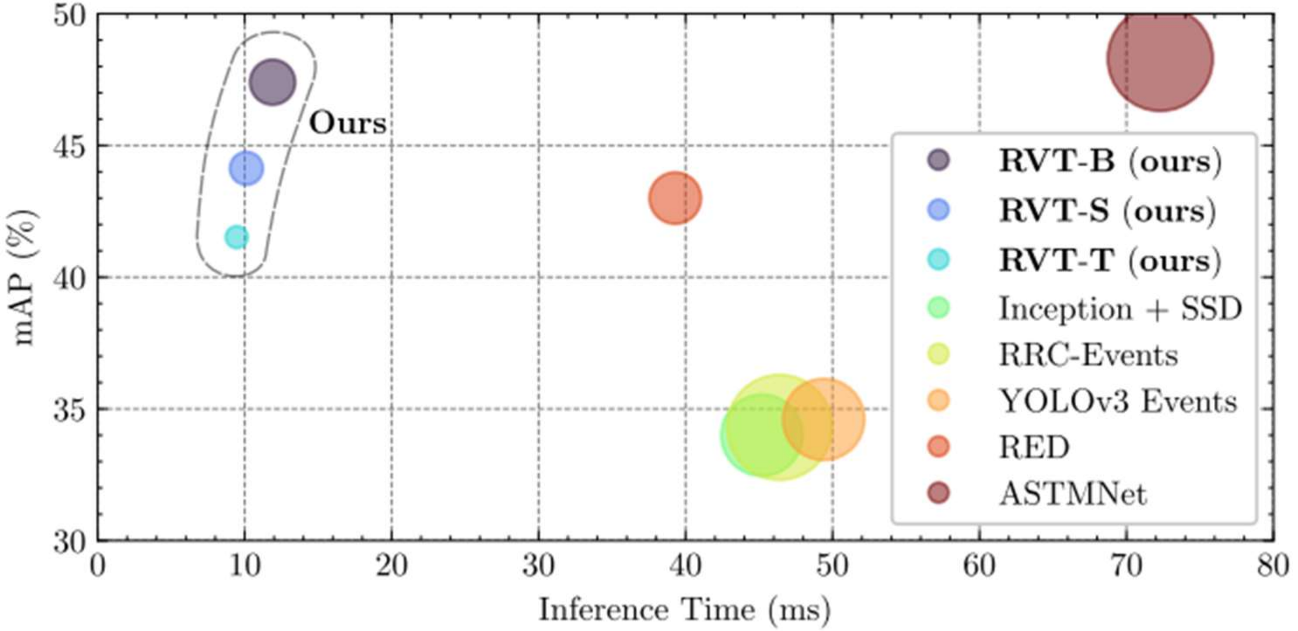
[1]

Experiments and Results: RVT

			Gen1		1 Mpx		
Method	Backbone	Detection Head	mAP	Time (ms)	mAP	Time (ms)	Params (M)
NVS-S [27]	GNN	YOLOv1 [40]	8.6	-	-	-	0.9
Asynet [34]	Sparse CNN	YOLOv1	14.5	-	-	-	11.4
AEGNN [43]	GNN	YOLOv1	16.3	-	-	-	20.0
Spiking DenseNet [10]	SNN	SSD [30]	18.9	-	-	-	8.2
Inception + SSD [19]	CNN	SSD	30.1	19.4	34.0	45.2	> 60*
RRC-Events [7]	CNN	YOLOv3 [41]	30.7	21.5	34.3	46.4	> 100*
MatrixLSTM [6]	RNN + CNN	YOLOv3	31.0	-	-	-	61.5
YOLOv3 Events [20]	CNN	YOLOv3	31.2	22.3	34.6	49.4	> 60*
RED [38]	CNN + RNN	SSD	40.0	16.7	43.0	39.3	24.1
ASTMNet [26]	(T)CNN + RNN	SSD	46.7	35.6	**48.3**	72.3	> 100*
RVT-B (ours)	Transformer + RNN	YOLOX [15]	**47.2**	**10.2 (3.7)**	47.4	**11.9 (6.1)**	18.5
RVT-S (ours)	Transformer + RNN	YOLOX	46.5	9.5 (3.0)	44.1	10.1 (5.0)	9.9
RVT-T (ours)	Transformer + RNN	YOLOX	44.1	9.4 (2.3)	41.5	9.5 (3.5)	4.4

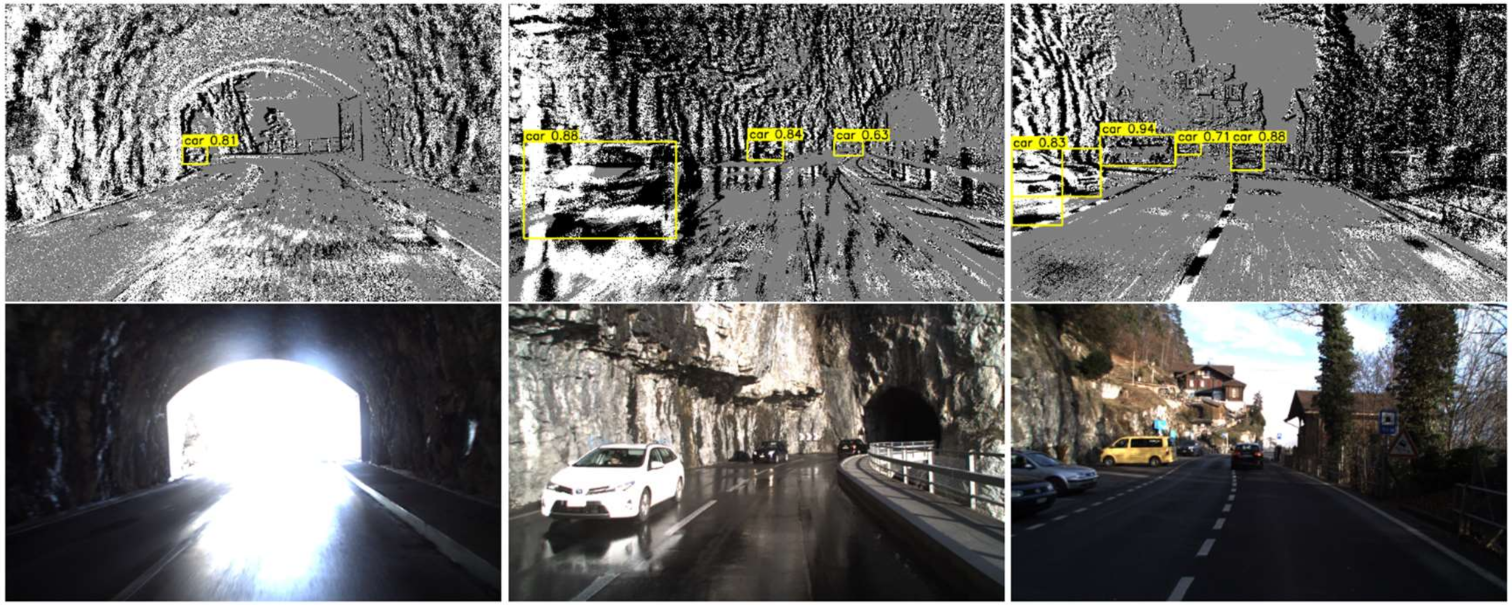
[2]

Experiments and Results: RVT



[2]

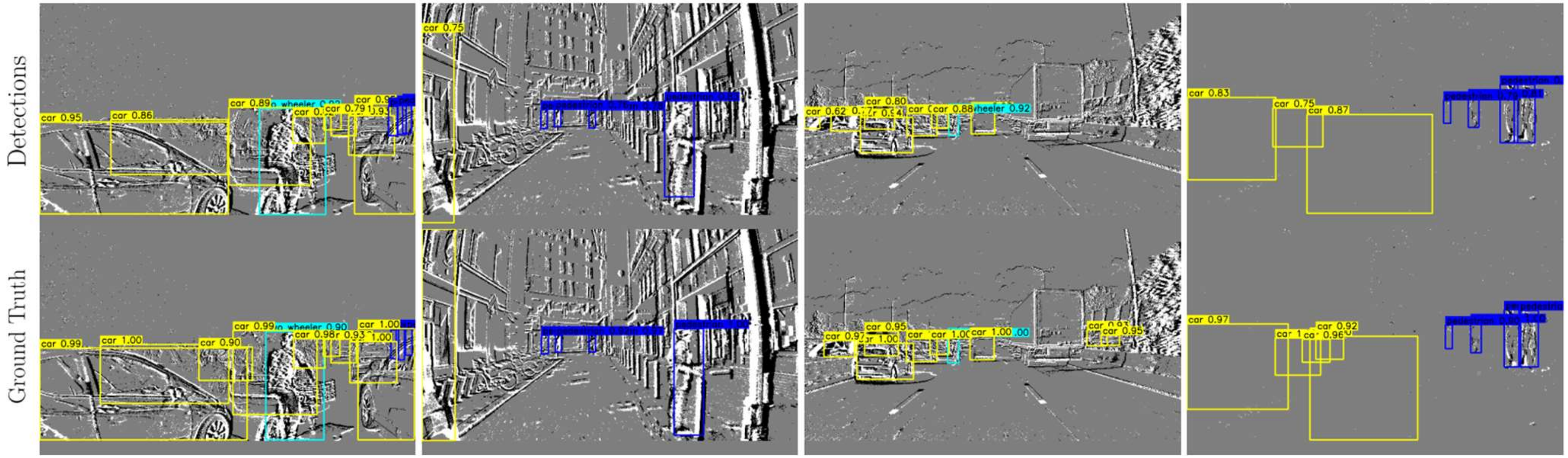
Experiments and Results: RVT



(a) (b) (c)

[2]

Experiments and Results: RVT



[2]

Personal Comments

- AEGNN
  - Despite one of the best performance in its class, less precise than dense NNs
  - Significant advantage in theoretical computational performance but not as hardware optimized as dense NNs
- RVT
  - Real-time capable (2-4 ms forward pass on RTX 3090 GPU)
  - State of the art accuracy and runtime despite using synchronous approach

Future Work

- Optimize AEGNN on specialized hardware (e.g., FPGAs, IPUs) for enhanced low-power performance [1]
- Fully leverage temporal structure of event data on RVT [2]
- Provide high quality frames to enrich information and overcome situations with no events available for longer time [2]
- Integrate event-based perception into a broader perception stack for more comprehensive real-time applications
- Label-efficient training on event data (e.g., LEOD [3] with RVT-S [2] could slightly outperform RVT-B with standard training) [3]

Summary

- Introduced the potential of event cameras in perception tasks
- Explored efficient methods for processing asynchronous data
- Presented and analyzed two distinct approaches: AEGNN and RVT
- Highlighted applications in real-time, difficult environments, and resource-limited scenarios
- Shown some potential research directions

References

1. Simon Schaefer, Daniel Gehrig, and Davide Scaramuzza. AEGNN: asynchronous event-based graph neural networks. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022, pages 12361–12371. IEEE, 2022.

2. Mathias Gehrig and Davide Scaramuzza. Recurrent vision transformers for object detection with event cameras. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023, pages 13884–13893. IEEE, 2023.

3. 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.