Recurrent Vision Transformers for Object Detection with Event Cameras

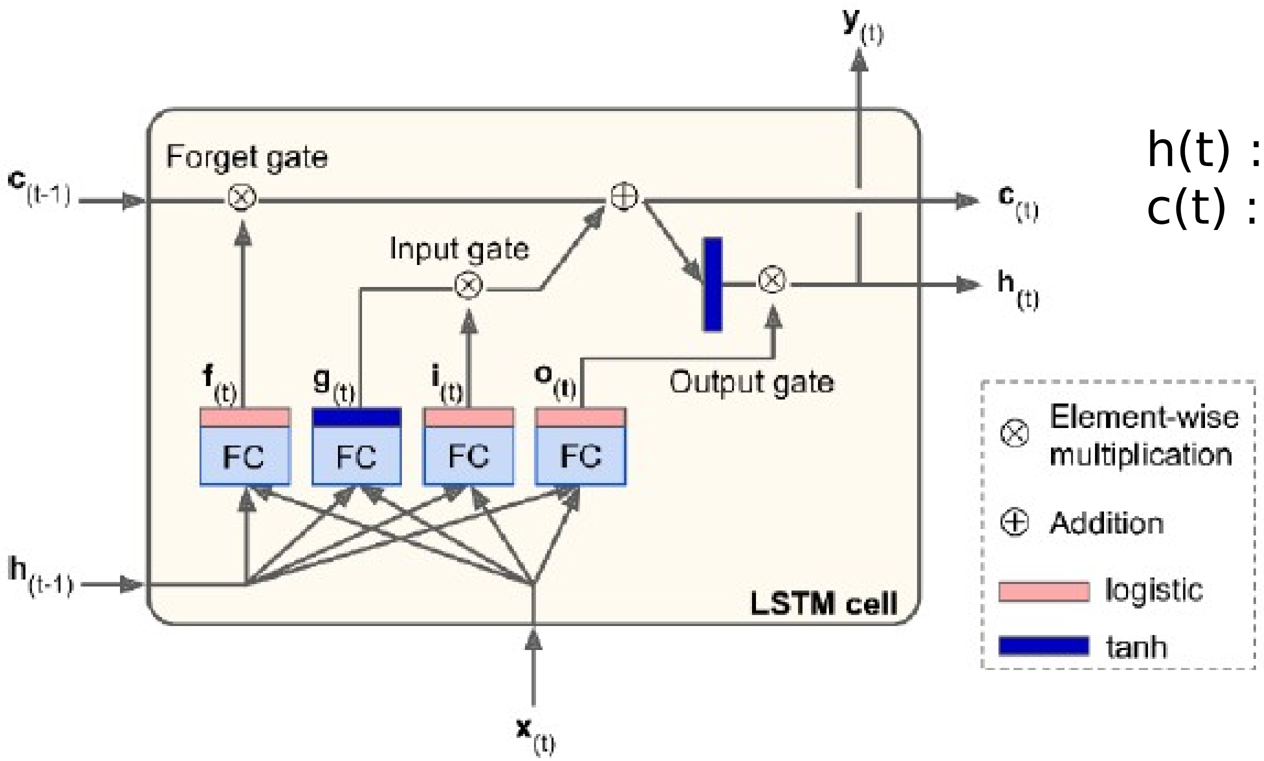
IT □□ □□□□□ 2021111183 □□□
IT □□ □□□□□ 2021114818 □□□

CONTENTS

01.
Before Reviewing…

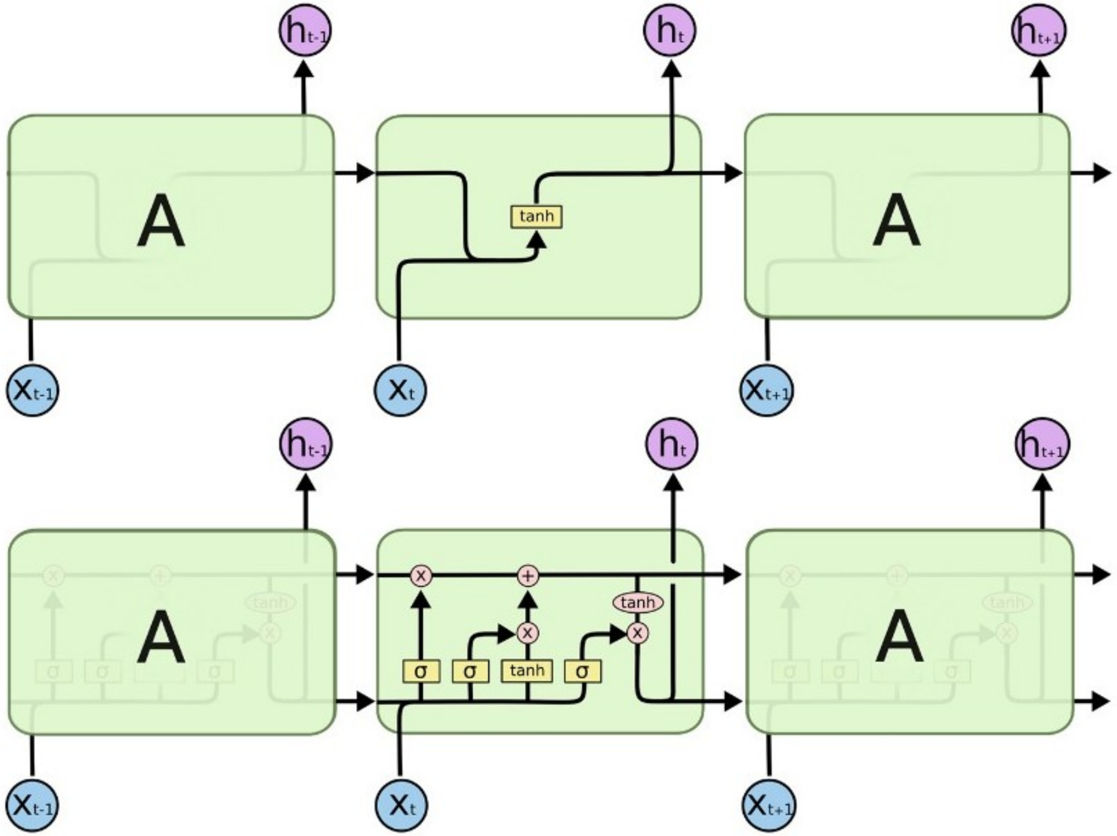
LSTM (Long short-term memory)

RNN □□ □ □□□ , gate 가□ □□ □□ □□ □□□□□ □□□□ Vanishing Gradient Problem □ □□□ .



h(t) : the short-term state
c(t) : the long-term state

LSTM (Long short-term memory)



Transformer

“Attention is All you Need”

Layer Type	Complexity per Layer	Sequential Operations	Maximum Path Length
Self-Attention	$O(n^2 \cdot d)$	$O(1)$	$O(1)$
Recurrent	$O(n \cdot d^2)$	$O(n)$	$O(n)$
Convolutional	$O(k \cdot n \cdot d^2)$	$O(1)$	$O(log_k(n))$
Self-Attention (restricted)	$O(r \cdot n \cdot d)$	$O(1)$	$O(n/r)$

N : sequence length
D : representation dimension
K : kernel size

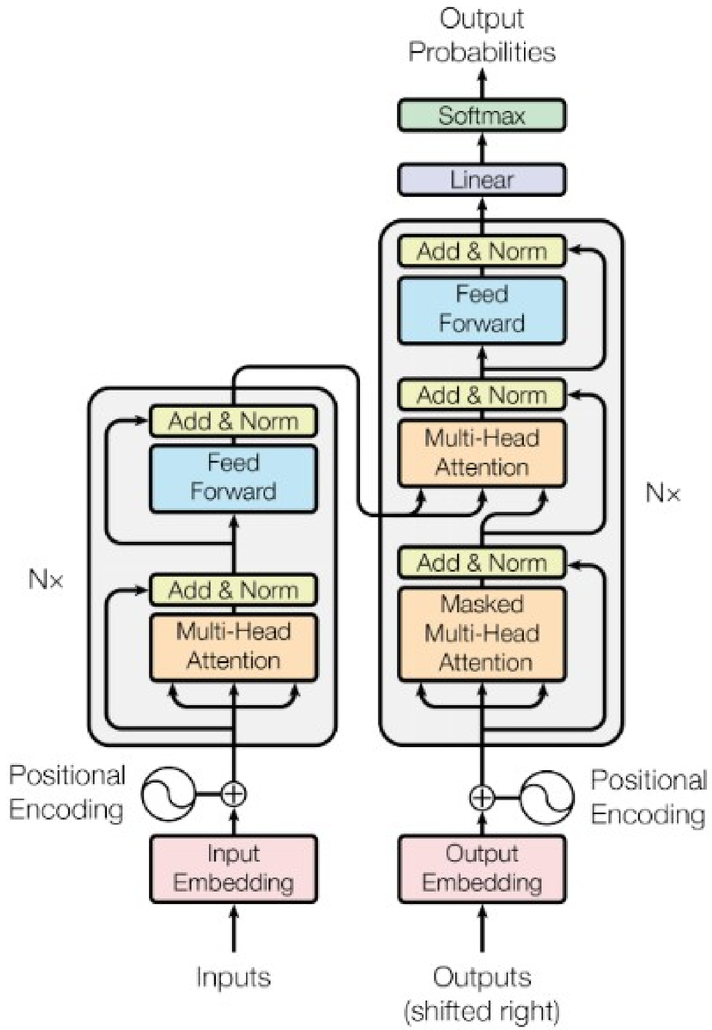
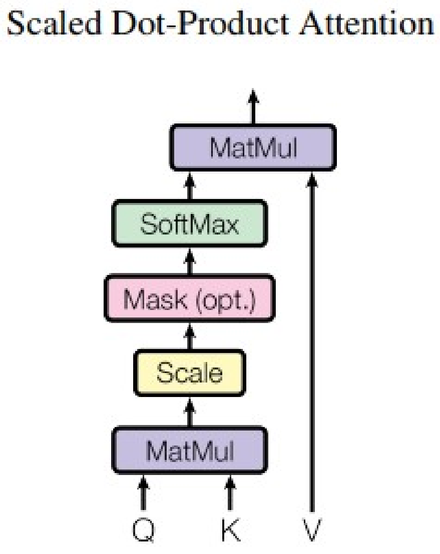
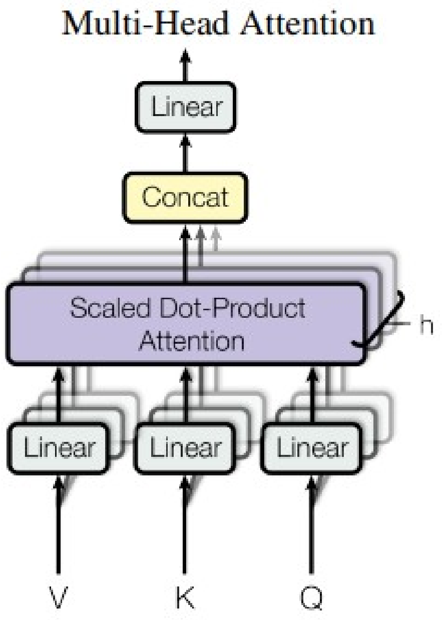


Figure 1: The Transformer - model architecture.

Transformer

Attention : can be described as mapping a query and a set of key-value pairs to an output





$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

$$\mathrm{MultiHead}(Q, K, V) = \mathrm{Concat}(\mathrm{head_1}, ..., \mathrm{head_h})W^O$$
$$\text{where } \mathrm{head_i} = \mathrm{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

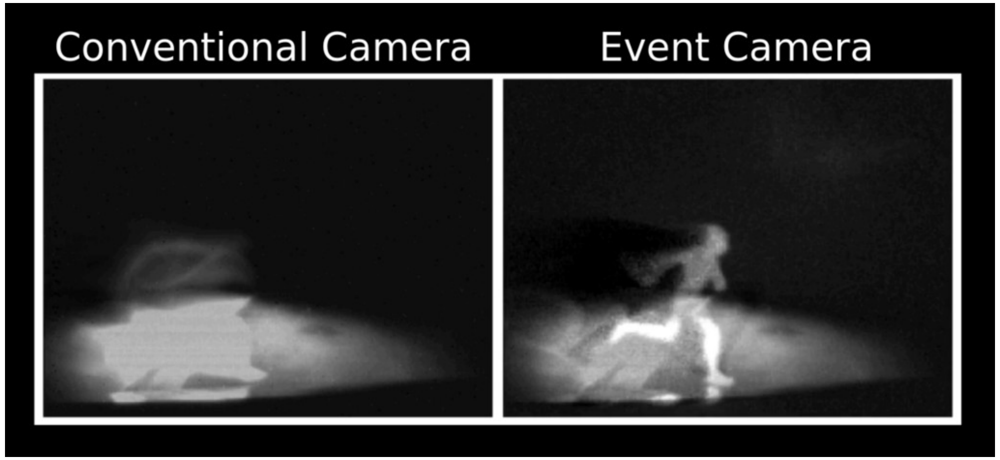
02.

Main Review

1. Introduction

What is **Event Camera**?

1. Low latency
2. High dynamic range
3. Strong robustness against motion blur



1. Introduction

Event Cameras vs **Traditional Camera**(a-frame-based-camera)

	Event Camera	Traditional Camara
absolute intensity information	↓	↑
Change in intensity	↑	↓
Reducing Latency	↑ (submillisecond latency)	↓

ue to latency, traditional camera may come at the cost of missing essential scene details in dynamic scene

1. Introduction

Event Camera □ □□□

□□, □□□ □□ □□□□□ binary event 가 □□
-> □□□ □□□□ detection □ □□□□ algorithm □ □□□ □□□□ □□.

□□ □□ (Related Work)

- GNN(Graph Neural Network)
□□□) □□ heavy □ backbone □□ & ConvLSTM □□ expensive □ cell □ □□
-> sparce neural network □ vision backbone □ design □□

1. Introduction

□ □□□□ inference time□ performance□ □□□ □□□□ □□ □□□ □□

1. Local and global self-attention
2. Preceding a simple convolution before attention
3. Conv-LSTM => plain LSTM

□□ □□

1. event-based pipelines □ □□□ design □□
2. Simple, composable □ state design
3. State-of-the-art object detection □□ □□□ □□□ □□□.

2. Related Work

Vision Transformation for Spatio-Temporal Data

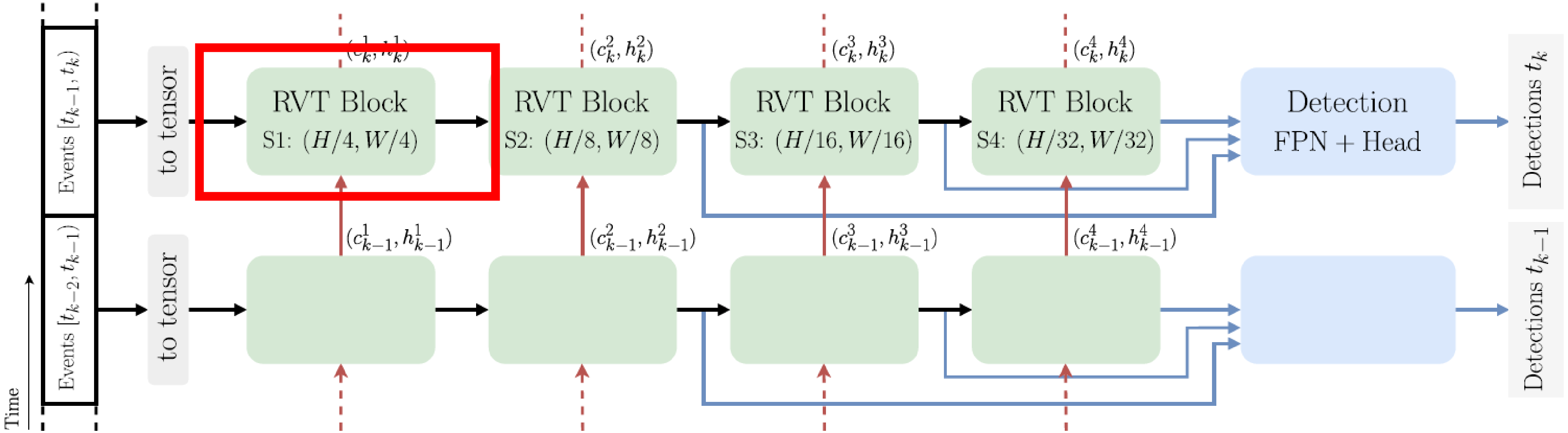
In event-based vision

- classification
- image reconstructure
- monocular depth estimation

=> Object detection has yet to be investigated

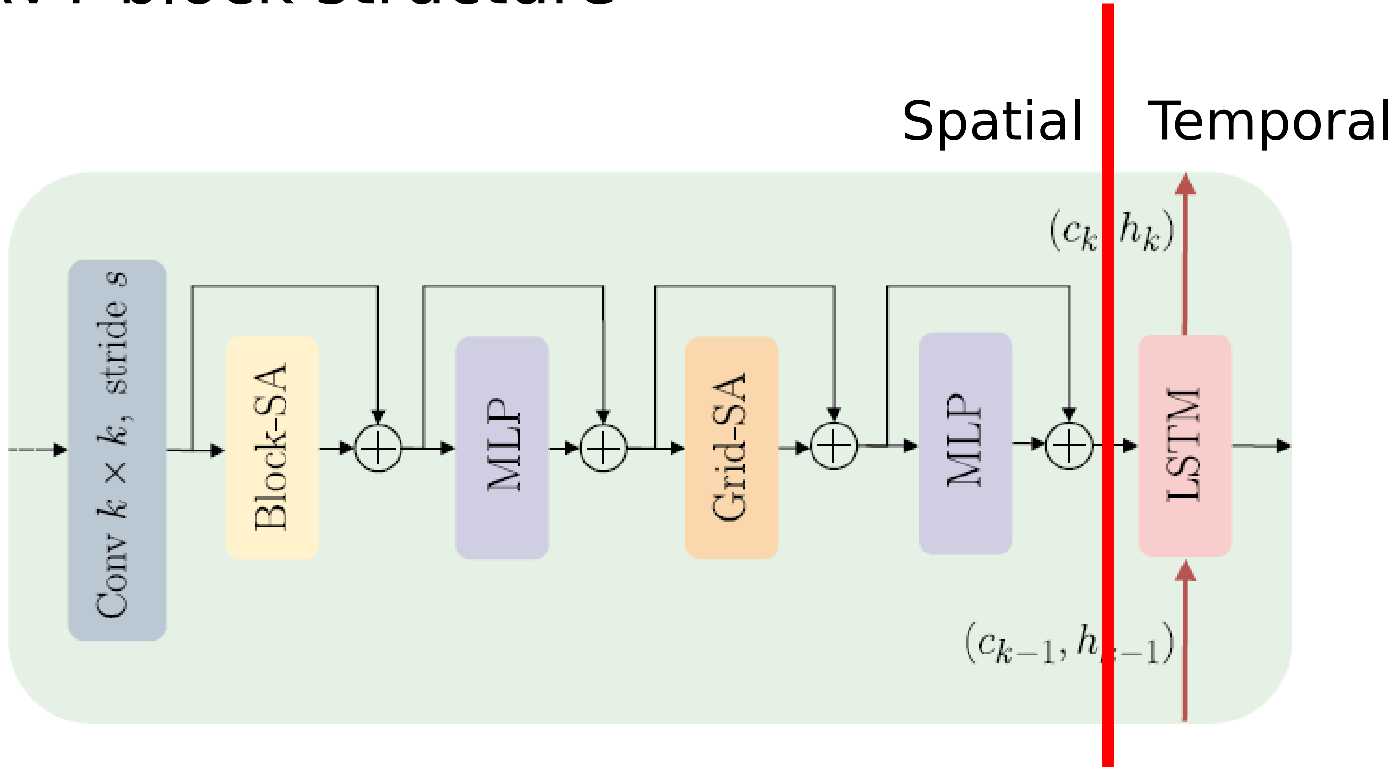
3. Method

Overall Structure



3. Method

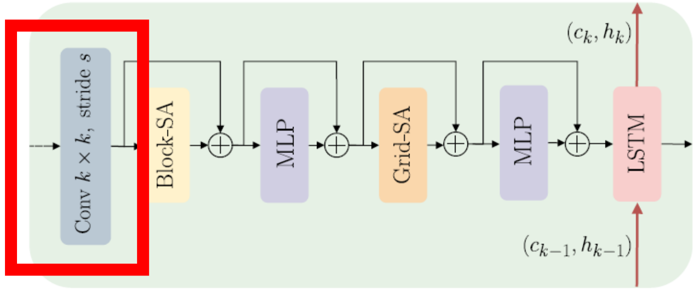
RVT block structure



1. Convolution (overlapping)
2. Block Self Attention
3. MLP(Multi Layer Perceptron)
4. Grid Self Attention
5. MLP
6. LSTM

※ Normalization and activation layers are omitted for conciseness

3. Method



Preprocessing step

Input data

$$E(p, \tau, x, y) = \sum_{e_k \in \mathcal{E}} \delta(p - p_k)\delta(x - x_k, y - y_k)\delta(\tau - \tau_k),$$

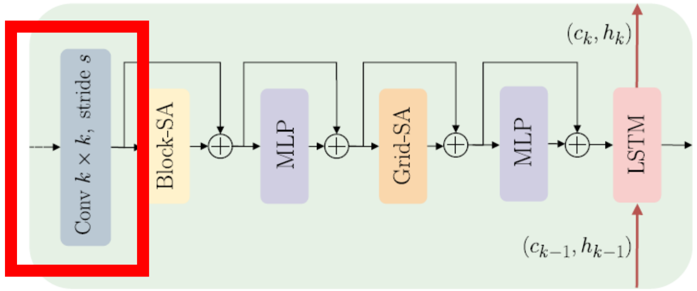
$$\tau_k = \left\lfloor \frac{t_k - t_a}{t_b - t_a} \cdot T \right\rfloor$$

(2T, H, W)

2D convolution □ □□□□ □□□ □□

x : width
y : height
p : polarity (□□□)
T : discretization steps of time

3. Method



Spatial Feature Extraction

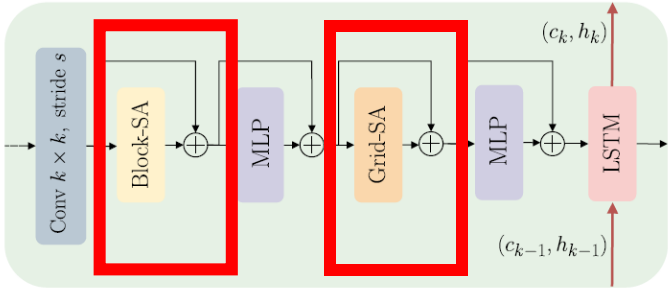
1. Convolution (overlapping)

Convolution with overlapping kernels

=> non-overlapping model □□ □□□ □□ □□□ □□□□.

3. Method

Multi-axis attention self-attention



2. Block Self Attention

Local feature interaction

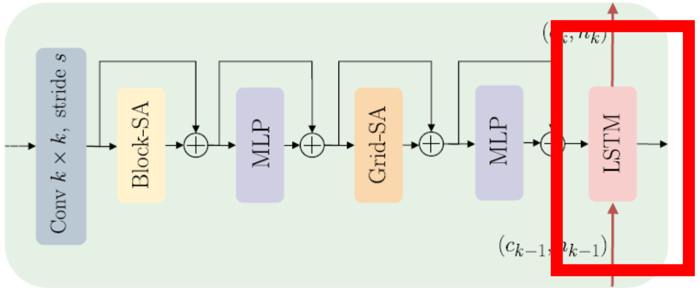
$(\frac{H}{P} X \frac{W}{P}, P X P, C)$ P X P : window size

4. Grid Self Attention

Global feature interaction

$(G X G, \frac{H}{G} X \frac{W}{G}, C)$

3. Method



Temporal Feature Extraction

Aggregation with LSTM

Parameter □ □□ □□□□□ □□ Conv-LSTM □□ Plain LSTM □□

More detail

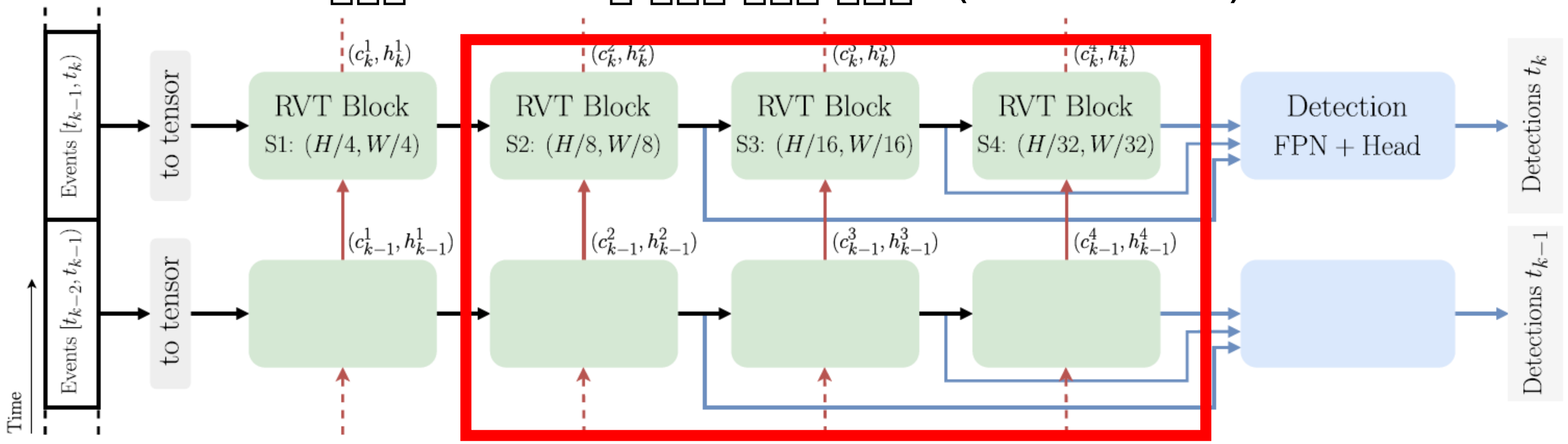
- □ attention □ MLP □□□ □□□□□ □, Layer Norm □ □□□.
- □ □□□□ Residual connection □□□.

3. Method

Hierarchical Multi-Stage Design

□□□ module □ □□□ □□□ □□□ . (LSTM state)

□□□□
Slice of events



2, 3, 4 □□ state □ hidden layer □ detection □ □□□

4. Experiments

				Channels		
Stage	Size	Kernel	Stride	RVT-T	RVT-S	RVT-B
S1	1/4	7	4	32	48	64
S2	1/8	3	2	64	96	128
S3	1/16	3	2	128	192	256
S4	1/32	3	2	256	384	512

□□ □□ :
1. **RVT-B (□□ □□):** □□ □□ .
2. **RVT-S (□□ □□):** RVT-B □ □□ □□ .
3. **RVT-T (□□ □□ □□):** RVT-B □ □□ □□ □□ .

4. Experiments

4.1 Setup – Implementation Details

1.□□ □□□ :

□ □□□□ LayerScale □ □□□ □□□□□ □□□□□ □□□.

2.□□ □□ :

1. □□□ ADAM □□□□□□ □□□□ 40□ □□ iteration □□ □□ □□□□ □□□□□.
2. OneCycle □□□ □□□ □□□□, □□ □□□□□ □□ □□□□□.

3.□□ □□ :

□□ □□ □□□ □□□□, □□□ □□□ □□ BPTT□ □□□□ □□ □□□□ Truncated BPTT□ □□□□□.

4. Experiments

4.1 Setup – Implementation Details

4. □□□ □□ :

random horizontal flipping, zooming in and zooming out □□□ □□□□□.

5. □□□ □□ :

50ms □□ □□ □□□□, □□ 10□□ □□(T=10)□□ □□□
□□□ □□□ □□□□□.

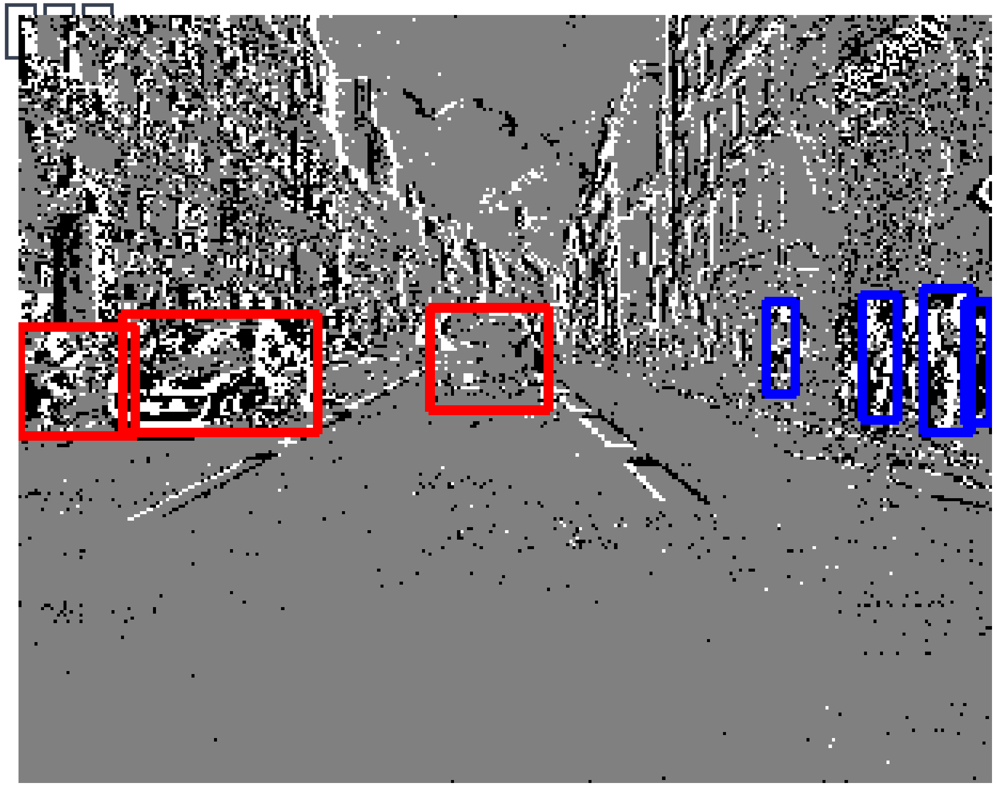
6. □□□□□ □ □□ □□ :

YOLOX □□□□□□ □□□□, IOU loss, class loss and regression loss 가
□ □□□ □□□□ □□ □ □□□ □□□ □□ □□□□□□.

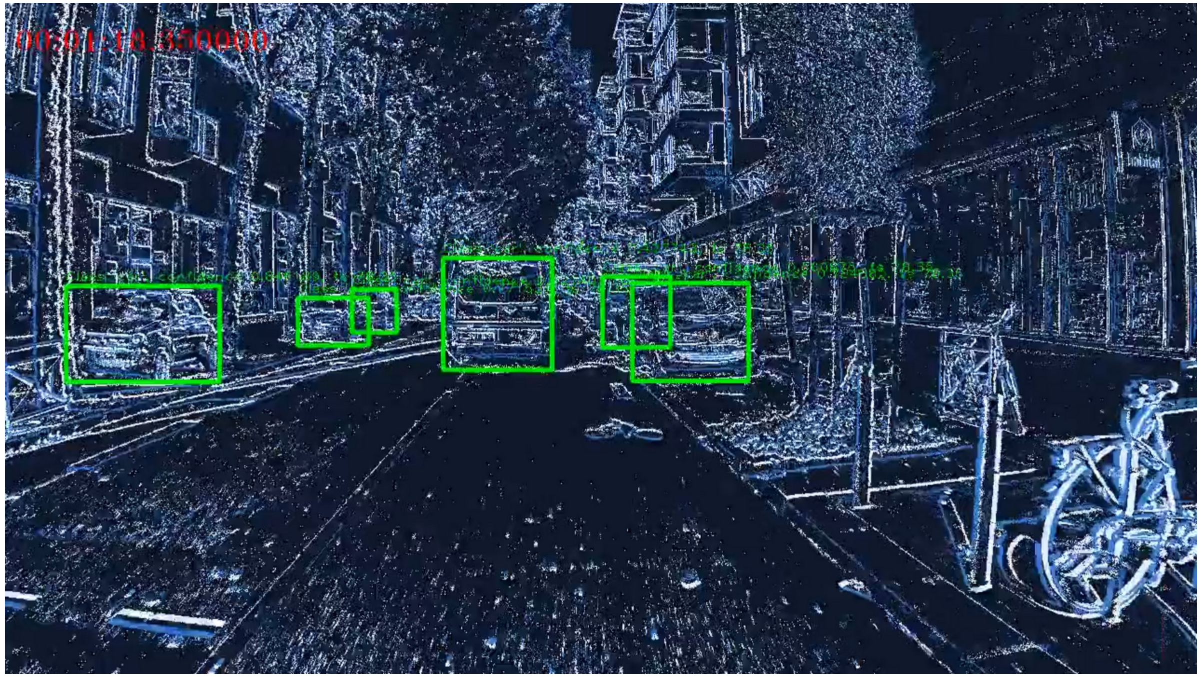
4. Experiments

4.1 Setup - Dataset

Gen1 Automotive Detection □□□□



1 MPx □□□□



4. Experiments

4.2 Ablation Studies – Model Components

	Gen1		1 Mpx		
Block-type	mAP	AP_{50}	mAP	AP_{50}	Params (M)
multi-axis	**47.6**	**70.1**	**46.0**	**72.3**	18.5
Swin	46.7	68.7	44.4	71.7	18.5
ConvNeXt	45.5	65.8	42.3	70.6	18.7

Table 2. **Spatial Aggregation**. Multi-axis attention leads to the best results on both the Gen1 and 1 Mpx dataset.

Conv. kernel type	mAP	AP_{50}	AP_{75}	Params (M)
overlapping	**47.6**	**70.1**	**52.6**	18.5
non-overlapping	46.1	68.6	50.5	**17.6**

Table 3. **Downsampling Strategy**. The usage of overlapping kernels leads to higher performance at the expense of a slight increase in the number of parameters.

4. Experiments

4.2 Ablation Studies – Model Components

LSTM kernel size	mAP	AP_{50}	AP_{75}	Params (M)
1 × 1	**47.6**	**70.1**	**52.6**	**18.5**
3 × 3	46.5	69.0	51.4	40.8
3 × 3 depth-sep	46.3	67.2	51.2	18.6

Table 4. **LSTM kernel size.** Conv-LSTM variants do not outperform the feature specific (1 × 1) LSTM.

S1	S2	S3	S4	mAP	AP_{50}	AP_{75}
				32.0	54.8	31.4
			✓	39.8	63.5	41.6
		✓	✓	44.2	68.4	47.5
	✓	✓	✓	46.9	70.0	50.8
✓	✓	✓	✓	**47.6**	**70.1**	**52.6**

Table 5. **LSTM placement.** LSTM cells contribute to the overall performance even in the early stages.

4. Experiments

4.2 Ablation Studies – Data Augmentations

h-flip	zoom-in	zoom-out	mAP	AP_{50}	AP_{75}
			38.1	59.5	41.1
✓			41.6	63.5	45.5
	✓		45.8	67.8	49.8
		✓	44.1	65.7	48.4
✓	✓	✓	**47.6**	**70.1**	**52.6**

Table 7. **Data Augmentation.** Data augmentation consistently improves the results.

4. Experiments

4.3 Benchmark Comparisons

Method	Backbone	Detection Head	Gen1 mAP	Gen1 Time (ms)	1 Mpx mAP	1 Mpx 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	<u>46.7</u>	35.6	**48.3**	72.3	> 100*
RVT-B (ours)	Transformer + RNN	YOLOX [15]	**47.2**	**10.2 (3.7)**	<u>47.4</u>	**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

4. Experiments

4.3 Benchmark Comparisons

- [illegible]:
    - Gen1 [illegible] **47.2 mAP**, 1 MPx [illegible] **47.4 mAP**[illegible]
- **ASTMNet:**
    - ASTMNet[illegible] backbone [illegible]
- **RED** [illegible]:
    - RED [illegible] Gen1 [illegible] **mAP**가 **7.2** [illegible],
    - 1 MPx [illegible] **4.4** [illegible]
- **Tiny** [illegible]:
    - Gen1 [illegible] RED [illegible] **4.1** [illegible] **mAP**[illegible] 5[illegible]

4. Experiments

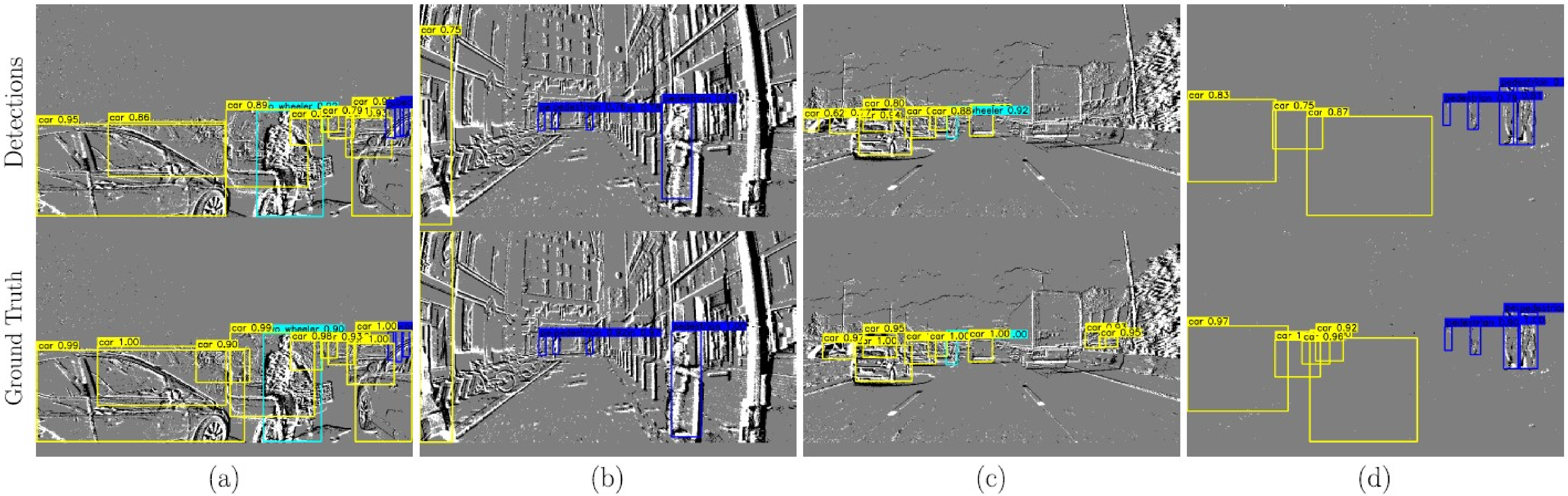


Figure 4. **Predictions on the 1 Mpx dataset**. All examples are thematically picked to illustrate the behaviour of the model in different scenarios. (d) shows a scenario in which the model can still partially detect objects in absence of events due the temporal memory.

6. Conclusion

[illegible] :

- [illegible] [illegible] [illegible] [illegible] [illegible] [illegible] Backbone [illegible] [illegible].
- [illegible] [illegible] [illegible] [illegible] [illegible] [illegible] [illegible] [illegible] [illegible] [illegible].

[illegible] [illegible] [illegible] :

Convolution prior
local- and sparse global attention
recurrent feature aggregation

[illegible] :

- RVT[illegible] [illegible] [illegible] [illegible] [illegible] [illegible] [illegible] [illegible] [illegible] [illegible] [illegible] [illegible] [illegible] [illegible] [illegible].

[illegible] [illegible] [illegible] :

- [illegible] [illegible] [illegible] [illegible] [illegible] [illegible] [illegible], [illegible] [illegible] [illegible] low -latency [illegible] [illegible] 가 [illegible].

03.
Relevance to the subject

Optimizer : Adam
Batch Strategy: BPTT , Truncated BPTT
Precision: MAP
Model LSTM, Transformer, CNN

Etc..
Overfitting □ □□□ □□ □□□ □□□□ □□
MLP □□
Residual □□ □□

□□ □□

https://m.hanbit.co.kr/channel/category/category_view.html?cms_code=CMS6074576268
https://wikidocs.net/31379
https://en.wikipedia.org/wiki/Event_camera
https://ko.wikipedia.org/wiki/%EB%AA%A8%EC%85%98_%EB%B8%94%EB%9F%AC

KNU KYUNGPOOK NATIONAL UNIVERSITY