Recurrent Vision Transformers for Object Detection with Event Cameras

IT대학 컴퓨터학부 2021111183 김은지
IT대학 컴퓨터학부 2021114818 김찬호

KNU

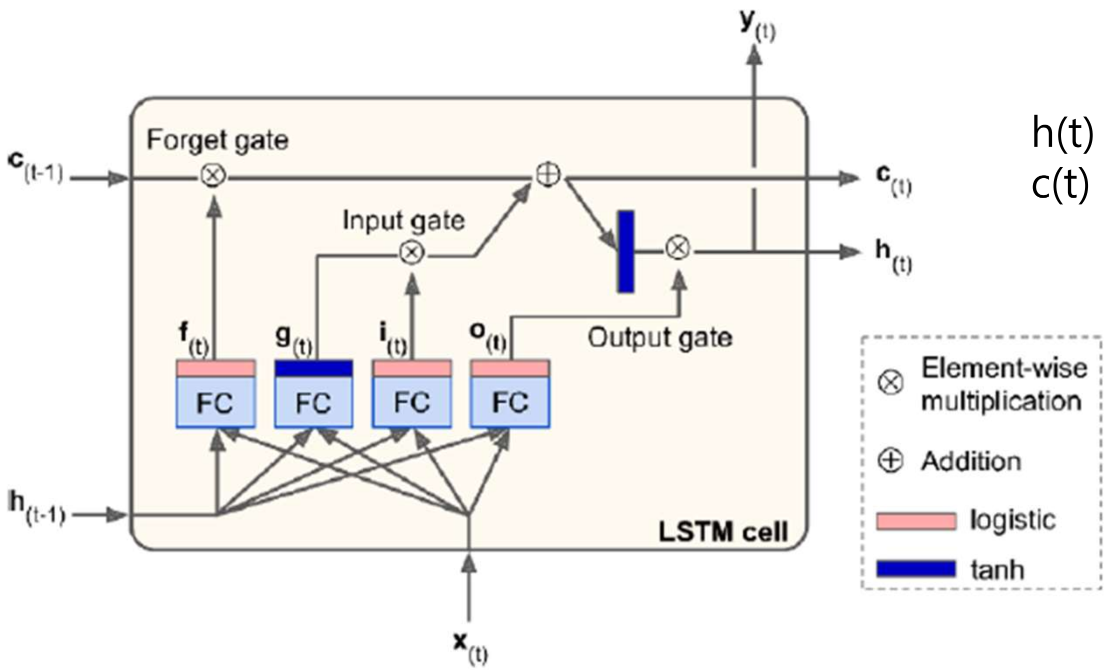
경북대학교

CONTENTS

01.
Before Reviewing...

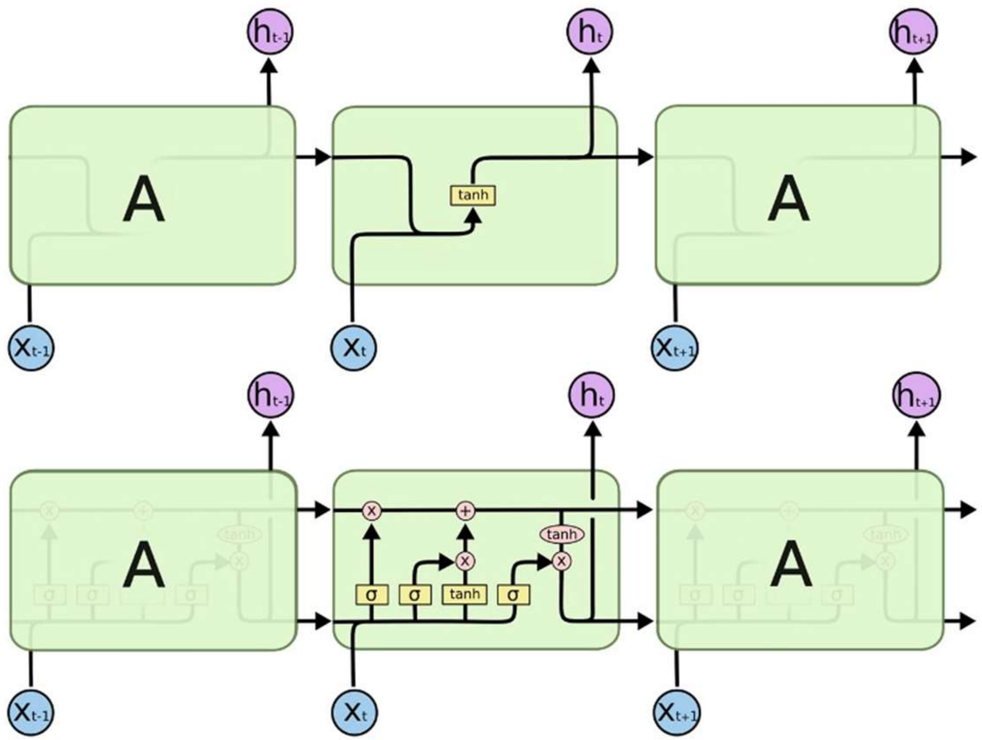
LSTM (Long short-term memory)

RNN 기법 중 하나로, gate를 추가하여 기존 순환 신경망에서 발생하는 Vanishing Gradient Problem을 해결함.



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

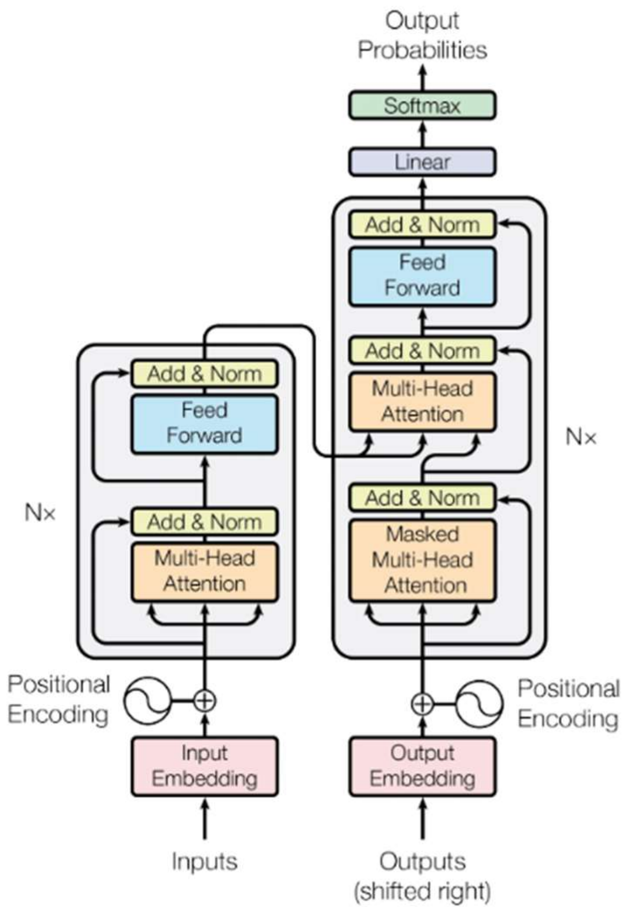
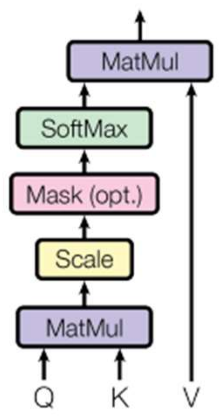


Figure 1: The Transformer - model architecture.

Transformer

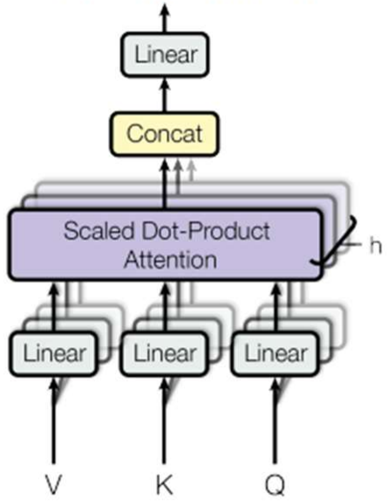
Attention : can be described as mapping a query and a set of key-value pairs to an output

Scaled Dot-Product Attention



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

Multi-Head Attention



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

02.

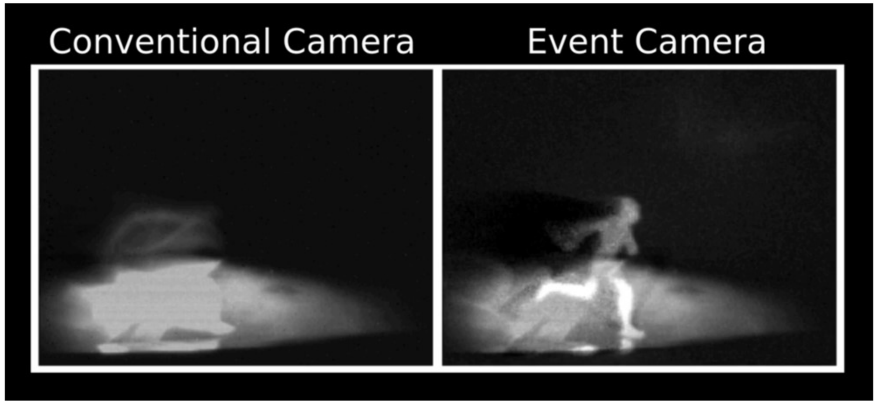
Main Review

1. Introduction

What is **Event Camera**?

픽셀 간의 밝기 변화를 비동기적으로 측정하고 기록하는 카메라

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)	↓

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

1. Introduction

Event Camera 의 한계점

시간, 공간에 따라 비동기적인 binary event가 발생
-> 시공간 영역에서 detection 을 진행하는 algorithm을 개발할 필요성이 있음.

선행 연구 (Related Work)

- GNN(Graph Neural Network)
한계점) 너무 heavy 한 backbone 사용 & ConvLSTM처럼 expensive한 cell을 사용
-> sparce neural network 로 vision backbone을 design하자

1. Introduction

본 연구에서 inference time과 performance간 균형을 유지하기 위해 사용한 기법

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

연구 요약

1. event-based pipelines 에 우세한 design 제작
2. Simple, composable 한 state design
3. State-of-the-art object detection 에서 우수한 성능을 도출함.

2. Related Work

Vision Transformation for Spatio-Temporal Data

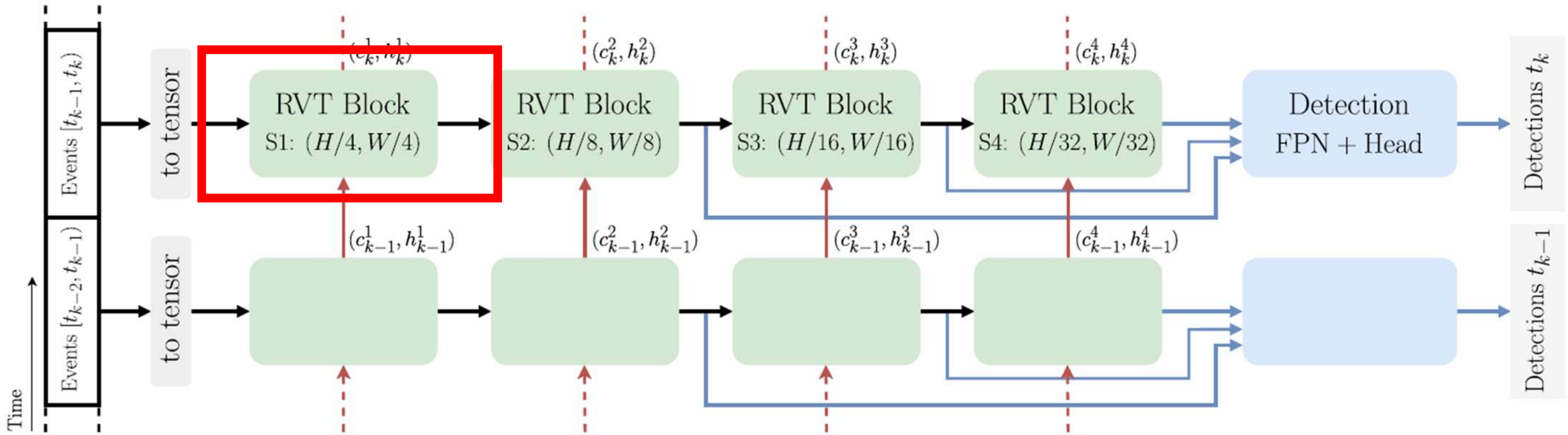
In event-based vision

- classification
- image reconstructure
- monocular depth estimation

=> Object detection has yet to be investigated

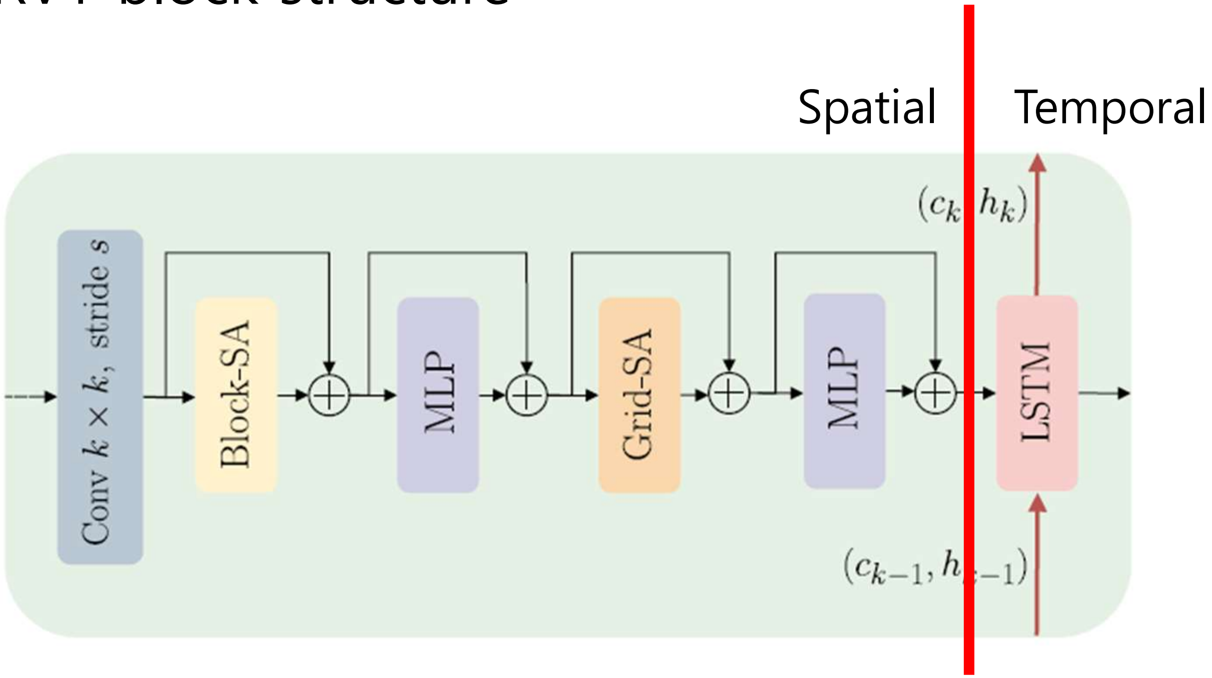
3. Method

Overall Structure



3. Method

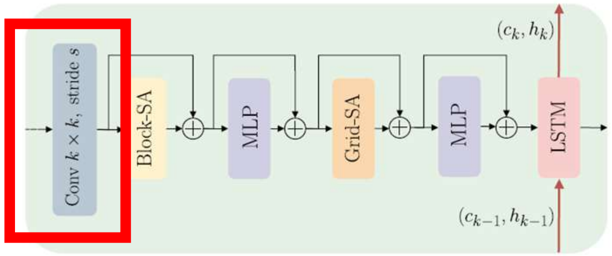
RVT block structure



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

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

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

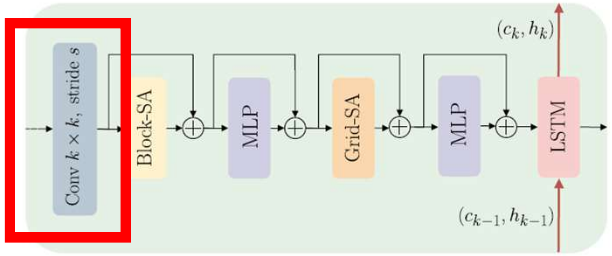
(2T, H, W)

2D convolution을 호환하는 형태로 변환

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

3. Method

Spatial Feature Extraction

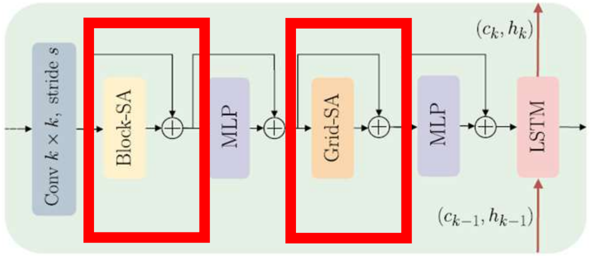


1. Convolution (overlapping)

Convolution with overlapping kernels

=> non-overlapping model 보다 약간의 성능 향상이 이루어짐.

3. Method



Multi-axis attention self-attention

2. Block Self Attention

Local feature interaction

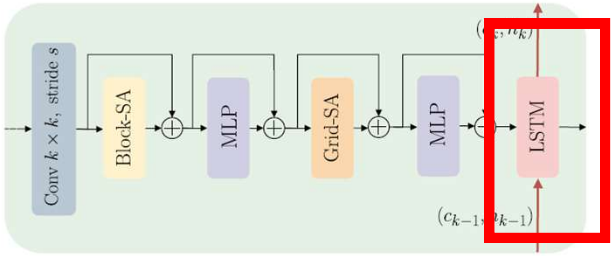
$(\frac{H}{P} X \frac{W}{P}, P X P, C)$ P X P : window size

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

Parameter의 수를 감소시키기 위해 Conv-LSTM 대신 Plain LSTM 사용

More detail

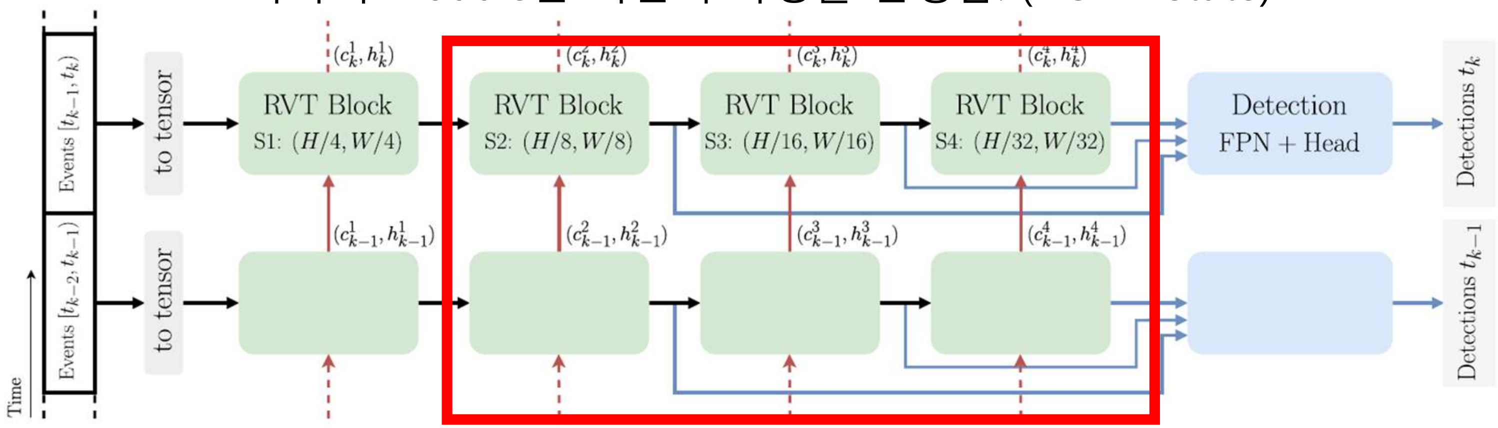
- 각 attention 과 MLP 모델이 이루어지기 전, Layer Norm을 적용함.
- 각 모듈마다 Residual connection 사용함.

3. Method

Hierarchical Multi-Stage Design

각각의 module은 이전의 특징을 반영함. (LSTM state)

시작부분
Slice of events



2, 3, 4번째 state를 hidden layer 로 detection을 진행함

4. Experiments

Stage	Size	Kernel	Stride	Channels		
				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

모델 변형:
1. **RVT-B (기본 모델):** 원본 모델.
2. **RVT-S (작은 모델):** RVT-B의 작은 변형.
3. **RVT-T (아주 작은 모델):** RVT-B의 아주 작은 변형.

4. Experiments

4.1 Setup – Implementation Details

1. 모델 초기화:

각 모듈에서 LayerScale를 제외한 파라미터는 랜덤값으로 초기화.

2. 학습 설정:

1. 모델은 ADAM 옵티마이저를 사용하여 40만 번의 iteration 동안 혼합 정밀도로 훈련됩니다.
2. OneCycle 학습률 일정을 사용하며, 최대 학습률에서 선형 감소합니다.

3. 배치 전략:

혼합 배치 전략을 사용하며, 배치의 절반에 대해 BPTT를 적용하고 다른 절반에는 Truncated BPTT를 적용합니다.

4. Experiments

4.1 Setup – Implementation Details

4. 데이터 증강:

random horizontal flipping, zooming in and zooming out 기법을 사용합니다.

5. 이벤트 표현:

50ms 시간 창을 고려하며, 이를 10개의 구간(T=10)으로 나누어
이벤트 표현을 구성합니다.

6. 프레임워크 및 손실 함수:

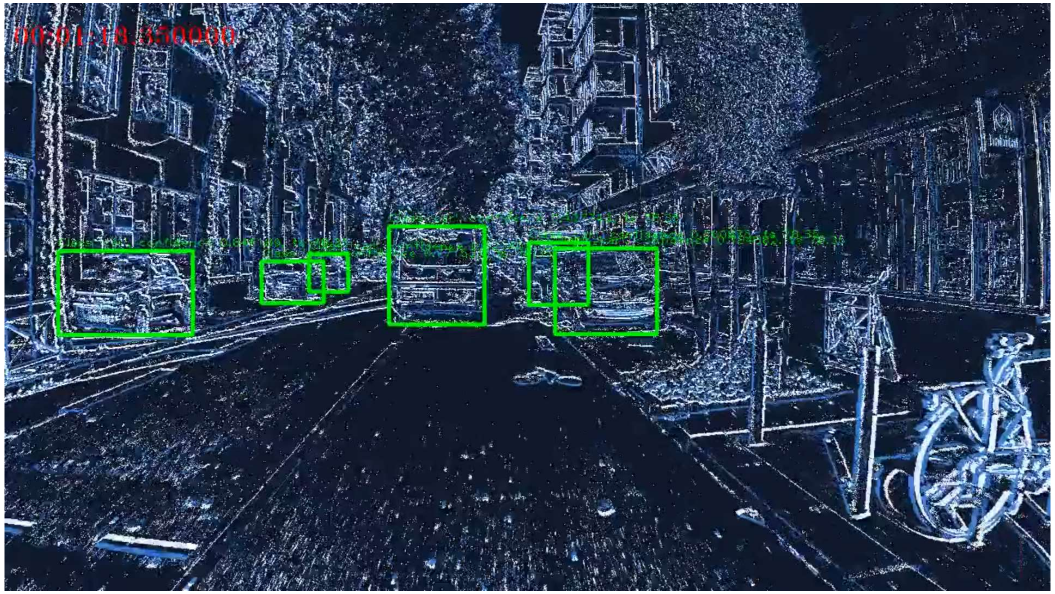
YOLOX 프레임워크를 사용하며, IOU loss, class loss and regression loss가
각 최적화 단계에서 배치 및 시퀀스 길이에 대해 평균화됩니다.

4. Experiments

4.1 Setup - Dataset

Gen1 Automotive Detection 데이터셋

1 MPx 데이터셋



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	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

4. Experiments

4.3 Benchmark Comparisons

•**베이스 모델 성능:**

- Gen1 데이터셋에서 **47.2 mAP**, 1 MPx 데이터셋에서 **47.4 mAP**로 새로운 최고 성능을 달성

•**ASTMNet:**

- ASTMNet은 더 큰 backbone과 증가된 추론 시간을 사용하면서 두 데이터셋에서 비슷한 결과

•**RED 모델:**

- RED 모델은 우리 모델에 비해 Gen1 데이터셋에서 **mAP가 7.2 낮고**,
- 1 MPx 데이터셋에서는 **4.4 낮은 성능**

•**Tiny 모델:**

- Gen1 데이터셋에서 RED 모델보다 **4.1 더 높은 mAP**를 달성하면서 파라미터는 5배 적게 사용

4. Experiments

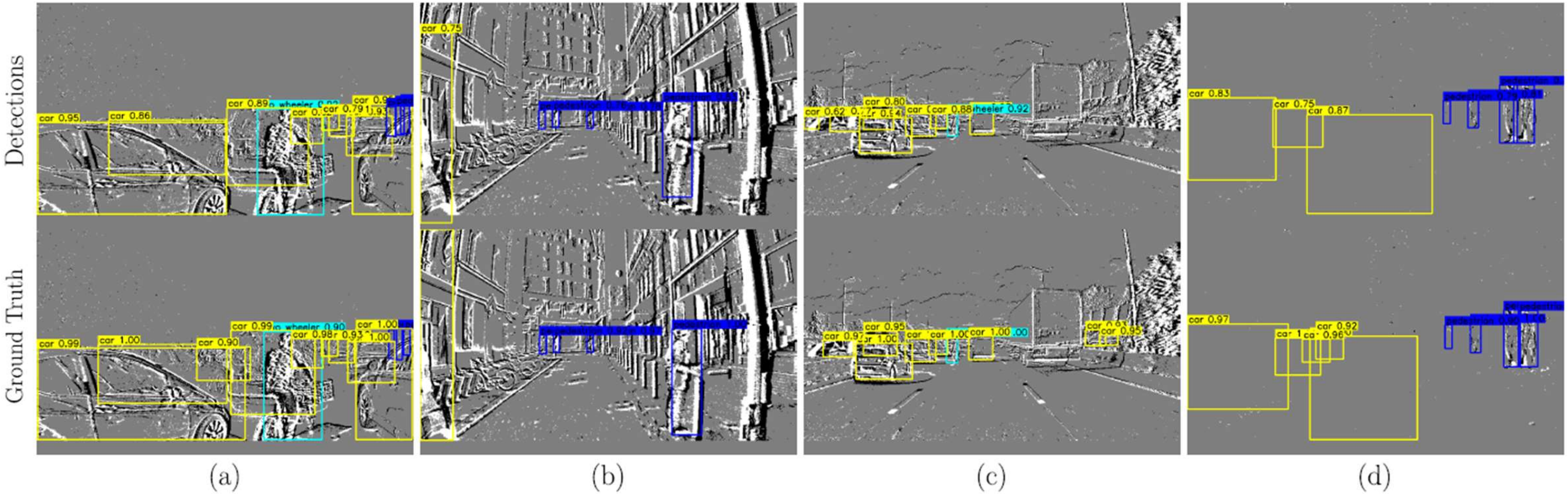


Figure 4. **Predictions on the 1 Mpx dataset**. All examples are thematically picked to illustrate the behaviour of the model in different scenarios. (d) shows a scenario in which the model can still partially detect objects in absence of events due the temporal memory.

6. Conclusion

아키텍처:

- 이벤트 카메라를 활용한 객체 감지에 새로운 Backbone 아키텍처를 소개.
- 다단계 계층적 신경망 형성을 위한 반복 적용되는 스테이지 디자인 도입.

스테이지 디자인 특징:

Convolution prior
local- and sparse global attention
recurrent feature aggregation

성능 :

- RVT는 이벤트 카메라 객체 감지에서 최첨단 성능을 처음부터 훈련하여 얻을 수 있음을 실험으로 확인.

결과 및 호환성:

- 표준 스테이지 디자인은 기존의 감지 프레임워크와 호환되며, 이벤트 카메라를 사용한 low -latency객체 감지를 표준 하드웨어에서 가능케 함.

03.
Relevance to the subject

Relevance to the subject

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

Etc..
Overfitting을 피하기 위해 데이터 증강기법 사용
MLP사용
Residual 기법 활용

참고 자료

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