Rethinking Scale-Aware Temporal Encoding for Event-based Object Detection

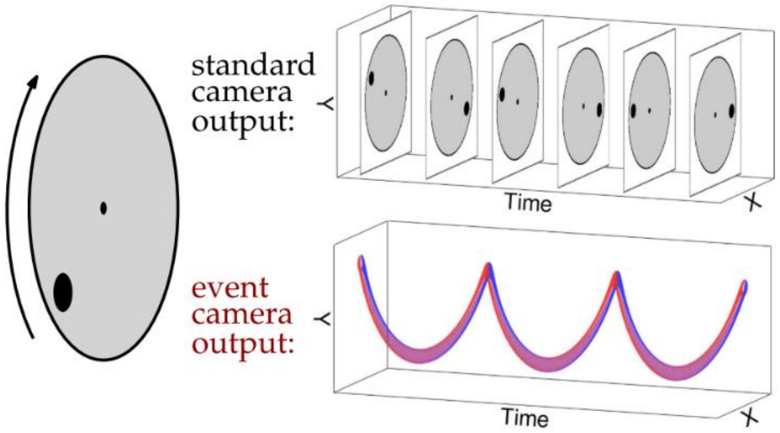
Lin Zhu [1], Tengyu Long [1], Xiao Wang [2], Lizhi Wang [3], Hua Huang [3]

[1] Beijing Institute of Technology, Beijing, China
[2] Anhui University, Anhui, China
[3] Beijing Normal University, Beijing, China

Introduction

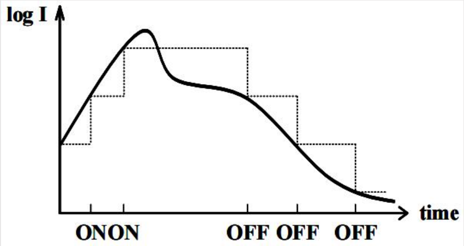
Event Camera

Records brightness changes at each pixel asynchronously once the change exceeds a preset threshold.



Advantages

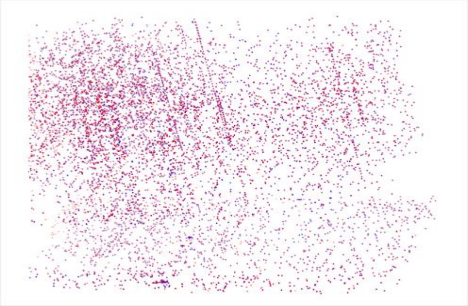
- ✓ High temporal resolution.
- ✓ Wide dynamic range.
- ✓ Low power consumption



- ✓ Enable robust detection in challenging scenarios (high-speed motion, low-light environments, etc).

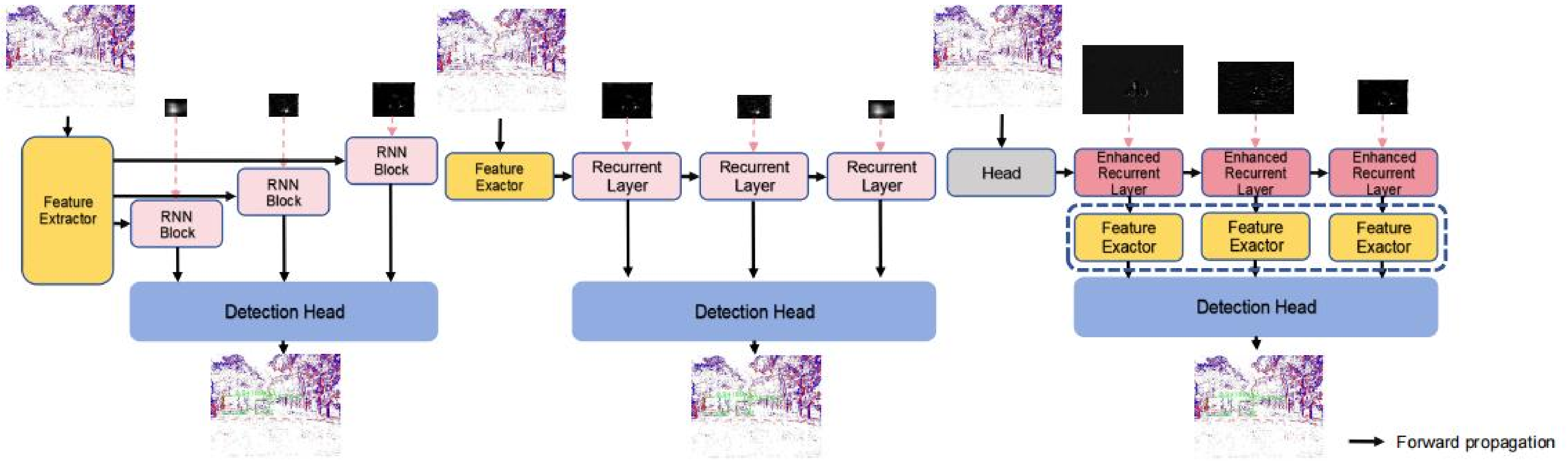
Challenge

- ☐ How to effectively modeling the temporal dynamics of the asynchronous event streams



Our thought

Existing CNN- or Transformer-based methods combined with recurrent modules comparison



(a) High-Level Parallel RNN-based network (b) High-Level Cascaded RNN-based network (c) Ours

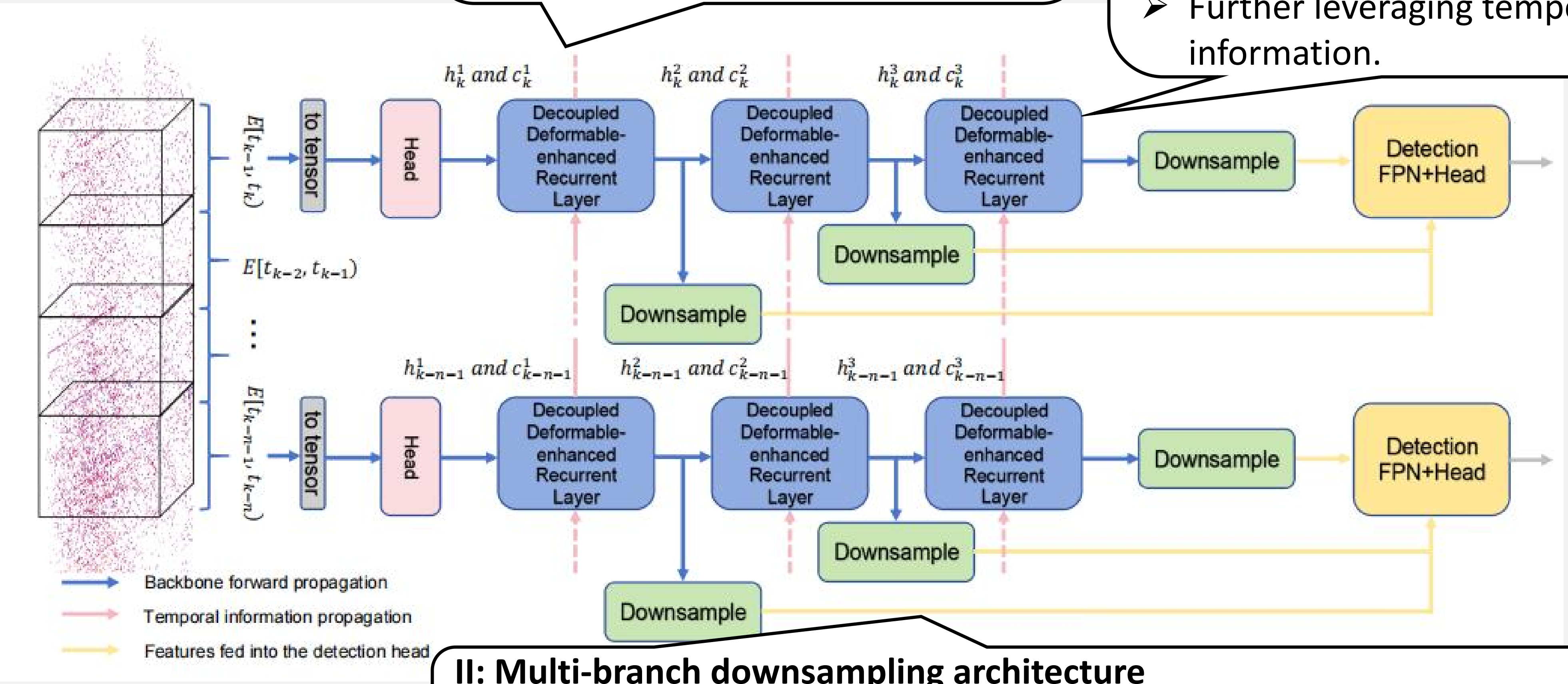
Motivated by the following characteristics of event data:

- **Sparse and noisy**: low-level features contain richer temporal cues but also more task-irrelevant noise.
- **Containing inherent motion information**: event cameras can capture the relative motion of objects.

Overall architecture

I: Temporal Modeling at Lower Scales
- Introduce our designed recurrent layer at lower scales.
  - ➢ Model fine-grained temporal structures.

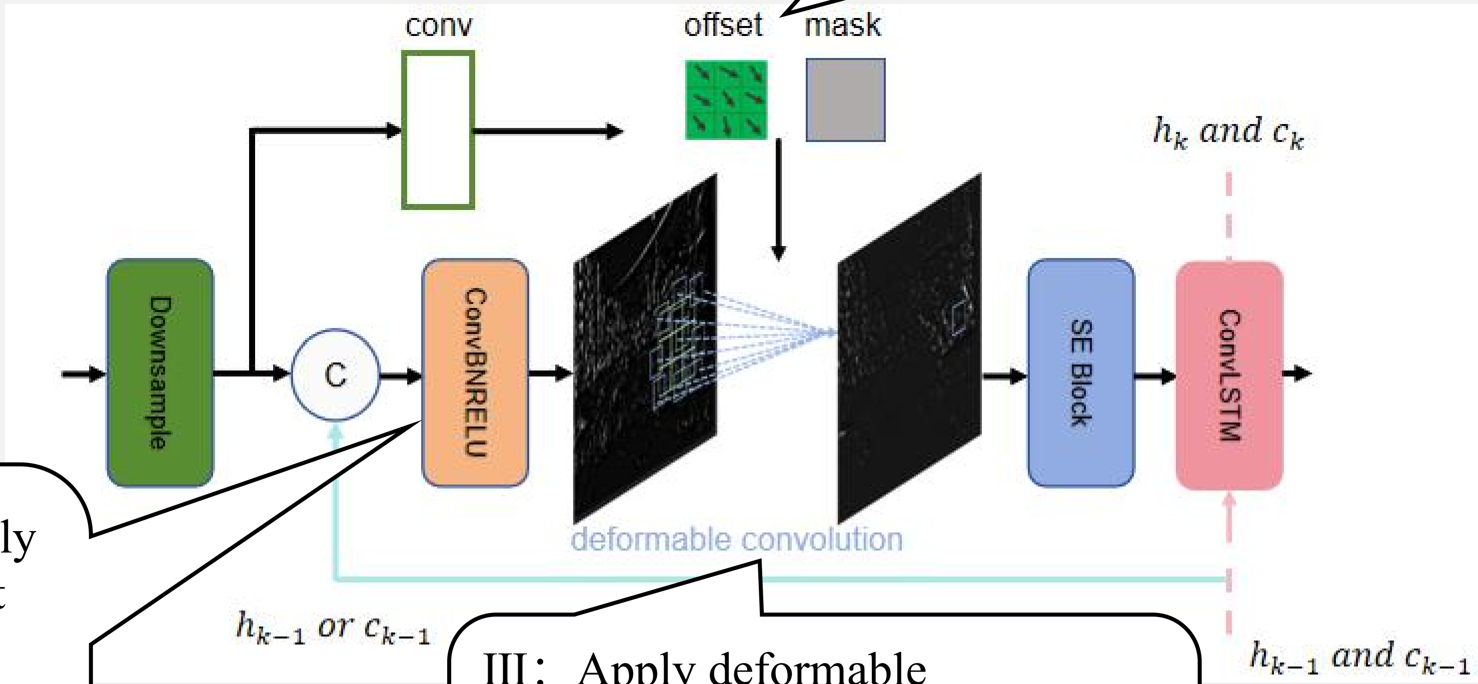
III: Decoupled Deformable-enhanced Recurrent Layer (DDRL)
- Adopt a divide-and-conquer strategy to decouple feature fusion and motion estimation.
  - ➢ Further leveraging temporal information.



II: Multi-branch downsampling architecture
- Perform independent spatiotemporal encoding across multiple scales.
  - ➢ Achieve scale-adaptive and hierarchical representation learning.

Architecture of DDRL

I： Learn the offsets and masks required for deformable convolution from the motion information of the current frame.



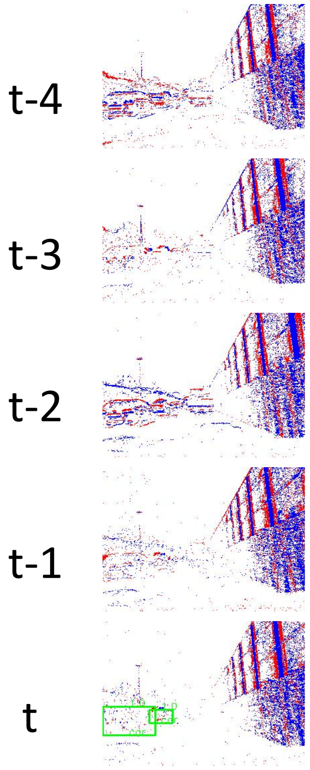
II： Preliminarily fuse the current frame features with historical spatiotemporal features.

III： Apply deformable convolution to adaptively adjust the receptive field and calibrate motion patterns.

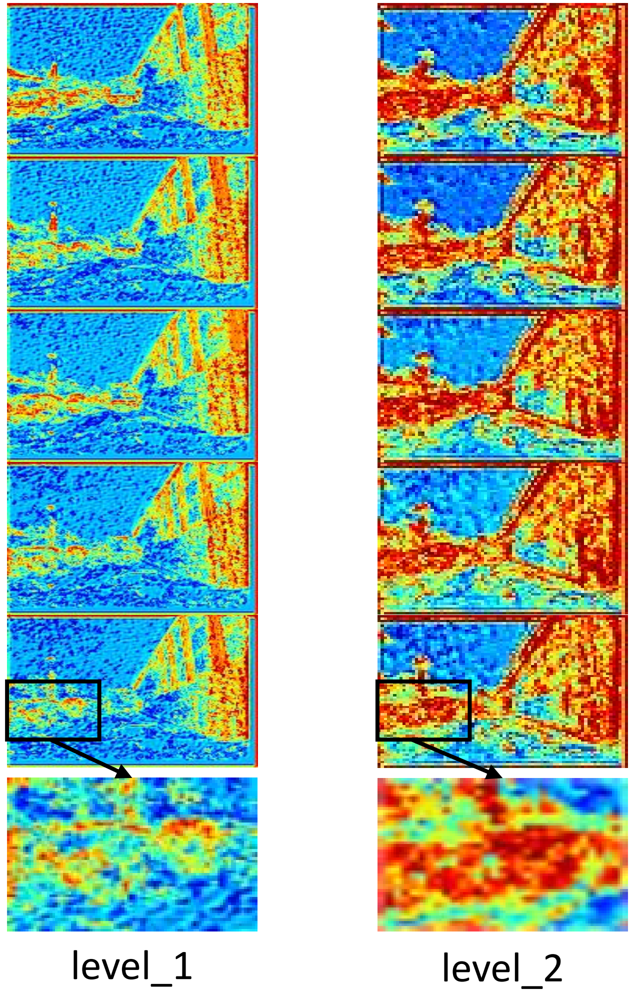
Not only improves feature alignment for moving objects but also filters out task-irrelevant noise .

Feature Visulizations

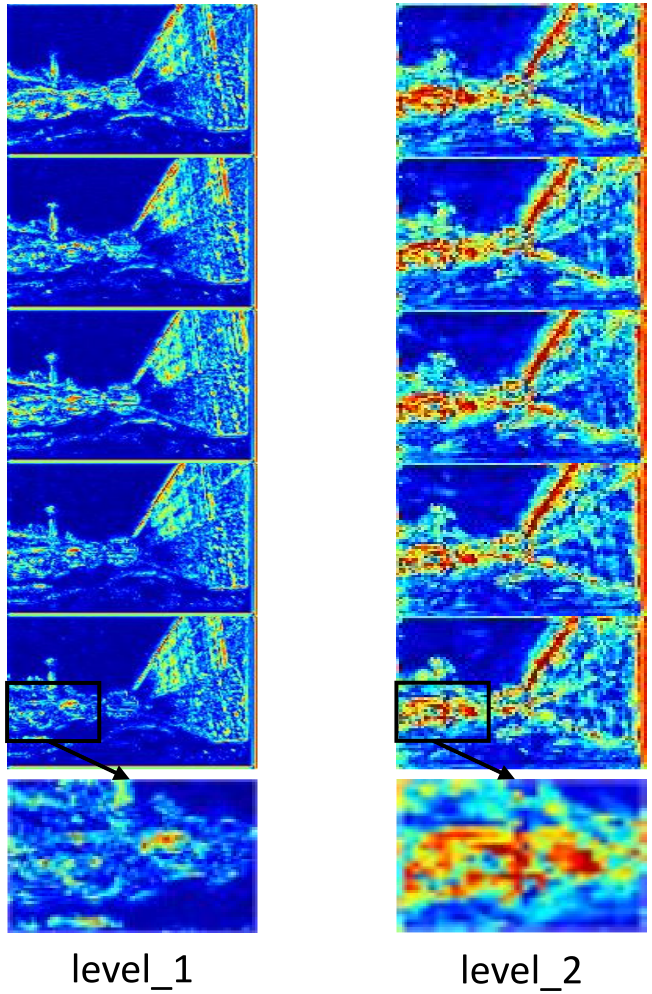
Inputs



Conventional Recurrent Layers



DDRL



* Here, level_n denotes the output features of the n-th recurrent layer.

Experiments

Results on Gen1 and 1 Mpx

Table 1: Comparison with state-of-the-art methods on Gen1 and 1 Mpx datasets.

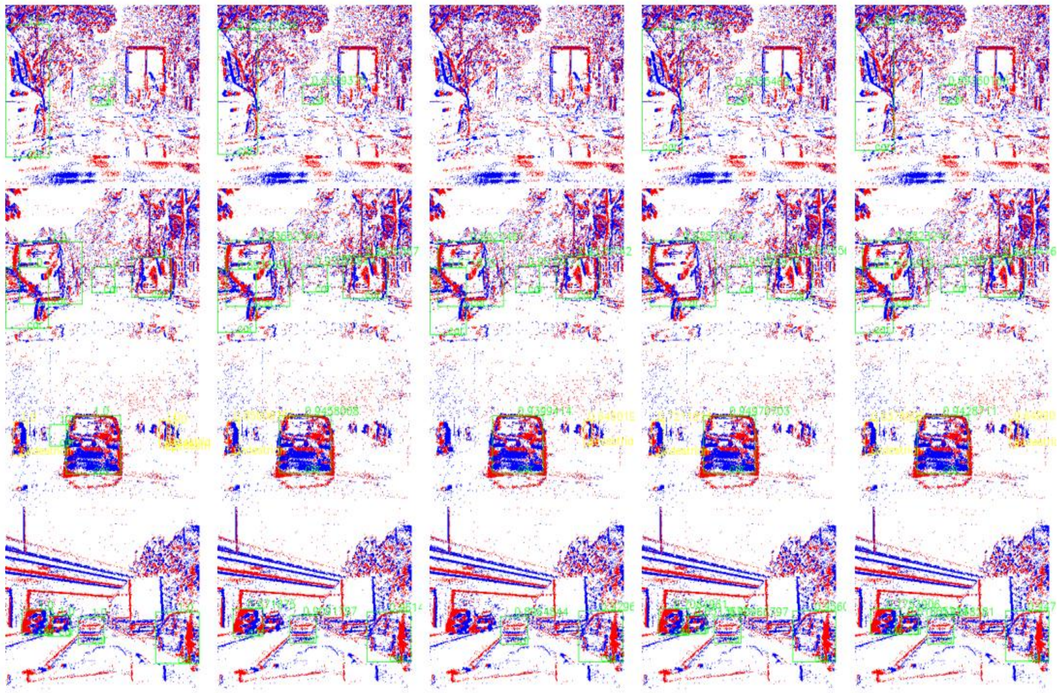
Method	Params	Backbone	Gen1 mAP	Time (ms)	1Mpx mAP	Time (ms)
Asynet	11.4	Sparse CNN	14.5	-	-	-
AEGNN	20.0	GNN	16.3	-	-	-
Spiking DenseNet	8.2	SNN	18.9	-	-	-
Inception + SSD	> 60*	CNN	30.1	19.4	34.0	45.2
RRC-Events	> 100*	CNN	30.7	21.5	34.3	46.4
MatrixLSTM	61.5	CNN + RNN	31.0	-	-	-
YOLOv3 Events	> 60*	CNN	31.2	22.3	31.6	49.4
RED	24.1	CNN + RNN	40.0	16.7	43.0	39.3
ASTMNet	> 100*	CNN + RNN	46.7	35.6	48.3	72.3
ERGO-12	59.6	Transformer	50.4	69.9	46.0	100.0
RVT-B	18.5	Transformer + RNN	47.2	10.2	47.4	11.9
Swin-T v2	21.1	Transformer + RNN	45.5	26.6	45.5	34.8
Nested-T	22.2	Transformer + RNN	46.3	20.6	46.0	33.5
GET-T	21.9	Transformer + RNN	47.9	16.8	48.4	18.2
SAST-CB	18.9	Transformer + RNN	48.2	22.7	48.7	23.6
S5-ViT-B	18.2	Transformer + SSM	47.7	**8.16**	47.8	**9.57**
Ours	26.4	CNN + RNN	**52.7**	8.80	**49.1**	13.3

Results on eTram

Table 2: Comparison with state-of-the-art methods on the traffic monitoring dataset eTram.

Method	Backbone	mAP	Time (ms)
RVT-B	Transformer + RNN	29.5	**10.88**
SAST-CB	Transformer + RNN	30.0	23.07
S5-ViT-B	Transformer + SSM	29.3	14.84
Ours	CNN + RNN	**33.0**	13.05

Result visualization



Thank you!

Contact:
https://github.com/BIT-Vision/SATE
linzhu@bit.edu.cn
longtengyu@bit.edu.cn