CVPR 2023

TriDet: Temporal Action Detection with Relative Boundary Modeling

Dingfeng Shi∗
VRLab, Beihang University, China
shidingfeng@buaa.edu.cn

Qiong Cao†
JD Explore Academy
mathqiong2012@gmail.com

Yujie Zhong
Meituan Inc.
jaszhong@hotmail.com

Lin Ma
Meituan Inc.
forest.linma@gmail.com

Jia Li†
VRLab, Beihang University, China
jiali@buaa.edu.cn

Dacheng Tao
JD Explore Academy
dacheng.tao@gmail.com

Citations 6

Table of Contents

Introduction

- **TAD** **Temporal action detection**

Introduction

- **TAD**$_{\textbf{Temporal action detection}}$ remains to be a very challenging task due to some unresolved problems.
  - TAD is that **action boundaries** are **usually not obvious**

Object Detection

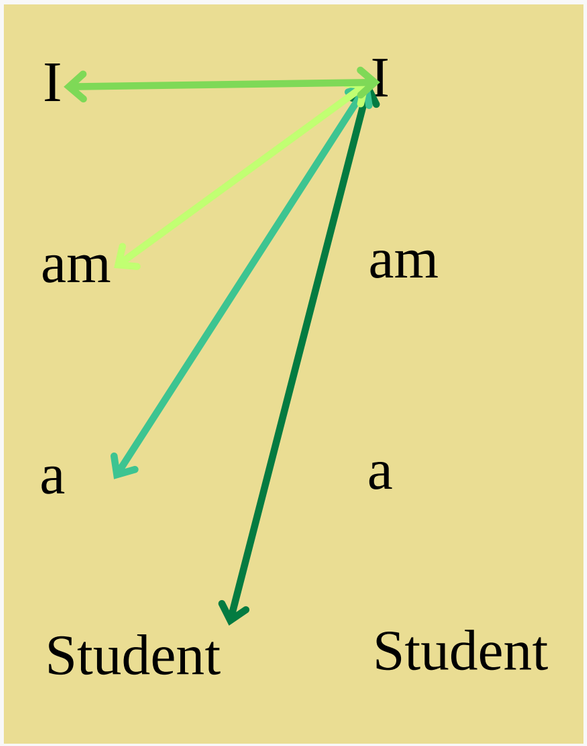
TAD

Typing on the keyboard ?

Writing the memo ?

Introduction

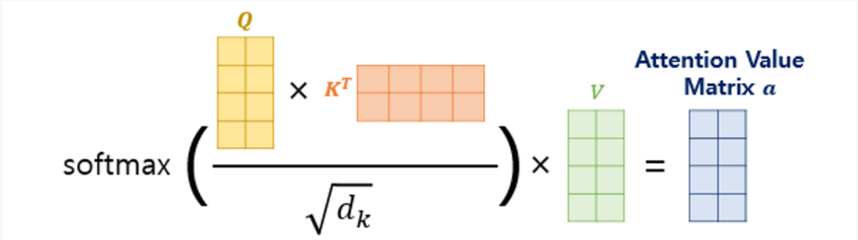
- **Self Attention**



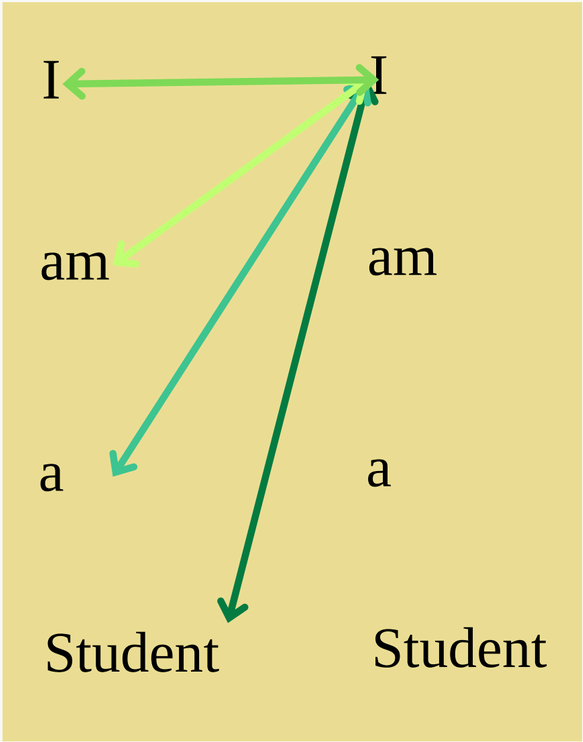
- I am a student
- **I > student**

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

Introduction



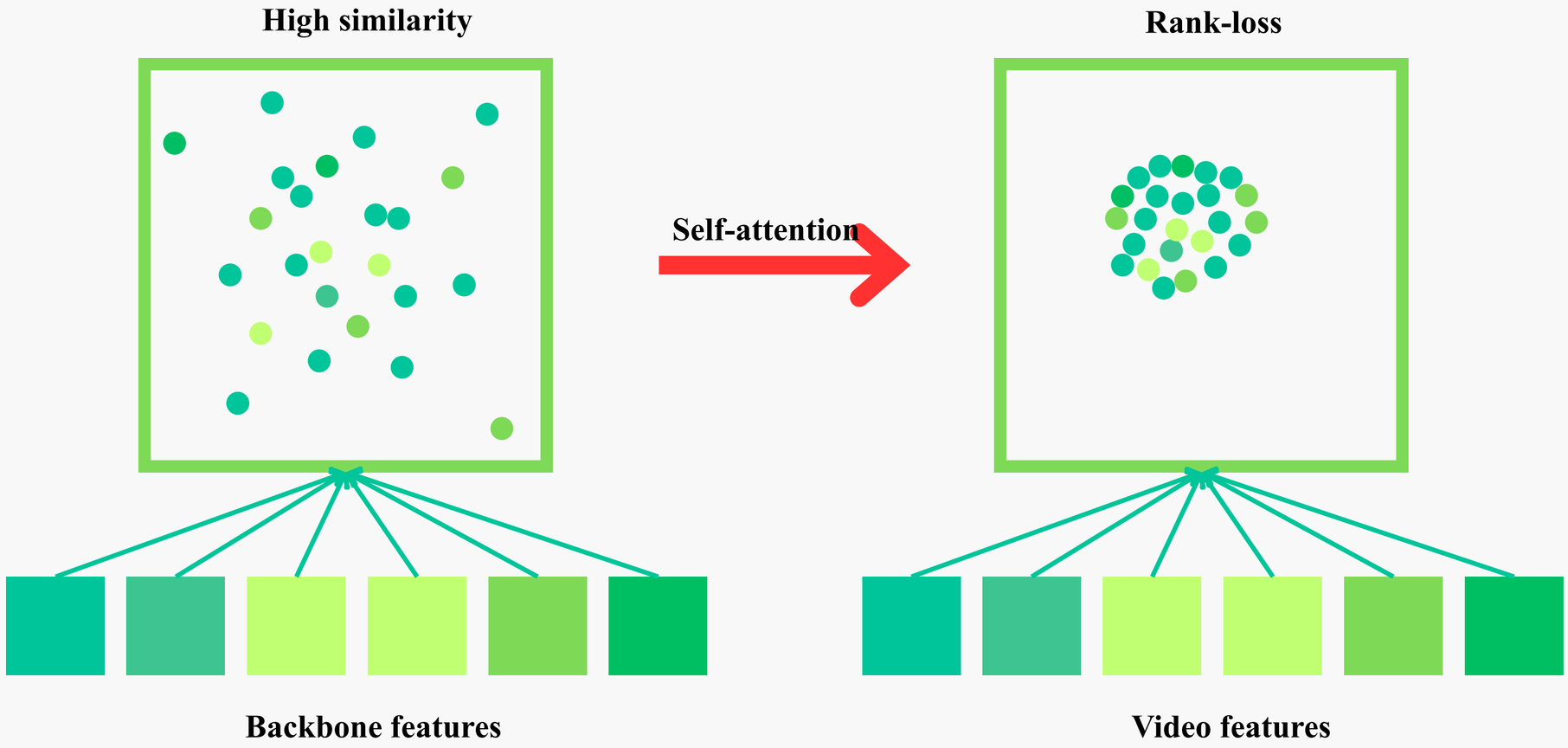
- **Self Attention**



	I	am	a	Student
I	**0.6**	**0.05**	**0.05**	**0.3**
am	**0.3**	**0.6**	**0.05**	**0.05**
a	**0.3**	**0.05**	**0.6**	**0.05**
Student	**0.3**	**0.05**	**0.05**	**0.6**

Introduction

- **Self attention**

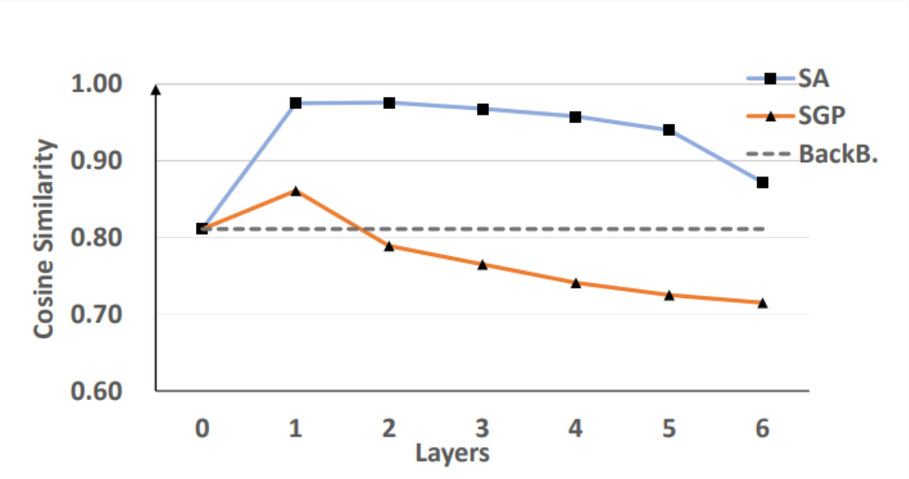


Introduction

Previous works

Transformer -base : Actionformer

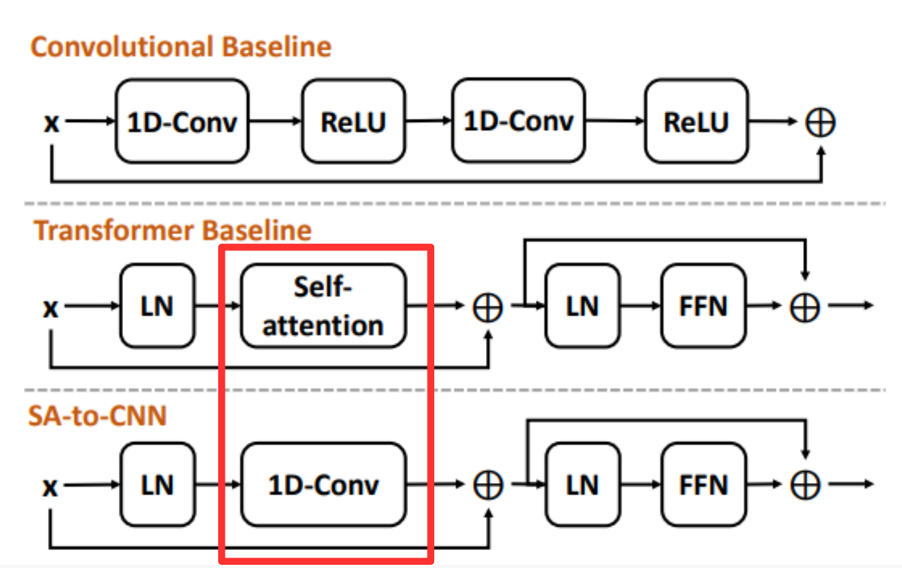
- Current Transformer-based methods for TAD tasks primarily rely on the macro-architecture of the Transformer rather than the self-attention mechanisms.
- We observe that the SA exhibits high similarity, indicating poor discriminability



HACS dataset and SlowFast backbone

[49] Chenlin Zhang, Jianxin Wu, and Yin Li. Actionformer: Localizing moments of actions with transformers. In Eur. Conf. Comput. Vis., 2022

Introduction



Method	Avg.	
CNN Baseline	58.7	-8.1%
Transformer Baseline	66.8	
SA-to-CNN	64.9	-1.9%

Introduction

- **Self Attention**
  - **Rank loss problem**
    - self attention (i.e. softmax(QK^T)) is non-negative and the sum of each row is 1
  - **High computational complexity**
    - Pairwise method

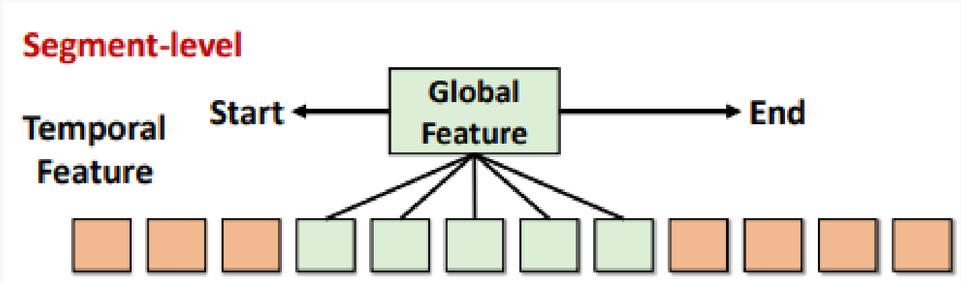
[13] Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. Attention is not all you need: Pure attention loses rank doubly exponentially with depth. In Int. Conf. Machine Learning, 2021

Related Work

Previous works

- **Segment-level**

- locate the boundaries based on the global feature of a predicted temporal segment [22,23,30,48,53], which may ignore detailed information at each instant.

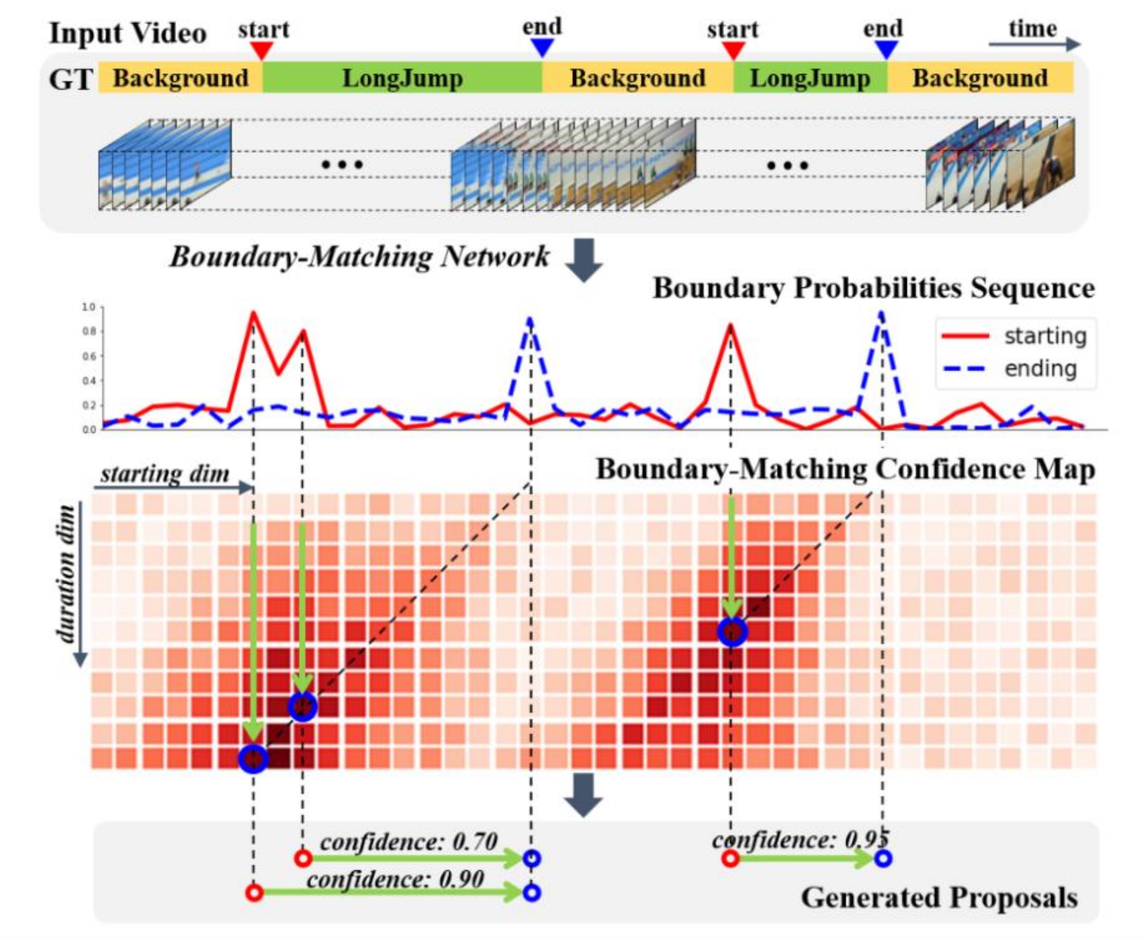


[22] Tianwei Lin, Xiao Liu, Xin Li, Errui Ding, and Shilei Wen. Bmn: Boundary-matching network for temporal action proposal generation. In Int. Conf. Comput. Vis., 2019
[23] Tianwei Lin, Xu Zhao, Haisheng Su, Chongjing Wang, and Ming Yang. Bsn: Boundary sensitive network for temporal action proposal generation. In Eur. Conf. Comput. Vis., 2018
[30] Fuchen Long, Ting Yao, Zhaofan Qiu, Xinmei Tian, Jiebo Luo, and Tao Mei. Gaussian temporal awareness networks for action localization. In IEEE Conf. Comput. Vis. Pattern Recog., 2019
[48] Runhao Zeng, Wenbing Huang, Mingkui Tan, Yu Rong, Peilin Zhao, Junzhou Huang, and Chuang Gan. Graph convolutional networks for temporal action localization. In Int. Conf. Comput. Vis., 2019
[53] Peisen Zhao, Lingxi Xie, Chen Ju, Ya Zhang, Yanfeng Wang, and Qi Tian. Bottom-up temporal action localization with mutual regularization. In Eur. Conf. Comput. Vis., 2020

Related Work

Previous works

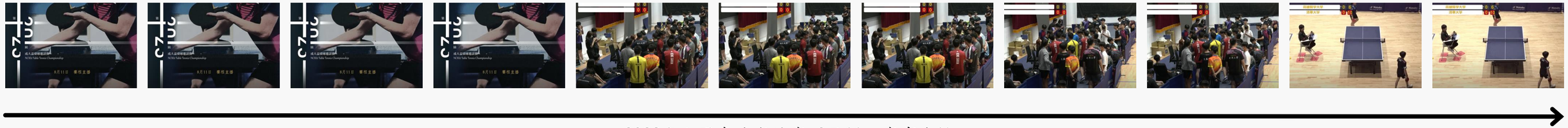
- **Segment-level**



[22] Tianwei Lin, Xiao Liu, Xin Li, Errui Ding, and Shilei Wen. Bmn: Boundary-matching network for temporal action proposal generation. In Int. Conf. Comput. Vis., 2019

background

multiple actions



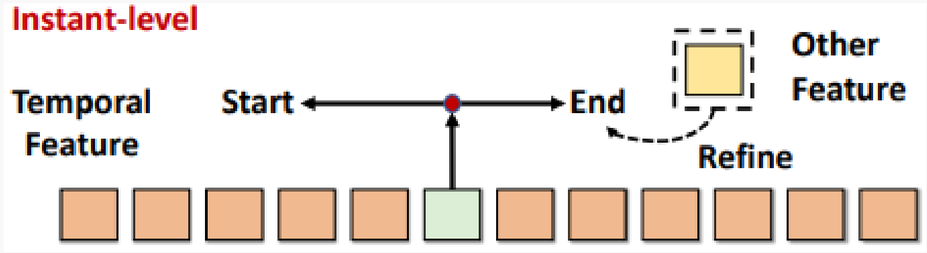
2023成大盃桌球邀請賽 8月11日賽事直播

- Too much information (activities, background, etc...)
- Each instant of detail information was discarded

Related Work

Previous works

- **Instant-level**
- directly regress the boundaries based on a single instant [33,49], potentially with some other features [21,34,51], which do not consider the relation between adjacent instants (e.g. the relative probability) around the boundary



[33] Sujoy Paul, Sourya Roy, and Amit K Roy-Chowdhury. W talc: Weakly-supervised temporal activity localization and classification. In Proceedings of the European Conference on Computer Vision (ECCV), pages 563–579, 2018
[49] Chenlin Zhang, Jianxin Wu, and Yin Li. Actionformer: Localizing moments of actions with transformers. In Eur. Conf. Comput. Vis., 2022
[21] Chuming Lin, Chengming Xu, Donghao Luo, Yabiao Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Yan wei Fu. Learning salient boundary feature for anchor-free temporal action localization. In IEEE Conf. Comput. Vis. Pattern Recog., 2021
[34] Zhiwu Qing, Haisheng Su, Weihao Gan, Dongliang Wang, Wei Wu, Xiang Wang, Yu Qiao, Junjie Yan, Changxin Gao, and Nong Sang. Temporal context aggregation network for temporal action proposal refinement. In IEEE Conf. Comput. Vis. Pattern Recog., 2021
[51] Chen Zhao, Ali K Thabet, and Bernard Ghanem. Video self stitching graph network for temporal action localization. In Int. Conf. Comput. Vis., 2021

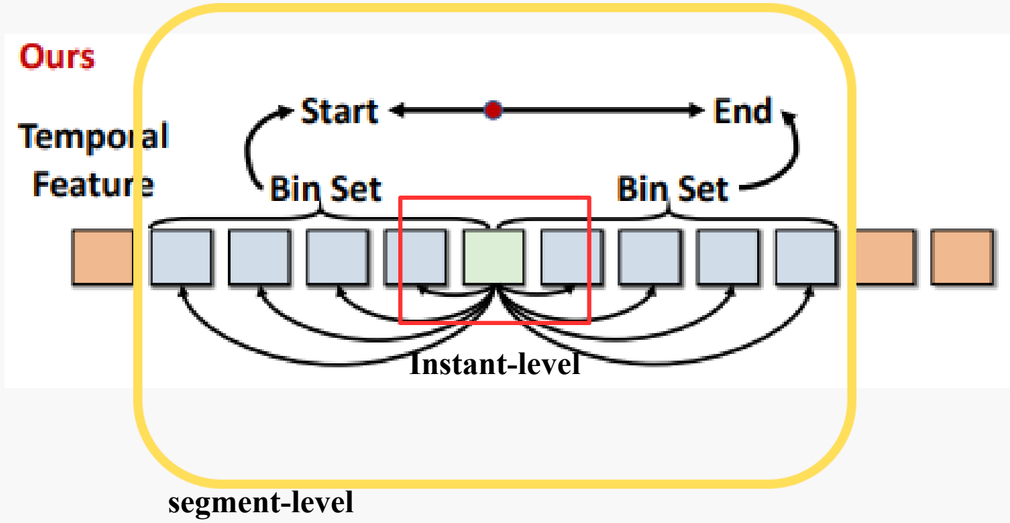
Related Work

- **Segment-level**
  - Global feature & More information
  - Ignore detailed information at each instant
  - Large receptive field
- **Instant-level**
  - More detailed feature
  - Do not consider the relation between adjacent instants
  - Small receptive field

Related Work

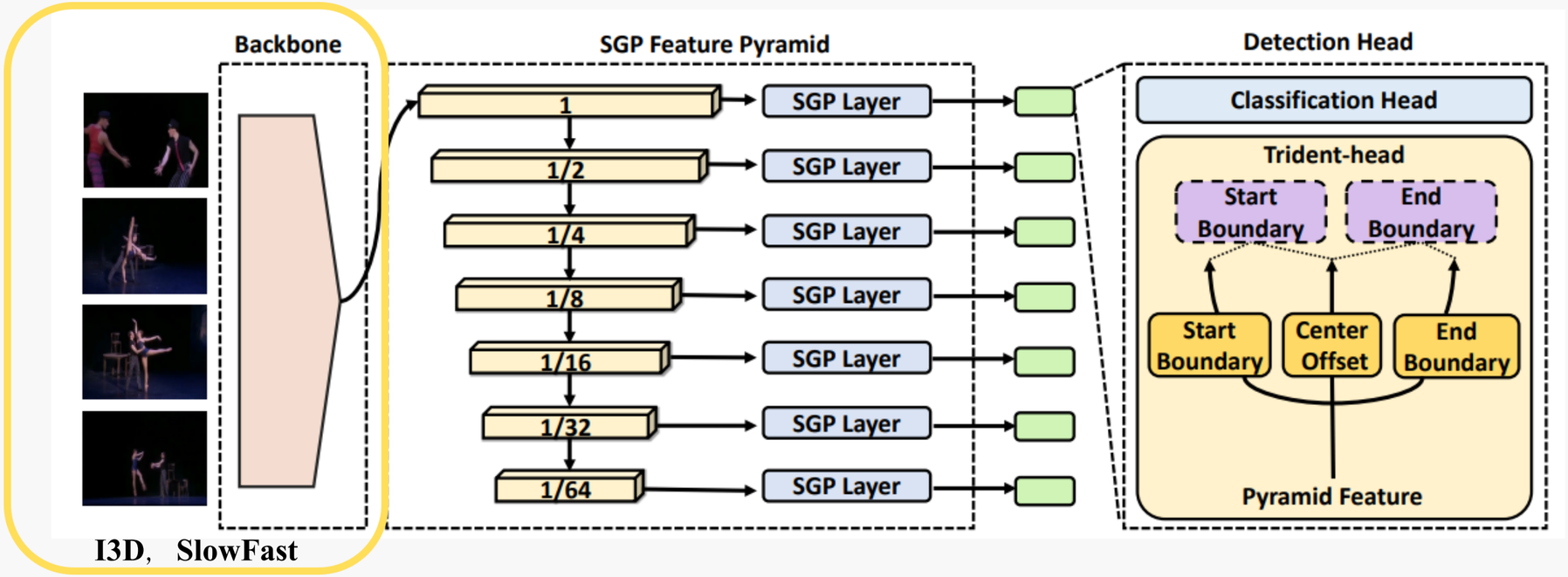
Author idea

- Segment-Level + Instant-Level = TriDet ?
  - the action boundaries are modeled via an estimated relative probability distribution of the boundary.



Method

TriDet Structure

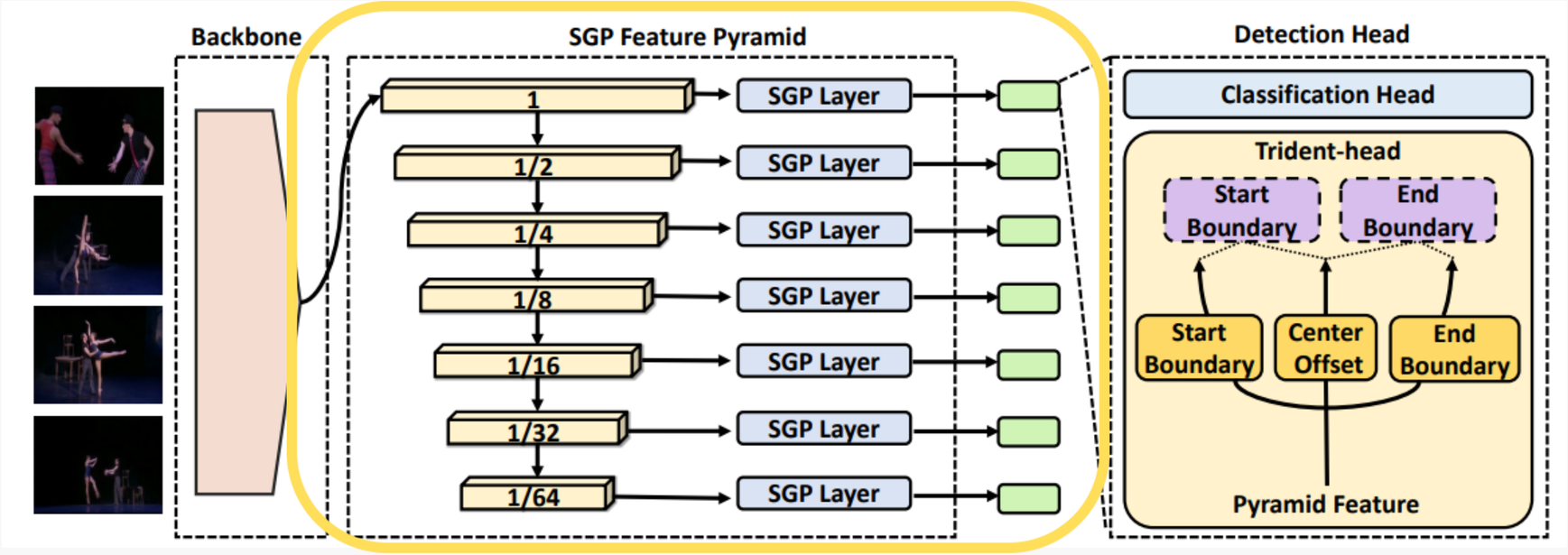


[8] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In IEEE Conf. Comput. Vis. Pattern Recog., 2017

[15] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In Int. Conf. Comput. Vis., 2019

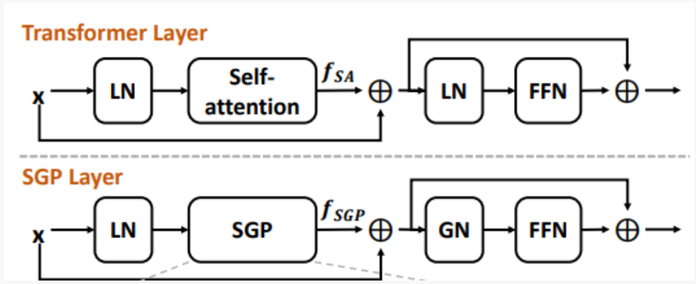
Method

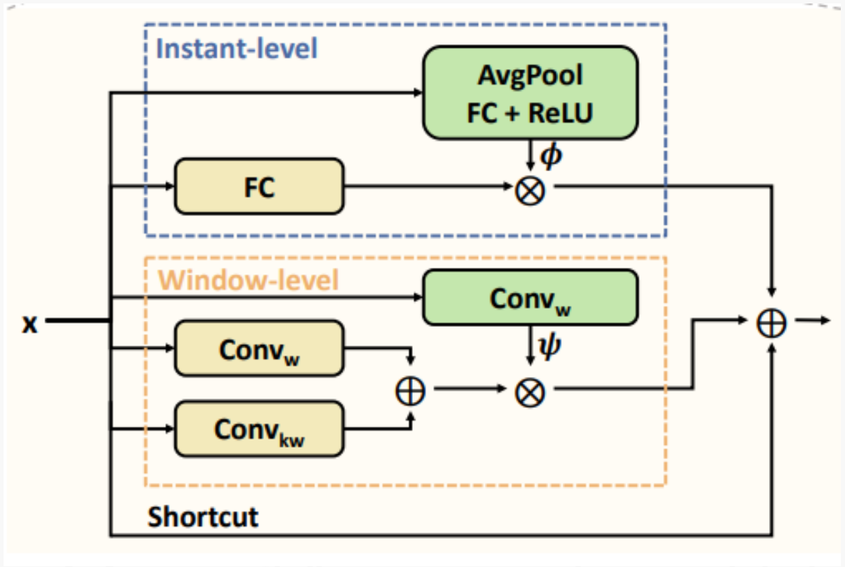
TriDet Structure



Method

TriDet Structure





- **Instant-level** **[element-wise]**
- Increase the feature discriminability between action and non-action instant by enlarging their feature distance with the video-level average feature.

$$f_{SGP} = \phi(x)FC(x)+\psi(x)(Conv_w(x)+Conv_{kw}(x))+x, \quad (1)$$

$$\phi(x) = ReLU(FC(AvgPool(x))), \quad (2)$$

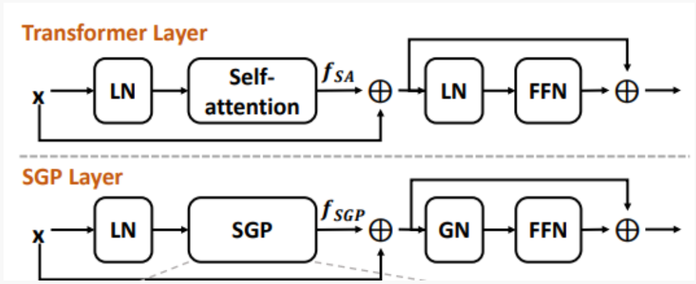
$$\psi(x) = Conv_w(x), \quad (3)$$

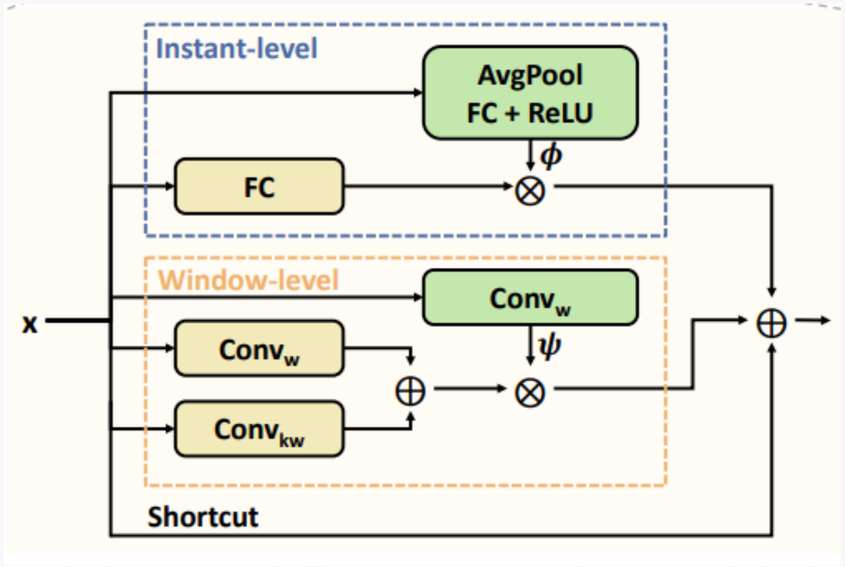
- **FC** : Fully-Connected layer
- **Convw** : 1-D depth-wise convolution layer

[11] Franc¸ois Chollet. Xception: Deep learning with depthwise separable convolutions. In IEEE Conf. Comput. Vis. PatternRecog., 2017

Method

TriDet Structure





- **Window-level**

- The window-level branch is designed to introduce the semantic content from a wider receptive field with a branch ψ to help dynamically focus on the features of which scale

$$f_{SGP} = \phi(x)FC(x)+\psi(x)(Conv_w(x)+Conv_{kw}(x))+x, \quad (1)$$

$$\phi(x) = ReLU(FC(AvgPool(x))), \quad (2)$$

$$\psi(x) = Conv_w(x), \quad (3)$$

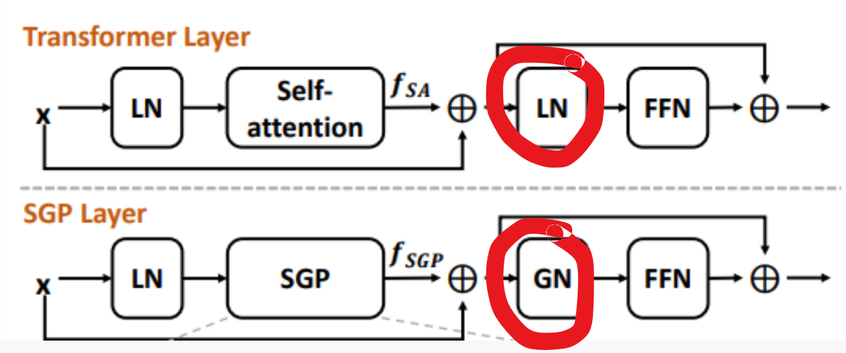
- **FC** : Fully-Connected layer
- **Convw** : 1-D depth-wise convolution layer

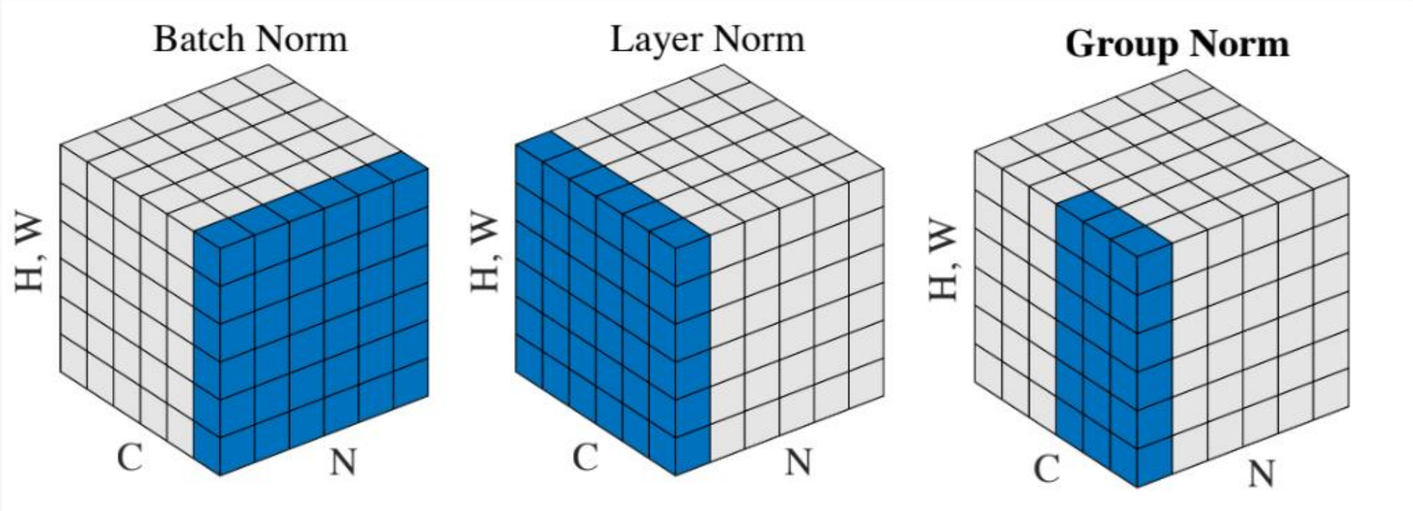
- k : scalable factor aiming at capturing a larger granularity of temporal information
- w : temporal dimension with window size w.

[11] Franc¸ois Chollet. Xception: Deep learning with depthwise separable convolutions. In IEEE Conf. Comput. Vis. PatternRecog., 2017

Method

TriDet Structure





- **We don't have to care about the batch size**

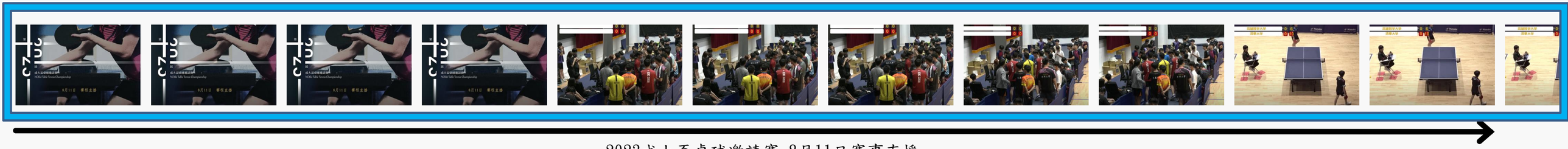
- **I love AI, <pad>, <pad>, <pad>.....**
- **I love AI, AI love me, we are a couple**

*** <pad> mean (x) std (x)**

[3] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hin?ton. Layer normalization. arXiv preprint arXiv:1607.06450, 2016.

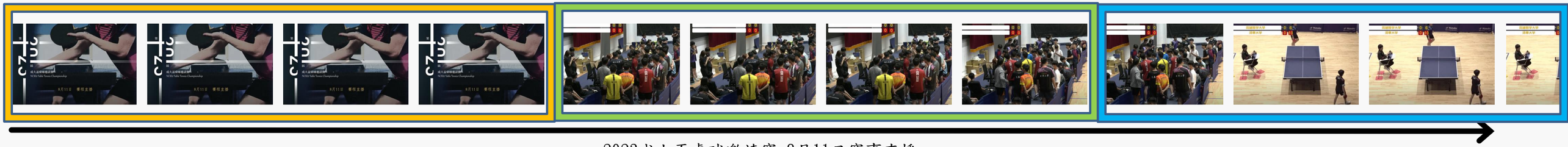
[43] Yuxin Wu and Kaiming He. Group normalization. In Eur. Conf. Comput. Vis., 2018

Method



2023成大盃桌球邀請賽 8月11日賽事直播

Layer normalization



2023成大盃桌球邀請賽 8月11日賽事直播

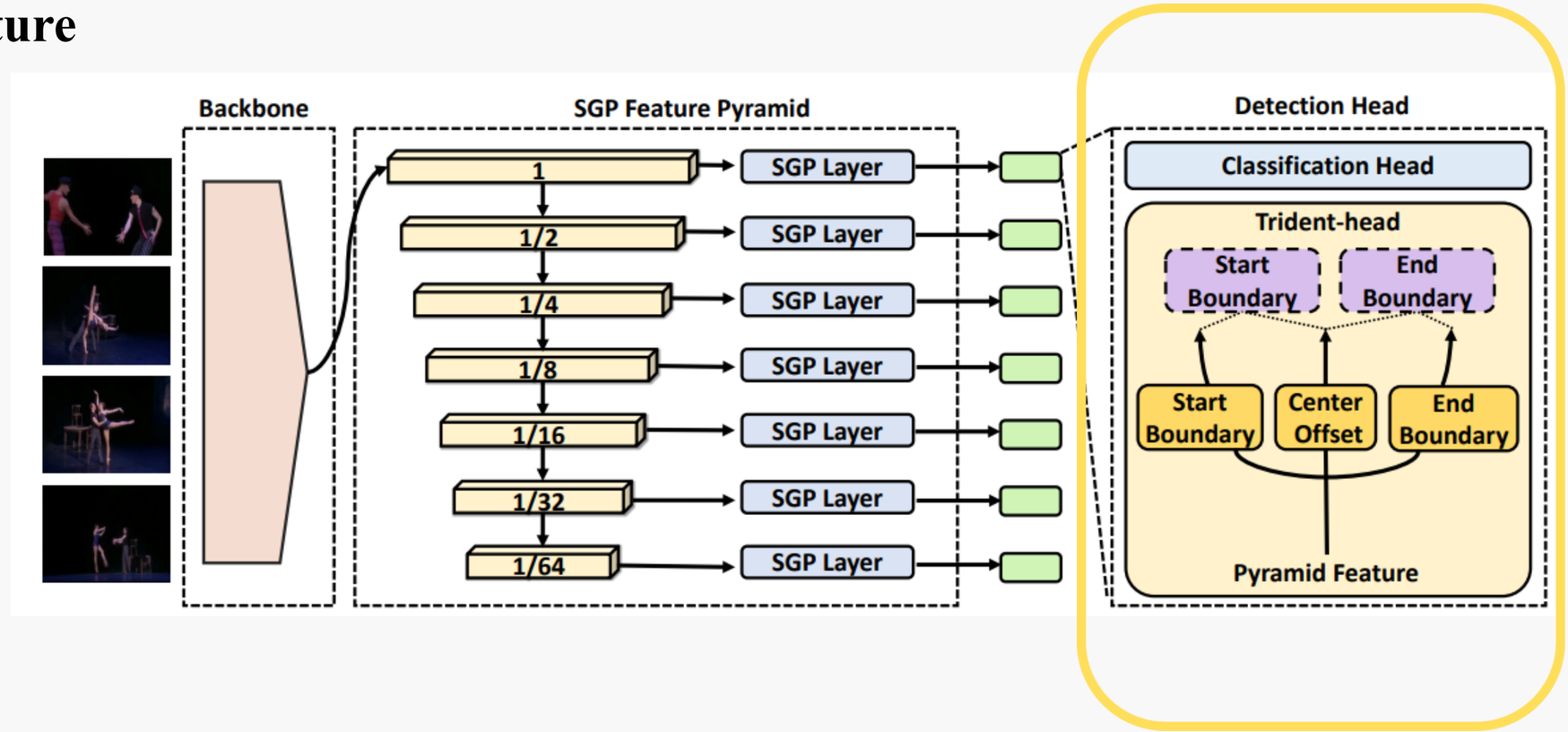
Group normalization

[3] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hin?ton. Layer normalization. arXiv preprint arXiv:1607.06450, 2016.

[43] Yuxin Wu and Kaiming He. Group normalization. In Eur. Conf. Comput. Vis., 2018

Method

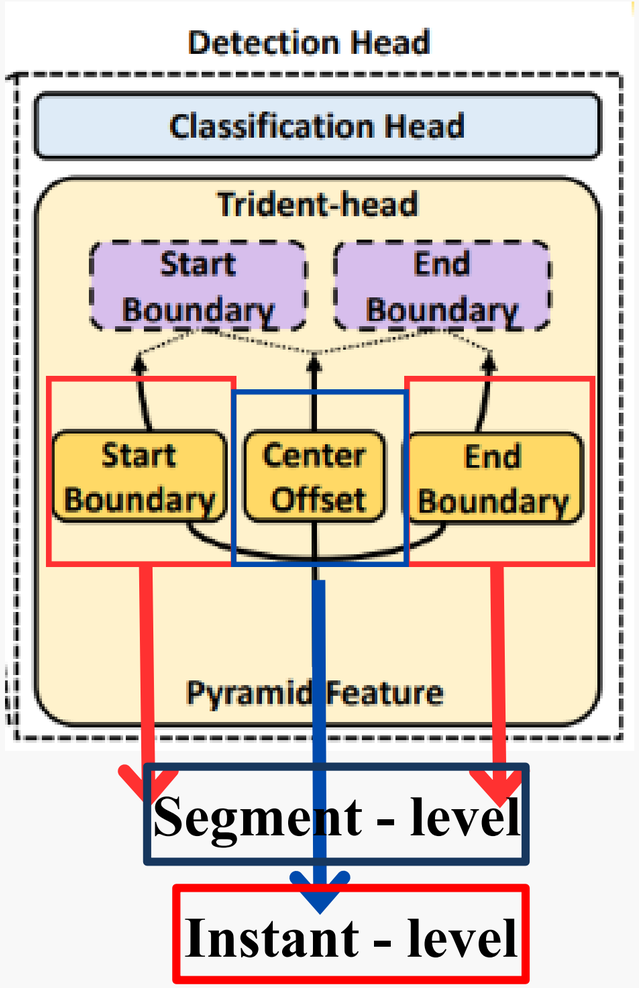
TriDet Structure



Method

TriDet Structure

- The Trident-head estimates the boundary offset based on a relative distribution predicted by three branches: **Start Boundary, End Boundary and Center Offset**
- Based on the relative boundary modeling, i.e. considering the relation of features in a certain period and obtaining the relative probability of being a boundary **for each instant** in that period



Method

TriDet Structure

3.

$$\widetilde{P}_{st} = Softmax(F_s^{[(t-B):t]} + F_c^{t,0}),$$

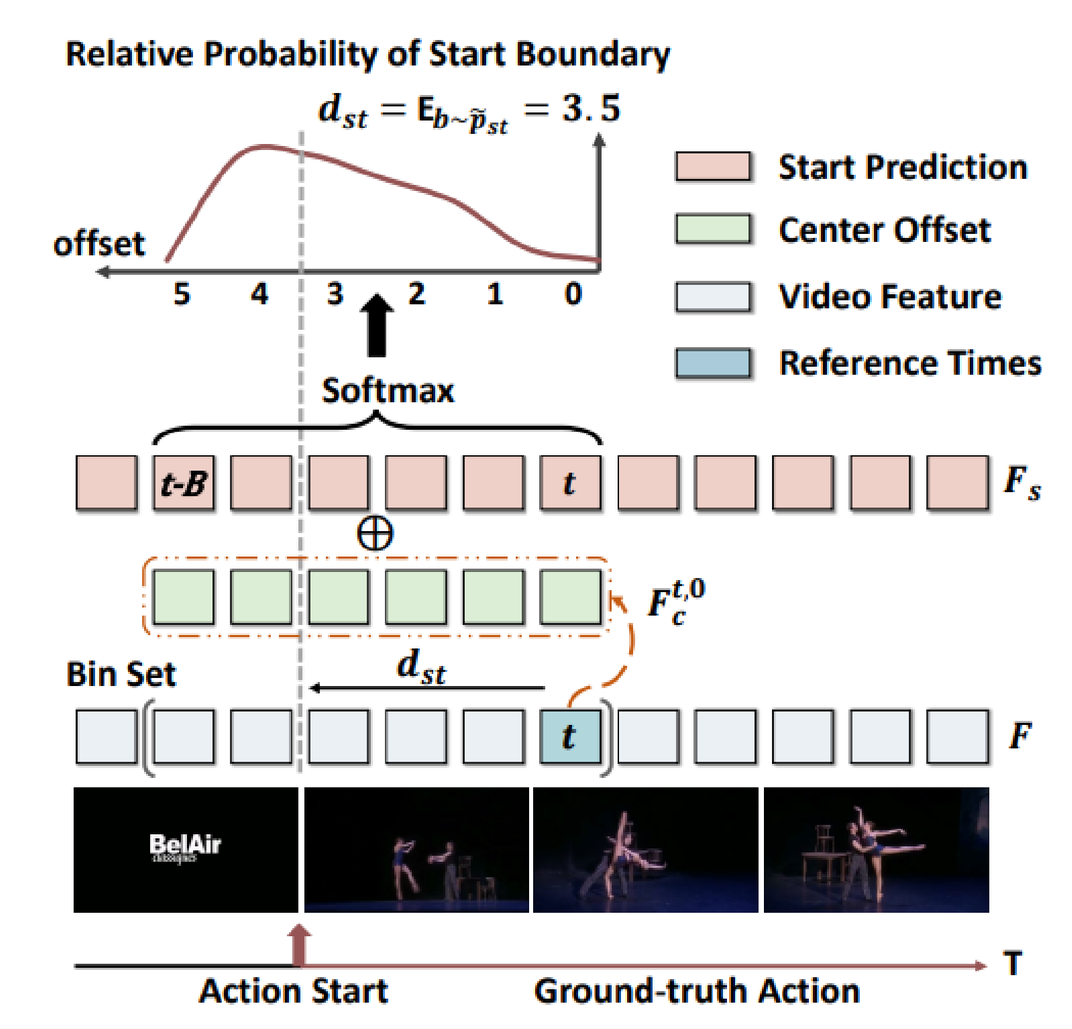
$$d_{st} = \mathbb{E}_{b\sim\widetilde{P}_{st}}[b] \approx \sum_{b=0}^{B}(b\widetilde{P}_{stb}),$$

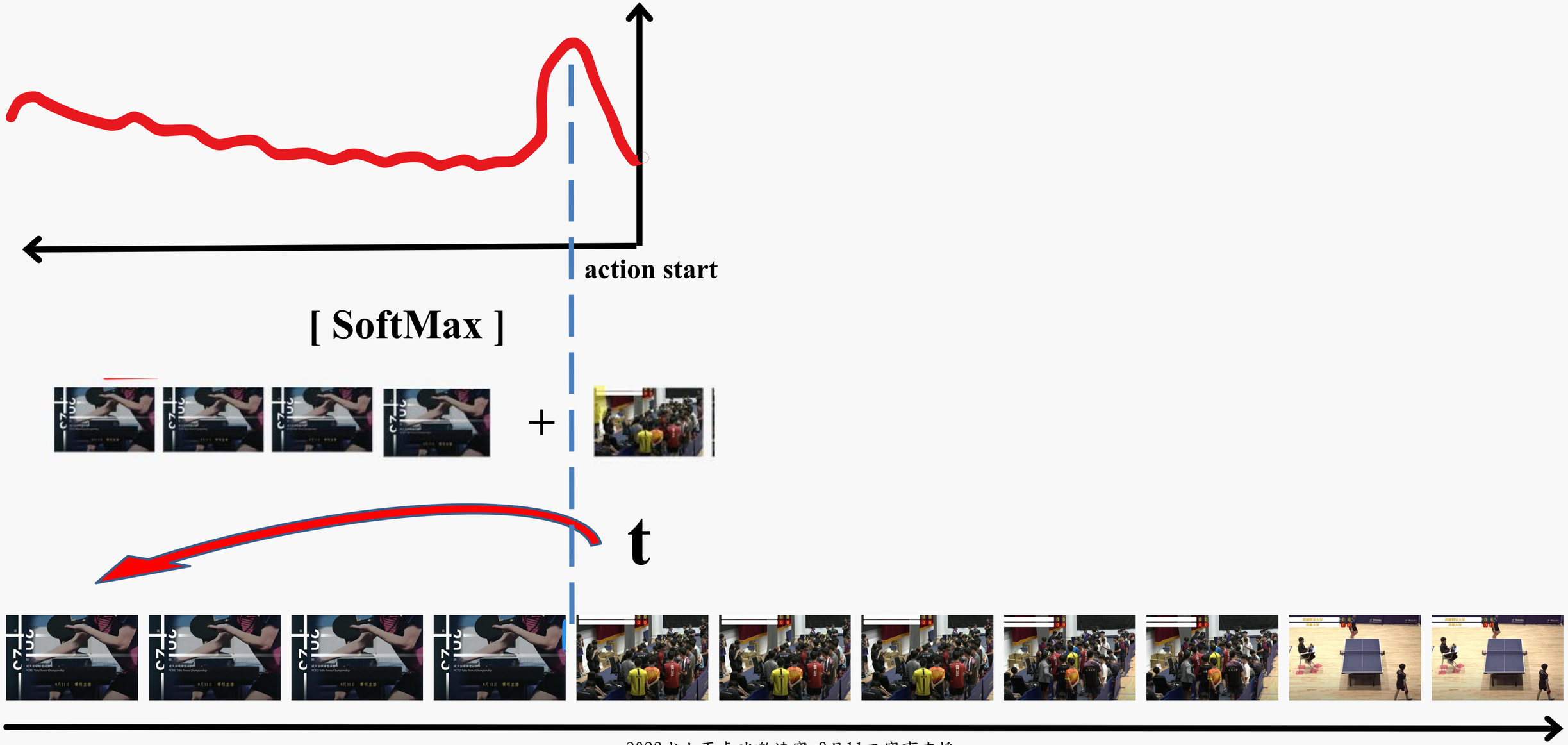
$$\widetilde{P}_{et} = Softmax(F_e^{[t:(t+B)]} + F_c^{t,1}),$$

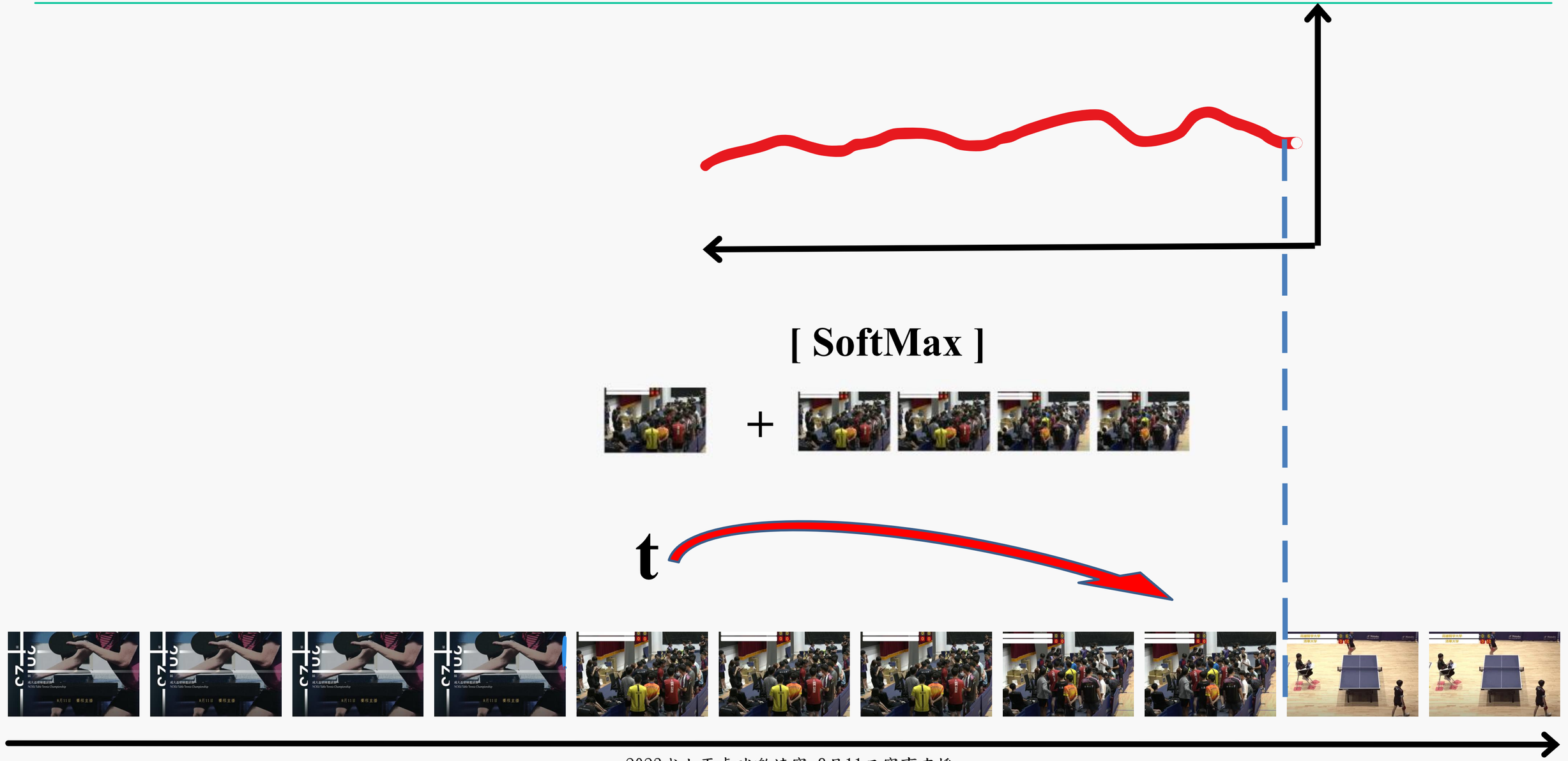
$$d_{et} = \mathbb{E}_{b\sim\widetilde{P}_{et}}[b] \approx \sum_{b=0}^{B}(b\widetilde{P}_{etb})$$

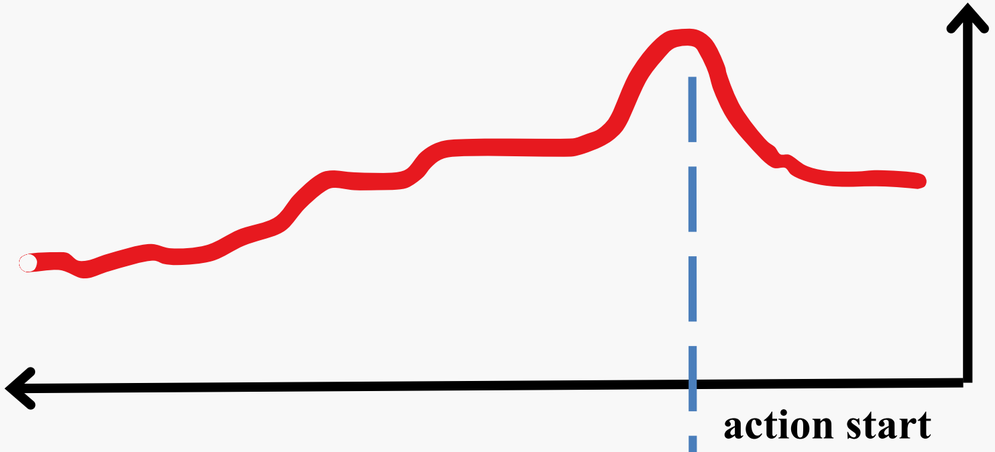
2. $F_s \in \mathcal{R}^T$, $F_e \in \mathcal{R}^T$ and $F_c \in \mathcal{R}^{T\times 2\times(B+1)}$

1. **B** is the number of bins for boundary prediction

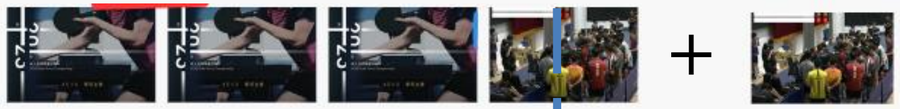



action start
[SoftMax]
t
2023成大盃桌球邀請賽 8月11日賽事直播

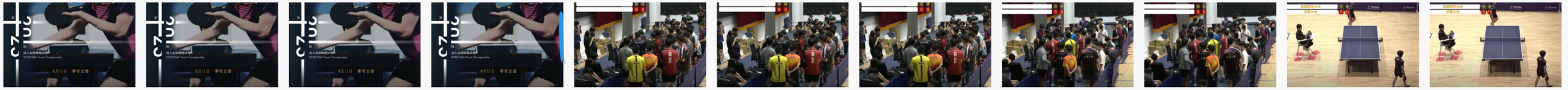

[SoftMax]
t
2023成大盃桌球邀請賽 8月11日賽事直播



[SoftMax]



t



2023成大盃桌球邀請賽 8月11日賽事直播


action start
action end
[SoftMax]
+
t
2023成大盃桌球邀請賽 8月11日賽事直播

Experiment

single NVIDIA A100 GPU

*** THUMOS14 consists of 20 sport action classes and it contains 200 and 213 untrimmed videos with 3,007 and 3,358 action instances on the training set and testing set**

Table 1. Comparison with the state-of-the-art methods on THUMOS14 dataset. *: TSN backbone. †: Swin Transformer backbone. Others: I3D backbone.

Method	0.3	0.4	0.5	0.6	0.7	Avg.
BMN [22]*	56.0	47.4	38.8	29.7	20.5	38.5
G-TAD [45]*	54.5	47.6	40.3	30.8	23.4	39.3
A2Net [46]	58.6	54.1	45.5	32.5	17.2	41.6
TCANet [34]*	60.6	53.2	44.6	36.8	26.7	44.3
RTD-Net [39]	68.3	62.3	51.9	38.8	23.7	49.0
VSGN [51]*	66.7	60.4	52.4	41.0	30.4	50.2
ContextLoc [55]	68.3	63.8	54.3	41.8	26.2	50.9
AFSD [21]	67.3	62.4	55.5	43.7	31.1	52.0
ReAct [36]*	69.2	65.0	57.1	47.8	35.6	55.0
TadTR [28]	74.8	69.1	60.1	46.6	32.8	56.7
TALLFormer [10]†	76.0	-	63.2	-	34.5	59.2
ActionFormer [49]	82.1	77.8	71.0	59.4	43.9	66.8
TriDet	**83.6**	**80.1**	**72.9**	**62.4**	**47.4**	**69.3**

Method	0.3	0.4	0.5	0.6	0.7	Avg.

[18] Y.-G. Jiang, J. Liu, A. Roshan Zamir, G. Toderici, I. Laptev, M. Shah, and R. Sukthankar. THUMOS challenge: Action recognition with a large number of classes, 2014.

Experiment

single NVIDIA A100 GPU

*** HACS is a large-scale datasets and consisting of 200 classes of action and it contains 37,613 videos for training and 5,981 videos for testing set**

Table 2. Comparison with the state-of-the-art methods on HACS dataset.

Method	Backbone	0.5	0.75	0.95	Avg.
SSN [54]	I3D	28.8	18.8	5.3	19.0
LoFi [44]	TSM	37.8	24.4	7.3	24.6
G-TAD [45]	I3D	41.1	27.6	8.3	27.5
TadTR [28]	I3D	47.1	32.1	10.9	32.1
BMN [22]	SlowFast	52.5	36.4	10.4	35.8
TALLFormer [10]	Swin	55.0	36.1	11.8	36.5
TCANet [34]	SlowFast	54.1	37.2	11.3	36.8
TriDet	I3D	**54.5**	**36.8**	**11.5**	**36.8**
TriDet	SlowFast	**56.7**	**39.3**	**11.7**	**38.6**

[52] Hang Zhao, Zhicheng Yan, Lorenzo Torresani, and Antonio Torralba. Hacs: Human action clips and segments dataset for recognition and temporal localization. arXiv preprint arXiv:1712.09374, 2019.

Experiment

single NVIDIA A100 GPU

*** The EPIC-KITCHEN 100 is a large-scale dataset in first-person vision, which have two sub-tasks: noun localization (e.g. door) and verb localization (e.g. open the door)**

It contains 495 and 138 videos with 67,217 and 9,668 action instances for training and test, respectively. The number of action classes for noun and verb are 300 and 97.

Table 3. Comparison with the state-of-the-art methods on EPIC-KITCHEN dataset. *V.* and *N.* denote the *verb* and *noun* sub-tasks, respectively.

	Method	0.1	0.2	0.3	0.4	0.5	Avg.
V.	BMN [22]	10.8	8.8	8.4	7.1	5.6	8.4
	G-TAD [45]	12.1	11.0	9.4	8.1	6.5	9.4
	ActionFormer [49]	26.6	25.4	24.2	22.3	19.1	23.5
	TriDet	**28.6**	**27.4**	**26.1**	**24.2**	**20.8**	**25.4**
N.	BMN [22]	10.3	8.3	6.2	4.5	3.4	6.5
	G-TAD [45]	11.0	10.0	8.6	7.0	5.4	8.4
	ActionFormer [49]	25.2	24.1	22.7	20.5	17.0	21.9
	TriDet	**27.4**	**26.3**	**24.6**	**22.2**	**18.3**	**23.8**

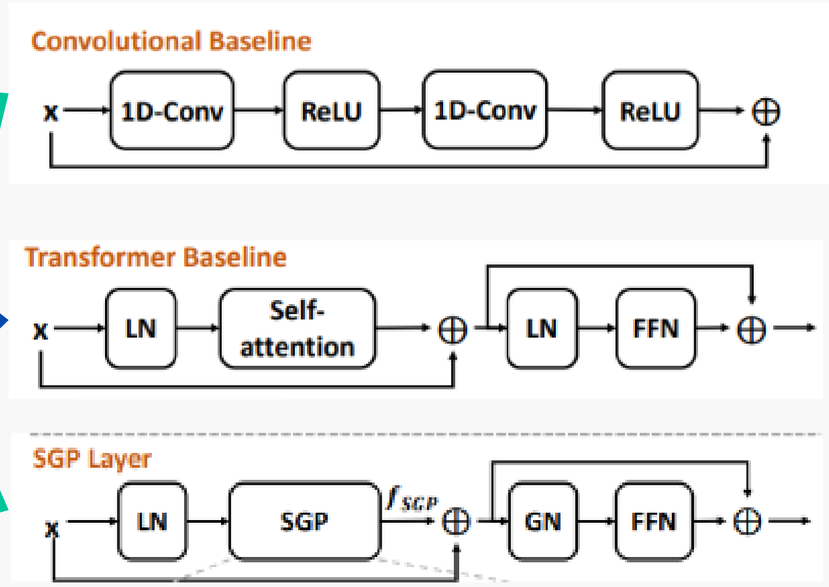
[12] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, , Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. Int. J. Com?put. Vis., 2022

Experiment

- **Ablation Study**
  - **Main components analysis**

Table 5. Analysis of the Effectiveness of three main components on THUMOS14.

	Method	SA	SGP	Trident	0.3	0.5	0.7	Avg.
-6.2%	1				77.3	65.2	40.0	62.1
-1.5%	2	√			82.1	71.0	43.9	66.8
Target	3		√		**83.6**	71.7	45.8	68.3
	4		√	√	**83.6**	**72.9**	**47.4**	**69.3**



Experiment

- **Ablation Study**
  - **Computational complexity**

Main : main architecture

Head : classification head and regression head

Table 6. Analysis of computation cost on THUMOS14. Main: All parts of the model except the detection head. *: Our method with a normal instant-level regression head.

Method	mAP			GMACs			Latency
	0.3	0.7	Avg.	Main	Head	All	(ms)
ActionFormer	82.1	43.9	66.8	30.8	**14.4**	45.3	224
TriDet*	**83.6**	45.8	68.3	**14.5**	**14.4**	**28.9**	**145**
TriDet	**83.6**	**47.4**	**69.3**	**14.5**	29.1	43.7	167

Experiment

- **Ablation Study**
  - **Ablation on the number of bins**

Table 8. Analysis of the number of bins.

Bin	THUMOS14				HACS			
	0.3	0.5	0.7	Avg.	0.5	0.75	0.95	Avg.
4	82.9	71.5	46.3	68.1	55.7	32.3	4.7	33.3
8	83.5	**72.9**	46.3	69.0	56.2	38.4	11.2	38.0
10	82.8	71.8	46.2	68.1	56.2	38.5	11.1	37.9
12	**83.6**	72.3	46.2	68.5	56.3	38.4	11.1	38.0
14	83.4	72.6	45.6	68.3	**56.7**	**39.3**	**11.7**	**38.6**
16	**83.6**	**72.9**	**47.4**	**69.3**	56.5	38.6	11.1	38.1
20	**83.6**	71.7	45.8	68.3	56.3	38.6	11.1	38.0

Conclusion

- **Improving the temporal action detection task with a simple one-stage convolutional based framework TriDet with relative boundary modeling**
  - **Achieves state-of-the art performance on the first three datasets (THUMOS14, HACS, EPIC KITCHEN)**

Thank you for listening!