Deeper Action

ECCV TEL AVIV 2022

A One-Stage Method for FineAction Localization from Multiple Views

Yepeng Tang[1,2†], Weining Wang[3], Chunjie Zhang[1,2], Jie Jiang[3,4],

Weitao Yuan[3,4], Sihan Chen[3,4], Jing Liu[3,4], Yao Zhao[1,2]

[1] Institute of Information Science, Beijing Jiaotong University

[2] Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing Jiaotong University

[3] Institute of Automation, Chinese Academy of Sciences

[4] School of Artificial Intelligence, University of Chinese Academy of Sciences

†Intern at the Institute of Automation, Chinese Academy of Sciences.

Deeper Action

Task Description

ECCV TEL AVIV 2022

Temporal action Localization task:

- Untrimmed videos
- locating the starting and ending time of action instances
- classifying the action instances

[1] Challenge and Workshop on Localized and Detailed Understanding of Human Actions in Videos. https://deeperaction.github.io/fineaction/

Deeper Action

Evaluation Metrics

ECCV TEL AVIV 2022

Temporal action localization

Interpolated Average Precision (AP) is used as the metric for evaluating the results on each activity category. Then, the AP is averaged over all the activity categories (mAP). Given a predicted action instance, the temporal intersection over union (tIoU) with a ground truth segment is calculated. If the tIoU is greater or equal to a given threshold, the predicted action instance is true positive.

The official metric used in this task is the average mAP, which is defined as the mean of all mAP values computed with tIoU thresholds between 0.5 and 0.95 with a step size of 0.05.

Temporal action proposal generation(sub task)

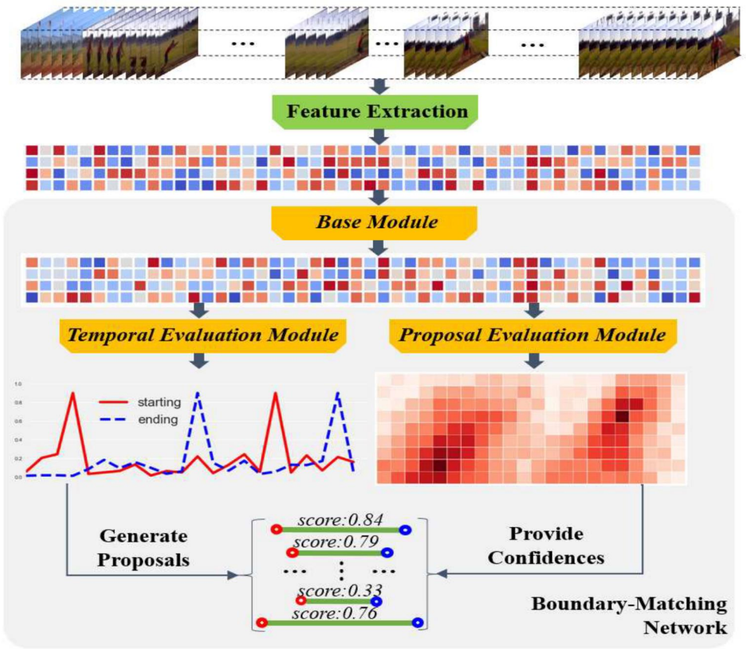
Average Recall (AR) calculated with multiple IoU thresholds is usually used as evaluation metric. To evaluate the relation between recall and proposals number, Average Recall (AR) with Average Number of proposals (AN) is evaluated, which is denoted as AR@AN. Moreover, area under the AR vs. AN curve (AUC) is also used as a metric, where AN often varies from 0 to 100.

Deeper Action

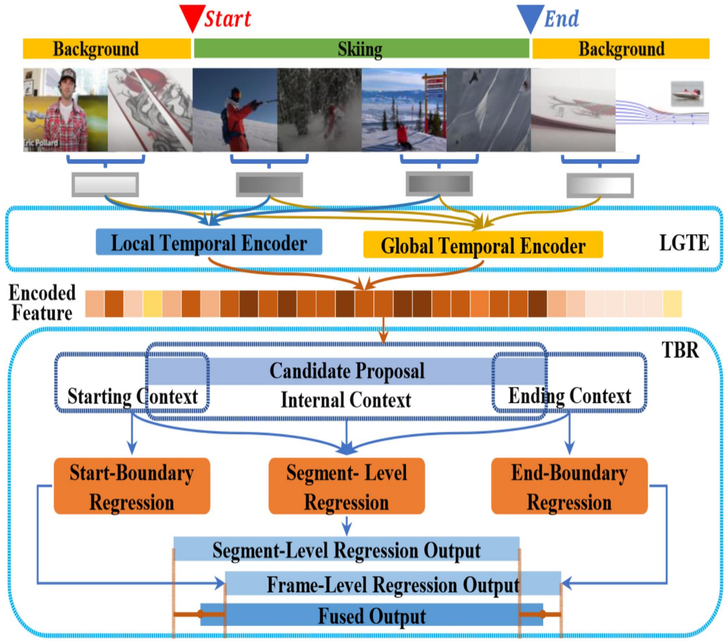
Mainstream Algorithms

ECCV TEL AVIV 2022

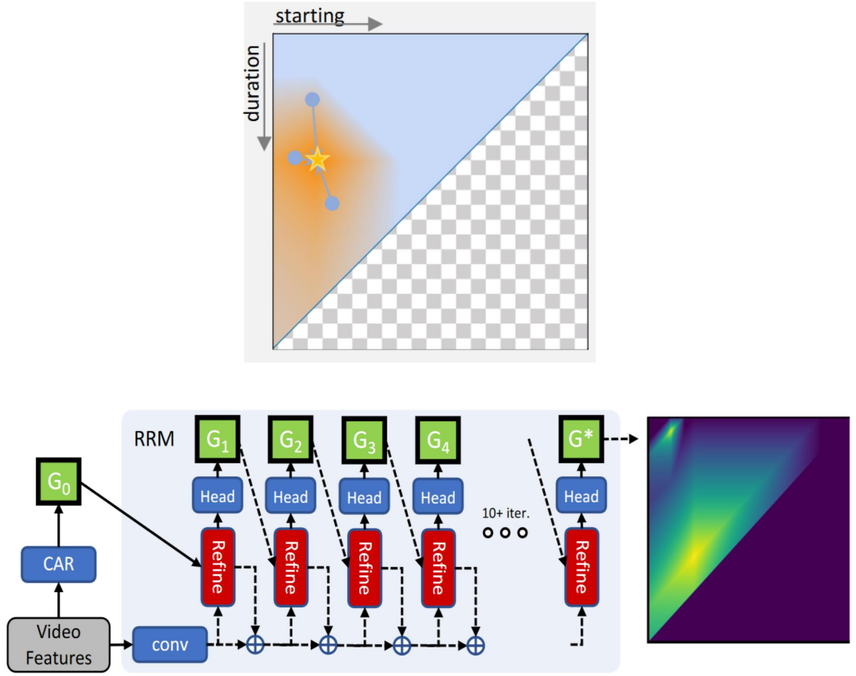
Two-Stage Methods



BMN



TCANet



RCL

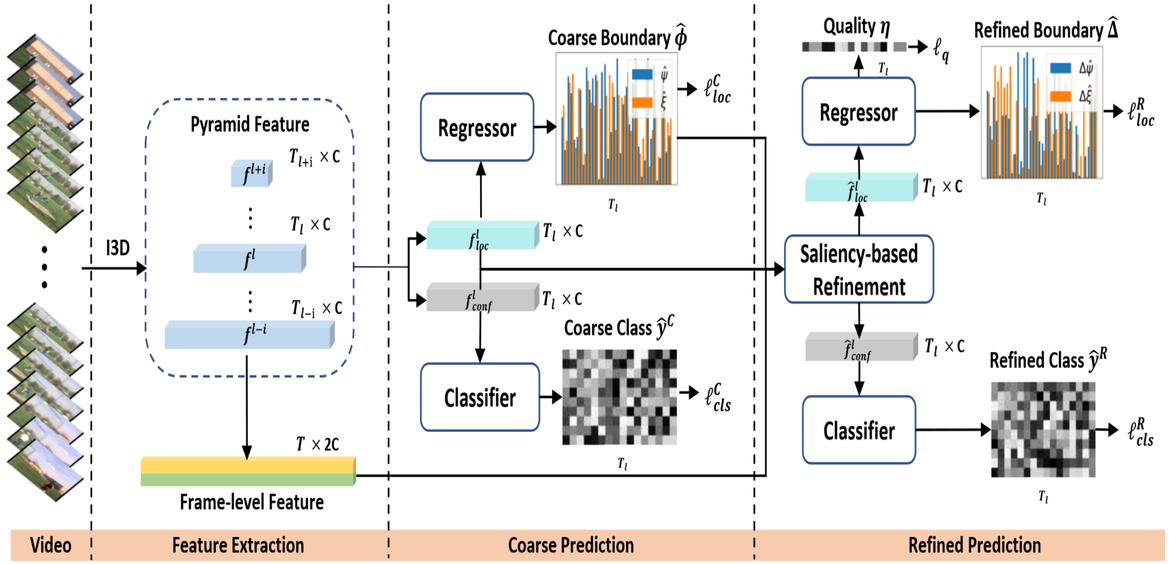
[2] Tianwei Lin, et al. "Bmn: Boundary-matching network for temporal action proposal generation." ICCV 2019.
[3] Zhiwu Qing, et al. "Temporal context aggregation network for temporal action proposal refinement." CVPR 2021.
[4] Qiang Wang, et al. "RCL: Recurrent Continuous Localization for Temporal Action Detection." CVPR 2022.

Deeper Action

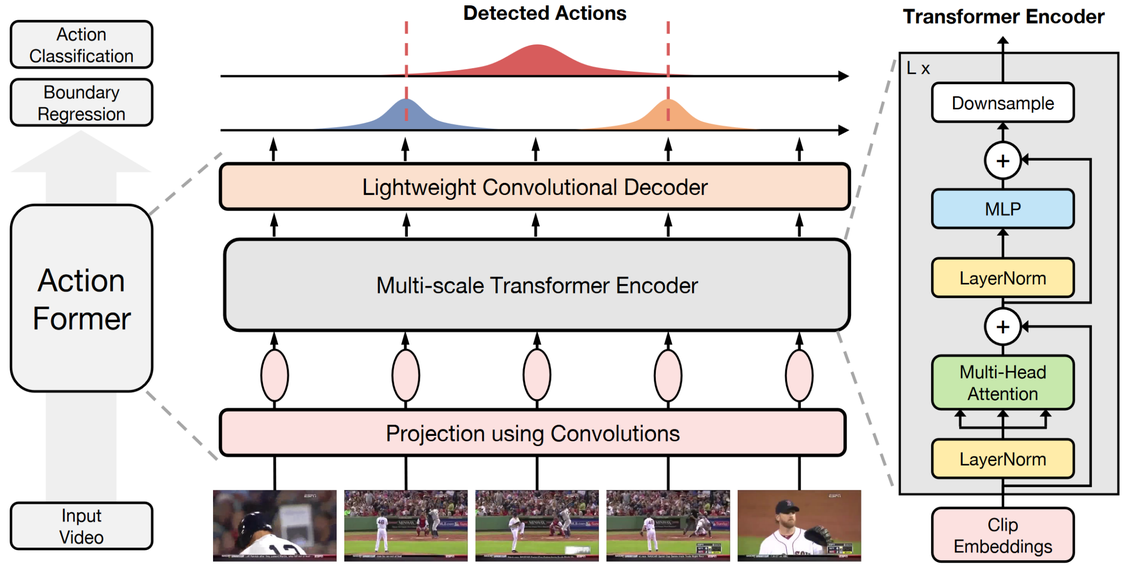
Mainstream Algorithms

ECCV TEL AVIV 2022

One-Stage Methods



AFSD



Actionformer

[5] Chuming Lin, et al. "Learning salient boundary feature for anchor-free temporal action localization." CVPR 2021.
[6] Chenlin Zhang, et al. "Actionformer: Localizing moments of actions with transformers." ECCV 2022.

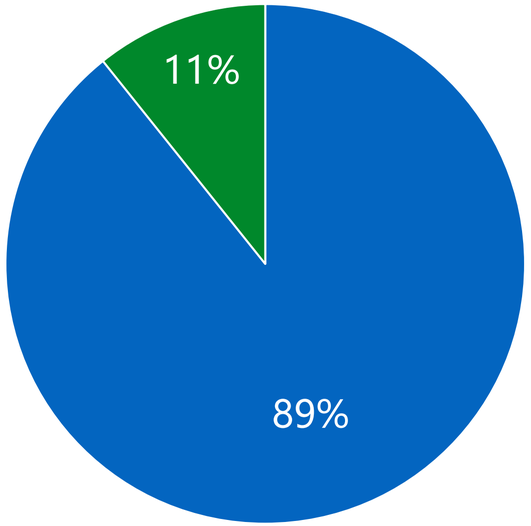
Deeper Action

Our Approach

ECCV TEL AVIV 2022

Dataset analysis

Database	Category	Video	Instance	Overlap	Duration	Action type
MPII Cooking	65	45	5,609	0.1%	11.1 m	kitchens
EPIC-Kitchens	4,025	700	89,979	28.1%	3.1 s	
FineGym V1.0	530	303	32,697	0.0%	1.7 s	sports
THUMOS14	20	413	6,316	17.5%	4.3 s	
ActivityNet	200	19,994	23,064	0.0%	49.2 s	daily events
HACS Segment	200	49,485	122,304	0.0%	33.2 s	
FineAction (Ours)	106	16,732	103,324	11.5%	7.1 s	

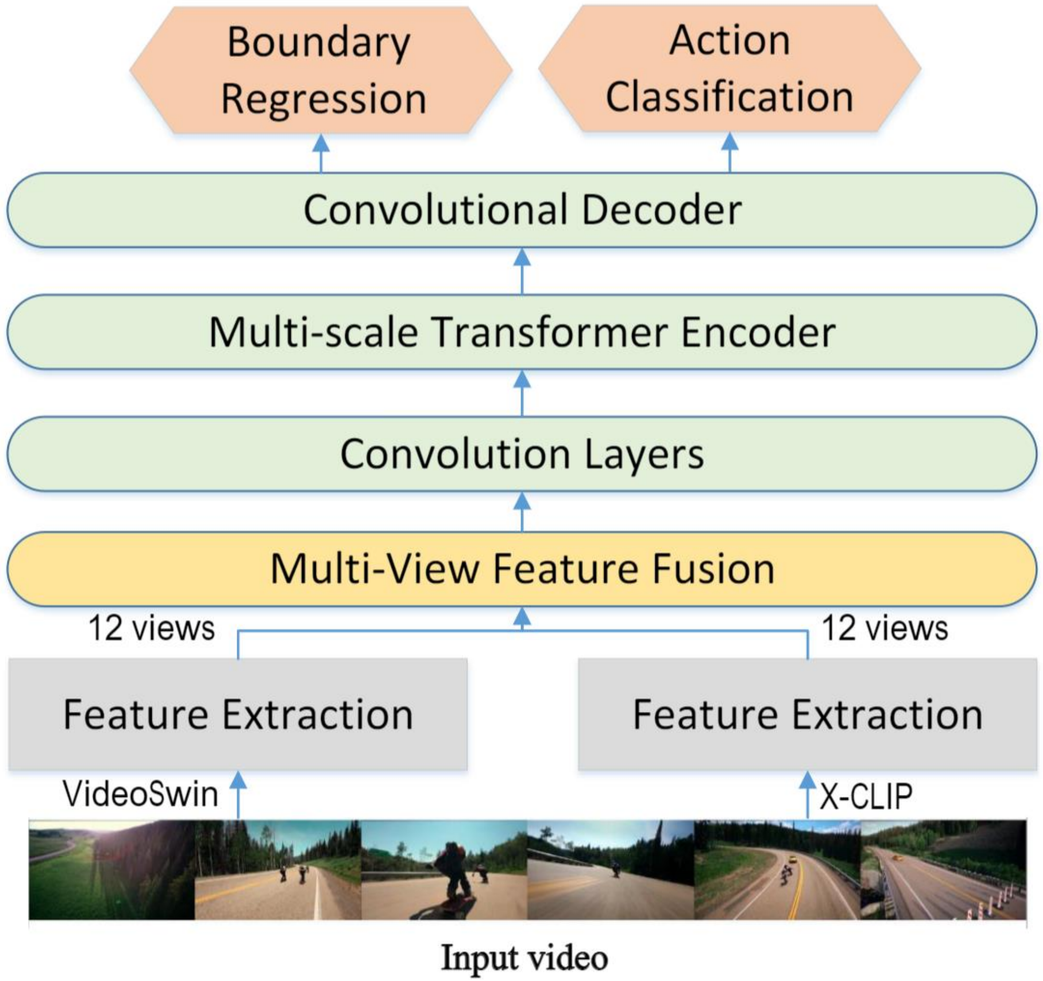


[7] Yi Liu, et al. "Fineaction: A fine-grained video dataset for temporal action localization."
[8] Chenglu Wu, et al. "Learning Efficient Feature Representation for Temporal Action Localization."

Deeper Action

Our Approach

ECCV TEL AVIV 2022



Pipeline:

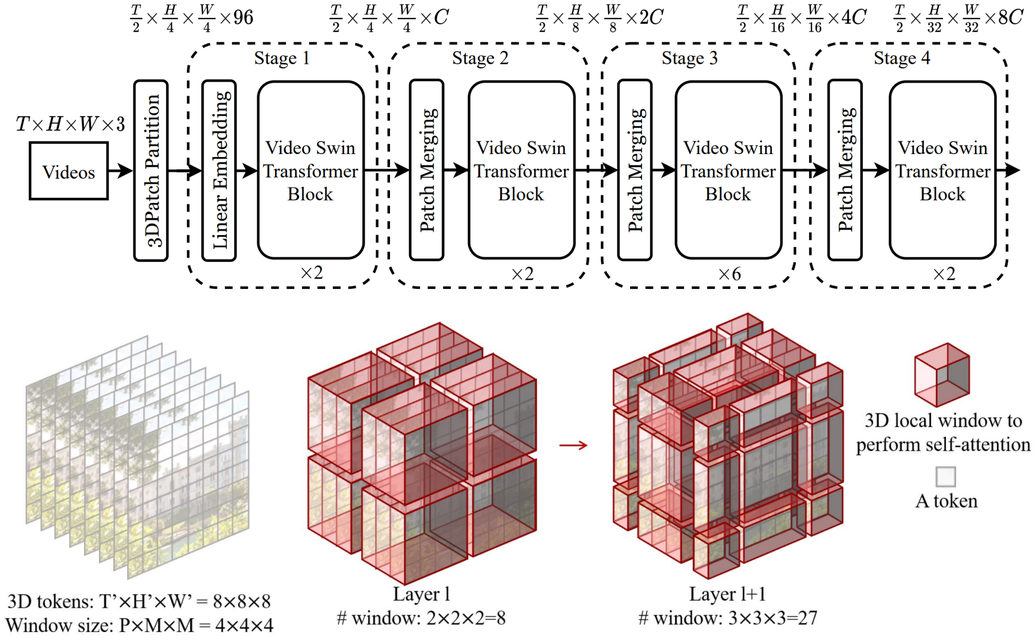
- Feature Extraction
- Multi-view Feature Fusion
- One-stage Action Detection

Deeper Action

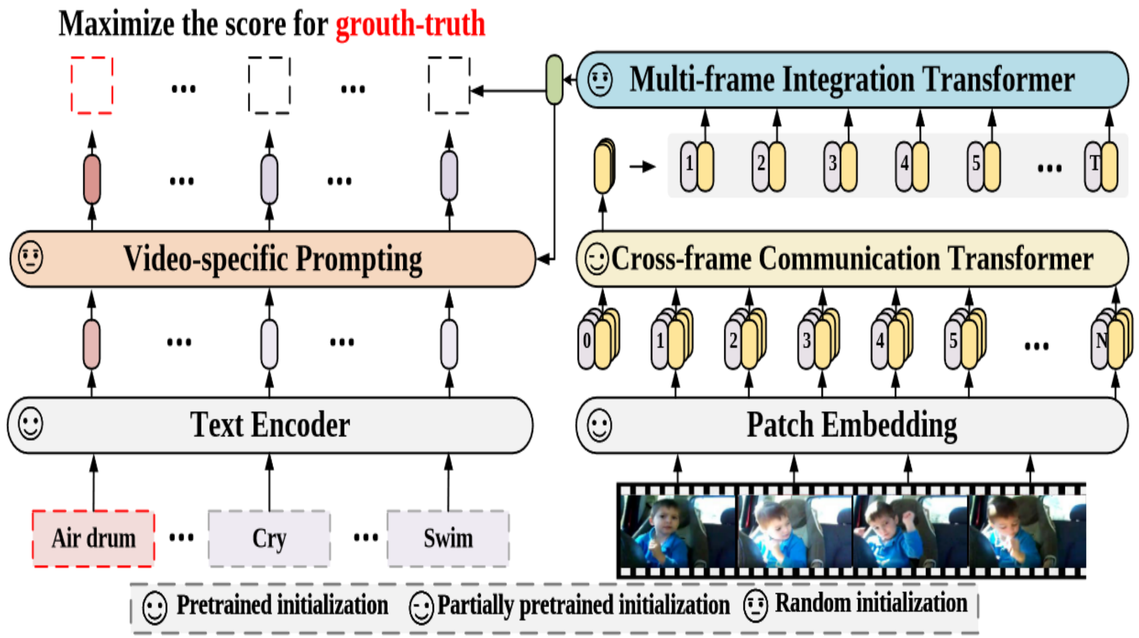
Our Approach

ECCV TEL AVIV 2022

Feature Extraction



Video Swin Transformer



X-CLIP

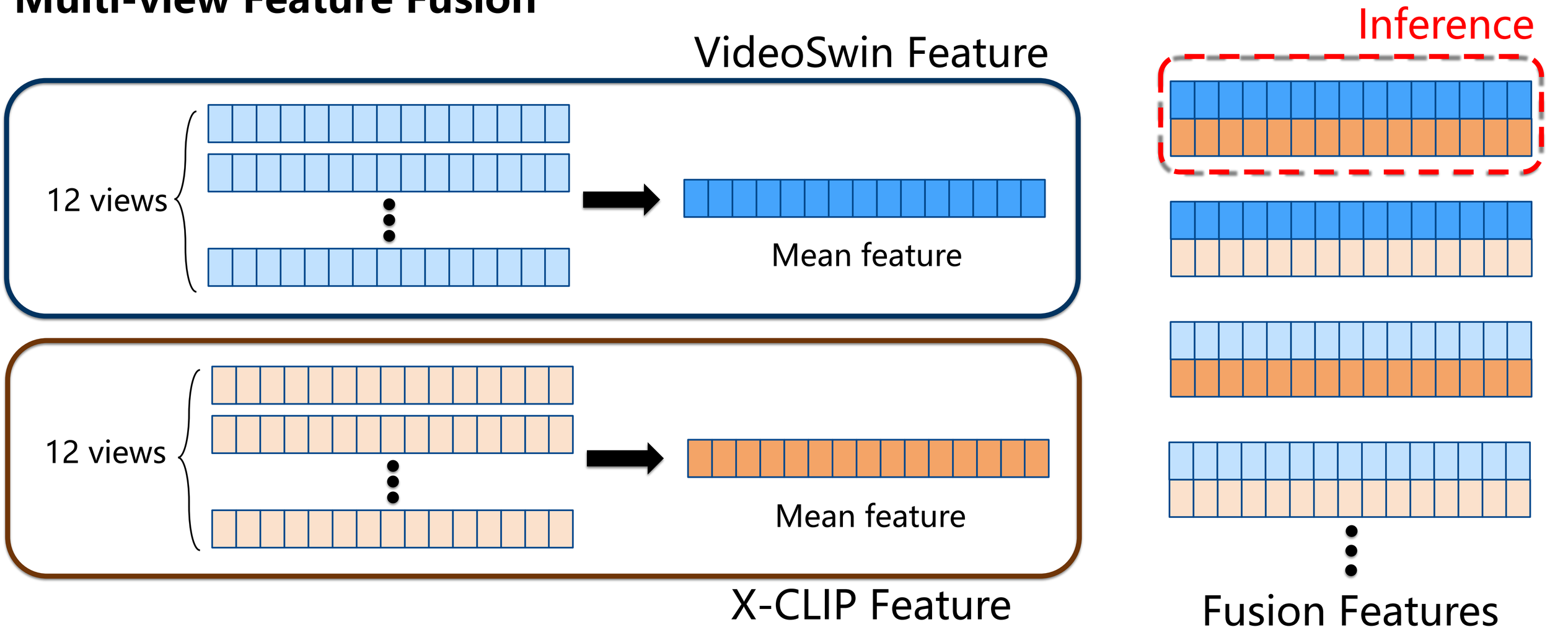
[9] Ze Liu, et al. "Video swin transformer." CVPR 2022.
[10] Bolin Ni, et al. "Expanding language-image pretrained models for general video recognition." ECCV 2022.

Deeper Action

Our Approach

ECCV TEL AVIV 2022

Multi-view Feature Fusion



Deeper Action

Our Approach

ECCV TEL AVIV 2022

One-stage Action Detection

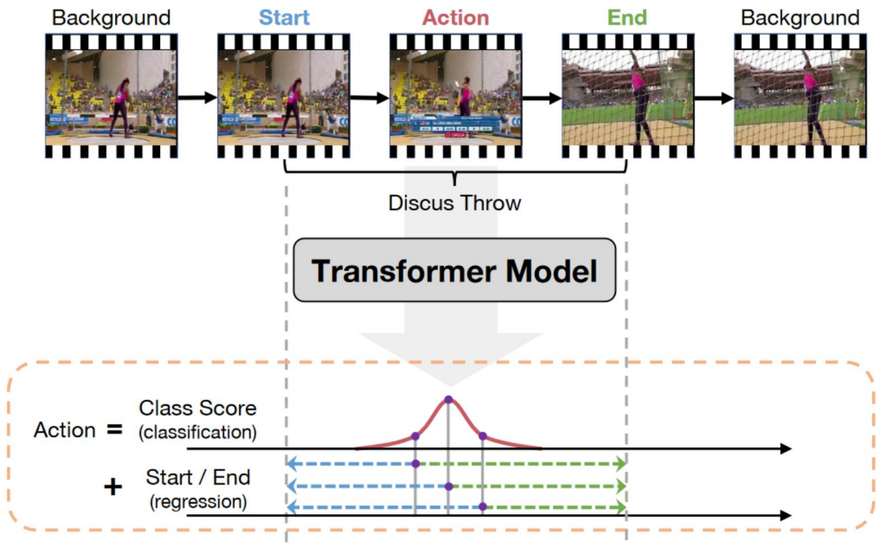
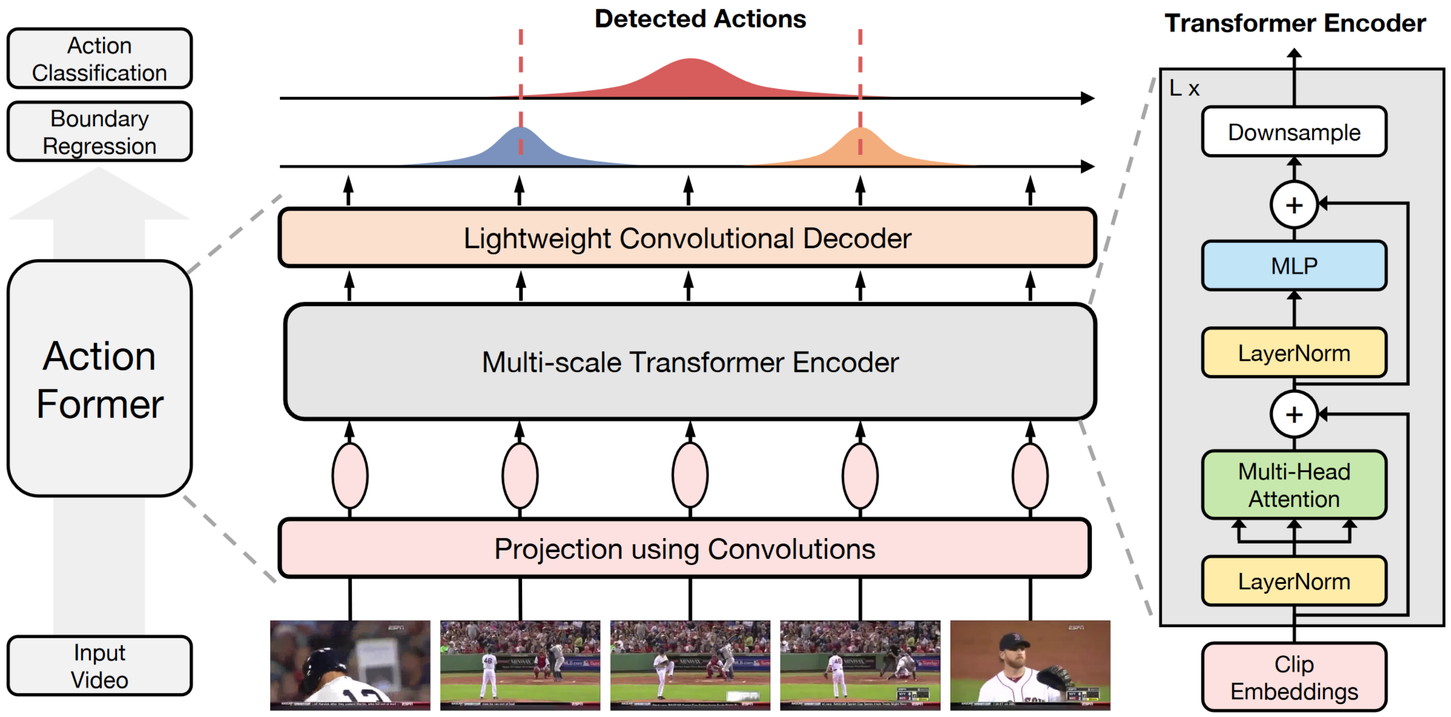


Fig. 1. An illustration of our ActionFormer.



Actionformer

Experiments

Detection Results

Feature	0.5	0.75	0.95	average
VideoSwin	35.46	20.22	3.43	21.00
X-CLIP	34.46	19.66	3.72	20.53
VideoSwin+X-CLIP	36.26	21.12	3.76	21.76
VideoSwin+X-CLIP (Multi-views)	**37.60**	**22.23**	**4.42**	**22.79**

Deeper Action

Conclusion

ECCV TEL AVIV 2022

- The models detecting action instances from multiple views perform better than the models using a single view.

- Using a suitable multi-view feature fusion strategy can improve the performance of temporal action localization.

- The one-stage temporal action detector without extra classifiers can achieve a good result on Fineaction dataset.

Deeper Action

ECCV TEL AVIV 2022

THANKS