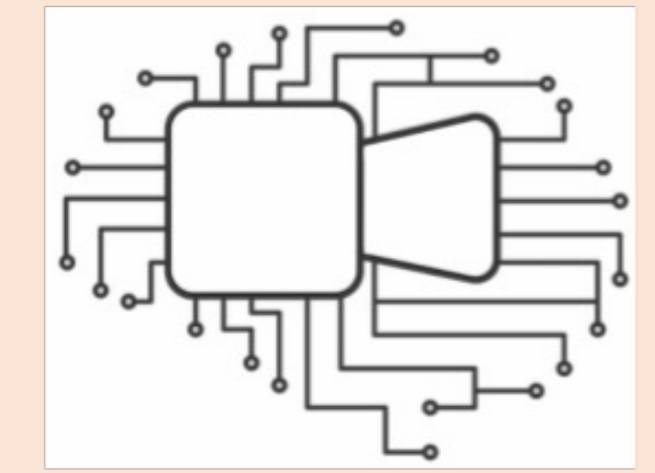




# Proposal-based Temporal Action Localization with Point-level Supervision

Yuan Yin, Yifei Huang, Ryosuke Furuta, Yoichi Sato  
Institute of Industrial Science, The University of Tokyo  
{yinyuan, hyf, furuta, ysato}@iis.u-tokyo.ac.jp



BMVC  
2023

## Problem

### Point-level Supervision

In the training phase, the low-cost yet efficient **point-level** supervision is utilized. It only provides one labeled frame for every action instance.



PlayingTennis (1'25)

PlayingVolleyball (1'44)

### Temporal Action Localization

In the inference phase, the trained model will predict the action classes and temporal duration of all action instances in the input video.



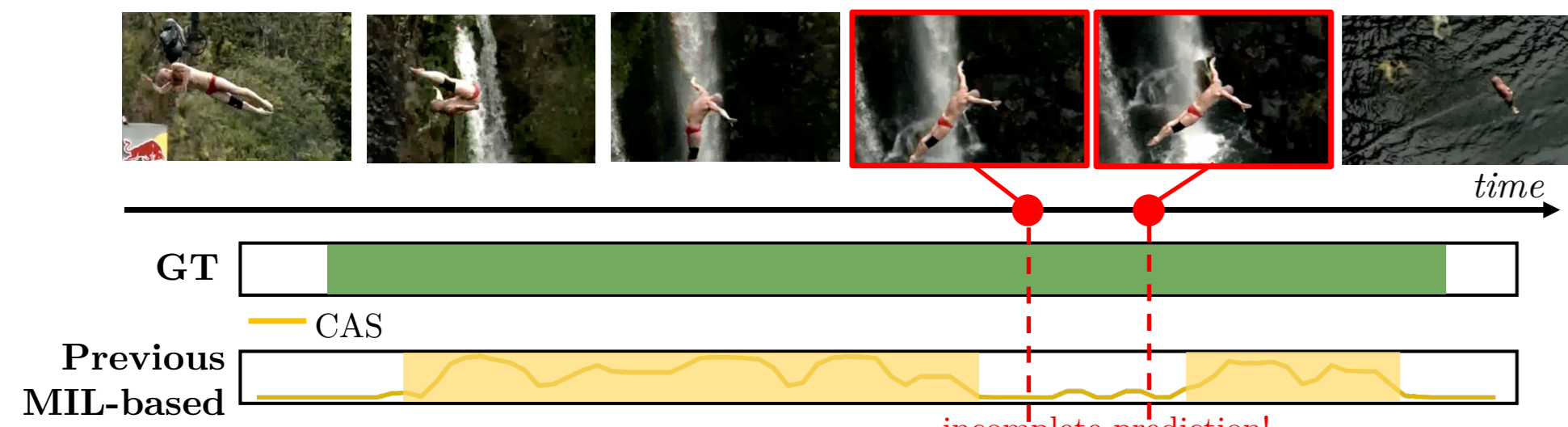
PlayingTennis (1'20 - 1'32)

PlayingVolleyball (1'43 - 1'46)

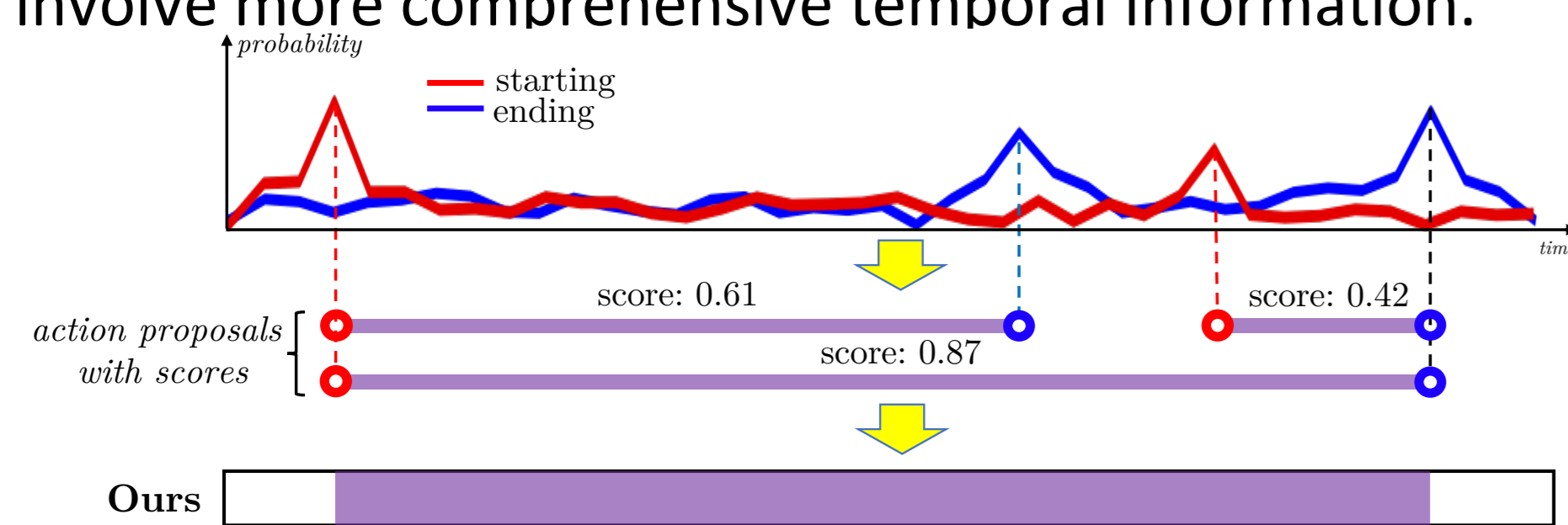
## Motivation

Previous methods based on multi-instance learning tend to output **incomplete** predictions when there are some hard-to-distinguish snippets.

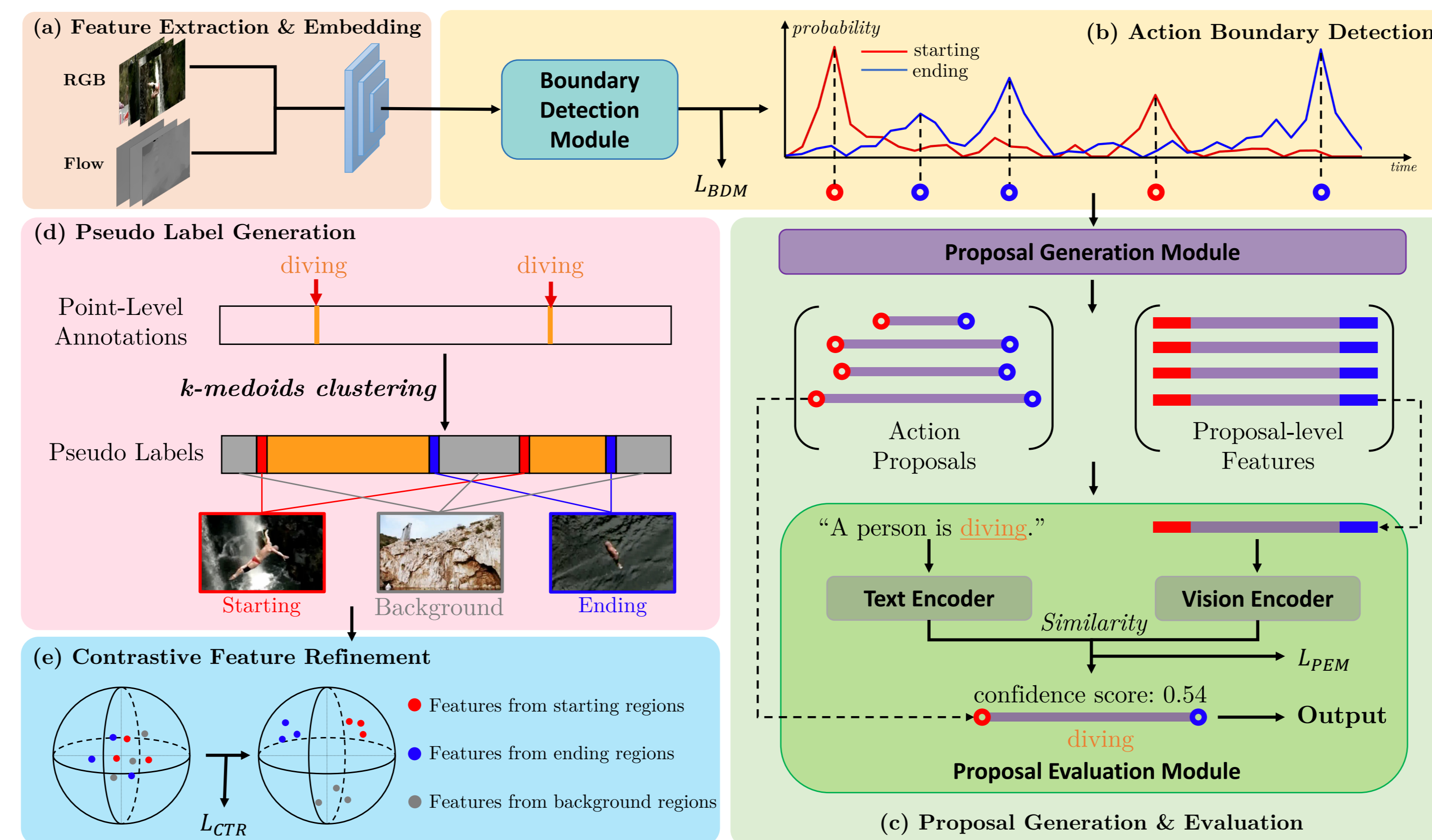
Hard-to-distinguish Snippets



To address this problem, we propose to localize actions by generating and evaluating **action proposals** of flexible duration that involve more comprehensive temporal information.



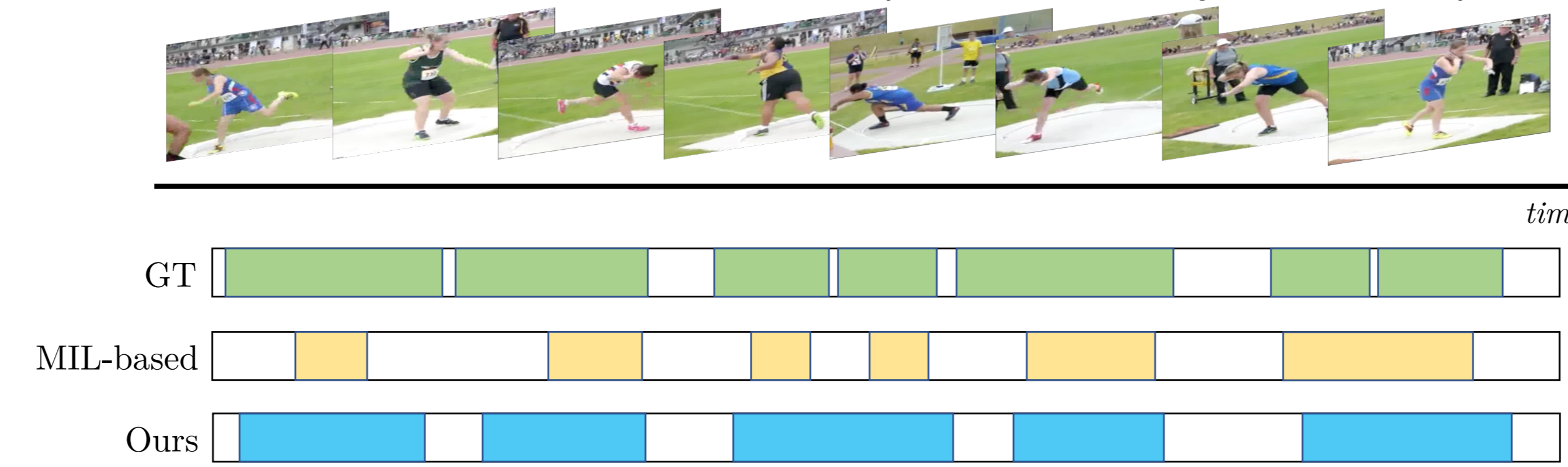
## Approach



- Given an input video, we extract features from its RGB and optical flow frames
- Boundary Detection Module** detects possible starting & ending time of action instances
- Proposal Generation Module** generates action proposals and use proposal-level feature to classify and evaluate action proposals
- To provide stronger supervision to the model, we design a k-medoids clustering algorithm to generate **dense pseudo labels**
- We further introduce a novel contrastive loss to refine the video features

### Qualitative results on THUMOS 14 dataset

The visualization results show that our method outputs more **complete** action predictions



## Experiments

Evaluation metrics: Mean Average Precision (mAP)

Our method significantly surpasses the SOTA methods under point-level supervision and video-level supervision meanwhile performing comparably with some fully-supervised methods

### Quantitative results on ActivityNet 1.3 dataset

Supervision	Method	mAP @ IoU(%)			AVG
		0.5	0.75	0.95	
Frame-level (Full)	BMN	50.1	34.8	8.3	33.9
	BSN	46.5	30.0	8.0	30.0
	G-TAD	50.4	34.6	9.0	34.1
	TAGS	56.3	36.8	9.6	36.5
Video-level (Weak)	FAC-Net	37.6	24.2	6.0	24.0
	ACM-Net	37.6	24.7	6.5	24.4
	FTCL	40.0	24.3	6.4	24.8
	ASM-Loc	41.0	24.9	6.2	25.1
Point-level (Weak)	LACP	40.4	24.6	5.7	25.1
	<b>Ours</b>	<b>48.3</b>	<b>27.8</b>	<b>7.0</b>	<b>29.1</b>

### Quantitative results on THUMOS 14 dataset

Method	mAP @ IoU(%)				AVG	AVG
	0.1	0.3	0.5	0.7		
SF-Net	68.3	52.8	30.5	12.0	51.2	31.6
DCST	72.3	58.2	35.9	12.8	55.6	35.4
LACP	75.5	64.6	<b>45.3</b>	<b>21.8</b>	62.7	<b>44.5</b>
<b>Ours</b>	<b>77.1</b>	<b>65.9</b>	44.9	20.2	<b>63.3</b>	43.9

### Quantitative results on GTEA and BEOID dataset

Method	GTEA				BEOID			
	mAP @ IoU(%)			AVG	mAP @ IoU(%)			AVG
	0.3	0.5	0.7	[0.1:0.7]	0.3	0.5	0.7	[0.1:0.7]
SF-Net	37.9	19.3	11.9	31.0	40.6	16.7	3.5	30.1
DCST	38.3	21.9	18.1	33.7	46.8	20.9	5.8	34.9
LACP	<b>55.7</b>	33.9	20.8	43.5	61.4	42.7	25.1	51.8
<b>Ours</b>	52.1	<b>37.3</b>	<b>22.2</b>	<b>45.1</b>	<b>65.3</b>	<b>45.1</b>	<b>26.6</b>	<b>54.2</b>