Precise Temporal Localization for Complete Actions with Quantified Temporal Structure

Chongkai Lu[1], Ruimin Li[2], Hong Fu[3], Bin Fu*[2], Yihao Wang[1], Wai-Lun Lo[4], and Zheru Chi[1]

[1]The Hong Kong Polytechnic University
[2]Xi'an Institute of Optics and Precision Mechanics of CAS, Xi'an
[3]The Education University of Hong Kong
[4]Chu Hai College of Higher Education

*Corresponding Author

Outline

I. Background & Motivation

II. Action Progression Networks

1) Progression labels generation
2) Action Progression Network
3) Complete Action Searching

III. Experimental Results

IV. Conclusion & Discussion

Background & Motivation

- Fine-grained **Temporal Action Localization**;
- Small Dataset;
- Complete Actions Recognition.

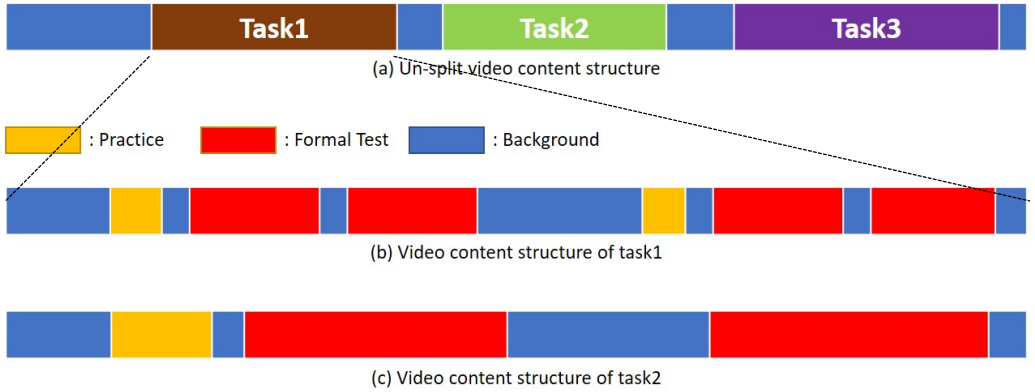


Fig 1. Distribution of content in a video of the DFMAD-70

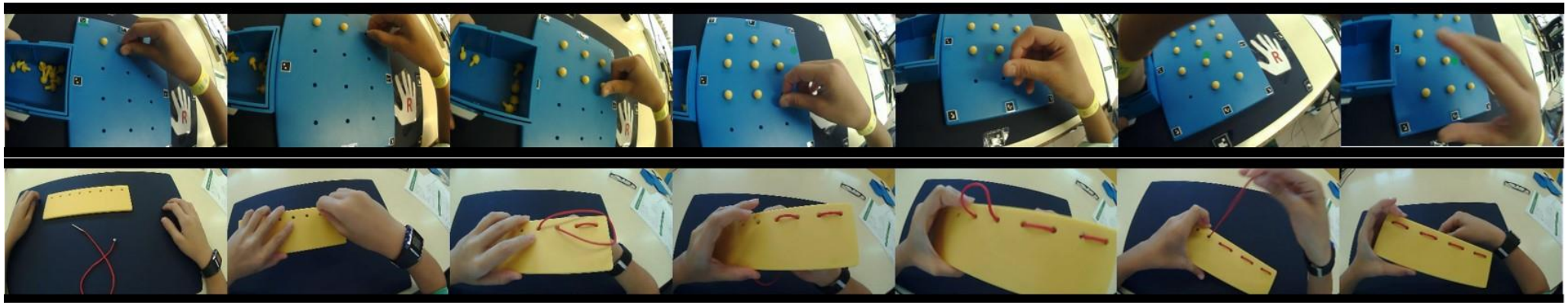


Fig 2. A complete action instance of task1 and task2

I. Background & Motivation

- Works on **action proposals**:
  - Based on other method: Complexity from two-stages workflow;
  - Based on sampling: Poor localization precision.
- Training at video level: Intensive computation, memory- and data-hungry;
- *Loc* based on *Cls*: Completeness agnostic problem.

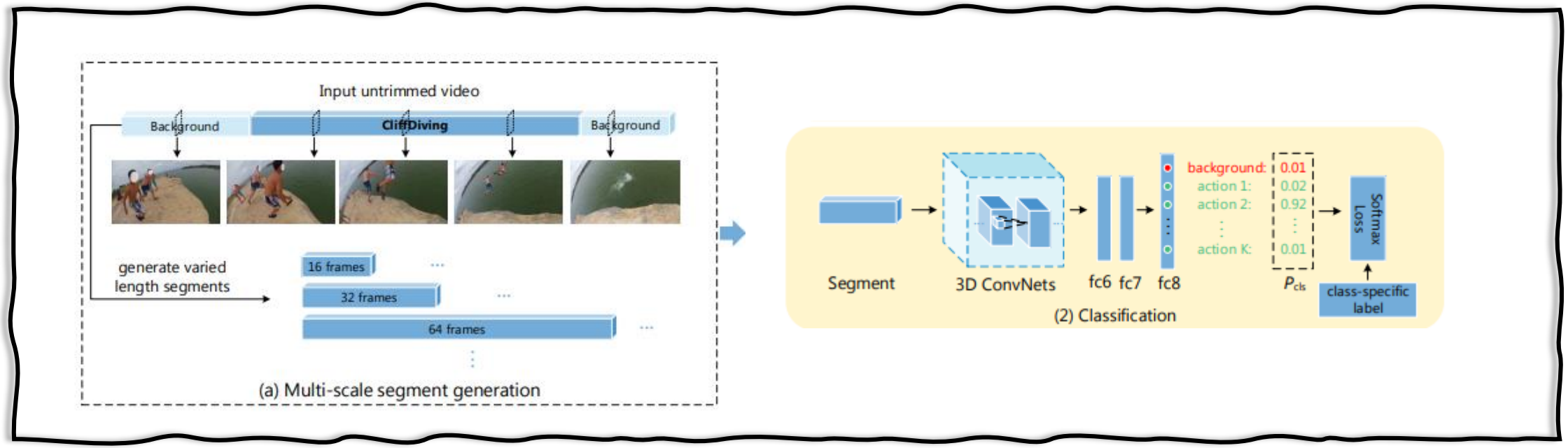


Fig 3. Conventional framework for Temporal Action Localization [1]

Action Progression Networks

Proposed methods: Neural network for predicting action progressions;

Central Point: Image-level training. Input is temporally local, but label is global.

Advantages: Fine-grained; Memory and data friendly; Recognize completeness;

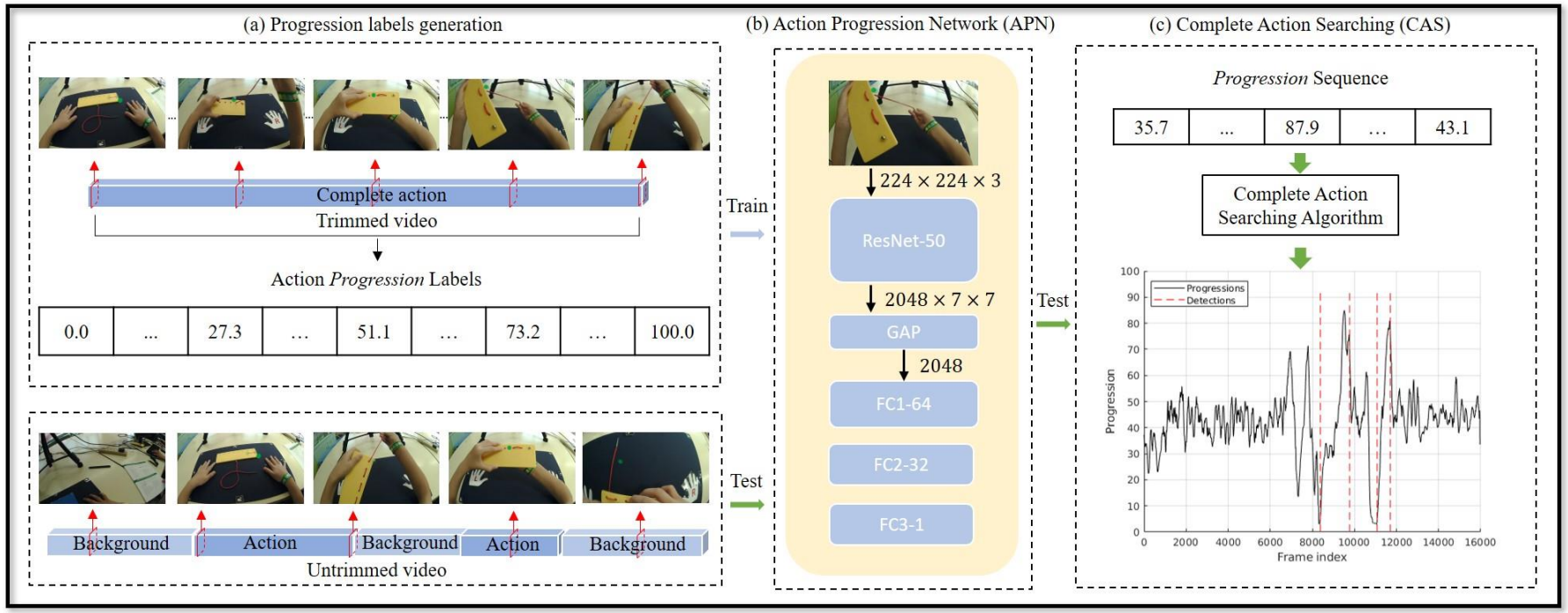


Fig 4. Overview of the framework of our proposed APN

Action Progression Networks

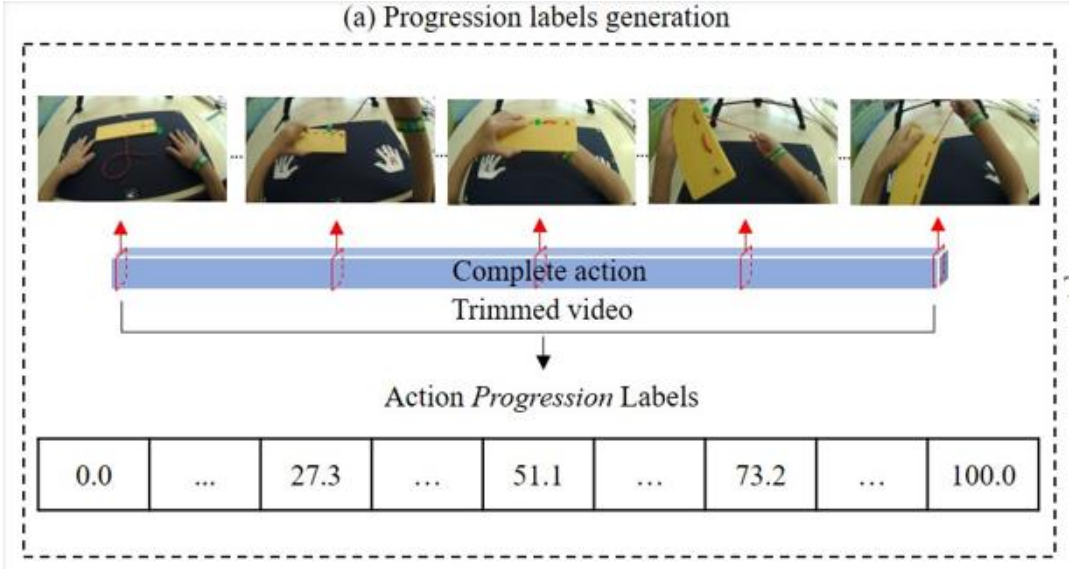
Progression labels generation:

Formula:

$$y_t^n = p_{min} + \frac{p_{max}}{e_n - s_n + 1} t,$$

y_t^n is the progression label assigned to the t-th frame in the n-th complete action instance; p_{min} and p_{max} are set as 0 and 100;

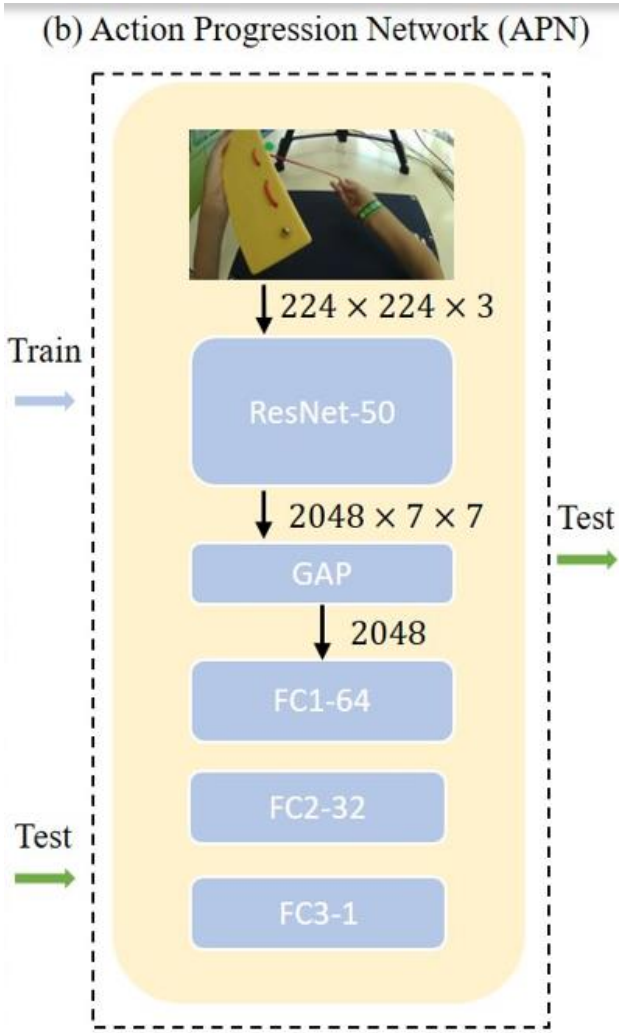
Since this is a kind of coarse labeling, we deemed our method to be most applicable in the domain of professional action localization, whose temporal pattern of action is almost fixed.



Action Progression Networks

Action Progression Network (APN)

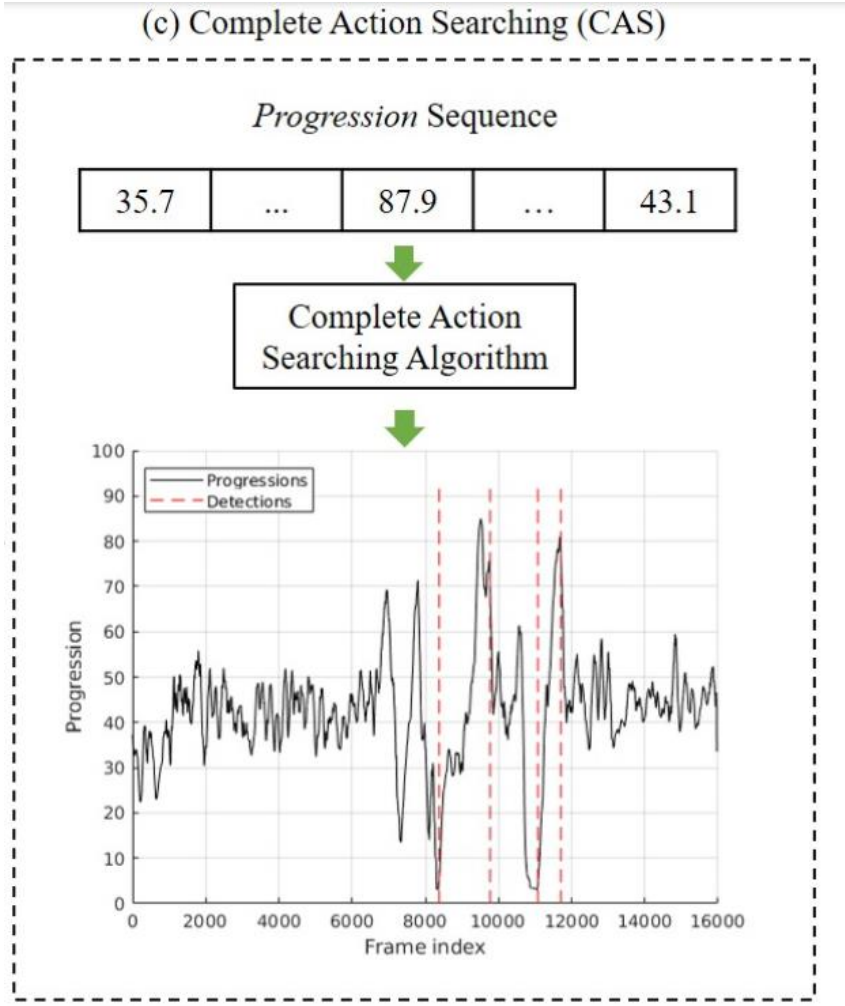
- Backbone: ResNet-50
- Input: single frame (224, 224, 3)
- Output: single scalar
- Target: single scalar – action progression of this frame
- Loss function: Mean Absolute Error (MAE)



Action Progression Networks

Complete Action Searching (CAS)

- Executed at test phase only
- Find intervals having pattern like *range(0, 100)*
- Completeness recognition
- Robust to noise frames.



Action Progression Networks

Experimental Results:

TABLE II
AVERAGE PRECISION FOR MULTI-TASK AND ACTION-SPECIFIC TRAINING

Method on Action		mAP @ IoU				
		0.5	0.6	0.7	0.8	0.9
multi-task	from sractch	49.5	48.2	40.4	16.9	2.1
	pre-trained	53.6	53.2	47.2	37.6	19.8
action-specific	from scratch	78.4	69.5	53.2	27.5	6.2
	pre-trained	**94.0**	91.8	83.5	56.1	12.1

TABLE III
ACTION DETECTION PERFORMANCE ON DFMAD-70

Method on Action		AP @ IoU				
		0.5	0.6	0.7	0.8	0.9
Action1	APN from sractch	96.9	95.0	65.4	18.6	1.2
	APN pre-trained	**100**	100	97.6	56.0	6.8
Action2	APN from scratch	59.8	44.0	40.9	36.3	11.1
	APN pre-trained	87.9	83.5	69.3	56.2	17.4
	Li et al. [29]	**90.6**	87.5	75.0	37.5	0.0

Action Progression Networks

Conclusion:

- We proposed a temporal action localization framework based on temporal analysis on quantified temporal structure of actions.
- Our framework can work on **small datasets**, **detect actions precisely** and **screen incomplete detections**.
- Our work bridge the models in image recognition and video recognition. Applying our method on videos, one can get the benefit of using the most advanced 2D deep architecture.

Discussion:

- Our framework are supposed to work on detecting actions that share similar temporal context in same category. Therefore, it's especially suitable for professional action detection.
- Our framework can be used for action classification, single action detection naturally without any change.
- Performance on the benchmark datasets (e.g., THUMOS14, ActivityNet) is under evaluating, preliminary experimental results display that our framework can get competitive precision in contrast to SOTA.

THANK YOU!

Q & A

Reference

[1] Z. Shou, D. Wang, and S.-F. Chang, “Temporal action localization in untrimmed videos via multi-stage cnns,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1049–1058, 2016.