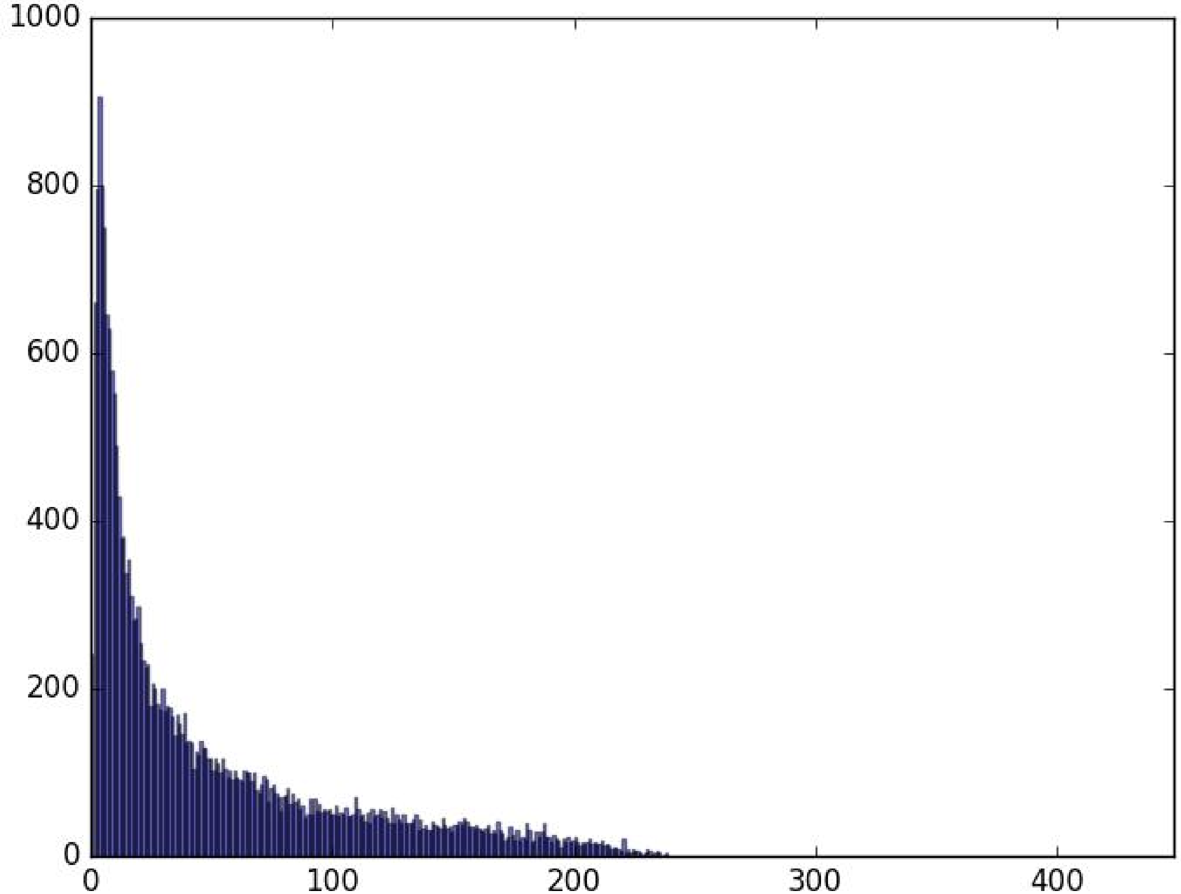
Temporal Convolution Based Action Proposal: Submission to ActivityNet 2017

Tianwei Lin[1], Xu Zhao[1], Zheng Shou[2]

{wzmsltw, zhaoxu}@sjtu.edu.cn, zheng.shou@columbia.edu

[1] Computer Vision Laboratory, Shanghai Jiao Tong University, China
[2] Columbia University, USA

ActivityNet Dataset Analysis

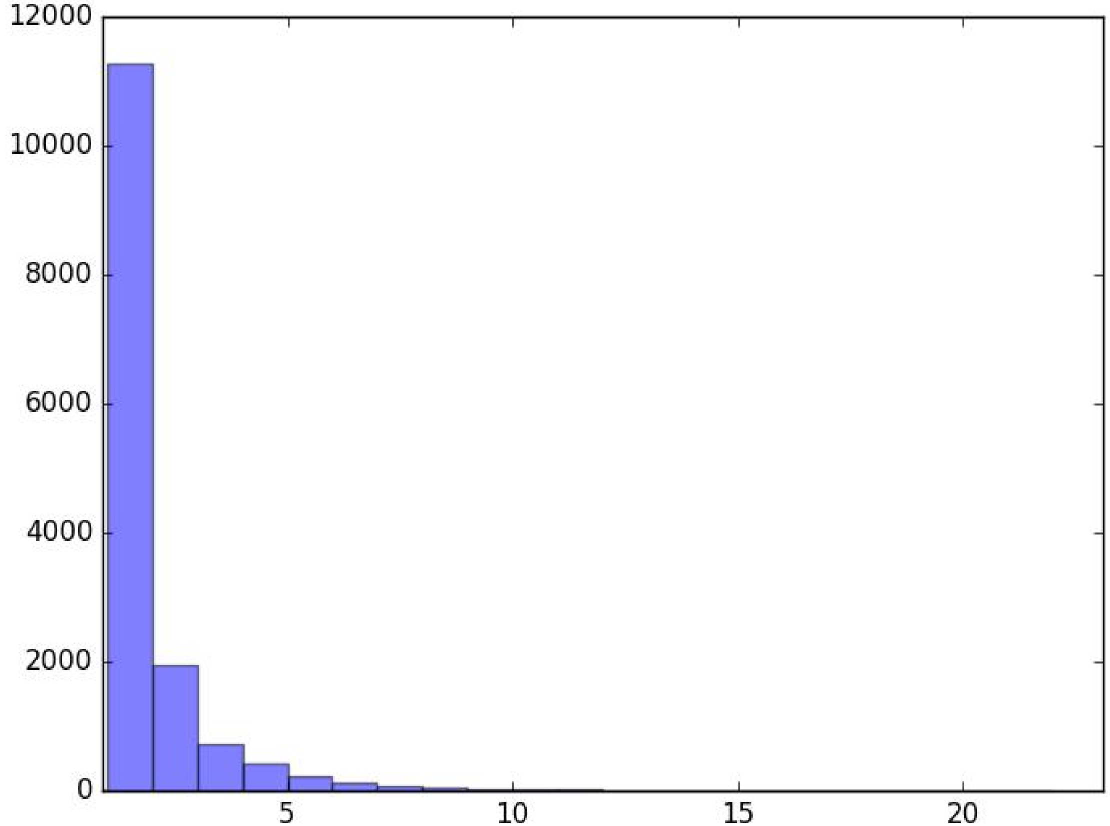


#videos

video length (second)

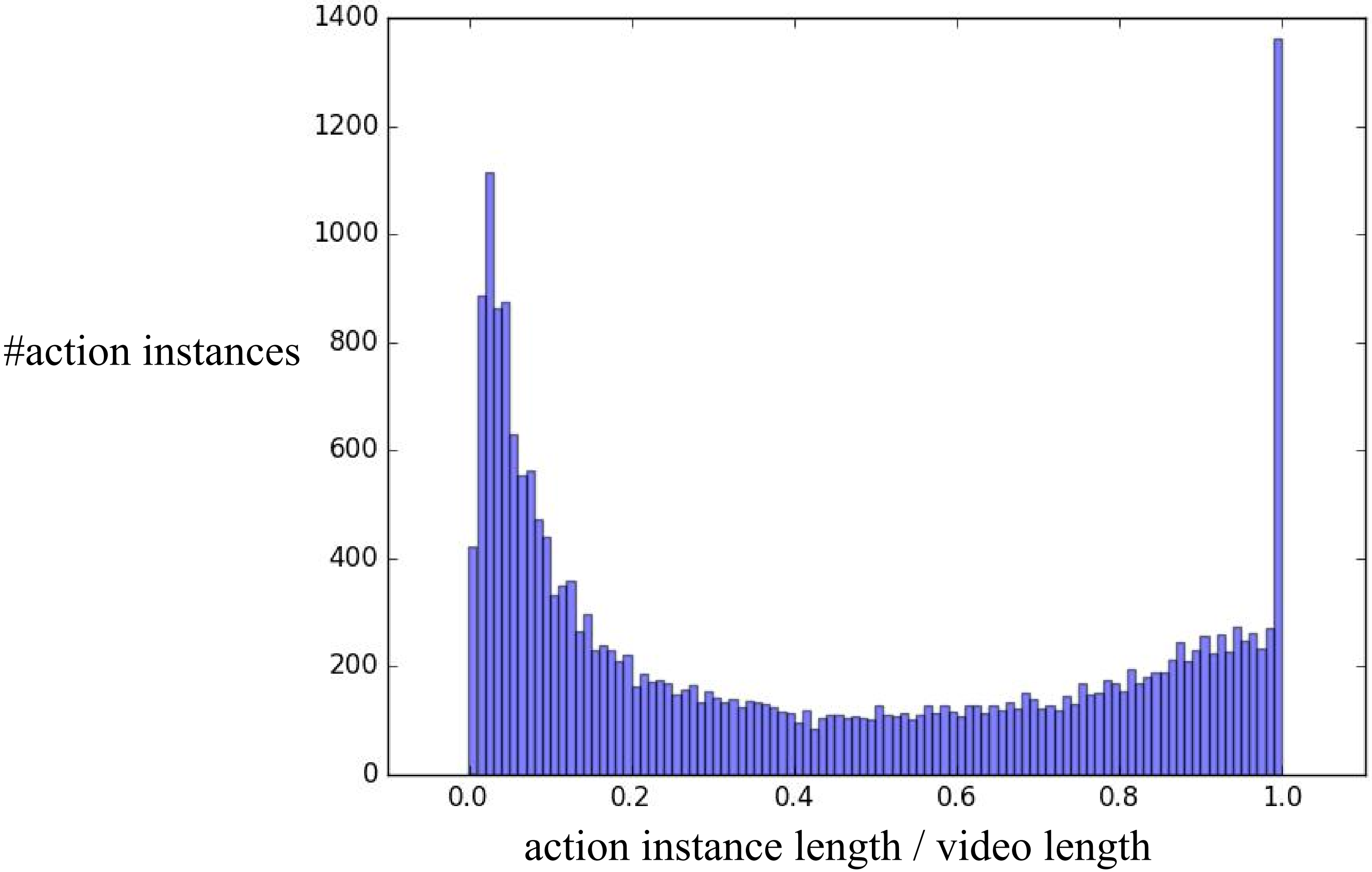
ActivityNet Dataset Analysis

#videos



#action instances contained in each video

ActivityNet Dataset Analysis



Problem Analysis

Problem Analysis

Proposal + **Classification** = **Localization**

Image: Classification

~3% error
on ILSVRC 2017

Video: Action Classification

~9% top-1 error on
ActivityNet 2017

The performance is close

Problem Analysis

Proposal + **Classification** = **Localization**

Image: Object Detection

~73% mAP
on ILSVRC 2017

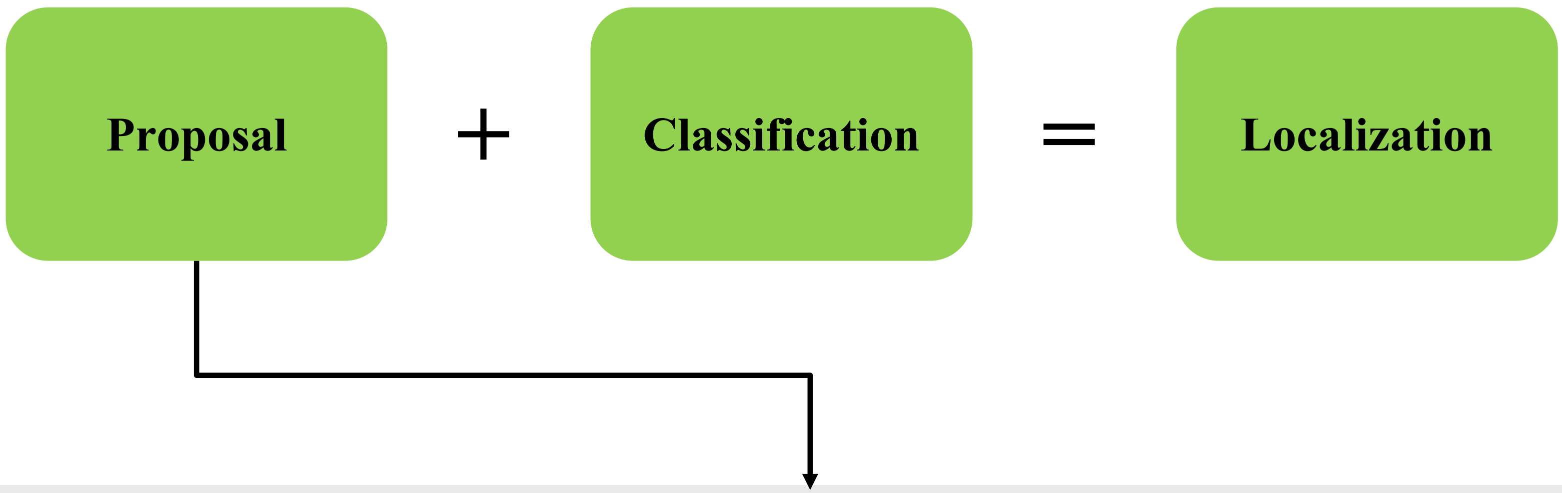
Video: Temporal Action Localization

~33% average mAP on
ActivityNet 2017

~30% mAP@0.5 on
THUMOS'14

The performance gap is huge!

Problem Analysis



Q: Why the performance of temporal action localization is much worse than object detection?

A:

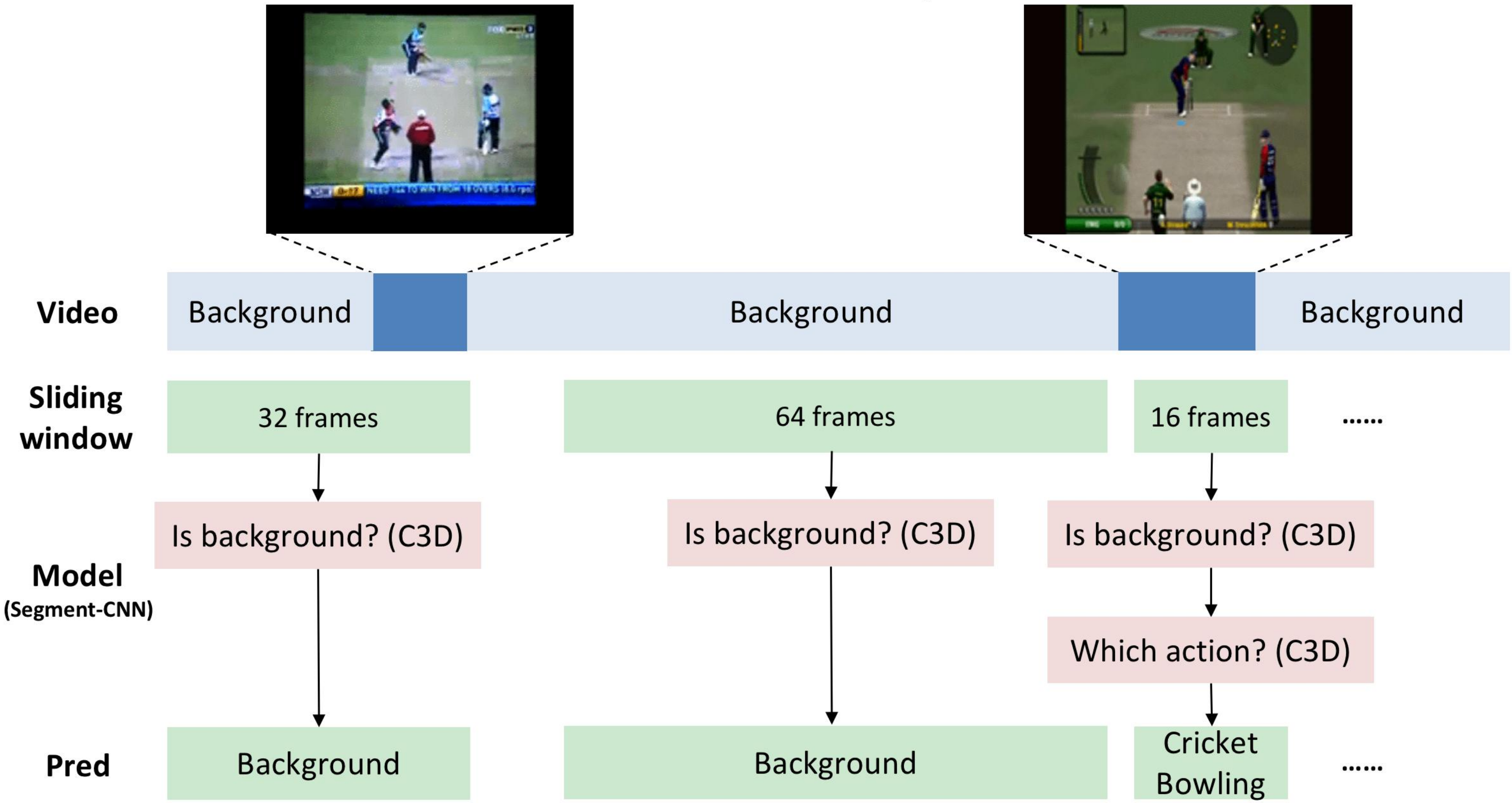
- **Main bottleneck -> the quality of temporal action proposal.**

- **Direction: mainly focus on the temporal action proposal task in this challenge.**

- **Problems to address:**

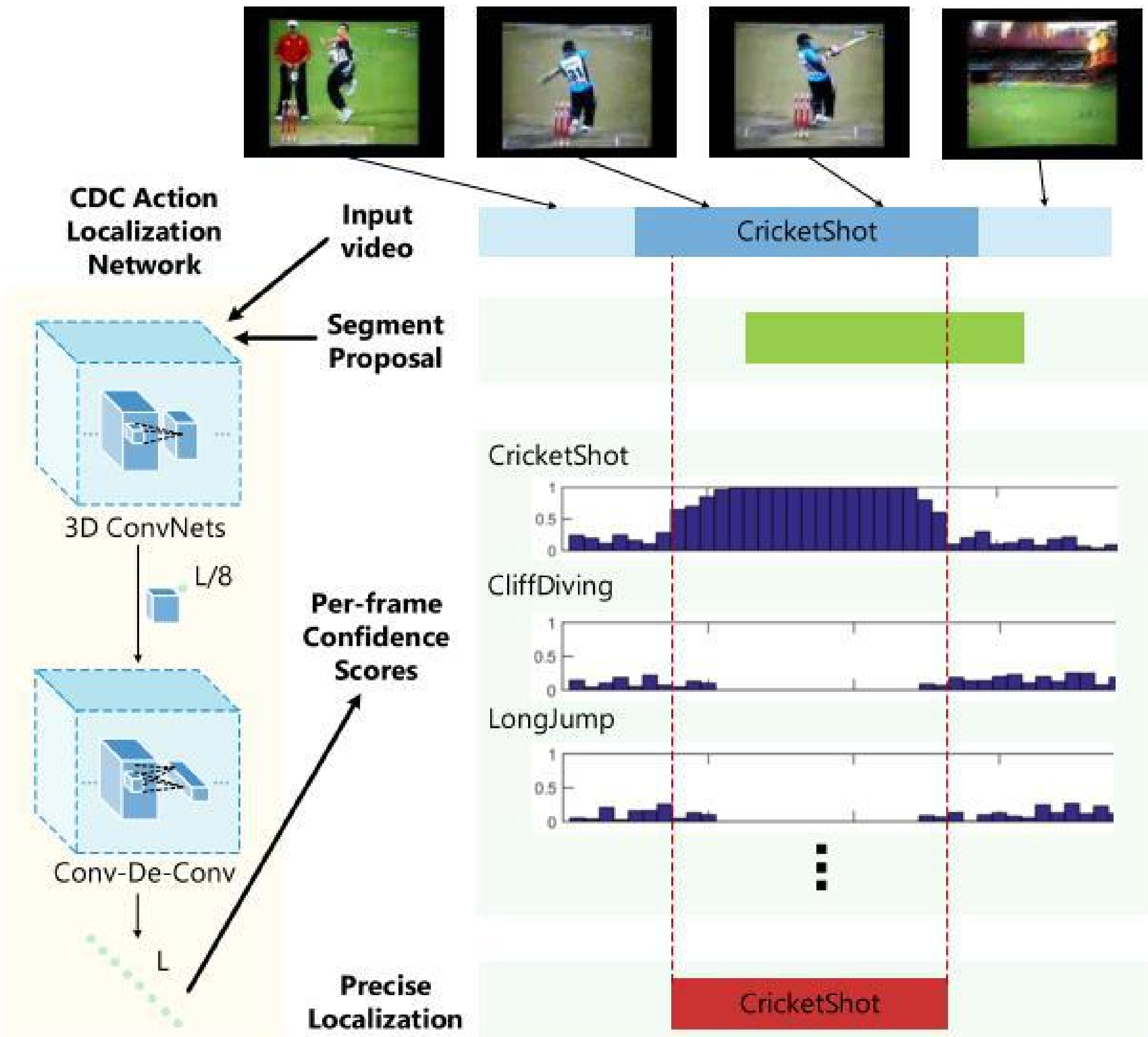
1. **Whether a proposal contains action or not. (confidence score)**
2. **Precisely locate the start and end time of proposal. (temporal boundaries)**

Previous Work: Conv-De-Conv Networks



Temporal Action Localization in Untrimmed Videos via Multi-stage CNNs. Z. Shou, D. Wang, and S.-F. Chang. In CVPR 2016.

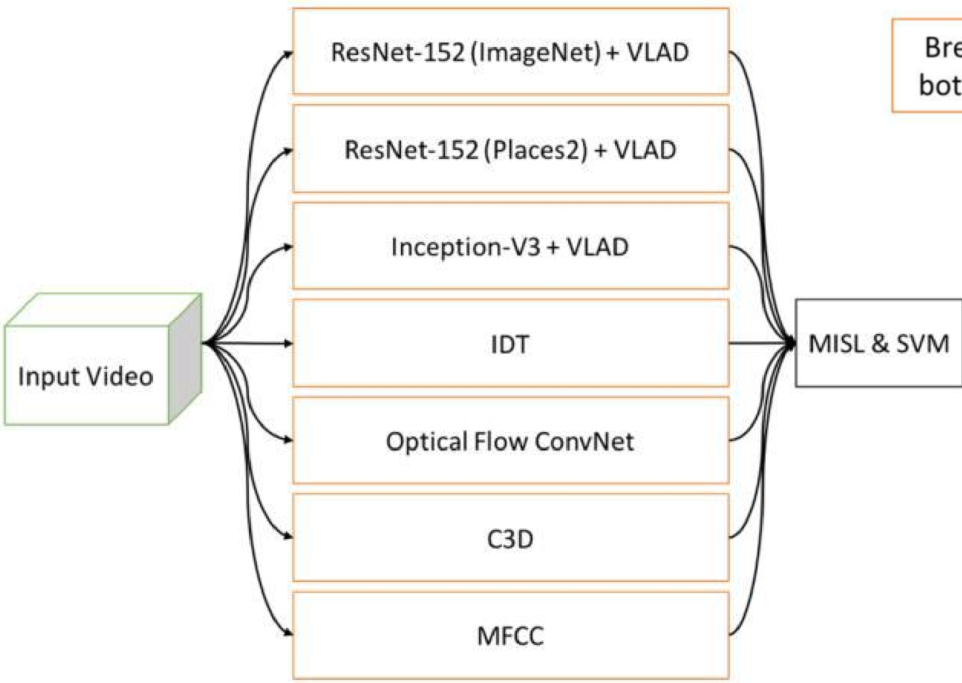
Previous Work: Conv-De-Conv Networks



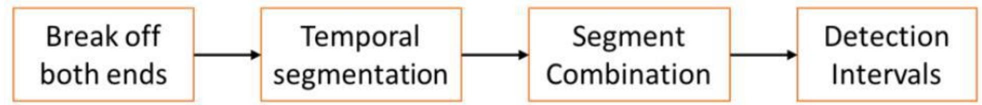
CDC: Convolutional-De-Convolutional Networks for Precise Temporal Action Localization in Untrimmed Videos. Z. Shou, J. Chan, A. Zareian, K. Miyazawa, and S.-F. Chang. In CVPR 2017. Oral.

Previous Work: UTS submission (last year winner)

Action classification framework



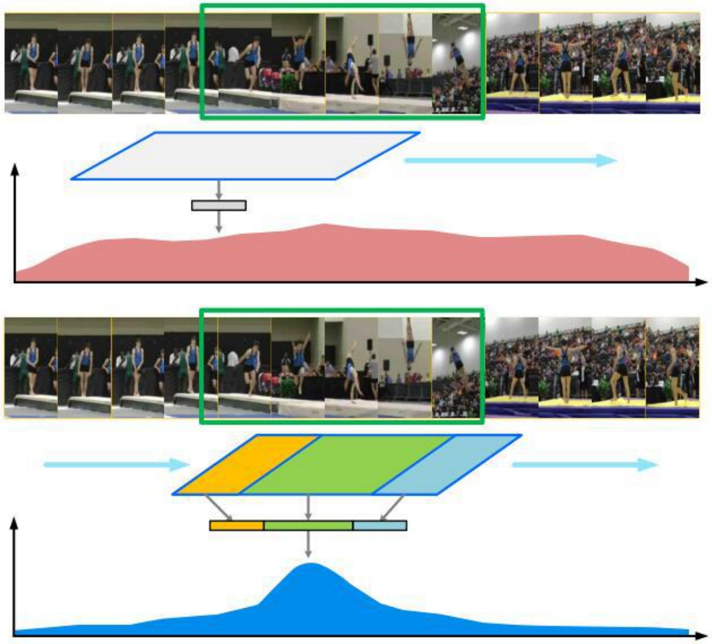
Detection pipeline



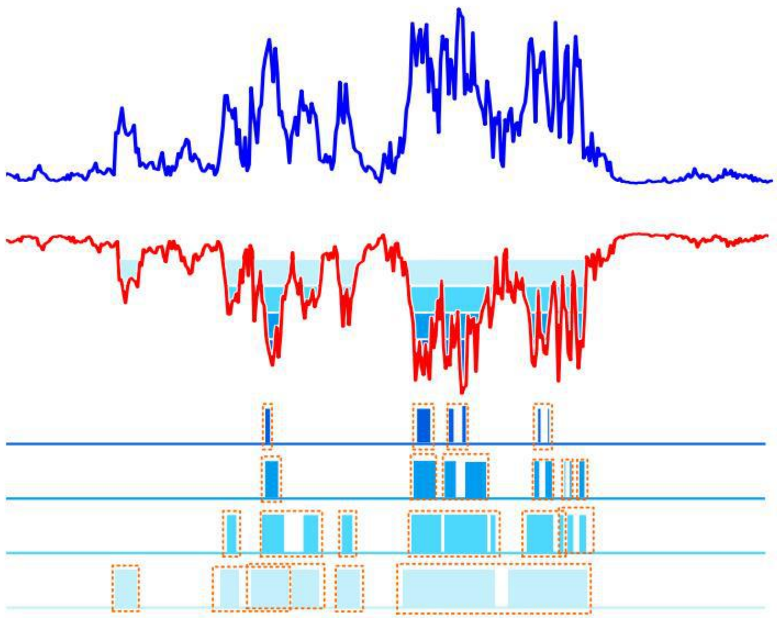
UTS at activitynet 2016.
Ruxin Wang, Dacheng Tao. In ActivityNet Challenge 2016.

Previous Work: SSN

Approach Overview



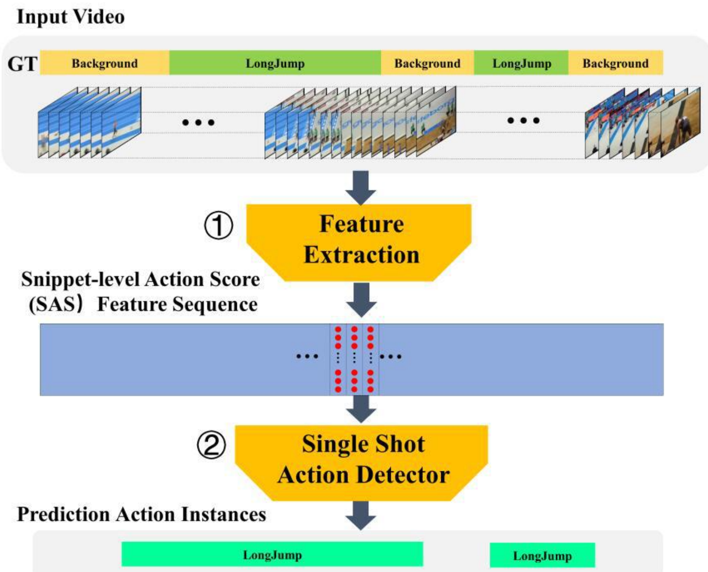
Temporal Actionness Grouping



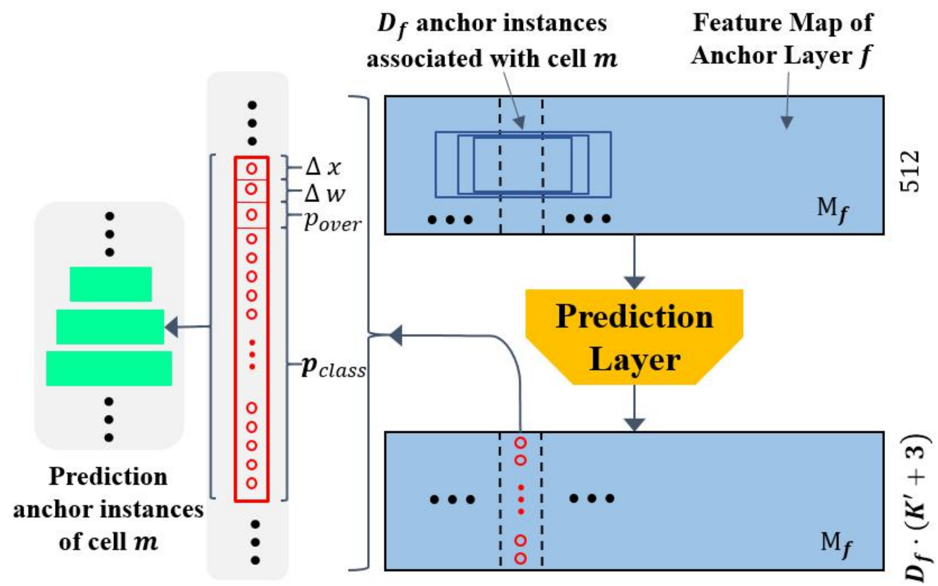
Temporal Action Detection with Structured Segment Networks.
Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Dahua Lin, Xiaoou Tang. arXiv 1704.06228

Previous Work: SSAD

Approach Overview



Anchor Mechanism of SSAD



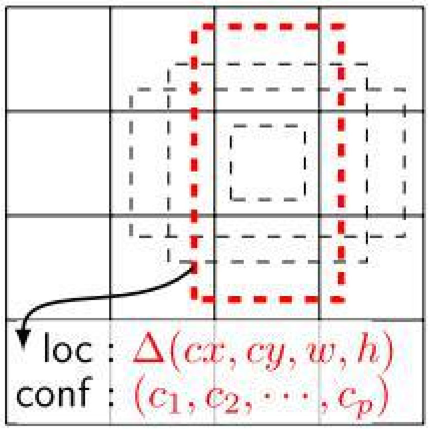
Single Shot Temporal Action Detection.
Tianwei Lin, Xu Zhao, Zheng Shou. In ACM Multimedia 2017.

Motivation

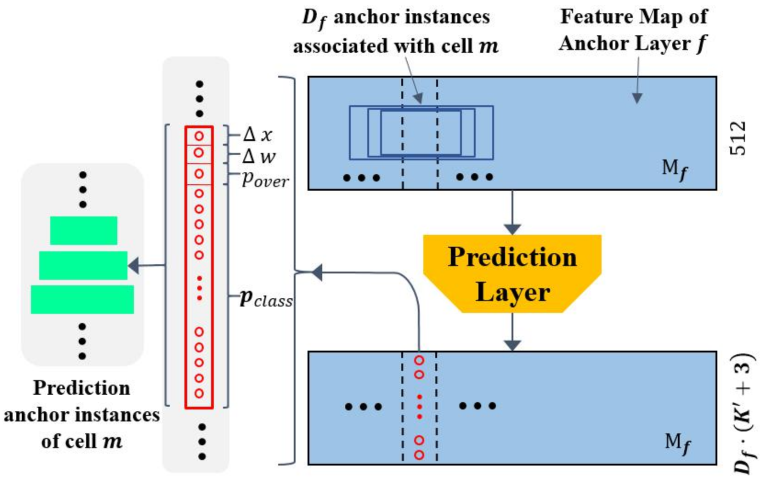
- Various scale at various position
- Likelihood of being action
- Locate precise temporal boundaries

Motivation

Anchor mechanism in object detection (SSD)



Anchor mechanism in temporal action detection (SSAD)



Advantages of anchor mechanism:

- can cover instances of various scale at various position
- can directly make localization using convolutional layer

SSD: Single Shot MultiBox Detector
Wei Liu, et. al. In ECCV 2016.

Single Shot Temporal Action Detection.
Tianwei Lin, Xu Zhao, Zheng Shou. In ACM Multimedia 2017.

Our Approach: Framework

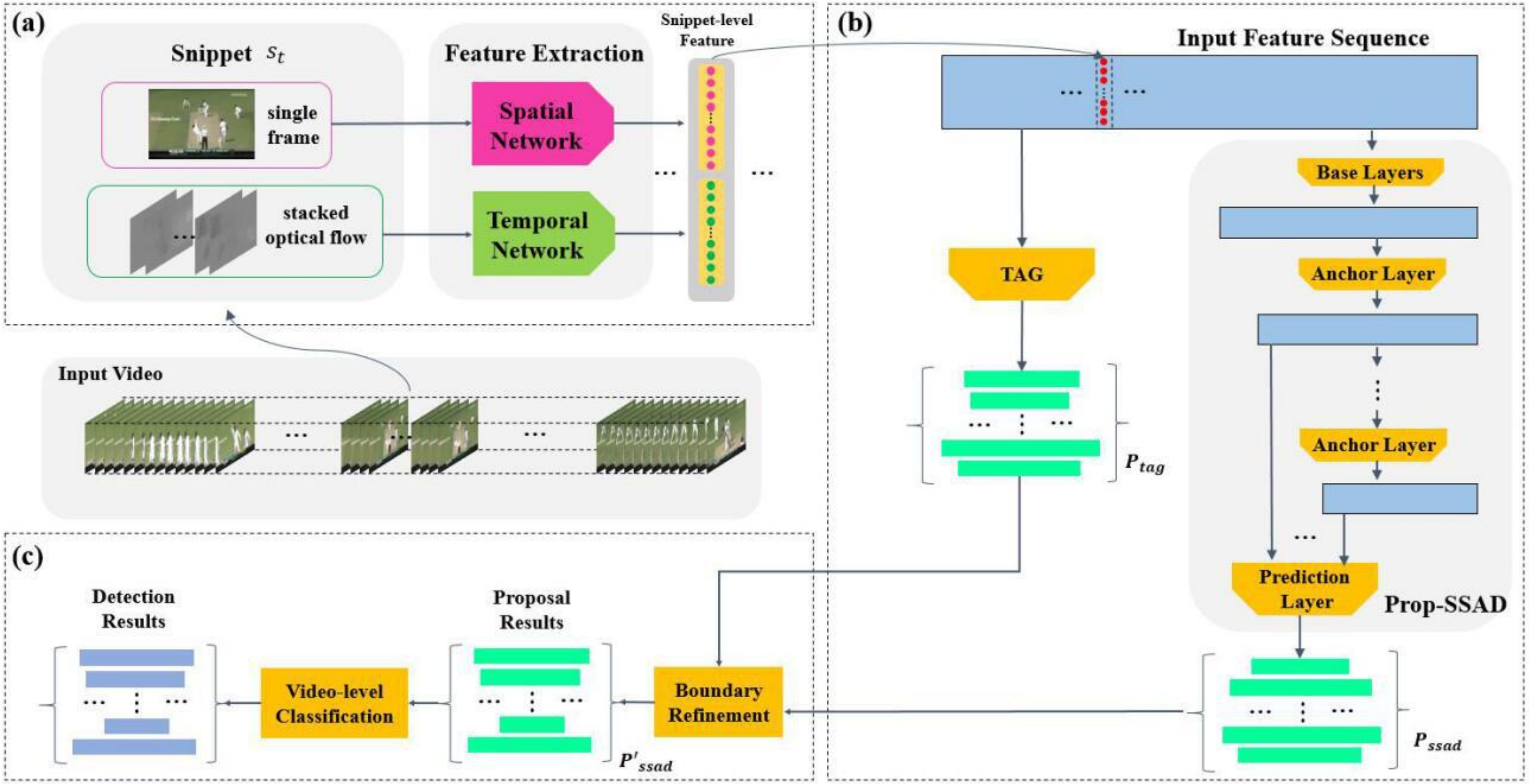
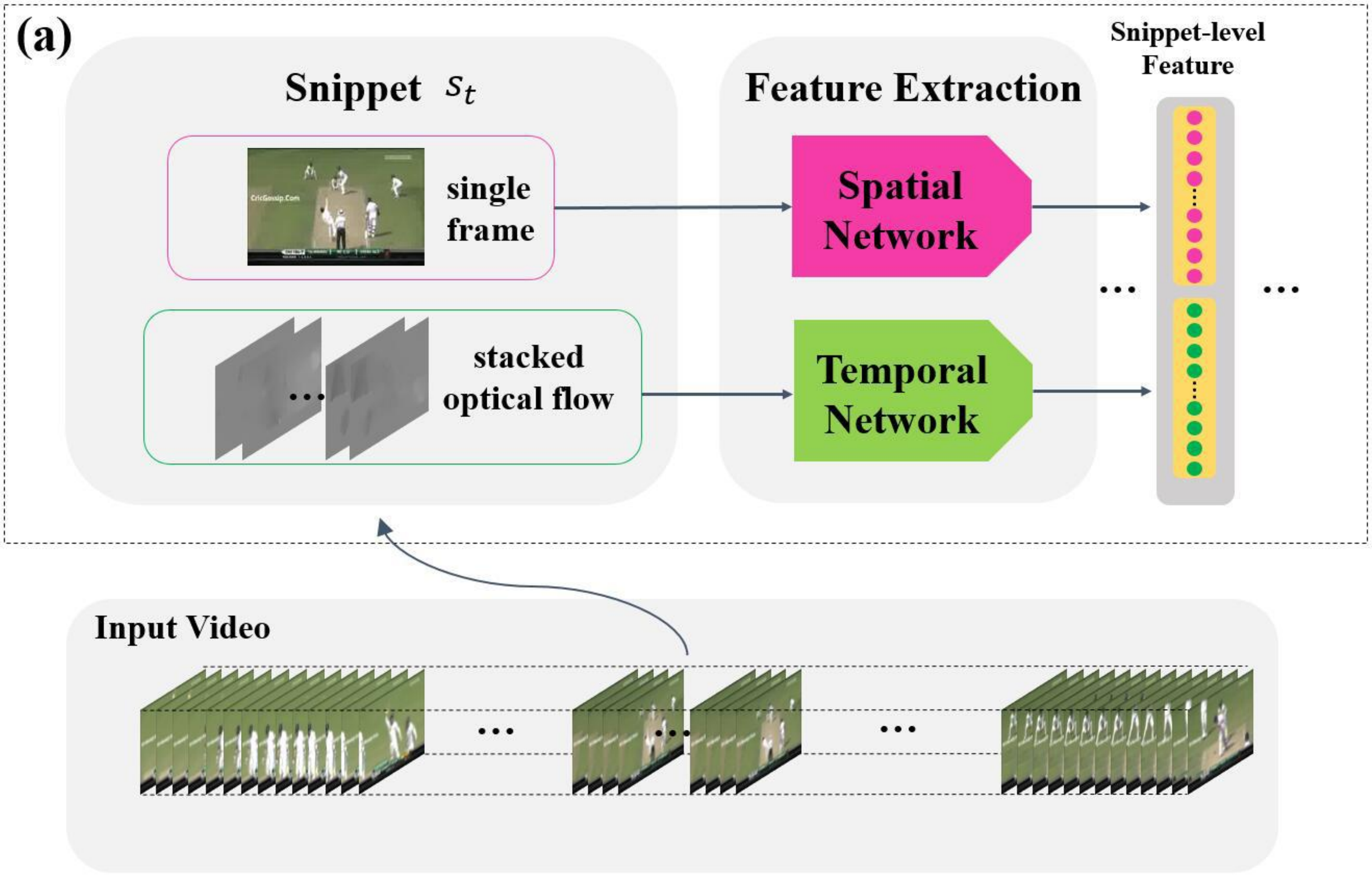


Figure 1: The framework of our approach. (a) Two-stream networks are used to extract snippet-level features. (b) Prop-SSAD model and TAG method are used for proposal generation separately. (c) Proposals generated by TAG are used for refining the boundaries of proposals generated by Prop-SSAD model. We use video-level action classification result as the category of temporal action proposals to get temporal action localization result.

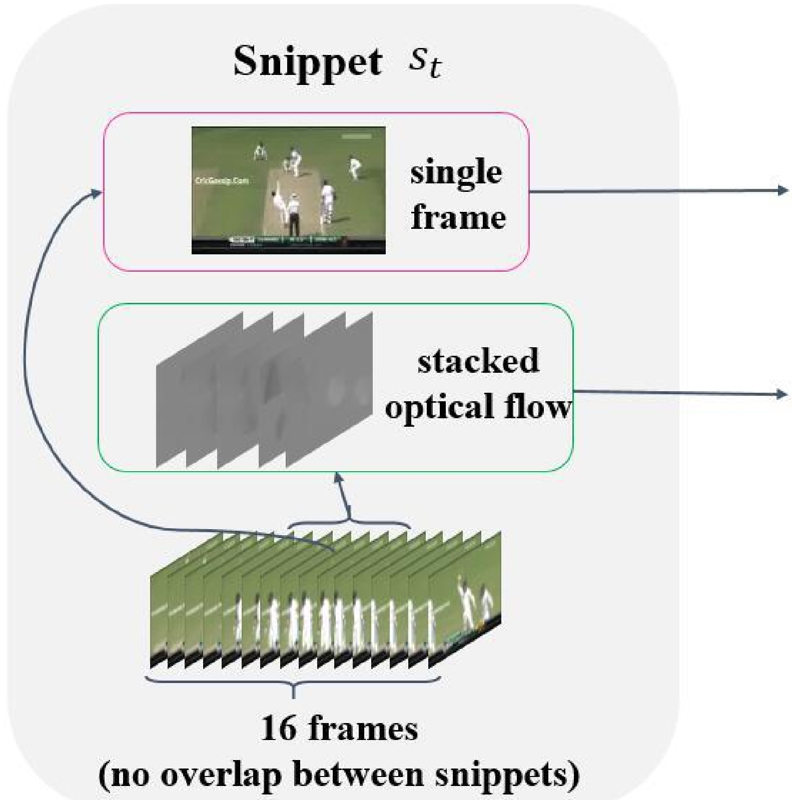
Our Approach: Feature Extraction

Feature extraction overview



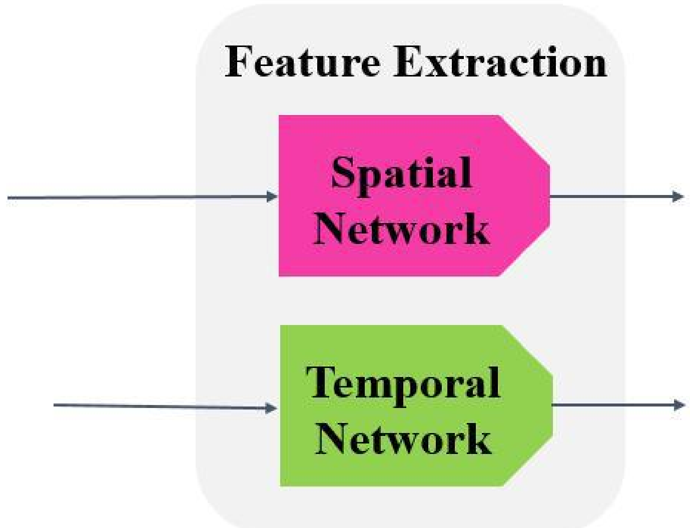
Our Approach: Feature Extraction

Definition of snippet



Our Approach: Feature Extraction

Two-stream network for feature extraction

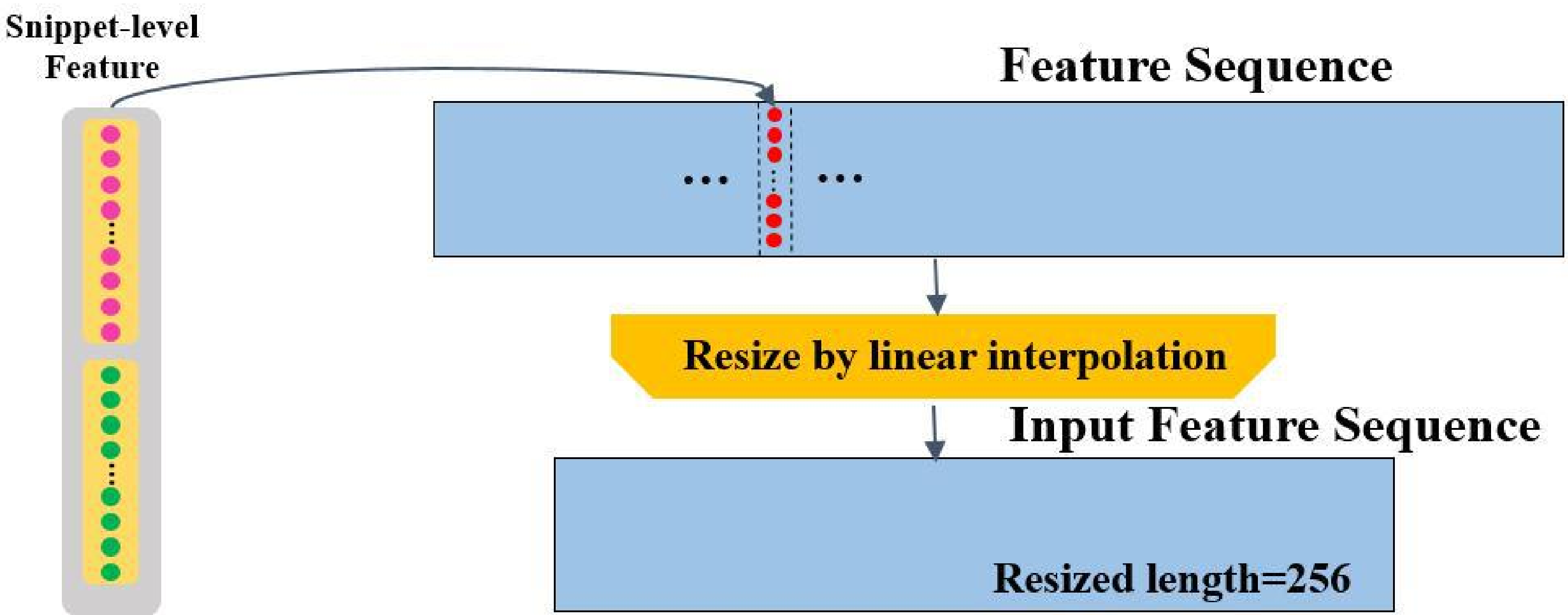


- **Two-stream network**
- Employ models from last year CUHK team, they are the winner of untrimmed action classification task of ActivityNet Challenge 2016.
- These models are trained on training set of ActivityNet dataset.

Cuhk & ethz & siat submission to activitynet challenge 2016.
Y. Xiong, L. Wang, Z. Wang, et. al.
***arXiv preprint. arXiv:1608.00797*, 2016**

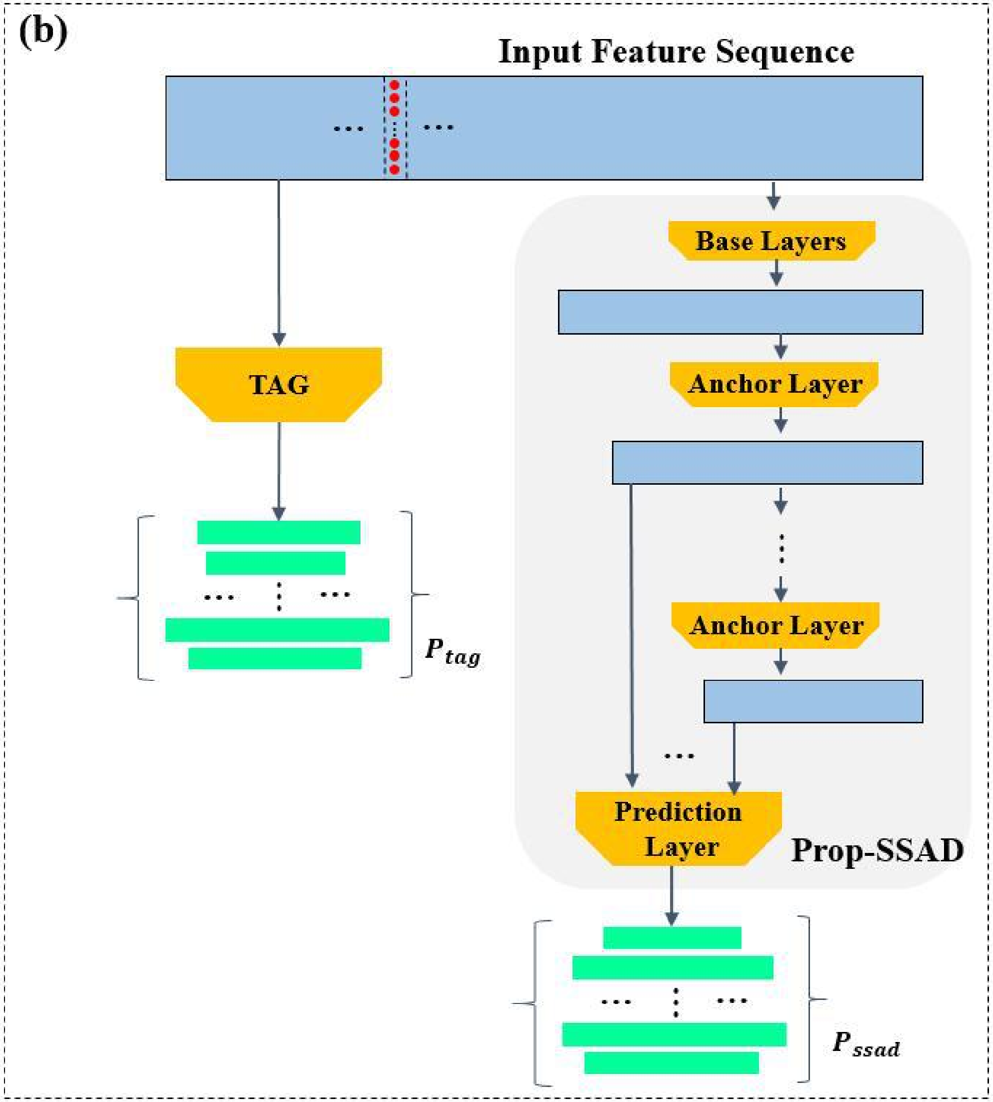
Our Approach: Feature Extraction

Temporal feature resize



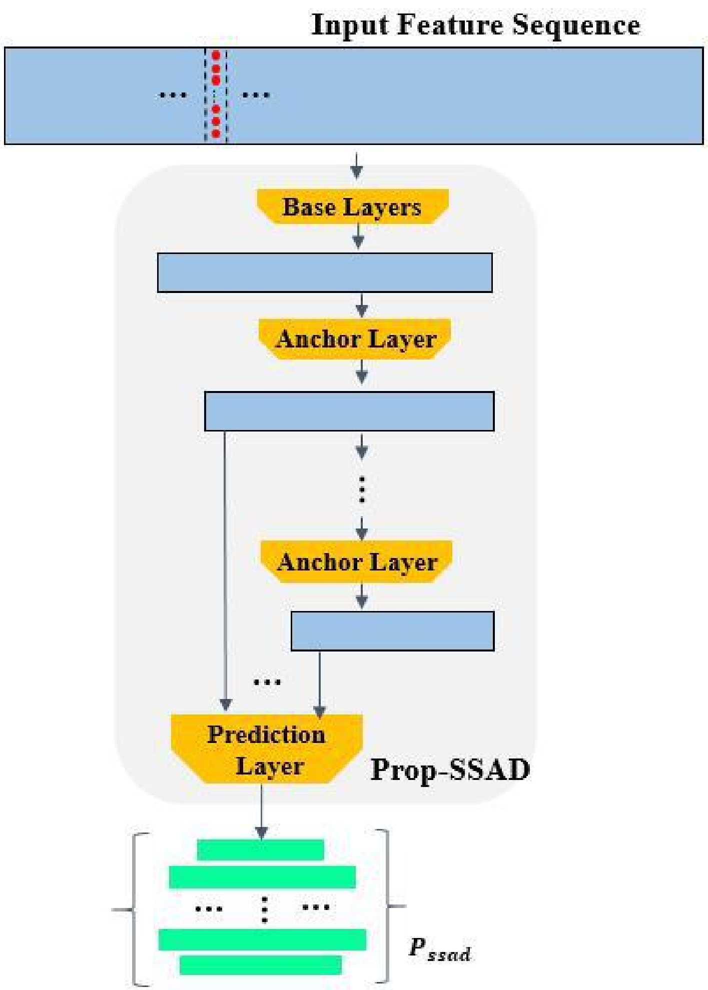
Our Approach: Proposal Generation

Proposal Generation Overview



Our Approach: Proposal Generation

Prop-SSAD method

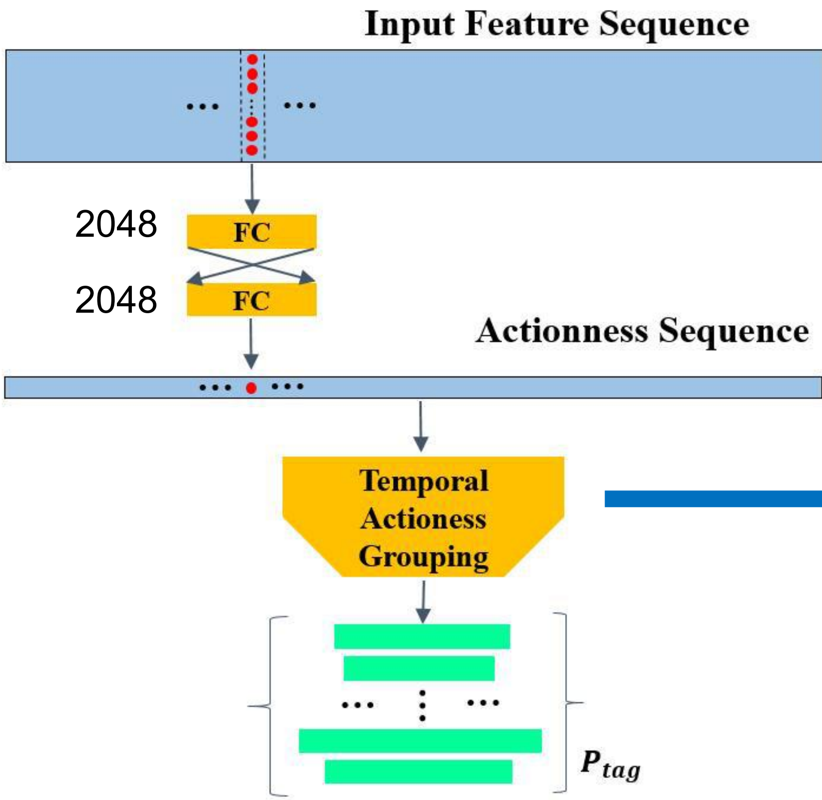


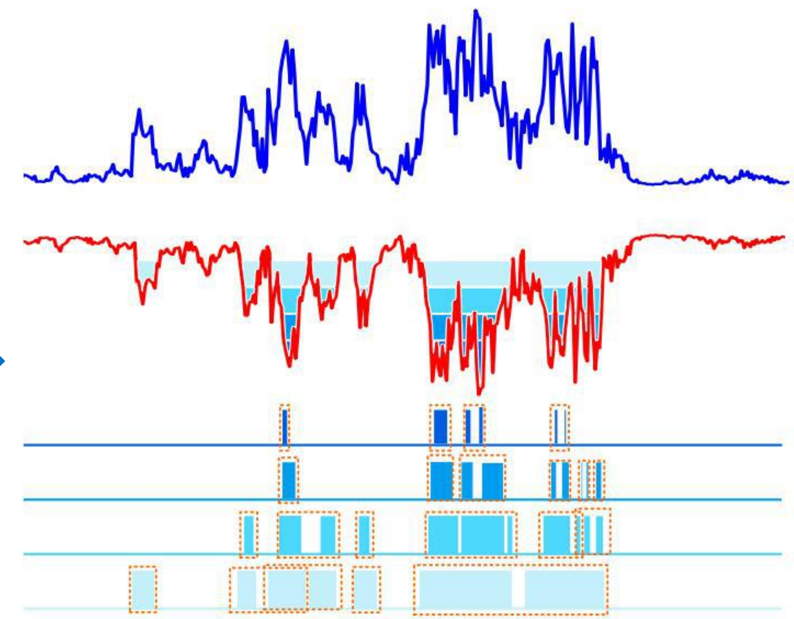
Key points of Prop-SSAD

- Anchor mechanism
- Only proposal, no action classification
- 7 anchor layers: 1, 2, 4, 8, 16, 32, 64 locations
- 4 scales: $1/\sqrt{2}$, 1, $\sqrt{2}$, 2 times base scale
- No boundary regression

Our Approach: Proposal Generation

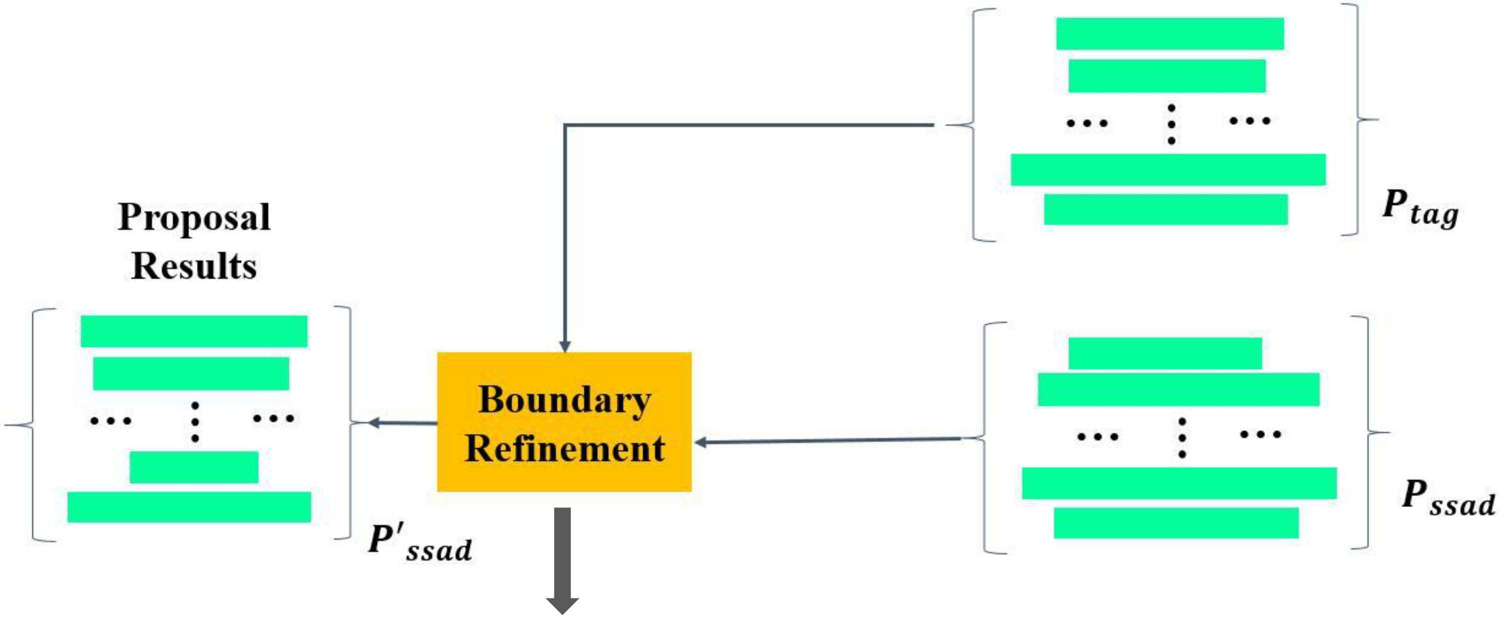
TAG method





Temporal Action Detection with Structured Segment Networks.
Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Dahua Lin, Xiaoou Tang. arXiv 1704.06228

Our Approach: Boundary Refinement



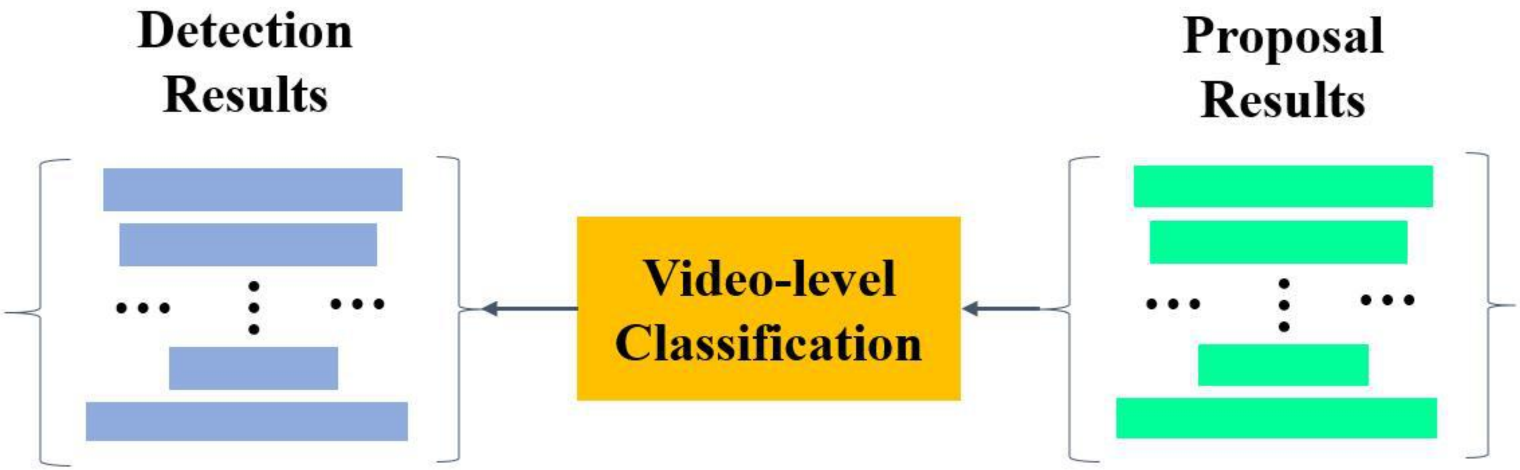
Algorithm 1 Boundary Refinement

Input: proposals generated by Prop-SSAD: $\boldsymbol{P}_{ssad}$;
proposals generated by TAG: $\boldsymbol{P}_{tag}$

Output: refined proposals: $\boldsymbol{P}'_{ssad}$

```
for p_t in P_tag do
    calculate IoU between p_t and all proposals in P_ssad
    if maximum IoU ≥ 0.75 then
        replace the boundaries of corresponding proposal p_s in P_ssad with boundaries of p_t
return P_ssad
```

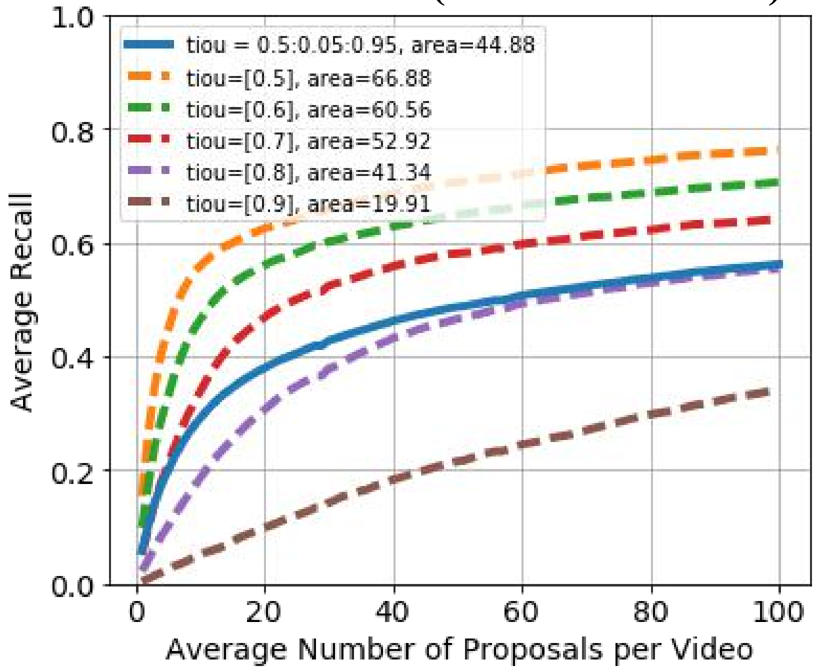
Our Approach: Action localization



Evaluation Metric: Temporal Action Proposal

- **Evaluation metric** is the area under the **Average Recall** vs. **Average Number of Proposals per Video (AR-AN)** curve.
- **AR** is defined as the mean of all recall values using tIoU thresholds between 0.5 and 0.95 (inclusive) with a step size of 0.05.
- **AN** is defined as the total number of proposals divided by the number of videos in the testing subset.

Baseline Method (Uniform Random)



Experiment: Temporal Action Proposal

Table 1: Proposal Results on validation set of ActivityNet.

Method	AR@10	AR@100	AR-AN
Uniform Random (baseline)	29.02	55.71	44.88
Prop-SSAD	50.44	69.54	61.52
Refined Prop-SSAD	52.50	73.01	64.40

Temporal Action Proposals **(testing set)**

Ranking	Username	Organization	Upload time	AUC
1	Tianwei Lin	Shanghai Jiao Tong University & Columbia University	2017-07-17 08:41:23	64.8084
2	Ting Yao	Multimedia Search and Mining Group, MSRA	2017-07-17 08:13:43	64.1807
3	TCN Dai	UMD	2017-07-16 09:22:25	61.5584
4	Cong Guo	University of Science and Technology of China	2017-06-22 19:17:23	58.804
5	Huijuan Xu	Boston University	2017-07-15 14:58:47	54.6138

Experiment: Temporal Action Proposal

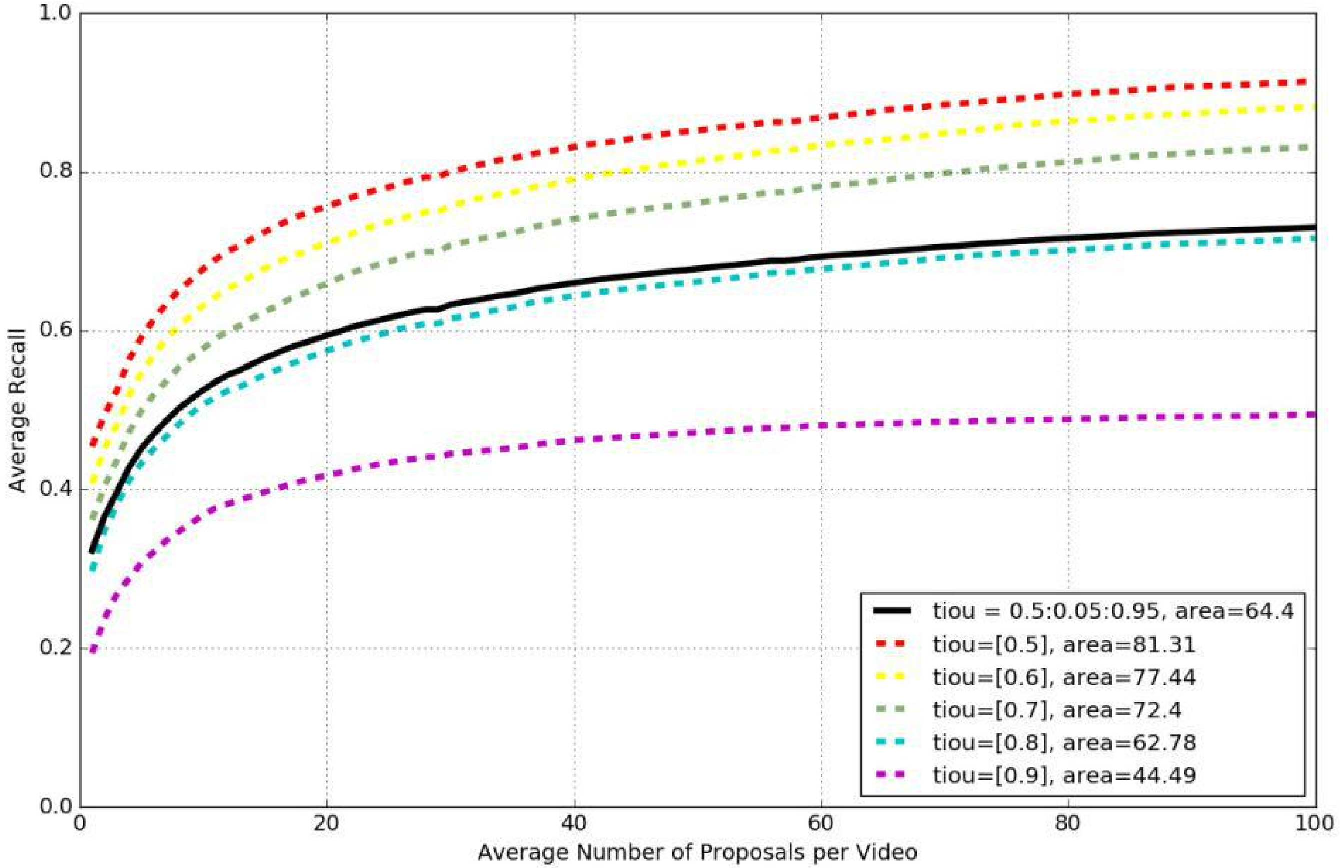


Figure 2: AR-AN curve of our proposal results in validation set. The area under black curve is the AR-AN score.

Evaluation Metric: Temporal Action Localization

- **Evaluation metric** is the **average mAP**.
- **mAP** is the mean AP over all the activity categories.
- **Average mAP** is the average of all mAP values computed with tIoU thresholds between 0.5 and 0.95 with a step size of 0.05.

Experiment: Temporal Action Localization

Table 2: Action localization results on validation set. Results are evaluated by mAP with different IoU thresholds α and average mAP of IoU thresholds from 0.5 to 0.95. Ours@n means first n proposals used for localization.

mAP	0.5	0.75	0.95	Average mAP
Wang et al. [13]	42.28	3.76	0.05	14.85
Shou et al. [10]	43.83	25.88	0.21	22.77
Xiong et. al. [15]	39.12	23.48	5.49	23.98
Ours@1	42.14	27.17	6.54	27.00
Ours@5	46.56	30.94	7.53	30.49
Ours@10	47.84	31.90	7.76	31.41
Ours@25	48.56	32.53	7.83	31.93
Ours@100	48.99	32.91	7.87	32.26

Experiment: Temporal Action Localization

Table 3: Action localization results on testing set. Only average mAP is provided in evaluation server, which is calculated with IoU thresholds from 0.5 to 0.95.

Method	Average mAP
Wang et. al. [13]	14.62
Xiong et. al. [15]	26.05
Zhao et. al. [16]	28.28
Ours result	33.40

Temporal Action Localization **(testing set)**

Ranking	Username	Organization	Upload time	Avg. mAP
1	Tianwei Lin	Shanghai Jiao Tong University & Columbia University	2017-07-17 09:32:21	0.33406
2	Yuanjun Xiong	CUHK	2017-07-17 09:08:37	0.31863
3	Yuxiang Zhou	IC	2017-07-17 10:08:08	0.31827
4	Yiming Lin	Imperial College London	2017-07-17 02:39:16	0.31761
5	TCN Dai	UMD	2017-06-30 16:53:32	0.23674

Take-home Message

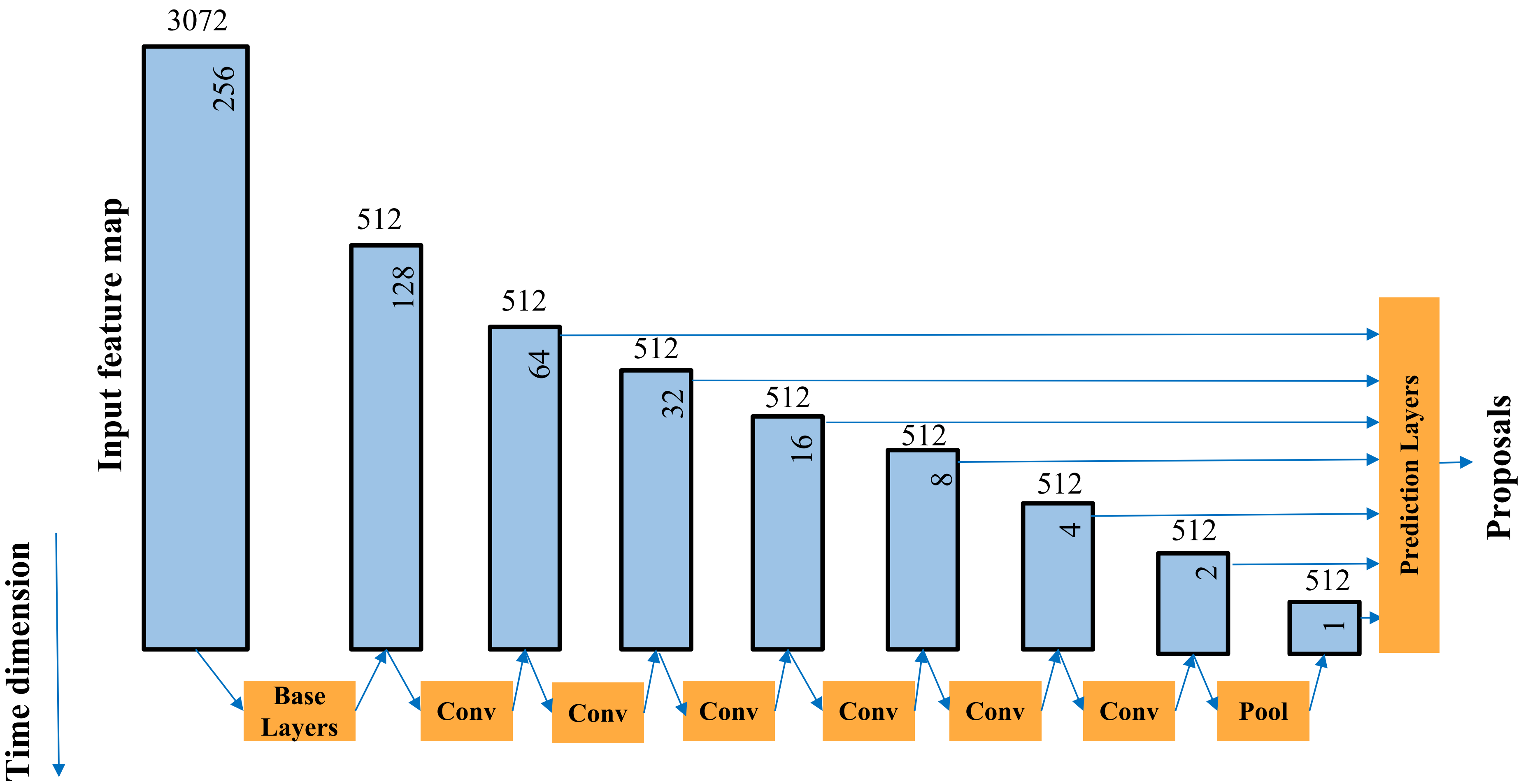
- Proposal is a very important for accurate localization
- Anchor mechanisms and temporal convolution can work well in temporal action proposal/localization task

Thank you!

More details in:

- Paper: https://arxiv.org/abs/1707.06750
- Homepage: https://wzmsltw.github.io
- E-mail: wzmsltw@sjtu.edu.cn

Appendix: Network Architecture of Prop-SSAD



Appendix: Model Training

Training of Prop-SSAD

- Loss function: L1 loss for IoU regression
- Training data: training set of ActivityNet dataset
- Training data proportion: [$IoU > 0.7$]: [$0.7 \geq IoU > 0.3$]: [$IoU \leq 0.3$]= 1:1:2
- Batch size: 16
- Learning rate: 0.0001
- Epoch: 10

Training of MLP in TAG

- Loss function: 2-class Softmax loss
- Training data: training set of ActivityNet dataset
- Training data proportion: action: not-action= 1:1
- Batch size: 16
- Learning rate: 0.001
- Epoch: 10