Weakly-supervised Action Localization with Background Modeling

ICCV2019 Submission #6496

Paper Summary

What are in this summary?

1. Motivations
2. High-level ideas of our architecture
3. Quantitative results summaries on THUMOS14
    a. Ours vs STPN
    b. Ours vs weakly-supervised methods
    c. Ours vs fully-supervised methods
4. Micro-videos as supplement training data
5. Visualizations
    a. Highly-confident detected action instances
    b. Failure modes.

Action localization + Weak supervision

“This video contains Gymnastics action.”

<u>Problem</u>: find the temporal locations of the action within an untrimmed video using only video-level labels.

This formulation is attractive because precise boundaries are difficult to obtain.

Background Gymnastics Background

Full supervision - exact boundaries are known

Background modeling is straightforward

“This video contains Gymnastics action.”

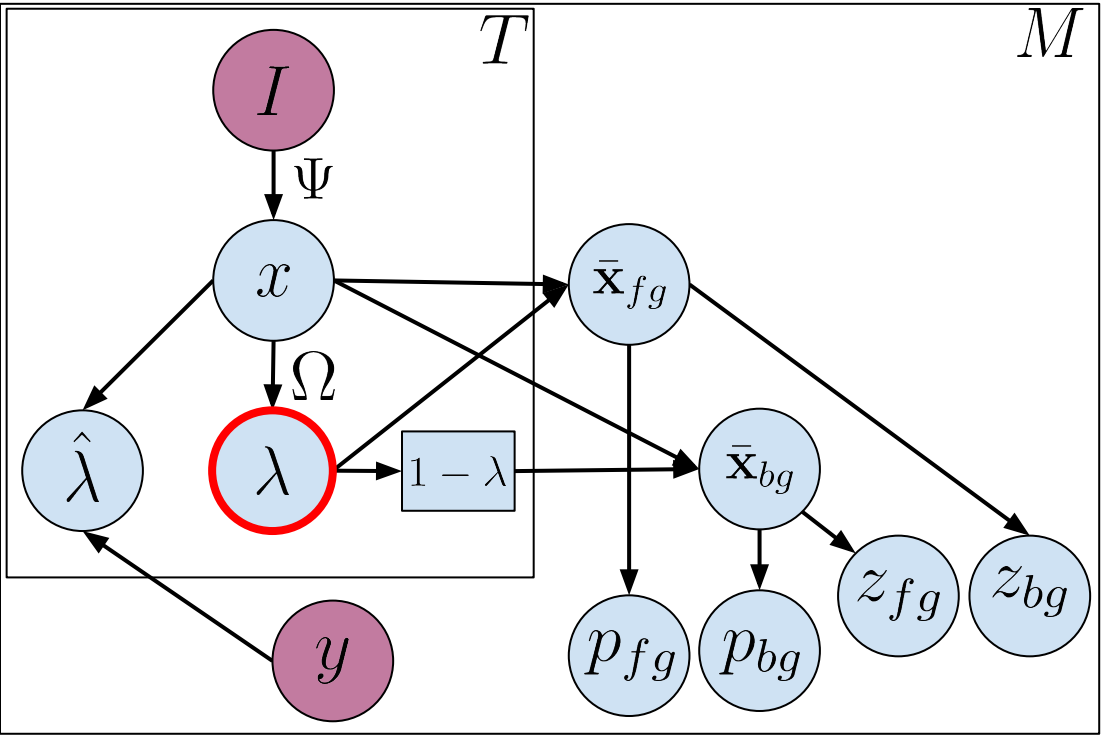
???background???

Weak supervision - boundaries are unknown

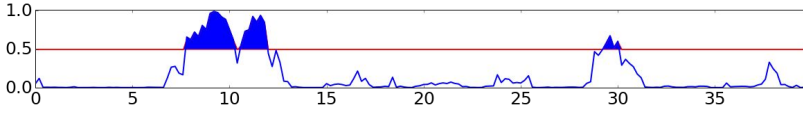
Background frames are often unmodeled.

In this paper, we show that models which explicitly accounts for background frames are substantially better.

Attention is key ingredient.

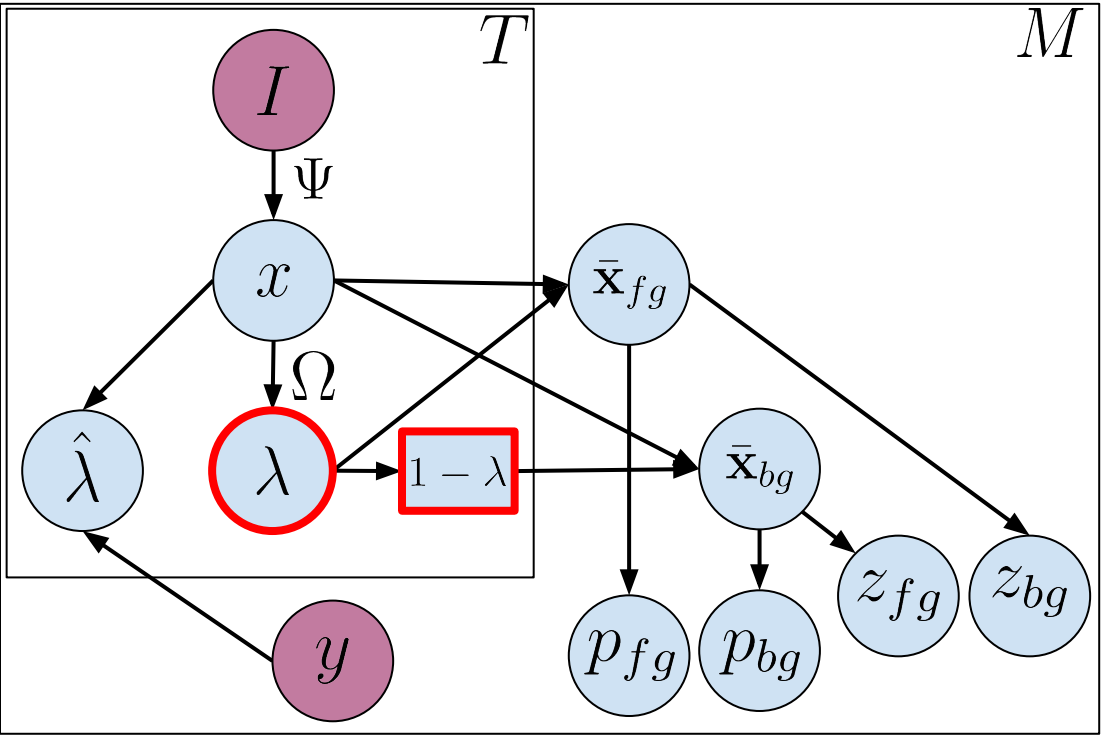


λ_t

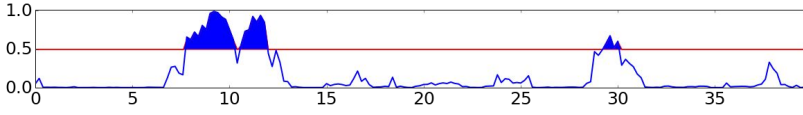


how likely of an action at timestamp t

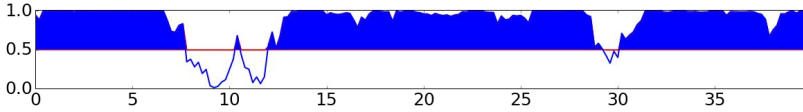
Attention is key ingredient.



λ_t



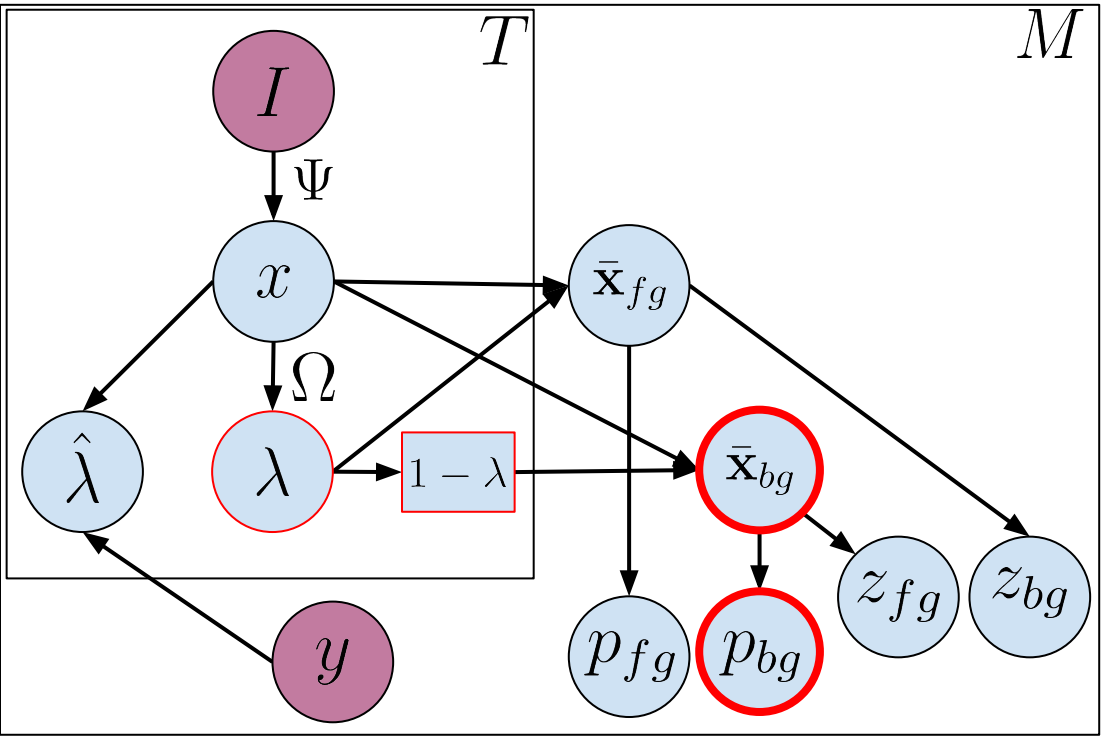
how likely of an action at timestamp t



$1 - \lambda_t$

probability of a non-event (background) at timestamp t

Background-awareness loss



$$\mathbf{x}_{\text{bg}} = \frac{1}{T}\sum_{t=1}^{T}(1-\lambda_t)\mathbf{x}_t$$

features for background concept

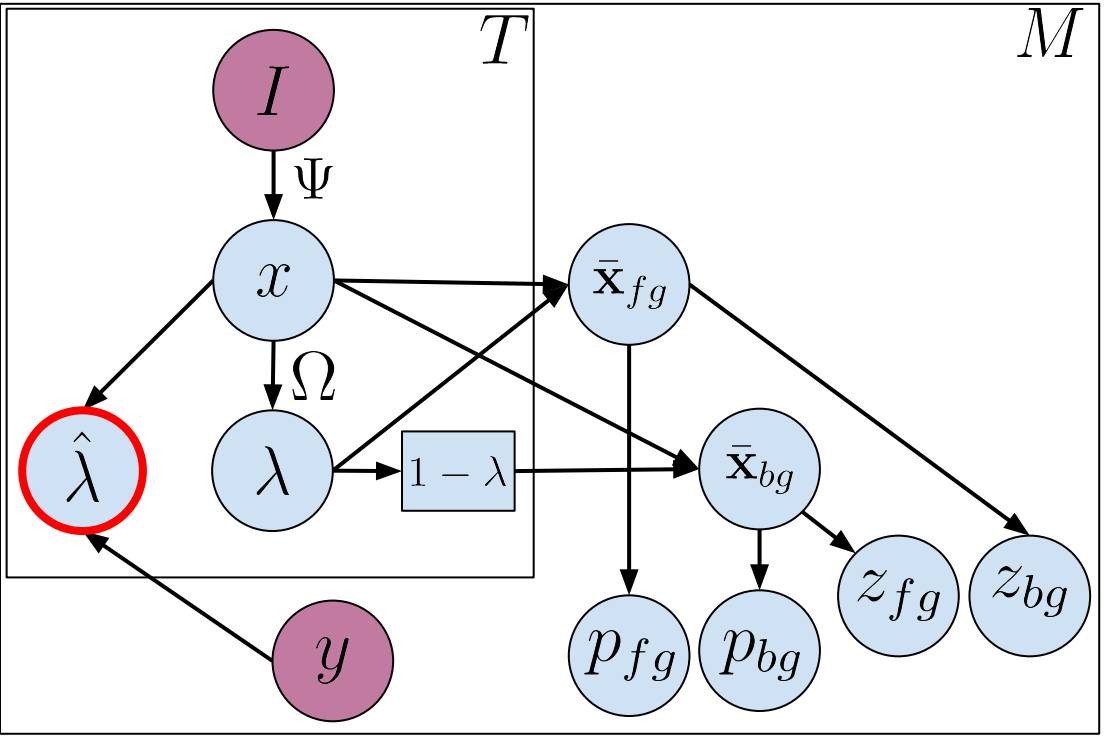
$$p_{\text{bg}}[c] = \frac{e^{w_c \cdot \mathbf{x}_{\text{bg}}}}{\sum_{i=0}^{C} e^{w_i \cdot \mathbf{x}_{\text{bg}}}}$$

softmax

$$\mathcal{L}_{\text{bg}} = -\log p_{\text{bg}}[0]$$

Background-awareness loss

Top-down attentional cues from T-CAM



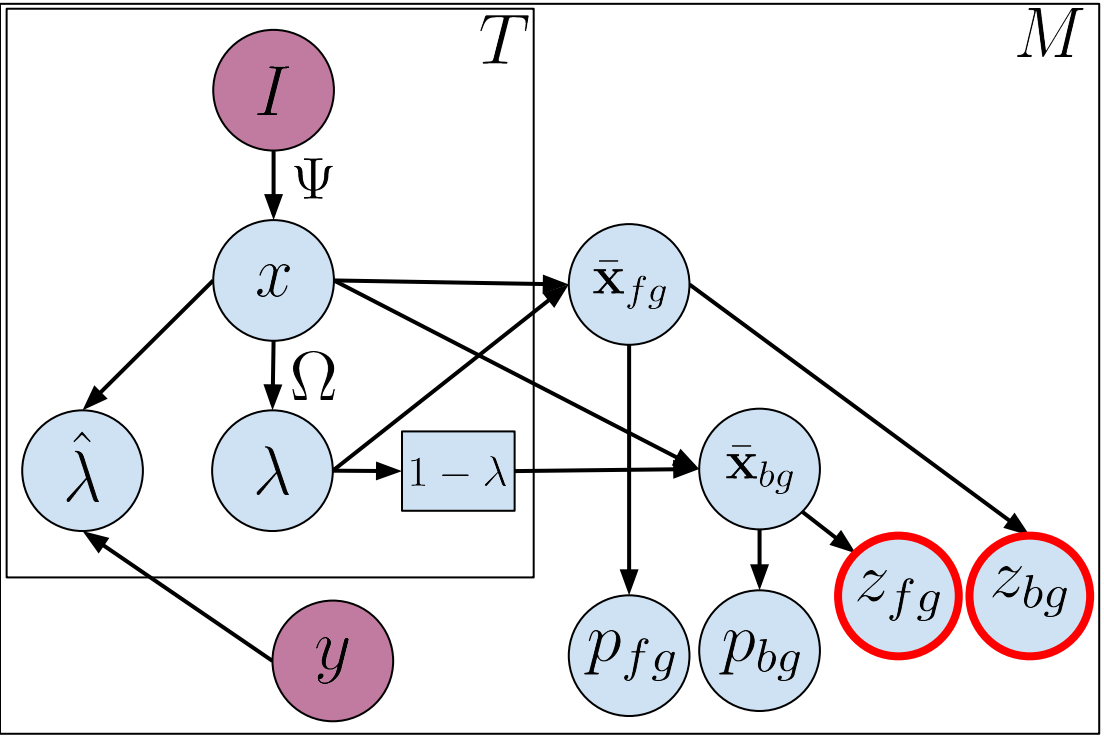
$$\hat{\lambda}_t^{\text{fg}} = G(\sigma) * \frac{e^{w_y \mathbf{x}_t}}{\sum_{i=0}^{C} e^{w_i \mathbf{x}_t}}$$

Smoothed Temporal Class Activation Mapping (T-CAM)

$$\mathcal{L}_{\text{guide}} = \frac{1}{T} \sum_t |\lambda_t - \hat{\lambda}_t^{\text{fg}}|$$

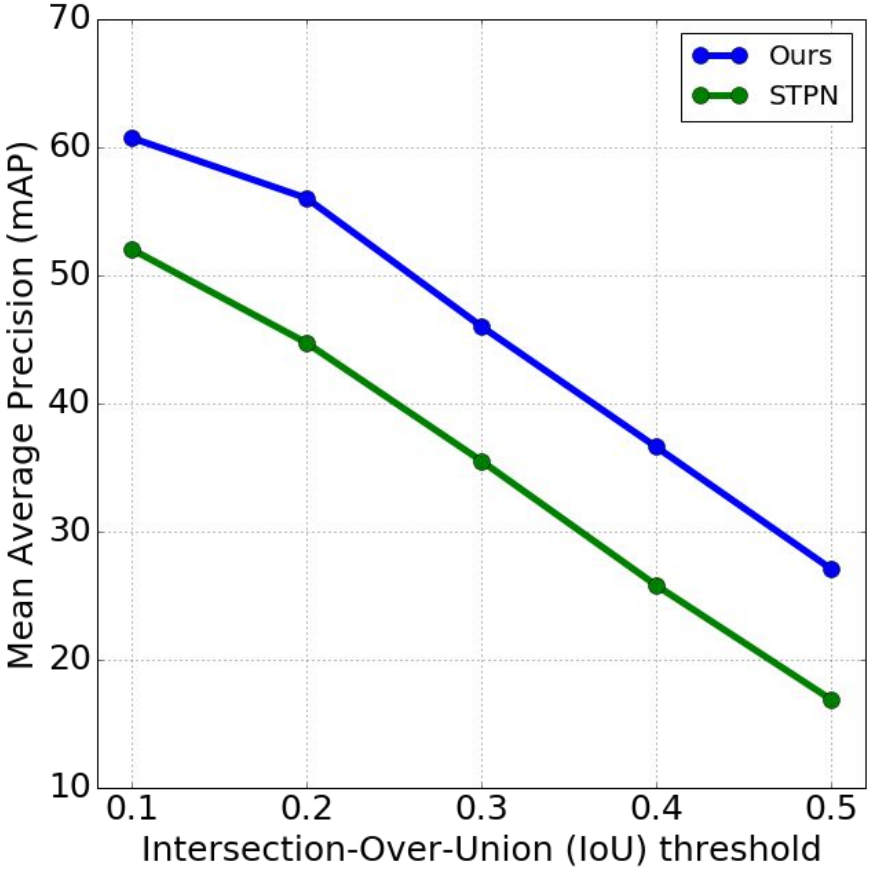
Self-guided Attention Loss

Foreground-Background Clustering Loss



- **Unsupervised loss**
- **Encourages foreground and background pooled features to be distinct.**

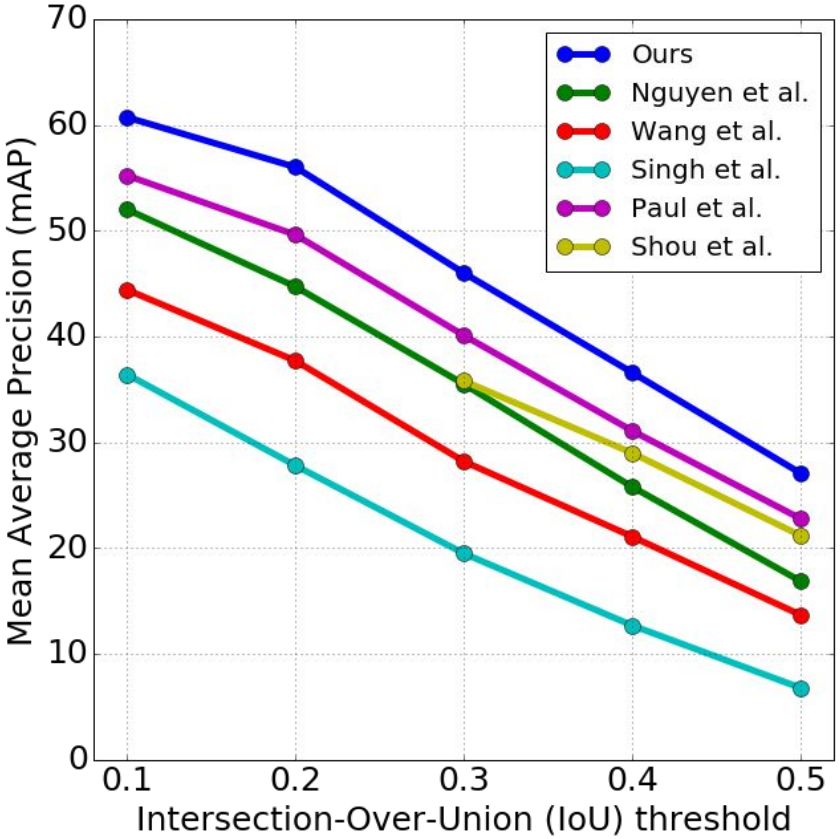
mAP@IoU - THUMOS14



The background modeling extension gives ~10% mAP improvement across all IoU thresholds

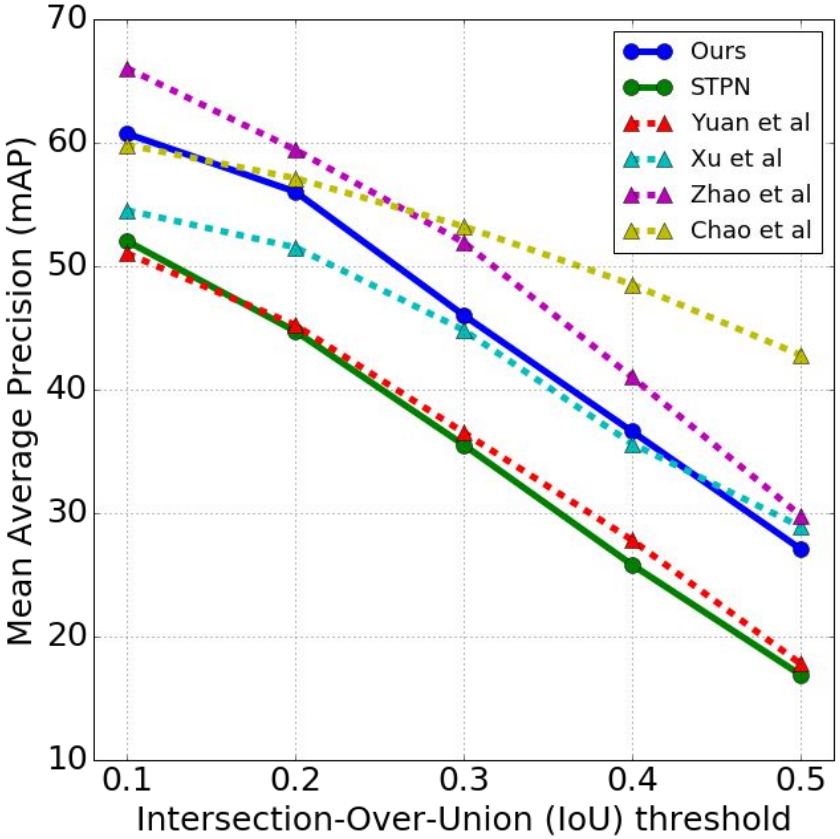
*higher is better

Compared with weakly supervised SoTA



We outperforms state-of-the-art for weakly supervised action localization.

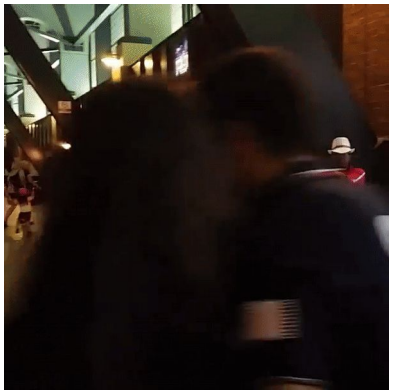
Compared with fully supervised SoTA



- Our methods are comparable with other fully supervised methods (dashline).
  - Especially at lower IoUs.
- Higher IoUs requires more accurate action boundary decisions.
  - which is difficult to do without actual boundary supervision.

Micro-videos as supplement training data

#hammerthrow



#baseballpitch

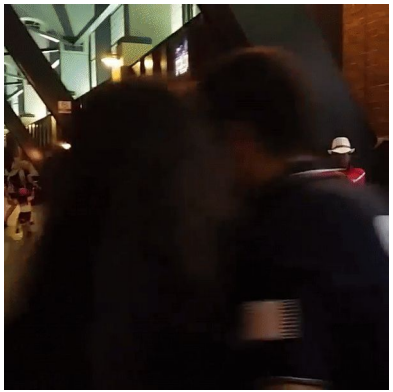
#basketballdunk

#javelinthrow

At the first glance, ideal source for weakly-superivsed training data.

Micro-videos as supplement training data

#hammerthrow



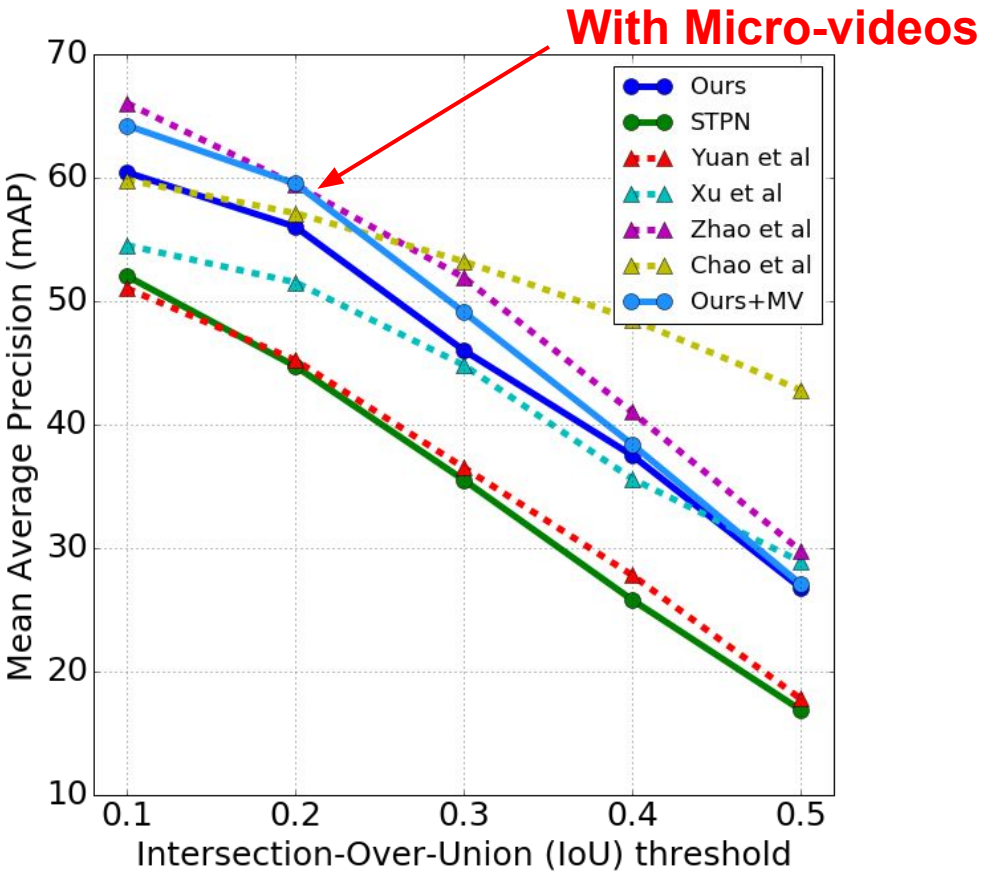
#baseballpitch

#basketballdunk

#javelinthrow

Would adding microvideos improve localization performance?

With Microvideos



- The addition of microvideos:
  - helps improve significantly at lower IoUs
  - Better recognize the action instance.
- It doesn't help with refining the boundaries of action instances
  - which is needed for better performance on higher IoUs.

High-Confident Detected Instances

BaseballPitch

BasetballDunk

Billiards

High-Confident Detected Instances

CliffDiving

GolfSwing

HammerThrow

Failure Modes (1) - Mini-action composition

Groundtruths

CleanAndJerk

Detections

Clean

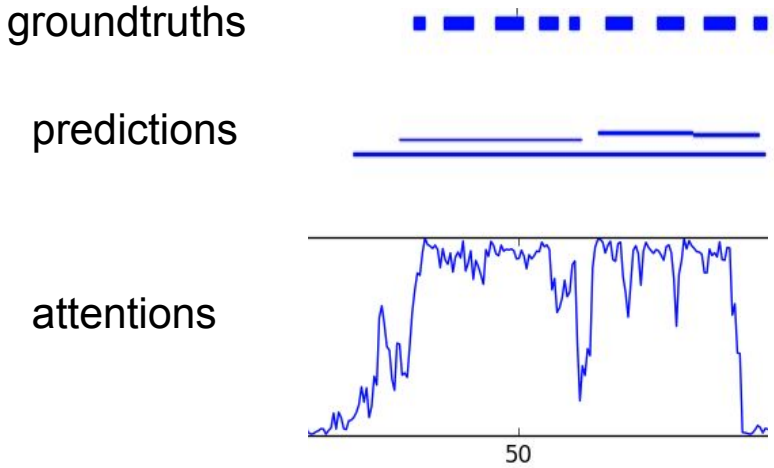
Background

Jerk

- ‘CleanAndJerk’ is composed of two mini-actions, ‘Clean’ and ‘Jerk’.
- The athlete often pauses in between, resembling background frames.
- Our model over-segments this single instance into two separate instances.

Failure Modes (2) - Quickly Repeated actions

Detections



- Each ‘TennisSwing’ instance is followed quickly by another ‘TennisSwing’ instance.
- There are little ‘background’ in between these swings.
- Our model often outputs one **large** detection containing these instances.

Failure Modes (3) - Human-agreed Boundaries

detection
(at regular speed)

groundtruth
(at 10% speed)

- ‘BasketballDunk’ action instances consists of
  - Run to the basket -> Jump -> Dunk the ball.
- Our model outputs the whole sequence as ‘BasketballDunk’ (left).
- Groundtruth segments only considers the last part as foreground (right).
- Intersection-Over-Union (IoU) < 10%
  - considered false positive.

Thanks!

Happy Reviewings :D!