# End-to-end Learning of Action Detection from Frame Glimpses in Videos

CVPR 2016

Serena Yeung, Olga Russakovsky, Greg Mori , Li Fei-Fei

Presenter: Wei-Jen Ko

# Action detection

- Predict which and when action occurs in the video.

# Related Work

Motion features: Dense Trajectories

Apearance features: CNN+SIFT+ COLOR

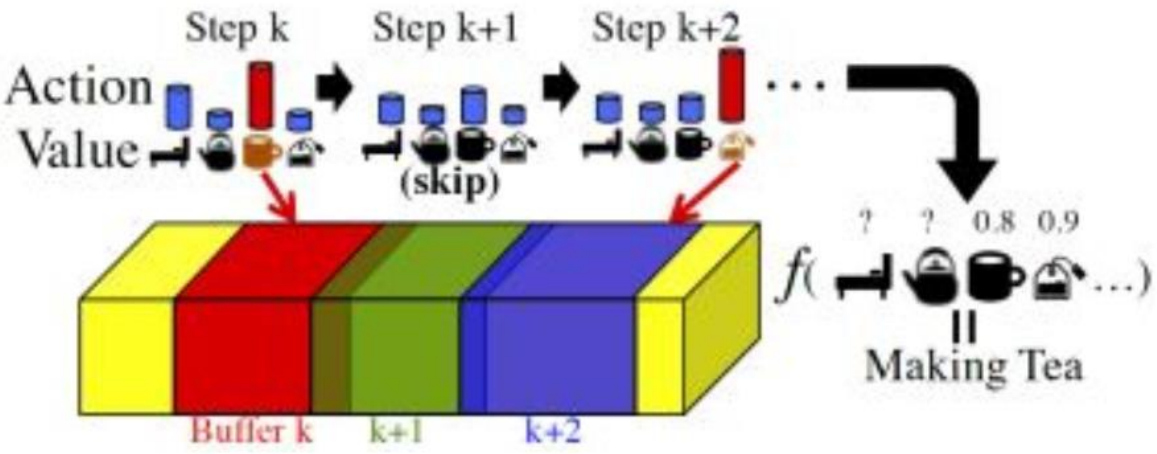
Audio features: MFCC+ASR

Classified by SVM over exhaustive segments with varying scale and temporal position.


D. Oneata, J. Verbeek, and C. Schmid. The lear submission at thumos 2014.
L. Wang, Y. Qiao, and X. Tang. Action recognition and detection by combining motion and appearance features
J. Yuan, Y. Pei, B. Ni, P. Moulin, and A. Kassim. Adsc submission at thumos challenge 2015

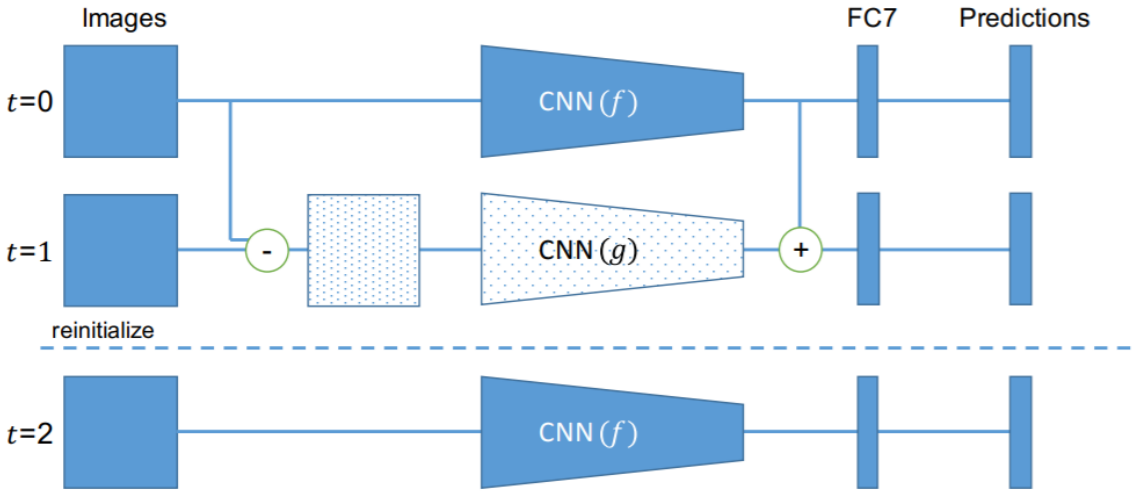
# Related Work

## Dynamic feature prioritization




Y-C. Su and K. Grauman. Leaving Some Stones Unturned: Dynamic Feature Prioritization for Activity Detection in Streaming Video , ECCV 2016.

## Predictive corrective networks




A. Dave, O. Russakovsky, D. Ramanan. *Predictive-Corrective Networks for Action Detection,* CVPR 2017.

# Proposed method

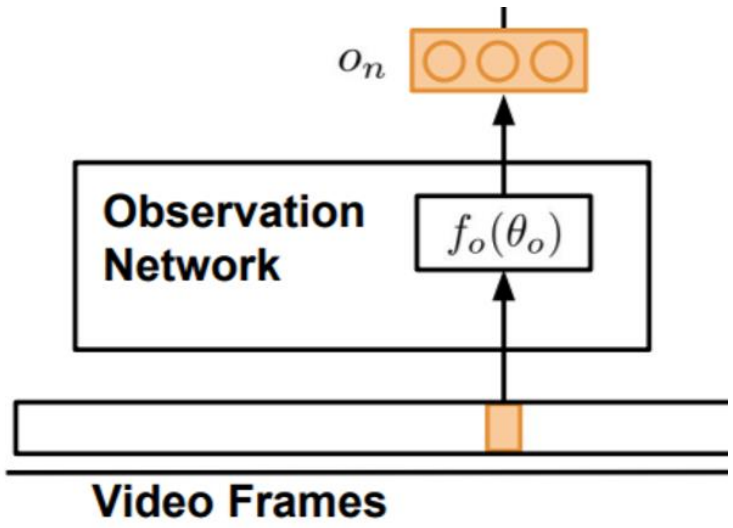
- Recurrent neural network-based end-to-end model

- Decides which frame to observe next and when to emit a prediction.

# Observation Network

Video frame $\boldsymbol{V_{In}}$ → VGG → Image fearure

Frame location $\boldsymbol{In}$ → FC → $\boldsymbol{On}$



# Recurrent Network

$s_n$ : start location of the action

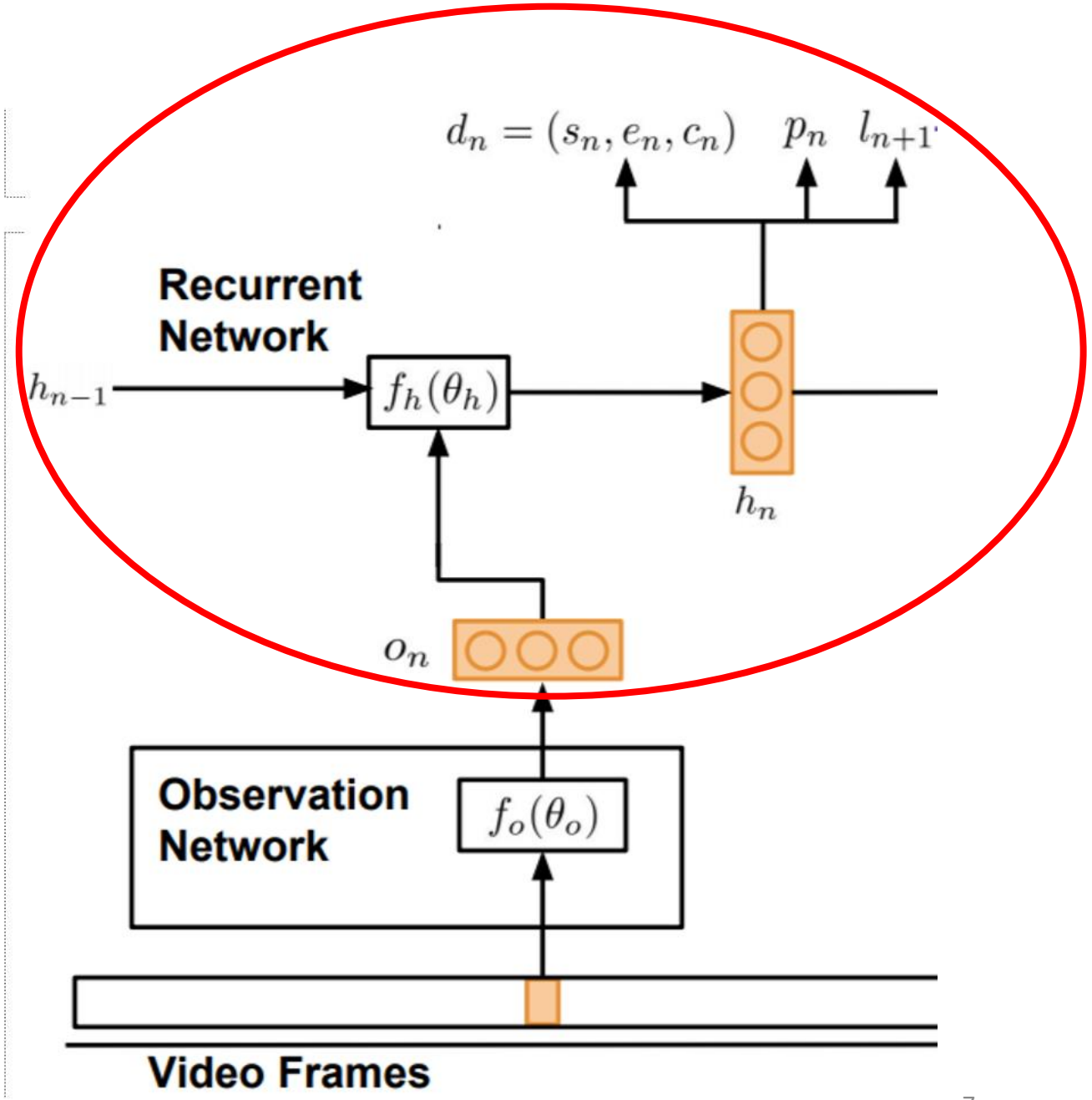
$e_n$ : end location of the action

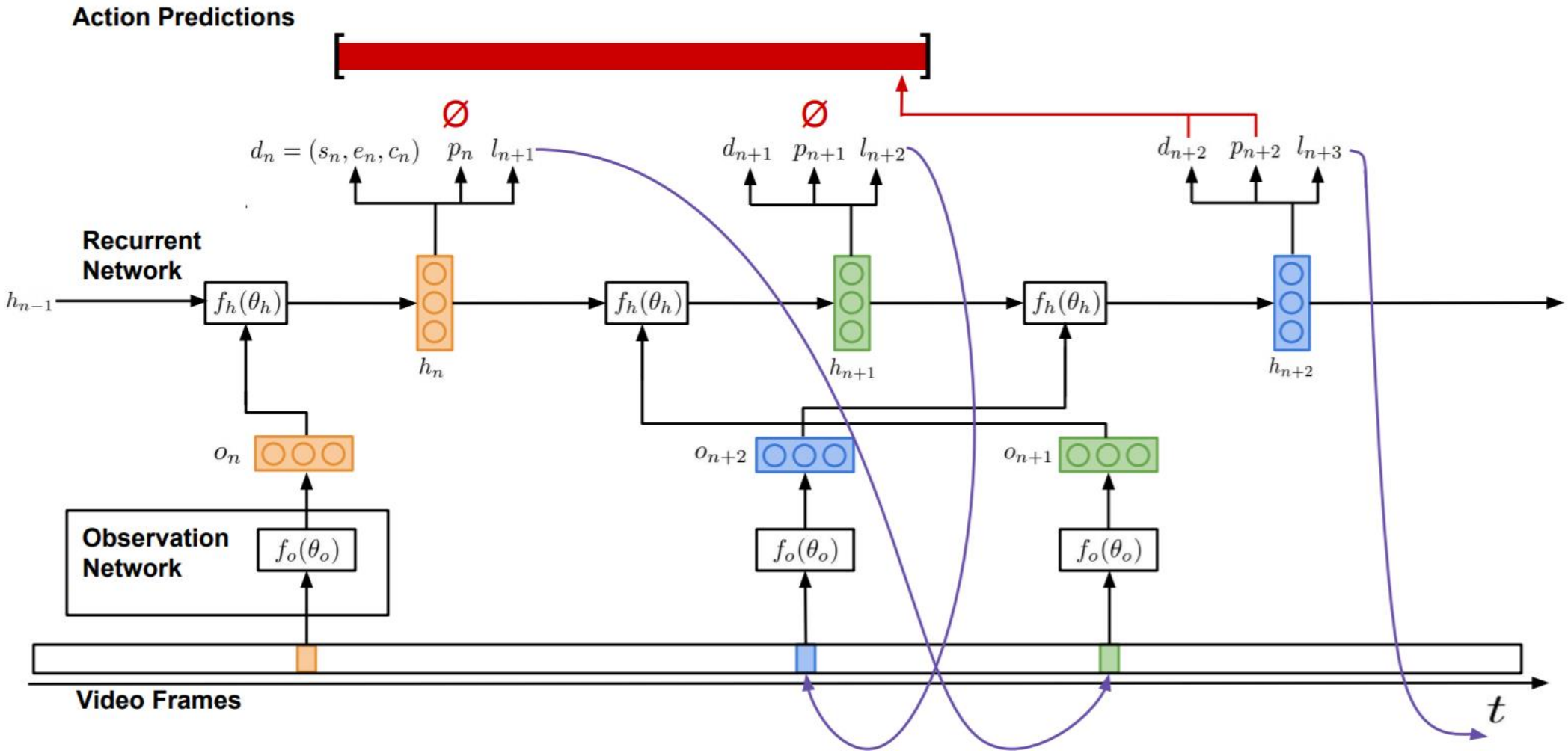
$l_{n+1}$: location of the video frame to observe next

$c_n$ : confidence level of the prediction

$P_n$ : prediction indicator

$S_n$, $e_n$, $l_{n+1}$ normalized to [0,1]




Action Predictions
∅
∅
$d_n = (s_n, e_n, c_n)$ $p_n$ $l_{n+1}$
$d_{n+1}$ $p_{n+1}$ $l_{n+2}$
$d_{n+2}$ $p_{n+2}$ $l_{n+3}$
Recurrent Network
$h_{n-1}$
$f_h(\theta_h)$
$f_h(\theta_h)$
$f_h(\theta_h)$
$h_n$
$h_{n+1}$
$h_{n+2}$
$o_n$
$o_{n+2}$
$o_{n+1}$
Observation Network
$f_o(\theta_o)$
$f_o(\theta_o)$
$f_o(\theta_o)$
Video Frames
$t$

# Loss function

$$L(D) = \sum_n L_{cls}(d_n) + \gamma \sum_n \sum_m \mathbb{1}[y_{nm} = 1] L_{loc}(d_n, g_m)$$

$L_{cls}(d_n)$: Cross-entropy loss on confidence ***Cn***

$L_{loc}(d_n, g_m)$ : L2- regression loss minimizing the distance

$$\|(s_n, e_n) - (s_m, e_m)\|$$

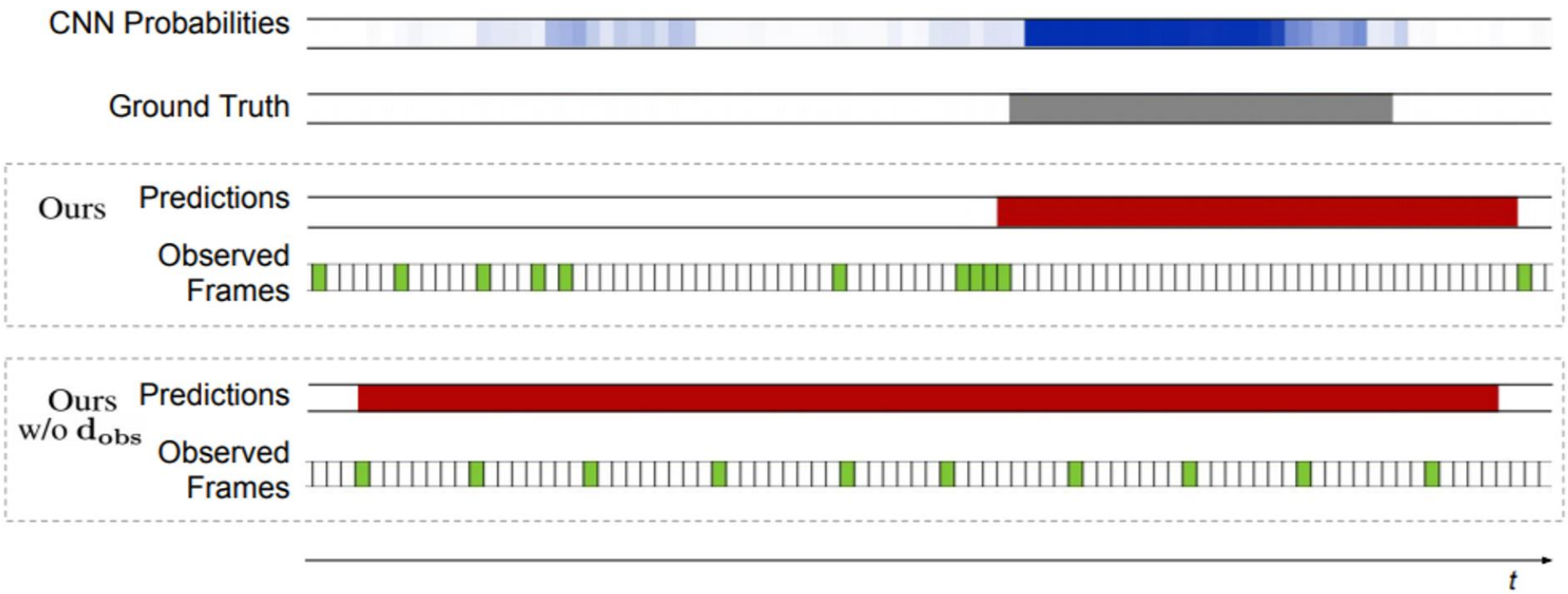
# $p_n$ and $l_{n+1}$ trained by REINFORCE

Reward function

$$r_N = \begin{cases} R_p & \text{negative reward if did not emit predictions for videos containg instances} \\ N_+ R_+ + N_- R_- & \text{otherwise} \end{cases}$$

# THUMOS'14 Results

| | $\alpha$=0.5 | $\alpha$=0.4 | $\alpha$=0.3 | $\alpha$=0.2 | $\alpha$=0.1 |
|---|---|---|---|---|---|
| Karaman et al. [13] | 0.9 | 1.4 | 2.1 | 3.4 | 4.6 |
| Wang et al. [39] | 8.3 | 11.7 | 14.0 | 17.0 | 18.2 |
| Oneata et al. [22] | 14.4 | 20.8 | 27.0 | 33.6 | 36.6 |
| **Ours (full)** | **17.1** | **26.4** | **36.0** | **44.0** | **48.9** |
| Ablation Experiments | | | | | |
| Ours w/o $d_{pred}$ | 12.4 | 19.3 | 26.0 | 32.5 | 37.0 |
| Ours w/o $d_{obs}$ | 9.3 | 15.2 | 20.6 | 26.5 | 31.2 |
| Ours w/o $d_{obs}$ w/o $d_{pred}$ | 8.6 | 14.6 | 20.0 | 27.1 | 33.3 |
| Ours w/o $loc$ | 5.5 | 9.9 | 16.2 | 22.7 | 27.5 |
| CNN with NMS | 6.4 | 9.6 | 12.8 | 16.7 | 18.5 |
| LSTM with NMS | 5.6 | 7.8 | 10.3 | 13.9 | 15.7 |




If observed frames are not be determined dynamically, it does not provide sufficient resolution to localize action boundaries.

# ActivityNet Results

| | [3] | Ours | | [3] | Ours |
|---|---|---|---|---|---|
| Archery | **34.7** | 5.2 | Long Jump | 41.1 | **56.8** |
| Bowling | 51.3 | **52.2** | Mountain Climb. | 31.0 | **53.0** |
| Bungee | 42.6 | **48.9** | Paintball | **31.2** | 12.5 |
| Cricket | 27.9 | **38.4** | Playing Kickball | 33.8 | **60.8** |
| Curling | 16.4 | **30.1** | Playing Volley. | 32.1 | **40.2** |
| Discus Throw | **26.2** | 17.6 | Pole Vault | **47.7** | 35.5 |
| Dodgeball | 26.6 | **61.3** | Shot put | 29.4 | **50.9** |
| Doing Moto. | 30.2 | **46.2** | Skateboarding | 21.3 | **34.4** |
| Ham. Throw | **22.2** | 13.7 | Start Fire | 25.3 | **38.4** |
| High Jump | **41.3** | 21.9 | Triple Jump | **36.4** | 16.1 |
| Javelin Throw | **48.1** | 35.7 | | | |
| **mAP** | | | | 33.2 | **36.7** |

Table 3: Per-class breakdown and mAP on the ActivityNet Sports subset, at IOU of $\alpha = 0.5$.

| | [3] | Ours | | [3] | Ours |
|---|---|---|---|---|---|
| Attend Conf. | 28.3 | **56.5** | Phoning | 34.7 | **52.1** |
| Search Security | 24.5 | **33.9** | Pumping Gas | **54.7** | 34.0 |
| Buy Fast Food | 34.4 | **45.8** | Setup Comp. | **37.4** | 30.3 |
| Clean Laptop Fan | 26.0 | **35.8** | Sharp. Knife | **36.3** | 35.2 |
| Making Copies | 18.2 | **41.7** | Sort Books | **29.3** | 16.7 |
| Organizing Boxes | **29.6** | 19.1 | Using Comp. | 37.4 | **50.2** |
| Organiz. Cabin. | 19.0 | **43.7** | Using ATM | 29.5 | **64.9** |
| Packing | 28.0 | **39.1** | | | |
| **mAP** | | | | 31.1 | **39.9** |

Table 4: Per-class breakdown and mAP on the ActivityNet Work subset, at IOU of $\alpha = 0.5$.

Strengths:

- First End-to-end training approach
- Select important frames to observe, no exhaustive searching
- Better results


CNN Probabilities
Ground Truth
Predictions
Observation Sequence
Frame
51
66
78
89
94 95
101
CNN Probabilities
Ground Truth
Predictions
Observation Sequence
Frame
51
62
77
82
92 93
101 104
90 SECS
GREG LOUGANIS
DOUBLE GOLD IN LOS ANGELES 1984
3M SPRINGBOARD