COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

DVMM lab
digital video | multimedia laboratory

Video Event Localization, Supervision, and Online Detection

Shih-Fu Chang

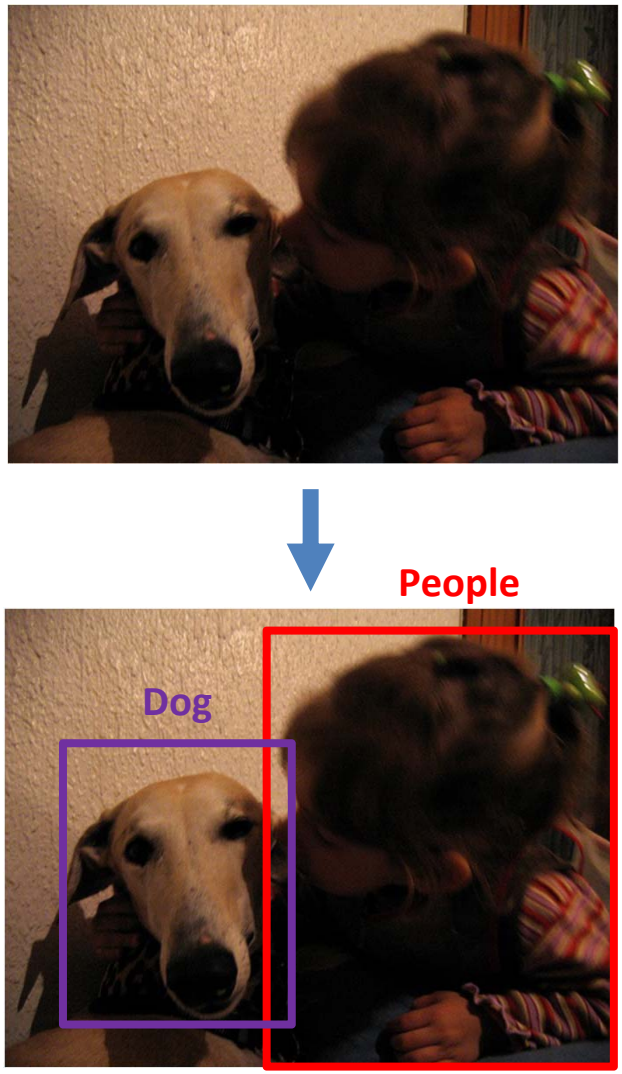
Columbia University
www.ee.columbia.edu/dvmm

6/22/2018

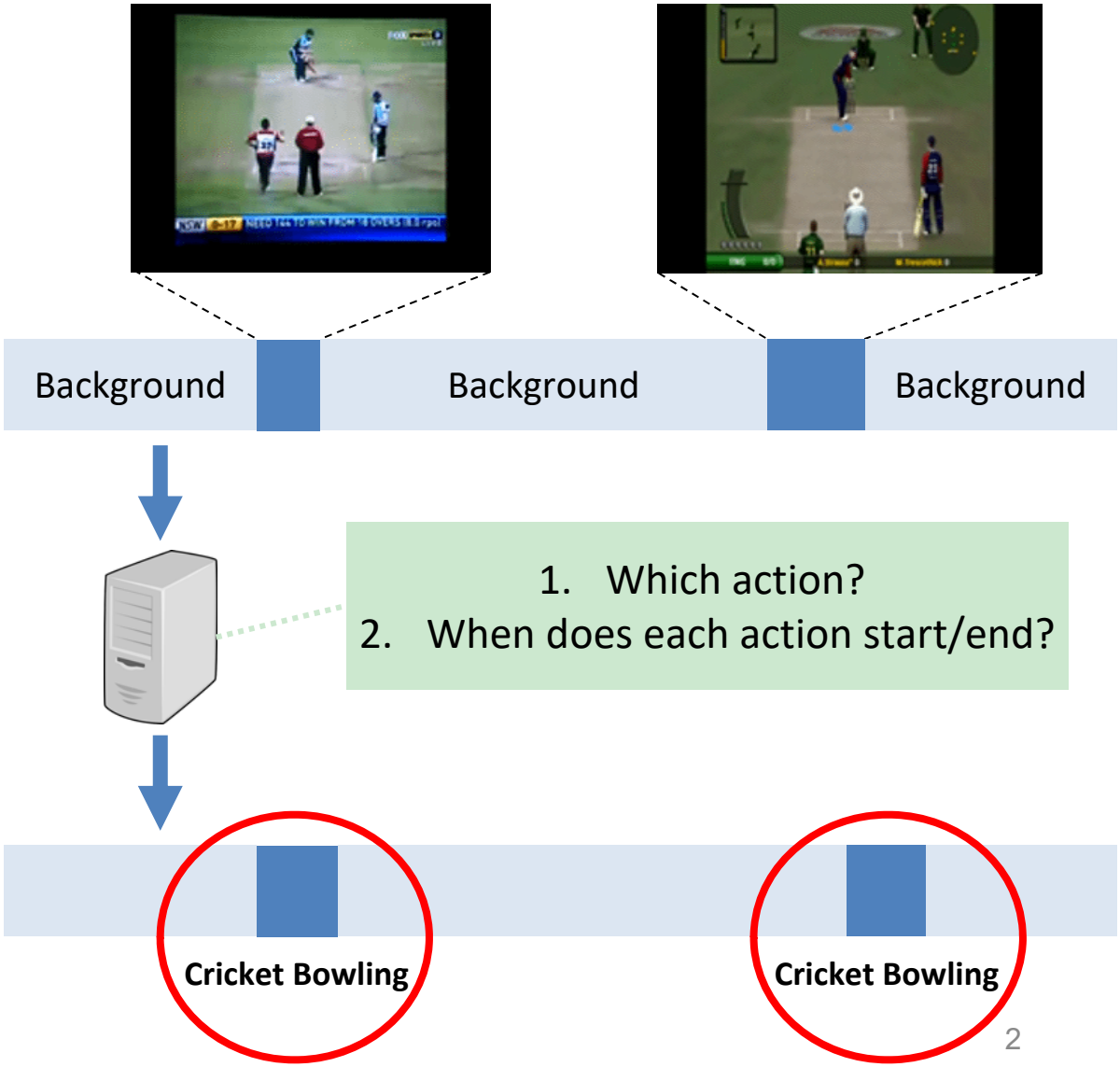
Joint work with
Zheng Shou, Víctor Campos, and Xavier Giro-i-Nieto

Localization: Analogy to Object Detection

Image: Object Detection

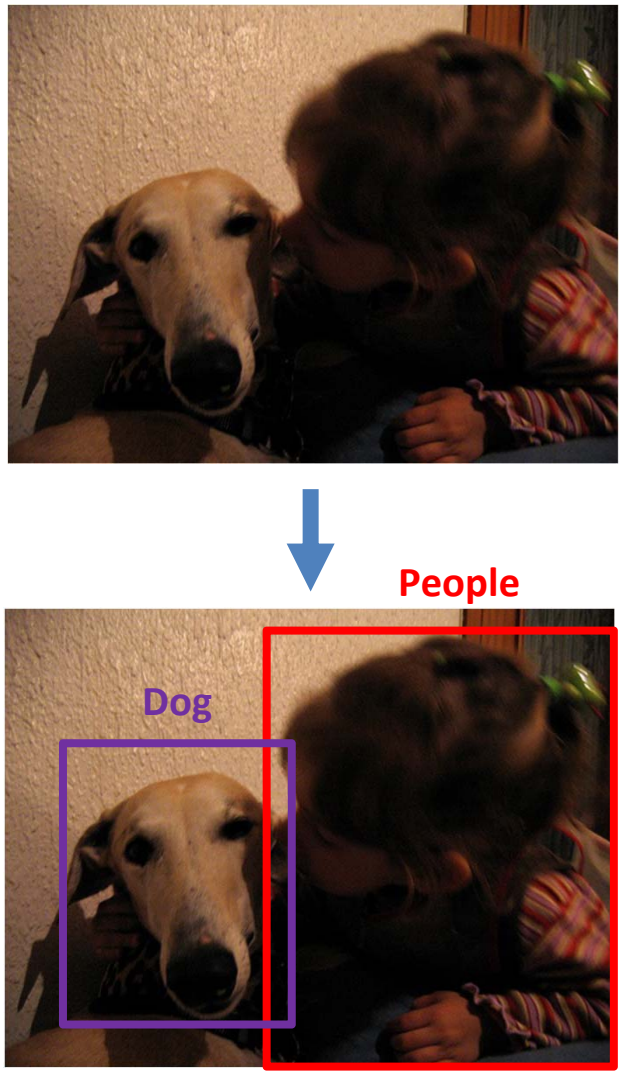


Video: Temporal Localization

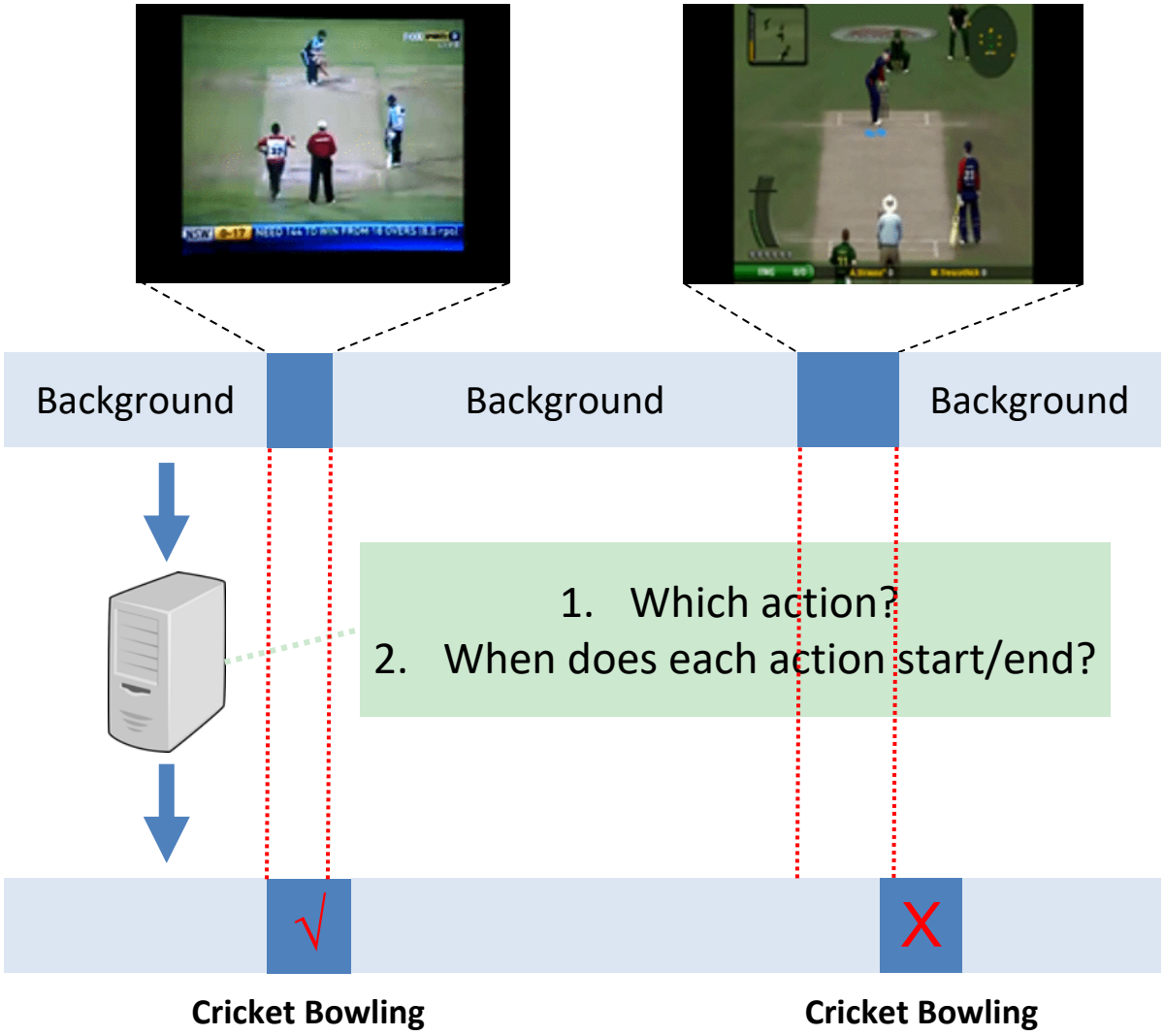


Goal: Temporal Precision

Image: Object Detection

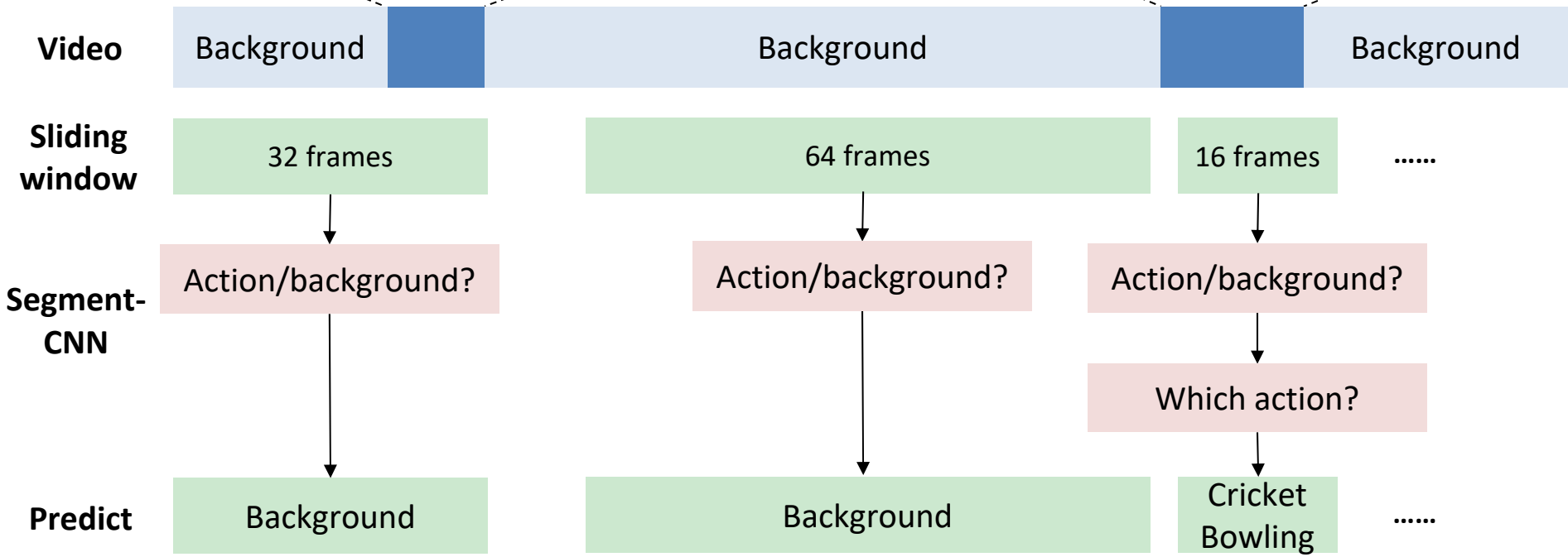


Video: Temporal Localization



Precision measured by IOU

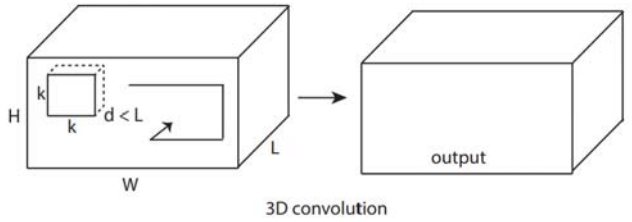
Baseline Approach: Segment-CNN



Temporal Action Localization in Untrimmed Videos via Multi-stage CNNs.
Zheng Shou, Dongang Wang, and Shih-Fu Chang. In CVPR'16.

Backbone Network

- Segment-CNN used C3D (Tran et al., ICCV'15):



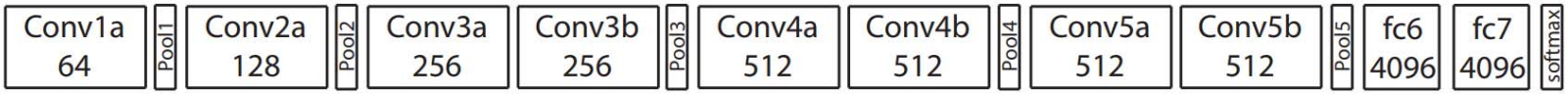


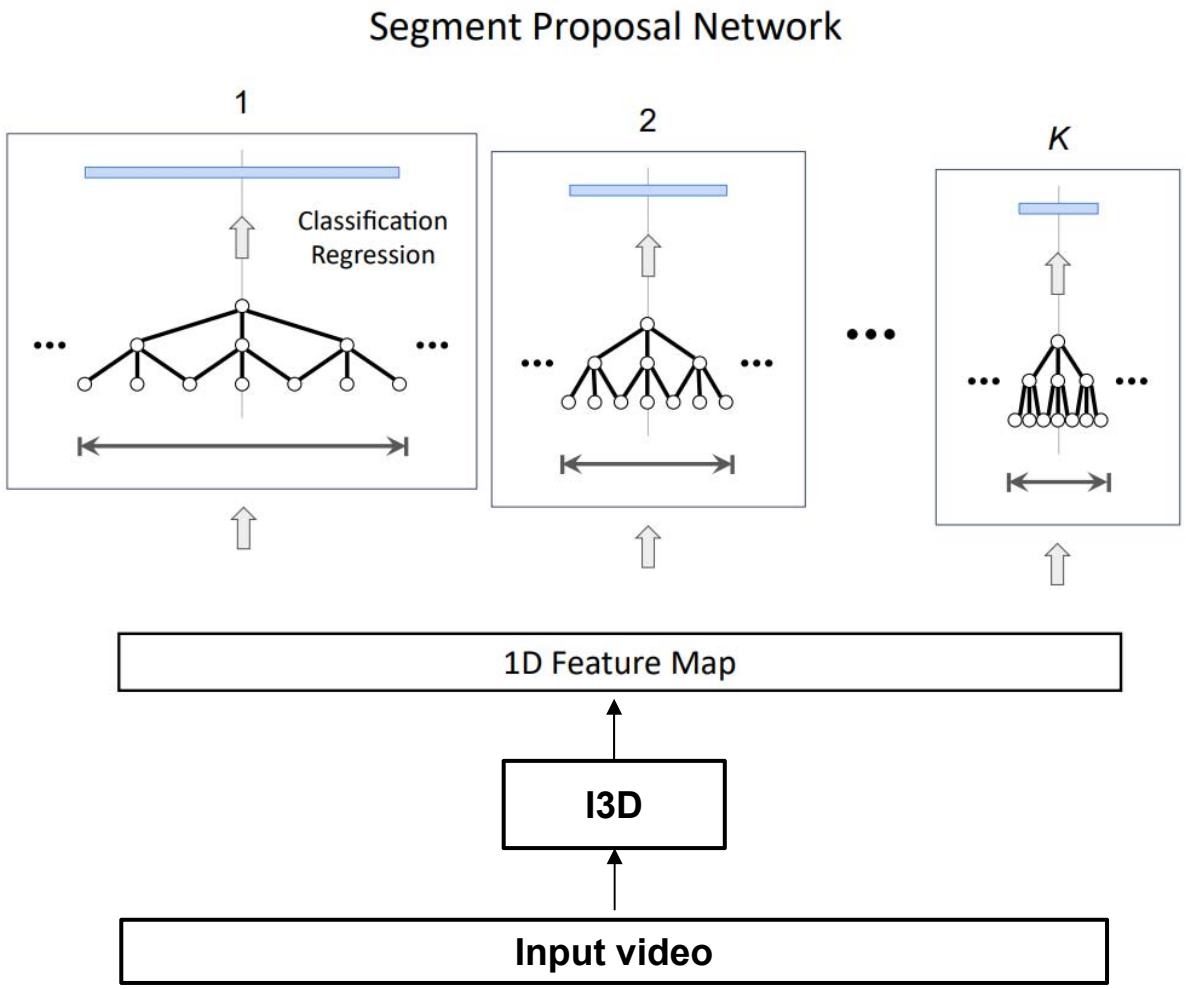
Figure 3. **C3D architecture**. C3D net has 8 convolution, 5 max-pooling, and 2 fully connected layers, followed by a softmax output layer. All 3D convolution kernels are $3 \times 3 \times 3$ with stride 1 in both spatial and temporal dimensions. Number of filters are denoted in each box. The 3D pooling layers are denoted from `pool1` to `pool5`. All pooling kernels are $2 \times 2 \times 2$, except for `pool1` is $1 \times 2 \times 2$. Each fully connected layer has 4096 output units.

- Other backbone networks:
  - I3D (Carreira and Zisserman, CVPR'17): Inflated 3D ConvNets
  - Res3D (Tran et al. 2017): resnet-like C3D
  - R(2+1)D (Tran et al. CVPR'18): replace 3D Conv by 2D Conv followed by 1D Conv
  - TSN (Wang et al. ECCV'16)

Heterogeneous Network Models (TAL-Net)

A Network is used to determine whether an input segment is background or not.

But unlike Segment-CNN, this work learns specific proposal network for each scale k.



Rethinking the Faster R-CNN Architecture for Temporal Action Localization.
Yu-Wei Chao, Sudheendra Vijayanarasimhan, Bryan Seybold, David A. Ross, Jia Deng, Rahul Sukthankar. In CVPR'18.

A Quick Comparison

- Temporal localization mAP on THUMOS'14.

IoU threshold	0.3	0.4	0.5	0.6	0.7
Heilbron et al. [3]	-	-	13.5	-	-
Escorcia et al. [4]	-	-	13.9	-	-
Oneata et al. [5]	28.8	21.8	15.0	8.5	3.2
Richard and Gall [6]	30.0	23.2	15.2	-	-
Yeung et al. [7]	36.0	26.4	17.1	-	-
Yuan et al. [8]	33.6	26.1	18.8	-	-
Yuan et al. [9]	36.5	27.8	17.8	-	-
S-CNN [1]	36.3	28.7	19.0	10.3	5.3
SST [10]	37.8	-	23.0	-	-
CDC [2]	40.1	29.4	23.3	13.1	7.9
Dai et al. [11]	-	33.3	25.6	15.9	9.0
SSAD [12]	43.0	35.0	24.6	-	-
TURN TAP [13]	44.1	34.9	25.6	-	-
R-C3D [14]	44.7	35.6	28.9	-	-
SS-TAD [15]	45.7	-	29.2	-	9.6
Gao et al. [16]	50.1	41.3	31.0	19.1	9.9
SSN [17]	51.9	41.0	29.8	19.6	10.7
TAL-Net [18]	**53.2**	**48.5**	**42.8**	**33.8**	**20.8**

Segment-CNN, CVPR'16

Heterogeneous proposal networks, CVPR'18

Insufficient precision caused by fixed segments

Video

Background

Background

Background

Sliding window

L

Backbone Network (e.g., C3D)

1

Coarse temporal boundaries

Pred

Cricket Bowling

Temporal Action Localization in Untrimmed Videos via Multi-stage CNNs.
Zheng Shou, Dongang Wang, and Shih-Fu Chang. In CVPR'16.

Coarse Fixed Segments

Video

Background

Background

Background

Sliding window

L

3D ConvNets

Down in **S** & **T**

Segment-CNN

L/8

FC

Down in **S** & **T**

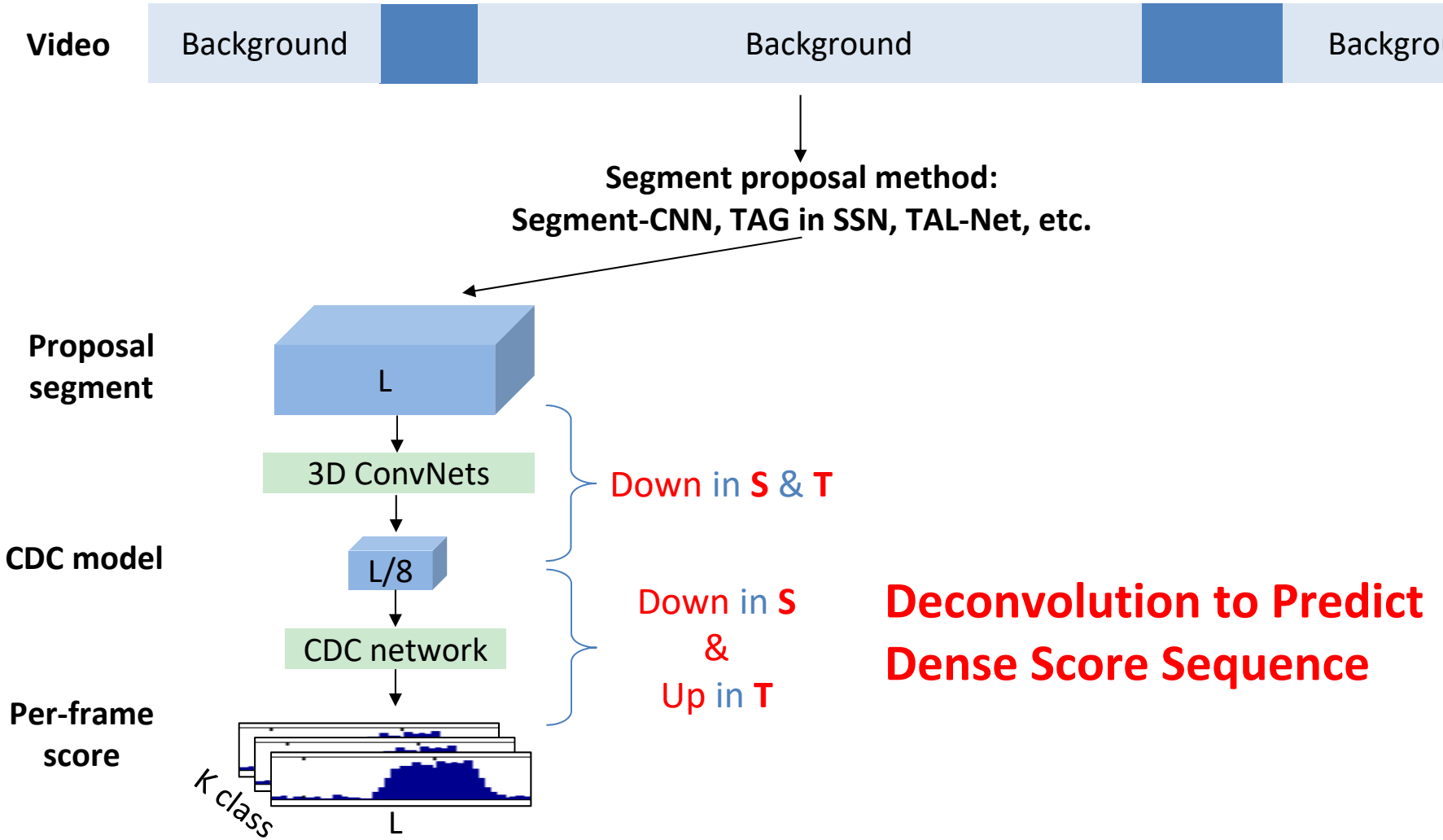
1

Pred

Cricket Bowling

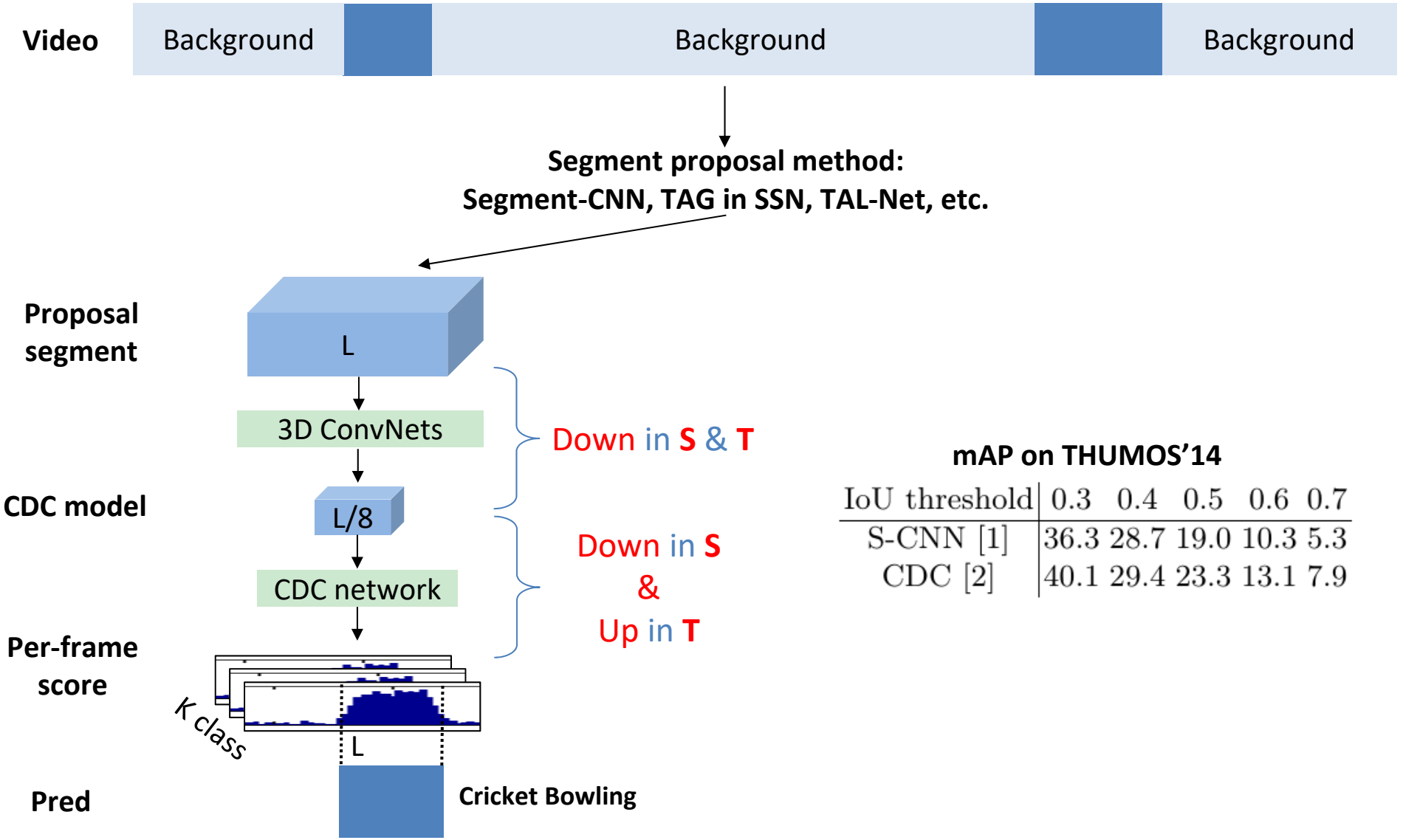
Temporal Action Localization in Untrimmed Videos via Multi-stage CNNs.
Zheng Shou, Dongang Wang, and Shih-Fu Chang. In CVPR'16.

Use Deconvolution to Predict Dense Scores



CDC: Convolutional-De-Convolutional Networks for Precise Temporal Action Localization in Untrimmed Videos.
Zheng Shou, Jonathan Chan, Alireza Zareian, Kazuyuki Miyazawa, and Shih-Fu Chang. In CVPR'17.

Boundary Refinement



mAP on THUMOS'14

IoU threshold	0.3	0.4	0.5	0.6	0.7
S-CNN [1]	36.3	28.7	19.0	10.3	5.3
CDC [2]	40.1	29.4	23.3	13.1	7.9

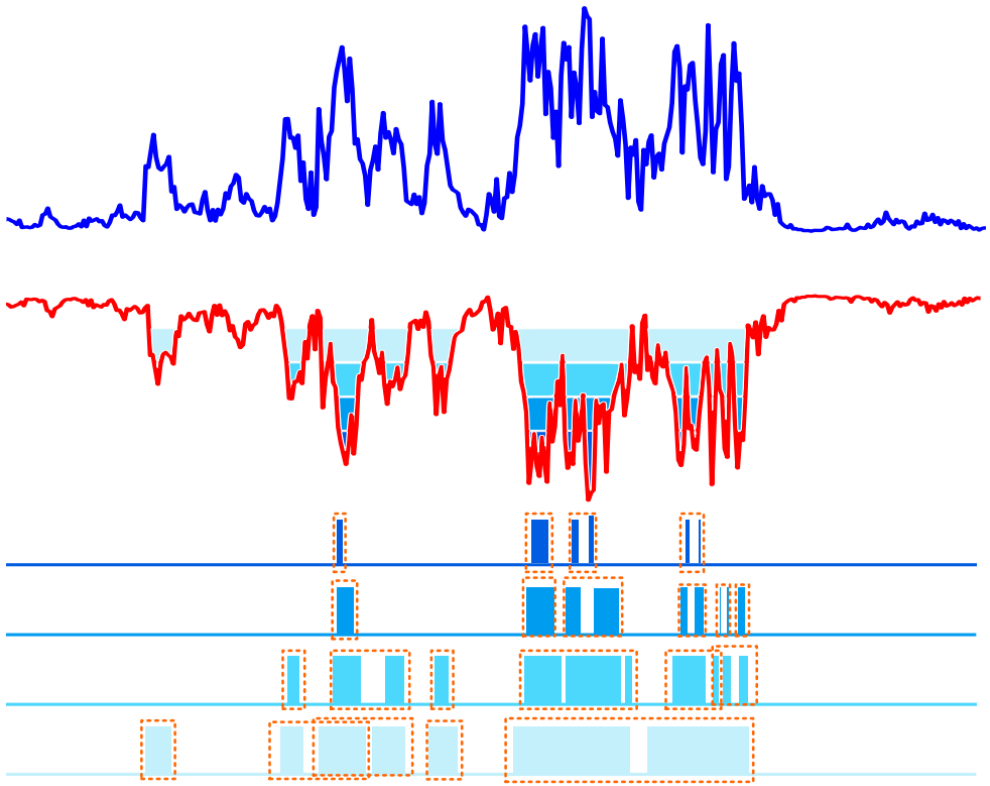
CDC: Convolutional-De-Convolutional Networks for Precise Temporal Action Localization in Untrimmed Videos. Zheng Shou, Jonathan Chan, Alireza Zareian, Kazuyuki Miyazawa, and Shih-Fu Chang. In CVPR'17.

Another Way of Improving Precision

Temporal Action Detection with Structured Segment Networks.
Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, Dahua Lin. In ICCV'17.

Slide action/background classifier over time to produce dense action score sequence

Use watershed algorithm to produce multi-scale proposals



A Quick Comparison

- Temporal localization mAP on THUMOS'14.

IoU threshold	0.3	0.4	0.5	0.6	0.7
Heilbron et al. [3]	-	-	13.5	-	-
Escorcia et al. [4]	-	-	13.9	-	-
Oneata et al. [5]	28.8	21.8	15.0	8.5	3.2
Richard and Gall [6]	30.0	23.2	15.2	-	-
Yeung et al. [7]	36.0	26.4	17.1	-	-
Yuan et al. [8]	33.6	26.1	18.8	-	-
Yuan et al. [9]	36.5	27.8	17.8	-	-
S-CNN [1]	36.3	28.7	19.0	10.3	5.3
SST [10]	37.8	-	23.0	-	-
CDC [2]	40.1	29.4	23.3	13.1	7.9
Dai et al. [11]	-	33.3	25.6	15.9	9.0
SSAD [12]	43.0	35.0	24.6	-	-
TURN TAP [13]	44.1	34.9	25.6	-	-
R-C3D [14]	44.7	35.6	28.9	-	-
SS-TAD [15]	45.7	-	29.2	-	9.6
Gao et al. [16]	50.1	41.3	31.0	19.1	9.9
SSN [17]	51.9	41.0	29.8	19.6	10.7
TAL-Net [18]	**53.2**	**48.5**	**42.8**	**33.8**	**20.8**

Watershed segmentation of dense score

Heterogeneous proposal network

Topic #2: Weak Supervision

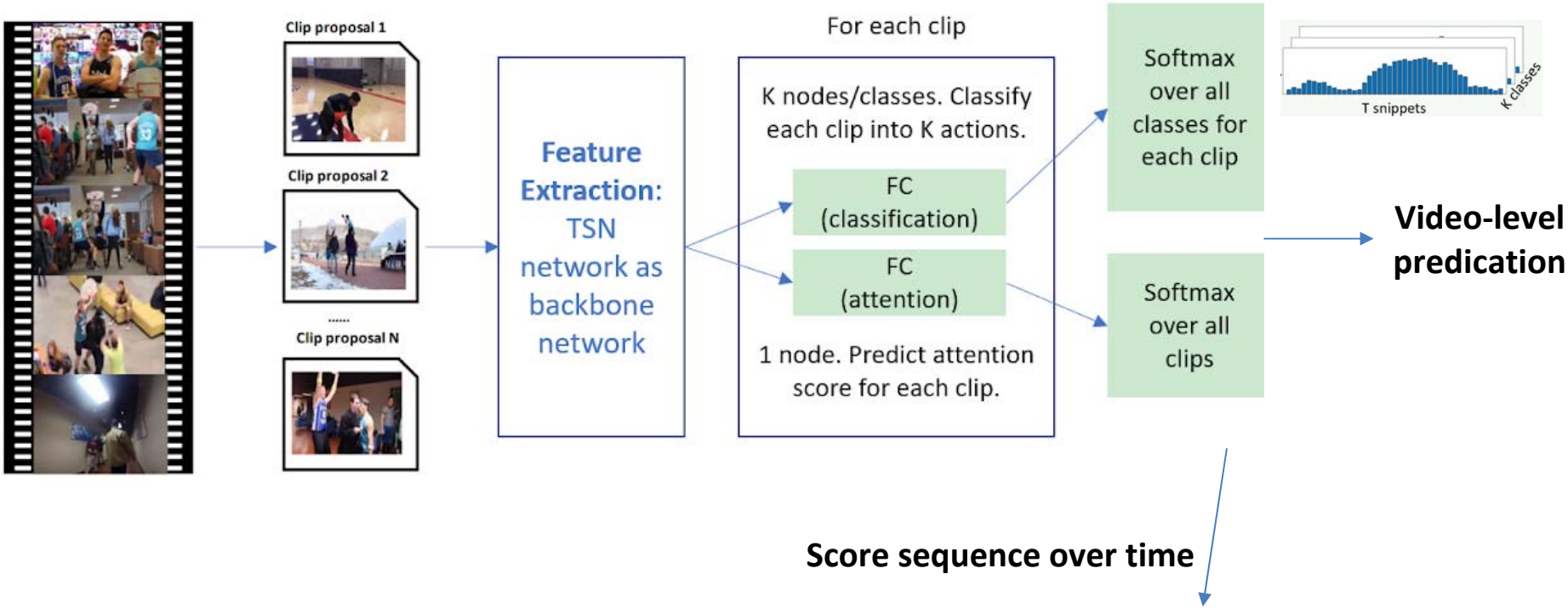
- Annotating boundaries is time-consuming -> often annotate video-level label only

Problem definition

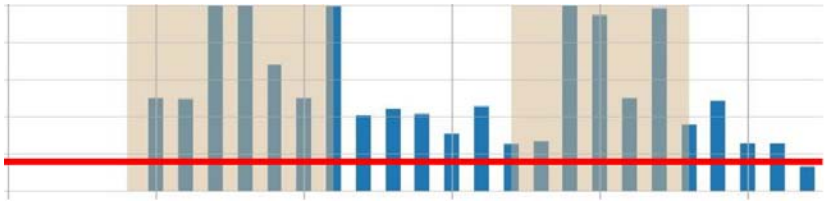
UntrimmedNets for Weakly Supervised Action Recognition and Detection.
Limin Wang, Yuanjun Xiong, Dahua Lin, and Luc Van Gool. In CVPR'17.

Weak Supervision

- Annotating boundaries is time-consuming -> only annotate video-level label

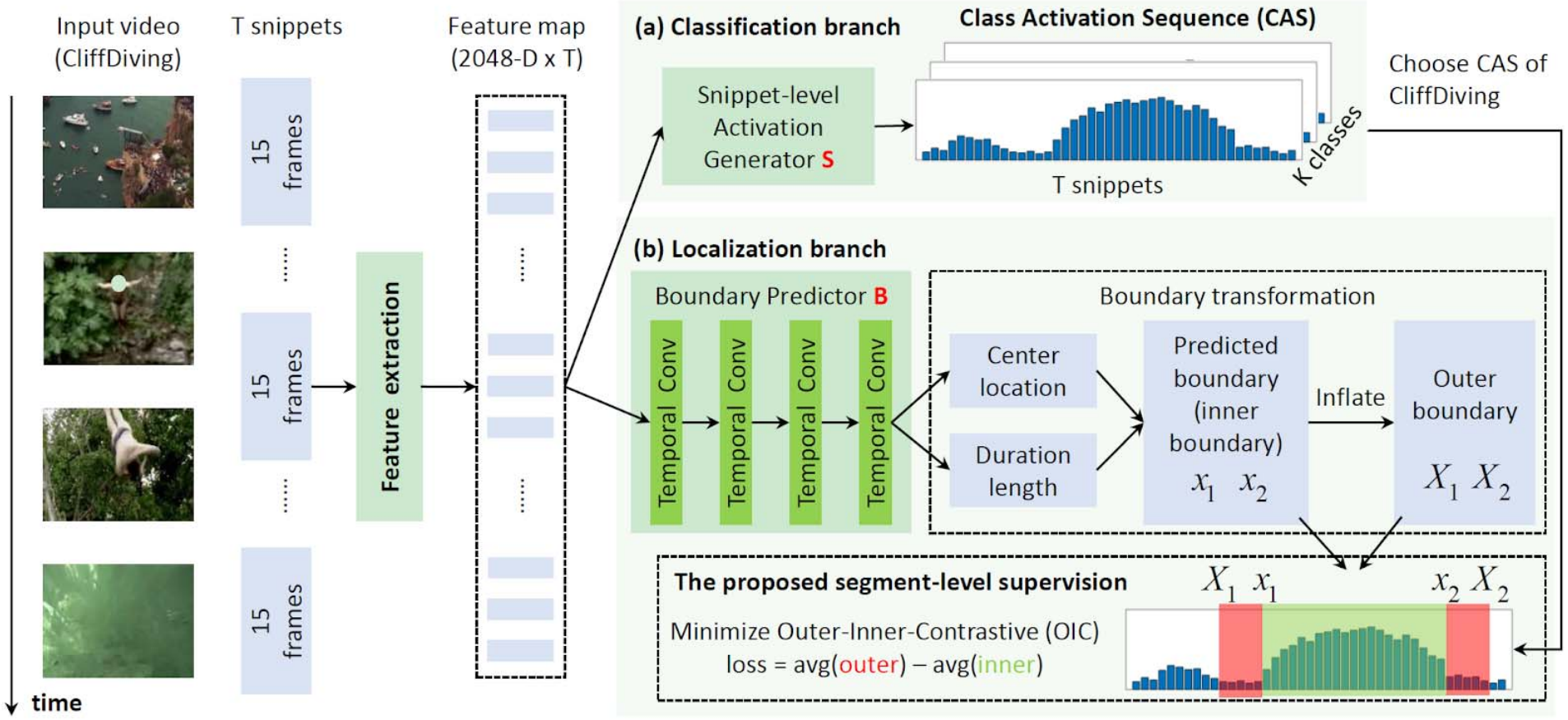


Conventional weakly-supervised methods: simple thresholding



UntrimmedNets for Weakly Supervised Action Recognition and Detection.
Limin Wang, Yuanjun Xiong, Dahua Lin, and Luc Van Gool. In CVPR'17.

Weak Supervision



AutoLoc: Weakly-supervised Temporal Action Localization in Untrimmed Videos.
Zheng Shou, Hang Gao, Lei Zhang, Kazuyuki Miyazawa, and Shih-Fu Chang. 2018.

Weak Supervision

- Temporal localization mAP on THUMOS'14. Compare with SOTA.

Supervision	IoU threshold	0.3	0.4	0.5	0.6	0.7
Full	Heilbron et al. [3]	-	-	13.5	-	-
Full	Escorcia et al. [4]	-	-	13.9	-	-
Full	Oneata et al. [5]	28.8	21.8	15.0	8.5	3.2
Full	Richard and Gall [6]	30.0	23.2	15.2	-	-
Full	Yeung et al. [7]	36.0	26.4	17.1	-	-
Full	Yuan et al. [8]	33.6	26.1	18.8	-	-
Full	Yuan et al. [9]	36.5	27.8	17.8	-	-
Full	S-CNN [1]	36.3	28.7	19.0	10.3	5.3
Full	SST [10]	37.8	-	23.0	-	-
Full	CDC [2]	40.1	29.4	23.3	13.1	7.9
Full	Dai et al. [11]	-	33.3	25.6	15.9	9.0
Full	SSAD [12]	43.0	35.0	24.6	-	-
Full	TURN TAP [13]	44.1	34.9	25.6	-	-
Full	R-C3D [14]	44.7	35.6	28.9	-	-
Full	SS-TAD [15]	45.7	-	29.2	-	9.6
Full	Gao et al. [16]	50.1	41.3	31.0	19.1	9.9
Full	SSN [17]	51.9	41.0	29.8	19.6	10.7
Full	TAL-Net [18]	**53.2**	**48.5**	**42.8**	**33.8**	**20.8**
Weak	Sun et al. [19]	8.5	5.2	4.4	-	-
Weak	Hide-and-Seek [20]	19.5	12.7	6.8	-	-
Weak	Wang et al. [21]	28.2	21.1	13.7	-	-
Weak	Ours - AutoLoc	**35.8**	**29.0**	**21.2**	**13.4**	**5.8**

Full supervision: segment-level boundary annotations during training

Weak supervision: video-level action class annotations during training

Comparable to some baselines trained with the full supervision

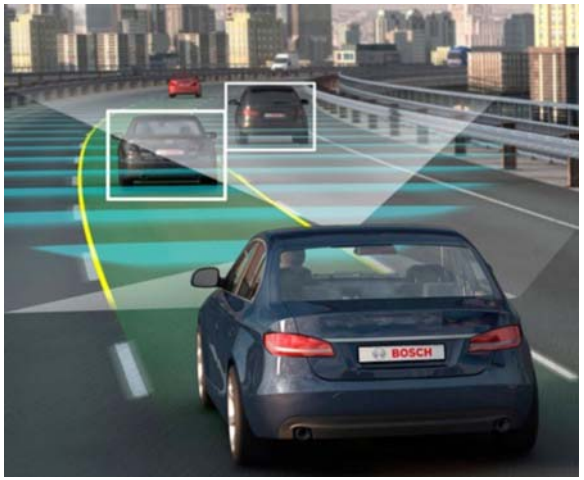
Achieve SOTA performance for weakly super. methods

AutoLoc: Weakly-supervised Temporal Action Localization in Untrimmed Videos.
Zheng Shou, Hang Gao, Lei Zhang, Kazuyuki Miyazawa, and Shih-Fu Chang. 2018.

Topic #3: Online Detection of Action Start

- Importance of detecting action start as soon as possible in streaming video.

Surveillance



Self-driving car

Robot

Online Detection of Action Start in Untrimmed, Streaming Videos.
Zheng Shou*, Junting Pan*, Jonathan Chan, Kazuyuki Miyazawa, Hassan Mansour, Anthony Vetro, Xavier Giro-i-Nieto, and Shih-Fu Chang. Arxiv 2018.

Example: Online Detection of Action Start

Start (Human Annotated)

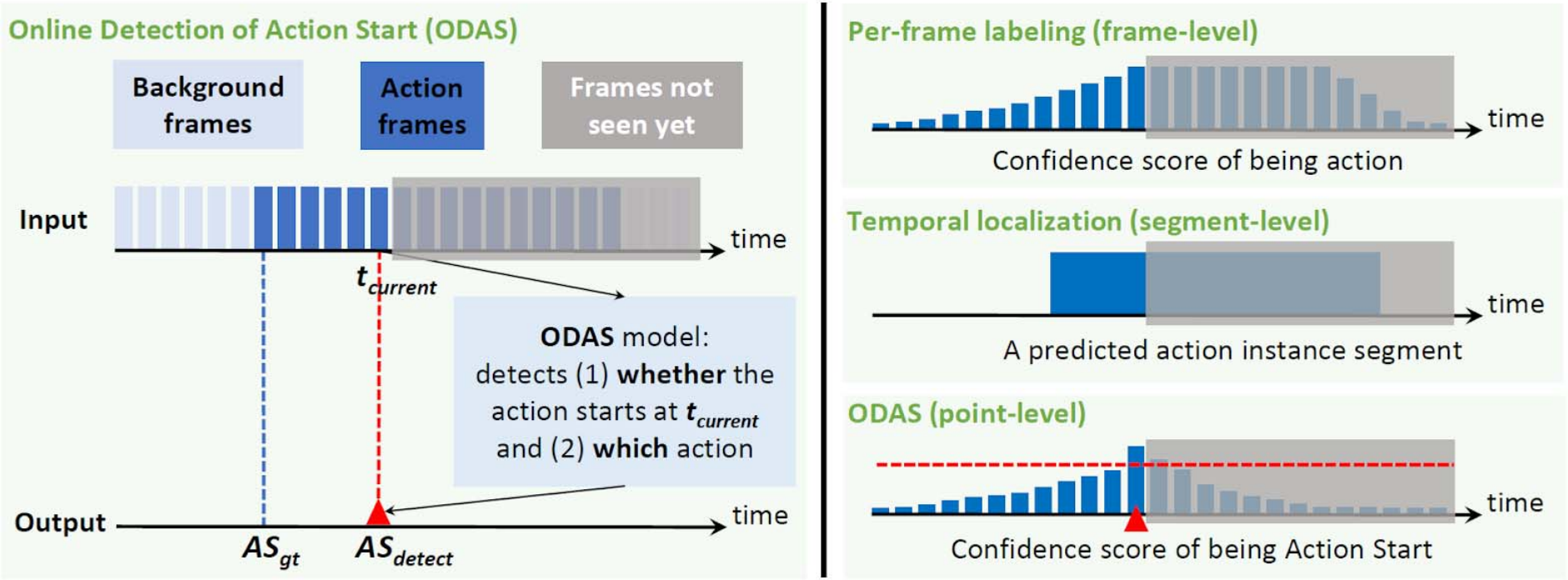
Detected Start (Our Method)

Detected Start (C3D Baseline)

Online Detection of Action Start in Untrimmed, Streaming Videos.
Zheng Shou*, Junting Pan*, Jonathan Chan, Kazuyuki Miyazawa, Hassan Mansour, Anthony Vetro, Xavier Giro-i-Nieto, and Shih-Fu Chang. Arxiv 2018.

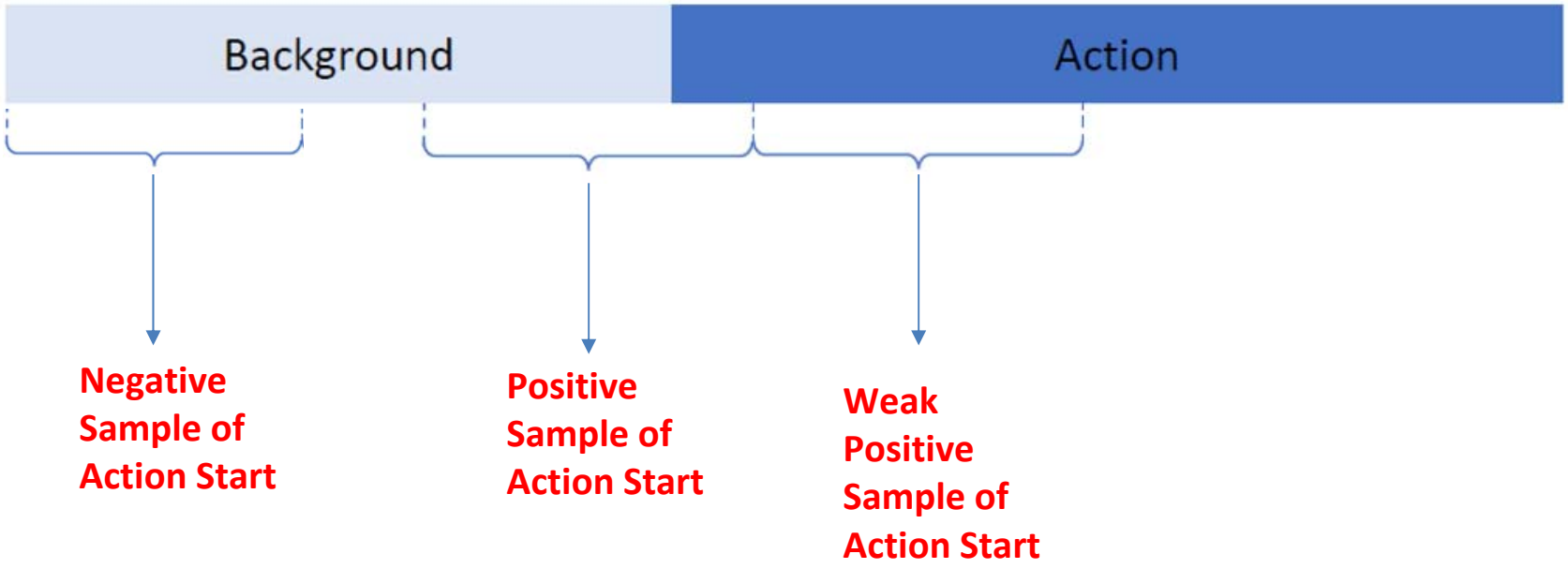
Online Detection of Action Start

- **Online detection:** only current and past frames, future frames not available
- Action Start (AS): background → action, or action 1 → action



Online Detection of Action Start in Untrimmed, Streaming Videos.
Zheng Shou*, Junting Pan*, Jonathan Chan, Kazuyuki Miyazawa, Hassan Mansour, Anthony Vetro, Xavier Giro-i-Nieto, and Shih-Fu Chang. Arxiv 2018.

Data Formulation for Action Start

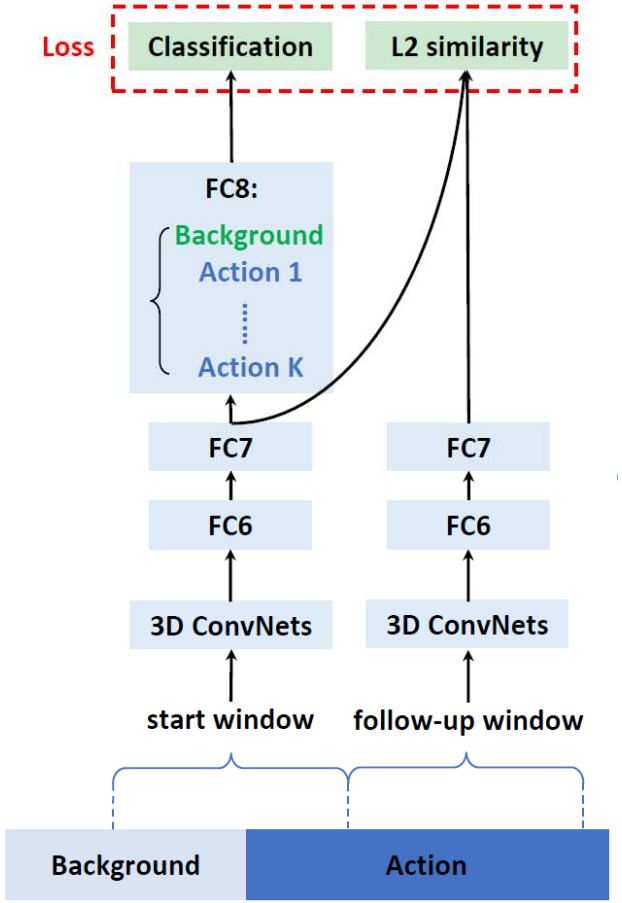


- Challenge:
  - Relation among temporal clips
  - Too Few Positive & Hard Negative

- Solutions:
  - Introduce temporal consistency
  - Use GAN to synthesize

Online Detection of Action Start

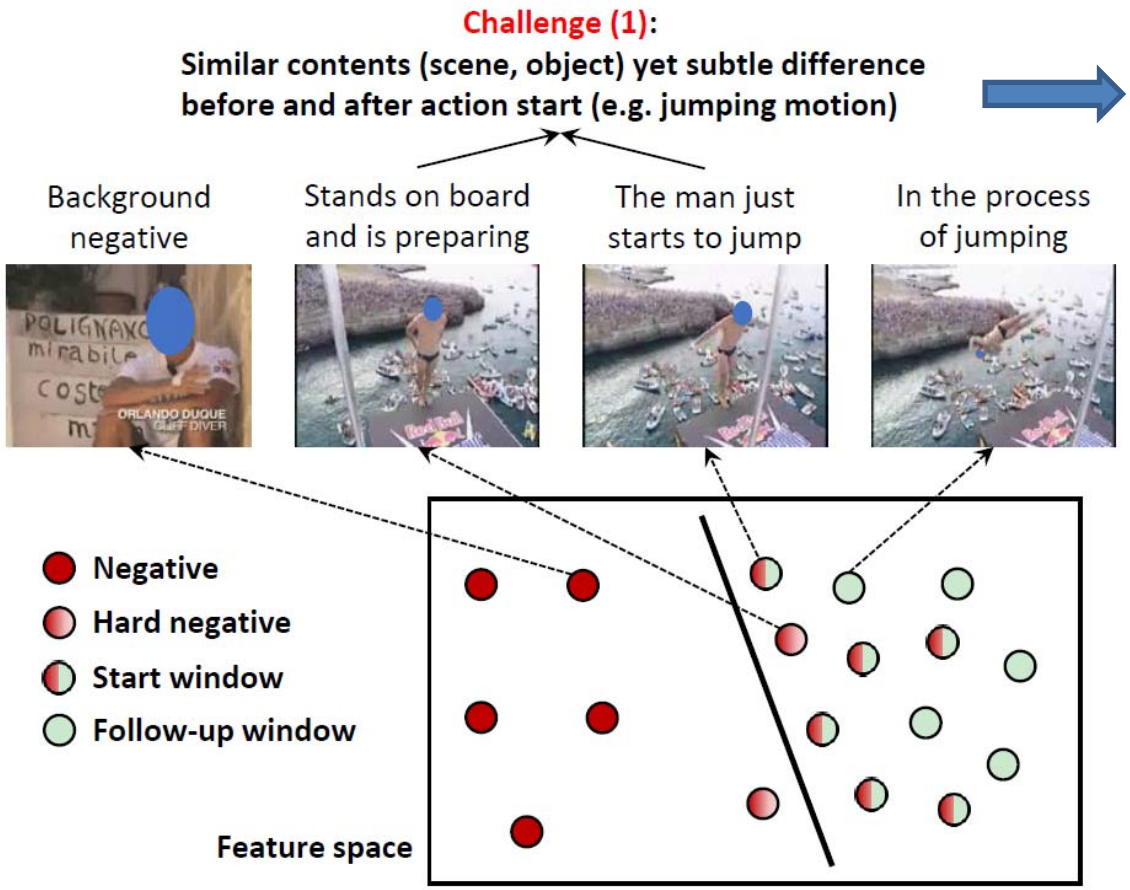
- Temporal Consistency



Model the temporal consistency

GAN for hard negative synthesis

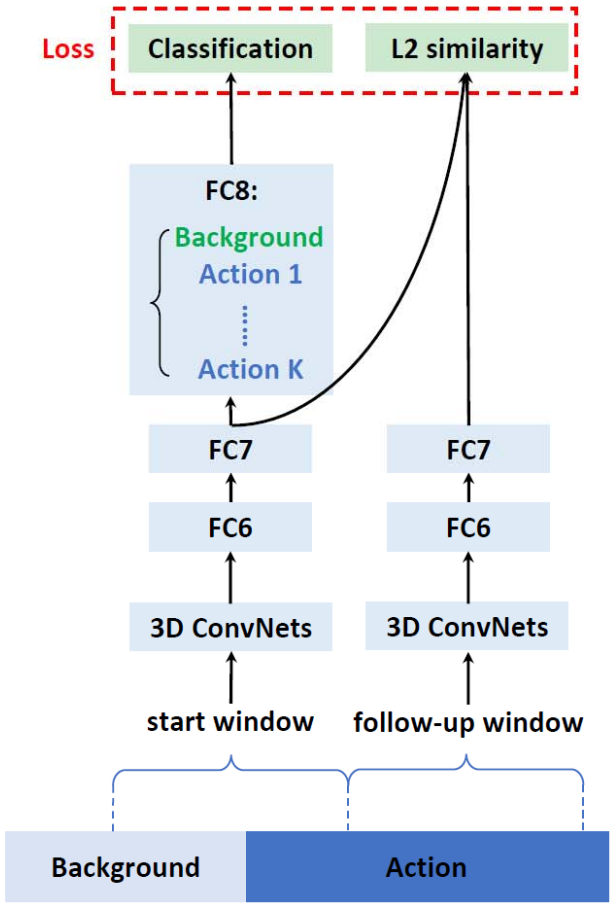
- Challenges



Online Detection of Action Start in Untrimmed, Streaming Videos.
Zheng Shou*, Junting Pan*, Jonathan Chan, Kazuyuki Miyazawa, Hassan Mansour, Anthony Vetro, Xavier Giro-i-Nieto, and Shih-Fu Chang. Arxiv 2018.

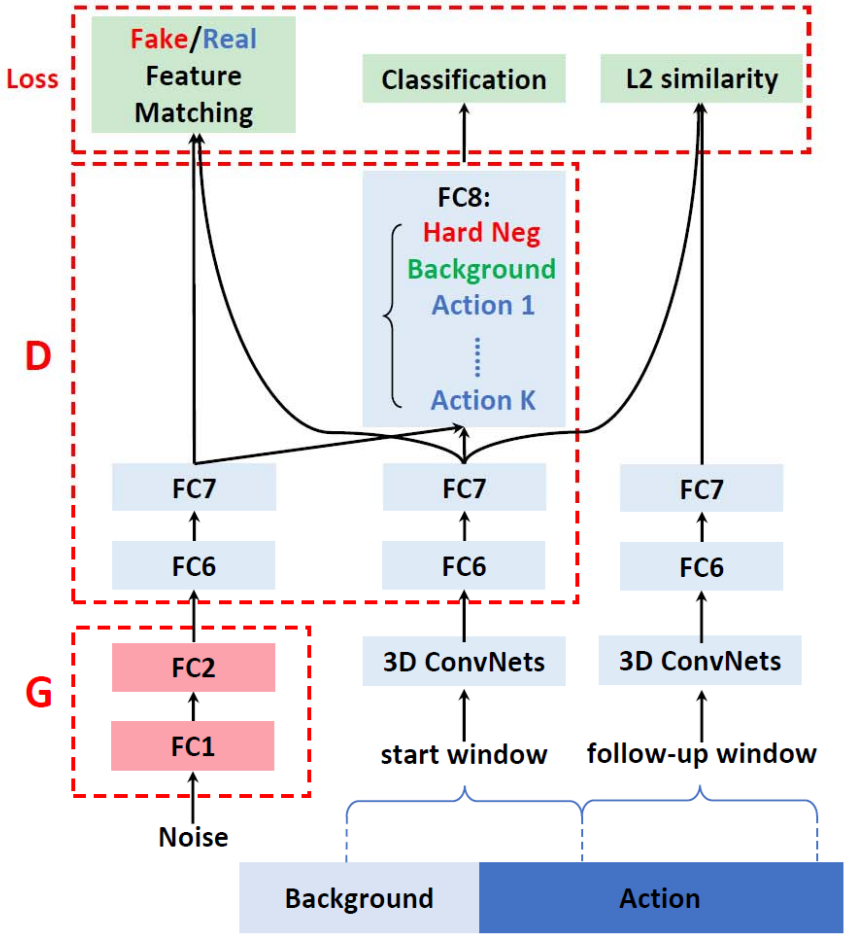
Online Detection of Action Start

- Temporal Consistency



Model the temporal consistency

- GAN to synthesize hard negatives



Generate hard negatives via GAN

Online Detection of Action Start

Start (Human Annotated)

Detected Start (Our Method)

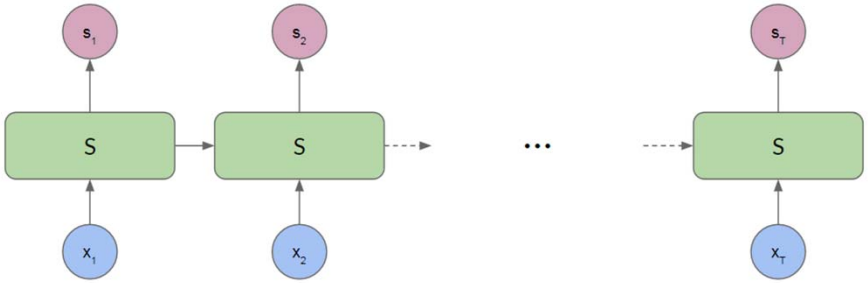
Detected Start (C3D Baseline)

offset threshold (s)	10	50	100
Random guess	0.06	0.14	0.17
SceneDetect	4.71	18.93	25.84
ShotDetect	6.10	24.35	33.76
TSN w/o ours	8.18	31.39	44.15
Our approach	**8.33**	**33.08**	**46.97**

mAP (%) on ActivitNet

Online Detection of Action Start in Untrimmed, Streaming Videos.
Zheng Shou*, Junting Pan*, Jonathan Chan, Kazuyuki Miyazawa, Hassan Mansour, Anthony Vetro, Xavier Giro-i-Nieto, and Shih-Fu Chang. Arxiv 2018.

Recurrent Models for Video?



Issues:

1. Slow inference
2. Difficulties capturing long term dependencies
3. Vanishing/exploding gradients during training

These issues are related to the resulting computational graphs being too long

Can we learn how to shorten them?

Vicor Campos, Jou, Giro-i-Nieto, Torres, Chang, ICLR 2018

Skip RNN: Learning to Skip State Updates in RNNs


Víctor Campos


Brendan Jou


Jordi Torres

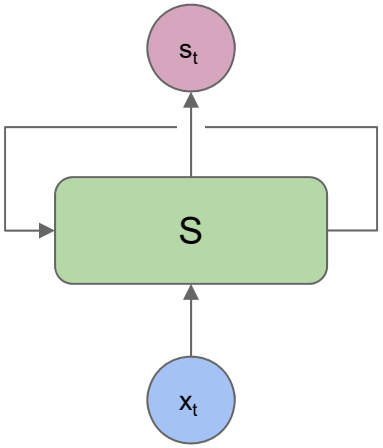

Xavier Giró-i-Nieto


Shih-Fu Chang

https://imatge-upc.github.io/skiprnn-2017-telecombcn

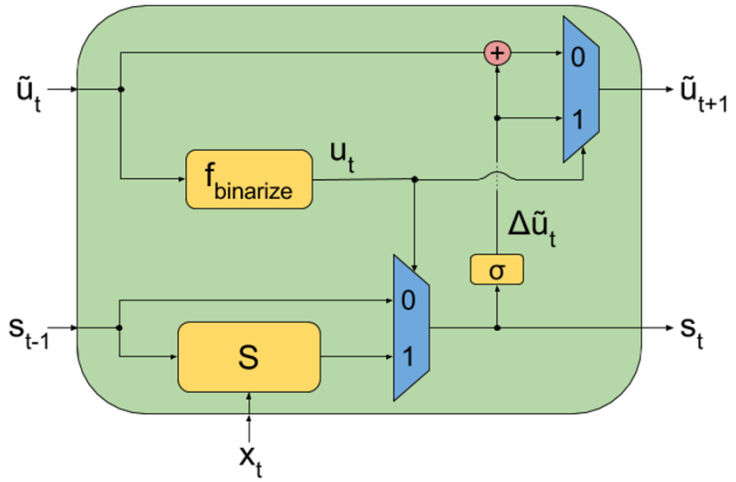
Model description

Intuition: introduce a binary *update state gate*, u_t, deciding whether the RNN state is updated or copied



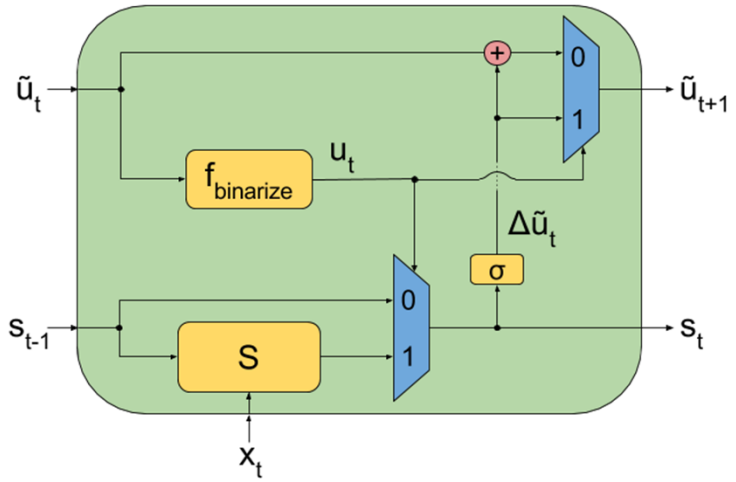
$$s_t = \begin{cases} S(x_t, h_{t-1}) & \text{if } u_t = 1 \quad \text{// update operation} \\ s_{t-1} & \text{if } u_t = 0 \quad \text{// copy operation} \end{cases}$$

Model description



$$u_t = f_{binarize}(\tilde{u}_t)$$
$$s_t = u_t \cdot S(s_{t-1}, x_t) + (1 - u_t) \cdot s_{t-1}$$
$$\Delta\tilde{u}_t = \sigma(W_p s_t + b_p)$$
$$\tilde{u}_{t+1} = u_t \cdot \Delta\tilde{u}_t + (1 - u_t) \cdot (\tilde{u}_t + \min(\Delta\tilde{u}_t, 1 - \tilde{u}_t))$$

Model description



$$u_t = f_{binarize}(\tilde{u}_t)$$

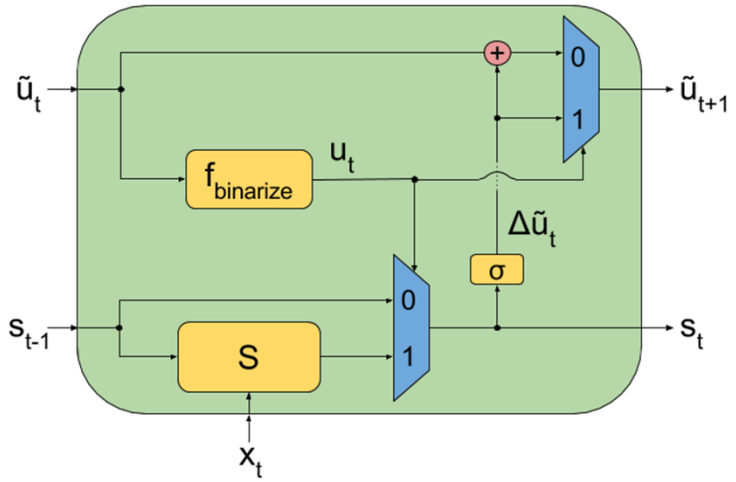
$$s_t = u_t \cdot S(s_{t-1}, x_t) + (1 - u_t) \cdot s_{t-1}$$

$$\Delta\tilde{u}_t = \sigma(W_p s_t + b_p)$$

$$\tilde{u}_{t+1} = u_t \cdot \Delta\tilde{u}_t + (1 - u_t) \cdot (\tilde{u}_t + \min(\Delta\tilde{u}_t, 1 - \tilde{u}_t))$$

Update state gate ∈ {0, 1}

Model description



$$u_t = f_{binarize}(\tilde{u}_t)$$

$$s_t = u_t \cdot S(s_{t-1}, x_t) + (1 - u_t) \cdot s_{t-1}$$

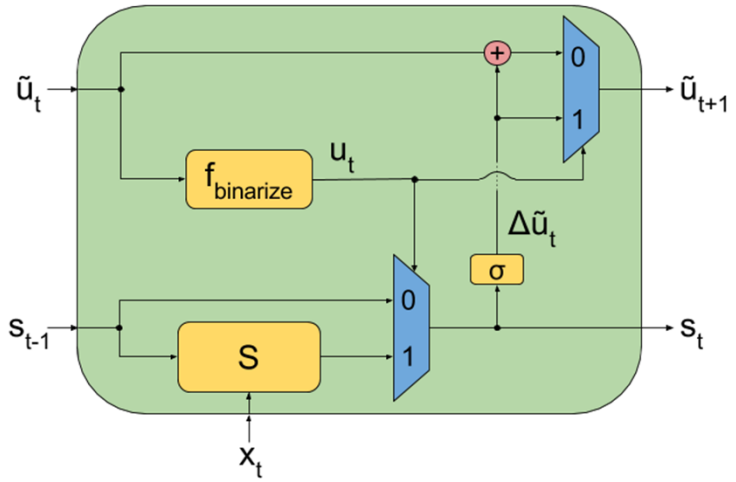
$$\Delta\tilde{u}_t = \sigma(W_p s_t + b_p)$$

$$\tilde{u}_{t+1} = u_t \cdot \Delta\tilde{u}_t + (1 - u_t) \cdot (\tilde{u}_t + \min(\Delta\tilde{u}_t, 1 - \tilde{u}_t))$$

Update state gate $\in \{0, 1\}$

Update state probability $\in [0, 1]$

Model description



$$u_t = f_{binarize}(\tilde{u}_t)$$

$$s_t = u_t \cdot S(s_{t-1}, x_t) + (1 - u_t) \cdot s_{t-1}$$

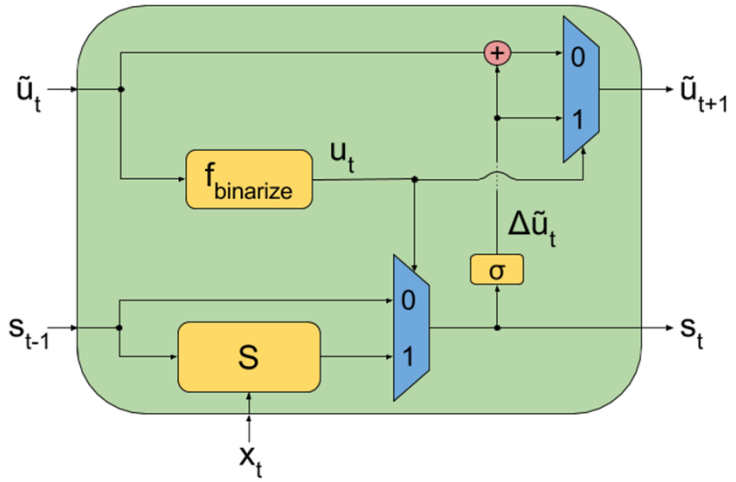
$$\Delta\tilde{u}_t = \sigma(W_p s_t + b_p)$$

$$\tilde{u}_{t+1} = u_t \cdot \Delta\tilde{u}_t + (1 - u_t) \cdot (\tilde{u}_t + \min(\Delta\tilde{u}_t, 1 - \tilde{u}_t))$$

Update state gate ∈ {0, 1}

Update state probability ∈ [0, 1]

Model description



$$u_t = f_{binarize}(\tilde{u}_t)$$

$$s_t = u_t \cdot S(s_{t-1}, x_t) + (1 - u_t) \cdot s_{t-1}$$

$$\Delta\tilde{u}_t = \sigma(W_p s_t + b_p)$$

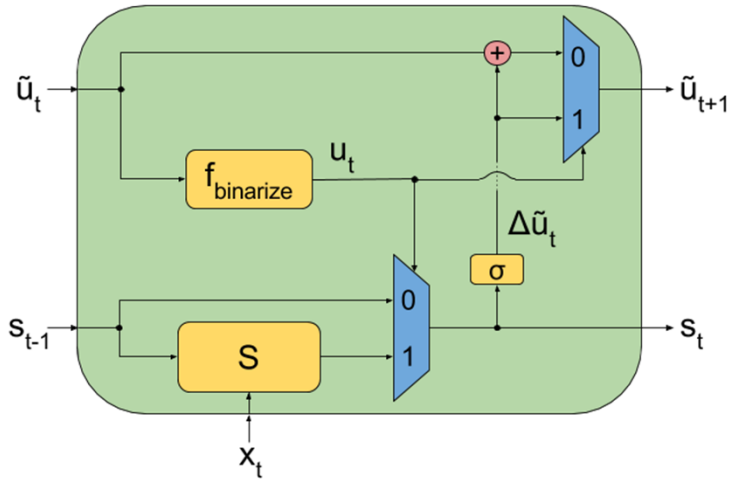
$$\tilde{u}_{t+1} = u_t \cdot \Delta\tilde{u}_t + (1 - u_t) \cdot (\tilde{u}_t + \min(\Delta\tilde{u}_t, 1 - \tilde{u}_t))$$

Update state gate ∈ {0, 1}

Update state probability ∈ [0, 1]

Increment for the update state probability

Model description



$$u_t = f_{binarize}(\tilde{u}_t)$$

$$s_t = u_t \cdot S(s_{t-1}, x_t) + (1 - u_t) \cdot s_{t-1}$$

$$\Delta\tilde{u}_t = \sigma(W_p s_t + b_p)$$

$$\tilde{u}_{t+1} = u_t \cdot \Delta\tilde{u}_t + (1 - u_t) \cdot (\tilde{u}_t + \min(\Delta\tilde{u}_t, 1 - \tilde{u}_t))$$

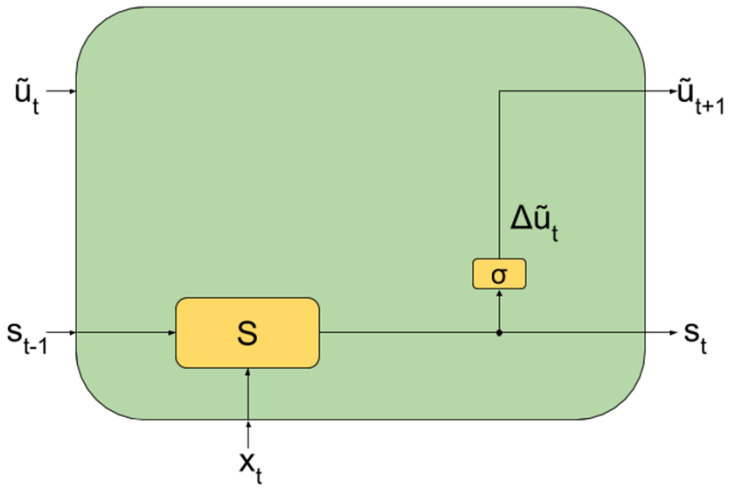
Update state gate ∈ {0, 1}

Update state probability ∈ [0, 1]

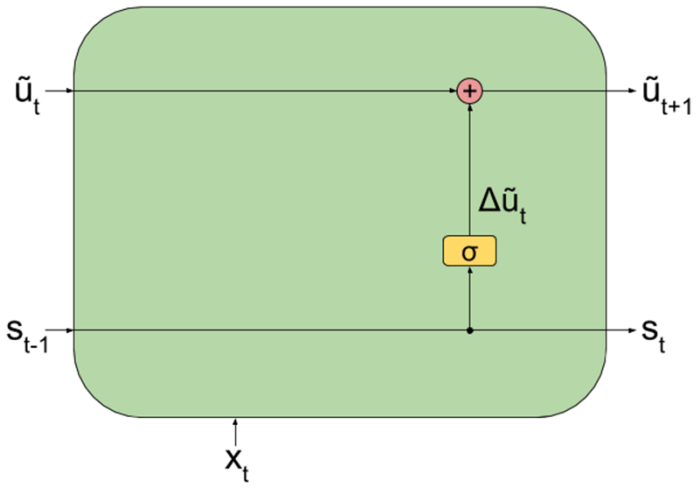
Increment for the update state probability

Model description

Update state ($u_t = 1$)

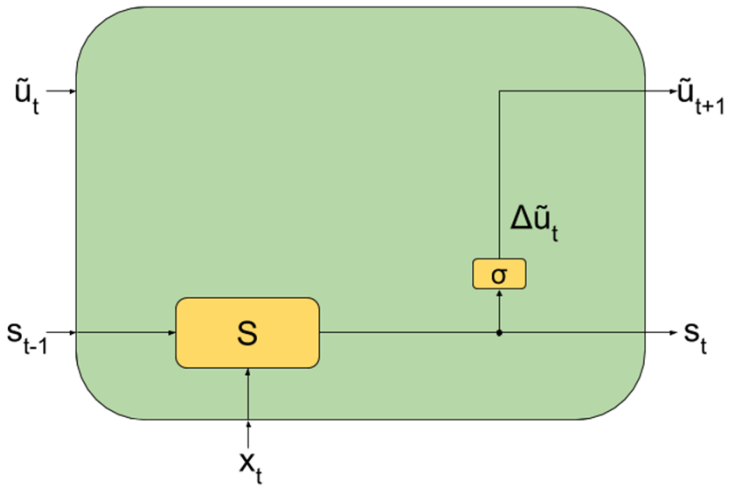


Copy state ($u_t = 0$)

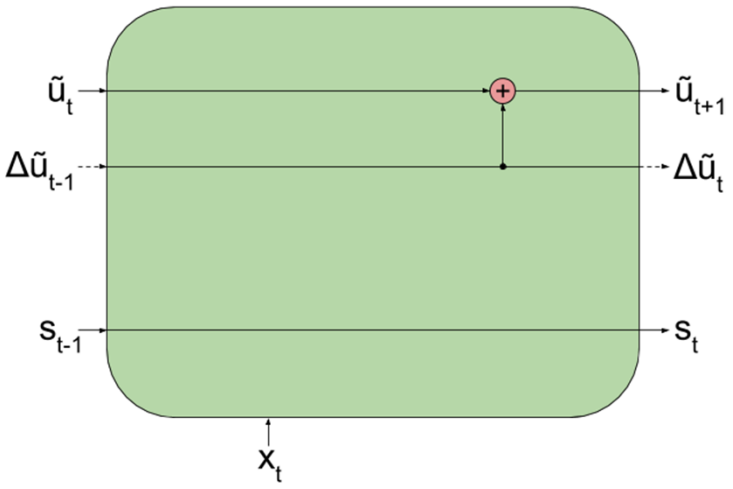


Model description

Update state ($u_t = 1$)



Copy state ($u_t = 0$)



$$N_{skip}(t) = \min\{n : n \cdot \Delta\tilde{u}_t \geq 0.5\} - 1$$

Limiting computation

Intuition: the network can be encouraged to perform fewer updates by adding a penalization when $u_t = 1$

$$L_{budget} = \lambda \cdot \sum_{t=1}^{T} u_t \longrightarrow \begin{cases} 1 & \text{if sample used} \\ 0 & \text{otherwise} \end{cases}$$

cost per sample

Applicable to many tasks

Skip RNN has been evaluated on

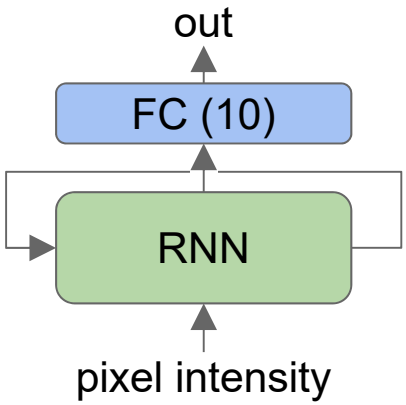
1. Adding task
2. Frequency discrimination task
3. Digit classification
4. Sentiment analysis
5. Action classification
6. Action localization

Applicable to many tasks

Skip RNN has been evaluated on

	Type of data
1. Adding task	synthetic (regression)
2. Frequency discrimination task	synthetic (classification)
3. Digit classification	image
4. Sentiment analysis	text
5. Action classification	video (many-to-one)
6. Action localization	video (many-to-many)

Skip RNN on Sequential MNIST



- Digit classification task with 10 classes, i.e. [0, 9]
- Traditionally addressed with CNNs, but it can be converted into a sequential task by flattening the images
  - Original images: 28x28
  - Flattened images: 784-d vectors
- The RNN is given 1 pixel at a time

Sequential MNIST: results

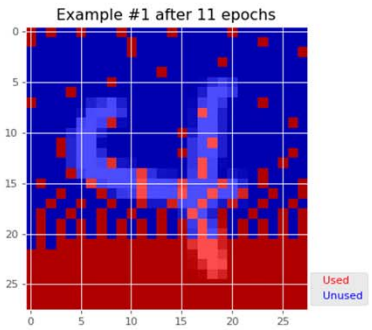
Model	Accuracy	State updates
LSTM	0.910 ± 0.045	784.00 ± 0.00
Skip LSTM, $\lambda = 10^{-4}$	0.973 ± 0.002	379.38 ± 33.09
GRU	0.968 ± 0.013	784.00 ± 0.00
Skip GRU, $\lambda = 10^{-4}$	0.976 ± 0.003	392.62 ± 26.48

Table 5.3: Accuracy and used samples on the test set of MNIST after 600 epochs of training. Results are displayed as $mean \pm std$ over four different runs.

Why can we improve accuracy and reduce computation at the same?

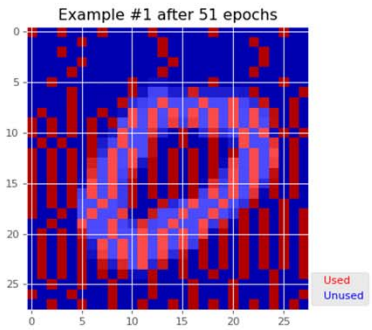
Sequential MNIST: examples

11 epochs



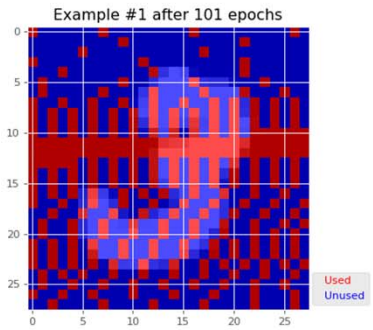
~30% acc

51 epochs



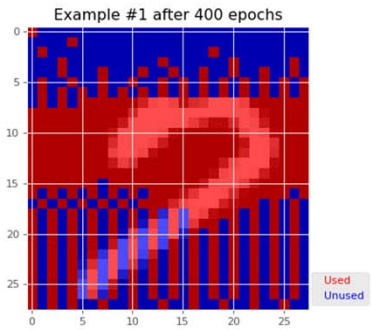
~50% acc

101 epochs



~70% acc

400 epochs

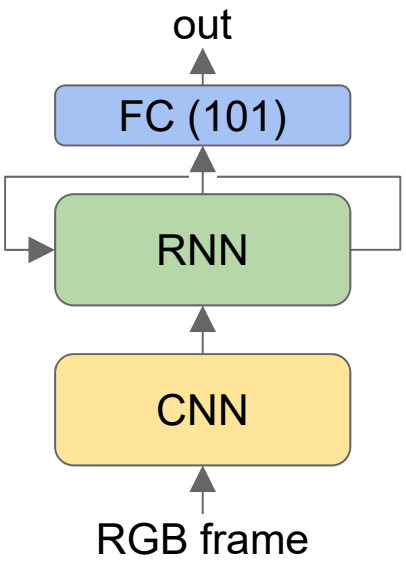


~95% acc

Epochs for Skip LSTM ($\lambda = 10^{-4}$)

Used
Unused

Skip RNN on Video Classification



- Short, trimmed videos
- 101 action classes
- 10s of video
  - Cropped longer videos
  - Padded shorter ones with empty frames
- Using original framerate: 25 fps
- Frame-level ResNet-50 GAP features

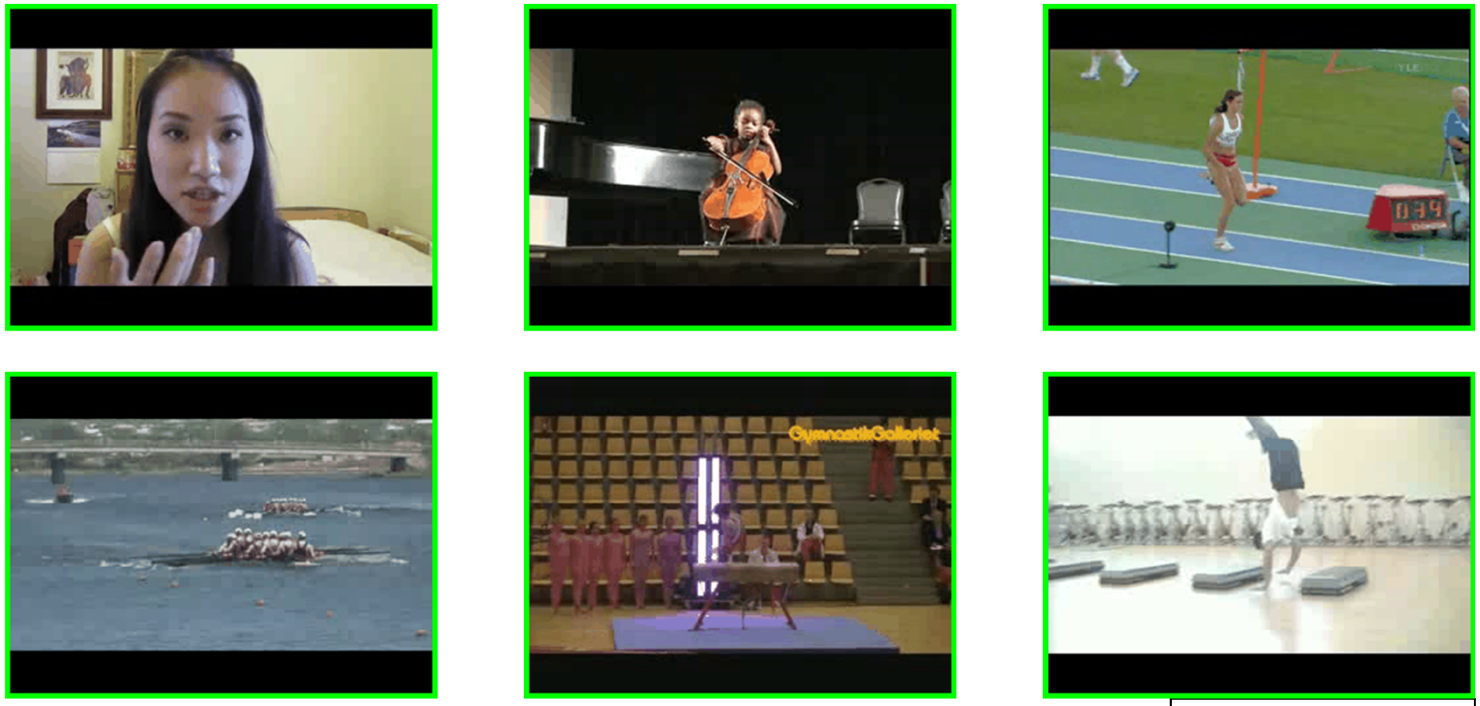
UCF-101: results

Model	Accuracy	State updates
LSTM	0.671	250.0
Skip LSTM, $\lambda = 0$	0.749	138.9
Skip LSTM, $\lambda = 10^{-5}$	0.757	24.2
Skip LSTM, $\lambda = 10^{-4}$	0.790	7.6
GRU	0.791	250.0
Skip GRU, $\lambda = 0$	0.796	124.2
Skip GRU, $\lambda = 10^{-5}$	0.792	29.7
Skip GRU, $\lambda = 10^{-4}$	0.793	23.7

Table 5.5: Accuracy and used samples on the validation set of UCF-101 (split 1).

Why can we improve accuracy and reduce computation at the same?

UCF-101: results (Skip LSTM, λ = 10-4)



Used
Unused

Video Classification (UCF 101 Activity) (Skip LSTM, λ = 10-4)

Ice Dancing

Rafting
(vs. rowing, or others)

Used Unused

Summary & Future Directions

- Methods for object detection can be borrowed for video
  - Proposal segment-CNN a decent baseline
  - Work on boundary refinement promising
  - Weakly supervised solutions start attracting attention

- Ongoing Efforts:
  - Online action detection, forecasting
  - Backbone models for video data
  - Reduce temporal complexity

- Other Interesting Topics:
  - Language vs. video grounding
  - Video QA, Video caption generation
  - Temporal causal modeling

Thank you!

Contact:

Prof. Shih-Fu Chang

shih.fu.chang@columbia.edu