Columbia University
Mitsubishi Electric
Changes for the Better

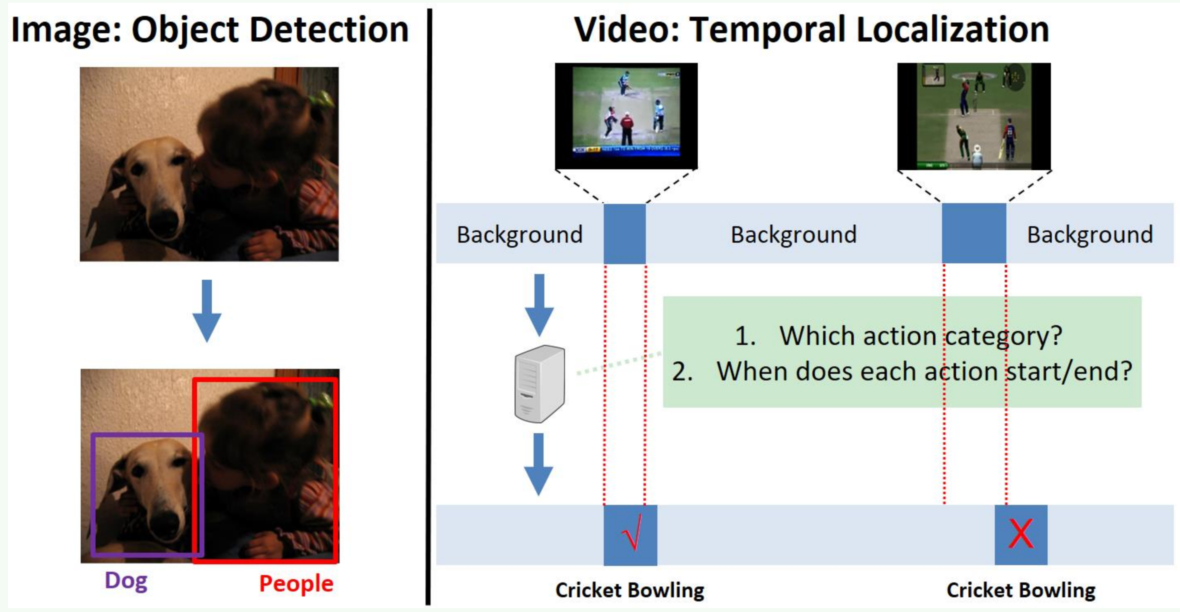
CDC: Convolutional-De-Convolutional Networks for Precise Temporal Action Localization in Untrimmed Videos

Zheng Shou[1], Jonathan Chan[1], Alireza Zareian[1], Kazuyuki Miyazawa[2], Shih-Fu Chang[1]
[1] Columbia University, [2] Mitsubishi Electric

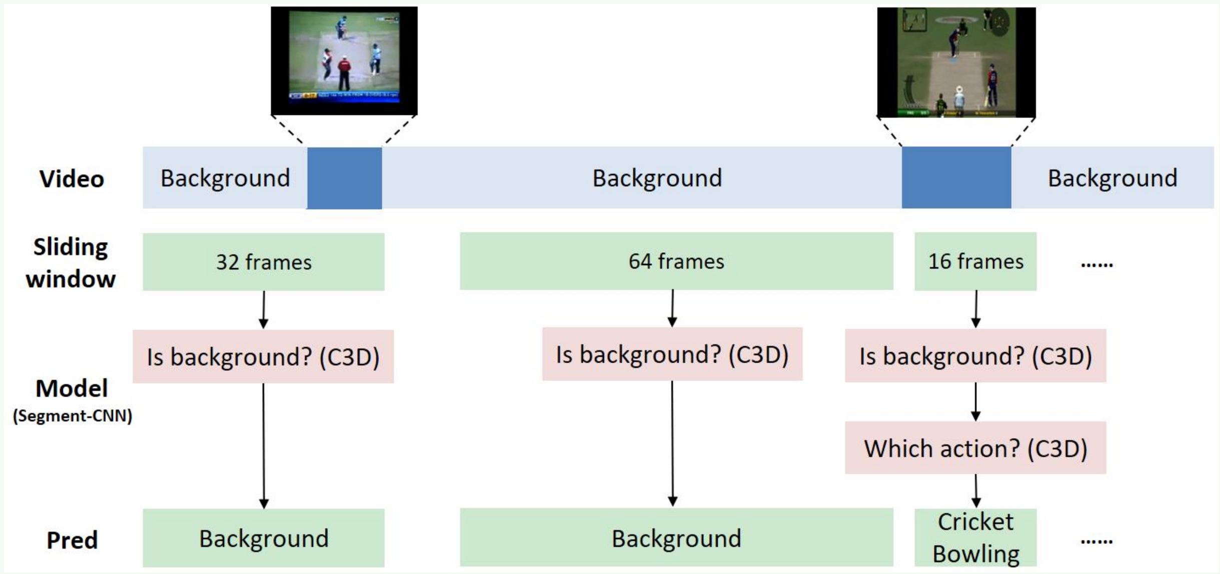
IEEE 2017 Conference on Computer Vision and Pattern Recognition
CVPR July 21-26 Honolulu 2017

1. Introduction:

➢ **Problem definition:** Temporal Localization



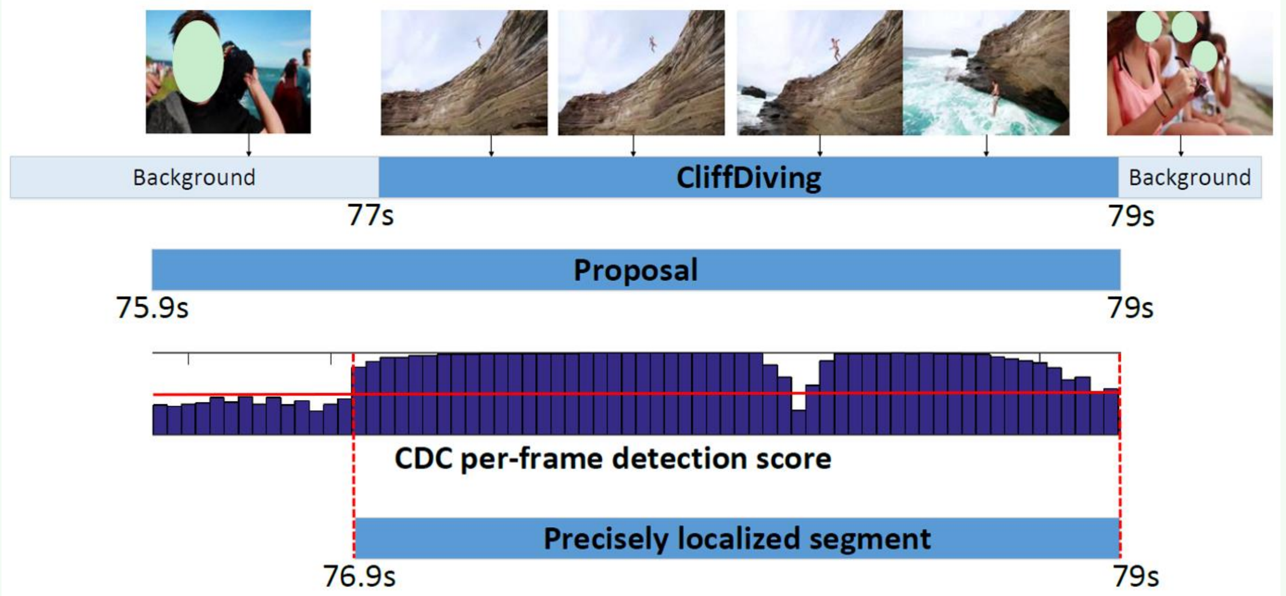
➢ **Related work:** Segment-CNN. Shou et al. CVPR'16.



2. How to make precise localization?

➢ **Motivation**:

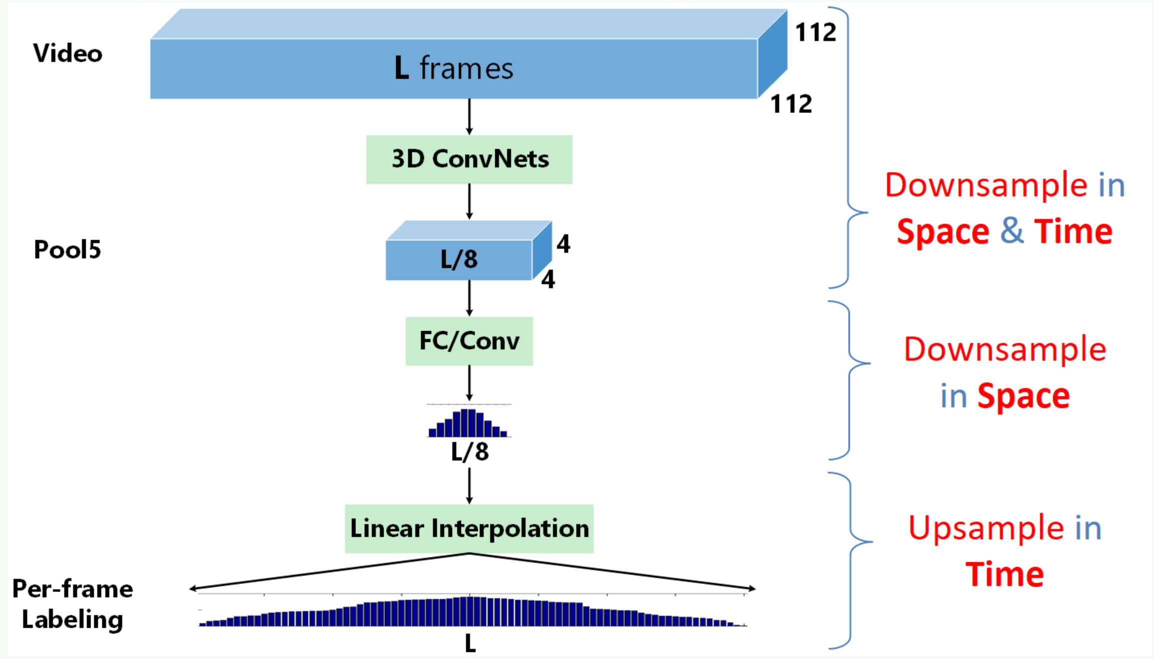
– SoA methods keep the pre-determined boundaries of input proposal windows, which might be inaccurate.
– We need to detect at a finer granularity in time (e.g. frame-level) and then refine boundaries of proposals.



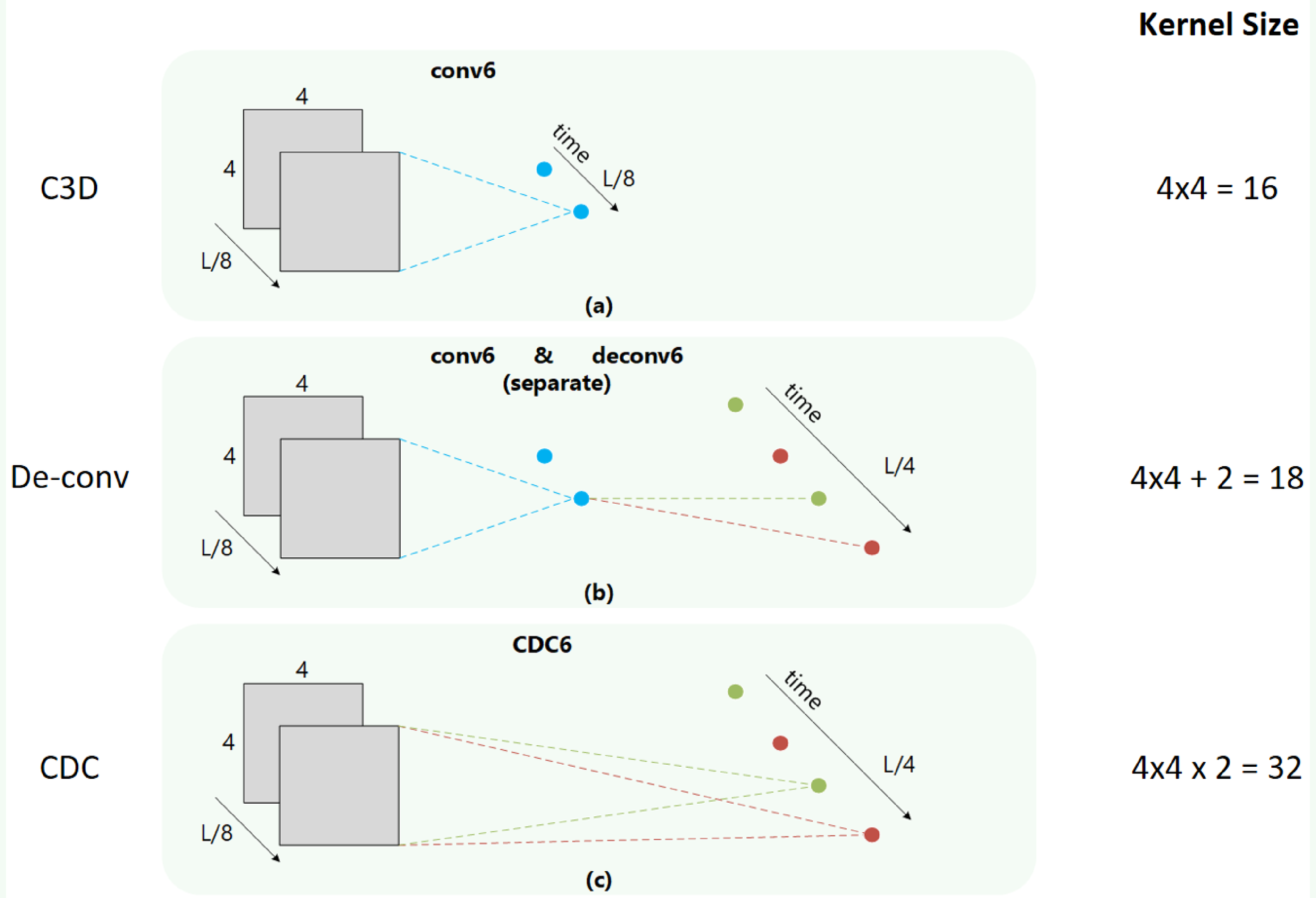
3. How to predict at the frame-level?

➢ **Need of down-sampling and up-sampling:**

– We build our work on C3D, which is SoA video classifier.
– From pool5 to the frame-level prediction, we need to perform down-sampling in space and up-sampling in time
– A simple baseline is Linear Interpolation:



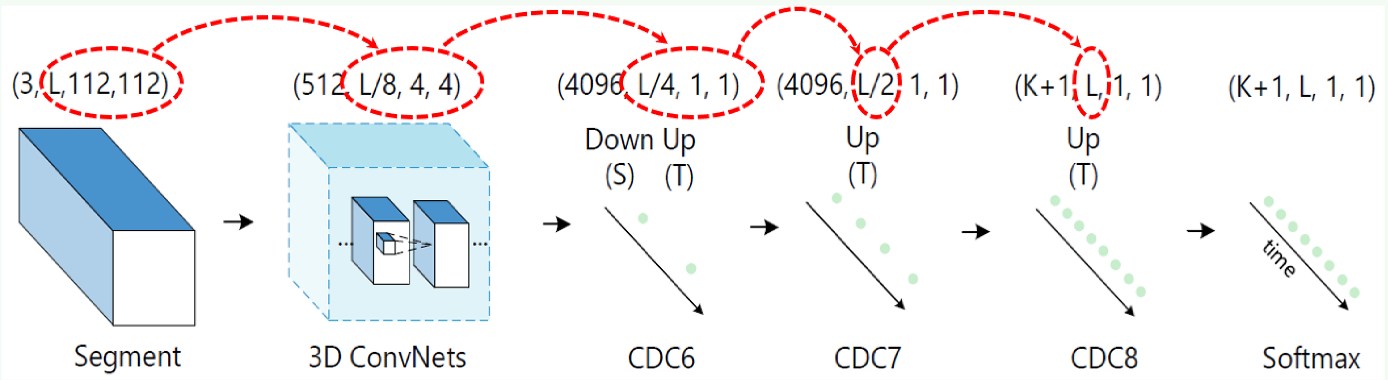
– For for up-sampling, learnable models (e.g. 3D de-conv [Du et al.]) are more effective than linear interpolation.
– In our case, a straightforward way is adding de-conv after conv, but this models down- and up-sampling separately.



➢ **Conv-De-Conv filter**: We propose a novel Conv-De-Conv (**CDC**) filter to simultaneously perform spatial down-sampling (for spatio-temporal semantic abstraction) and temporal up-sampling (for precise temporal localization).

4. Details of CDC network:

➢ **Network architecture:**



➢ **Training data construction:**

– CDC network can operate on videos of variable lengths.
– From the temporal boundary annotations, we know the label of every frame. To prevent including too many background frames for training, we only keep windows that have at least one frame belonging to actions.

➢ **Loss function:** frame-wise softmax loss

$$\mathcal{L} = \frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{L}\left(-\log\left(P_n^{(z_n)}[t]\right)\right) \qquad \frac{\partial \mathcal{L}}{\partial O_n^{(i)}[t]} = \begin{cases} \frac{1}{N}\cdot\left(P_n^{(z_n)}[t]-1\right) & \text{if } i = z_n \\ \frac{1}{N}\cdot P_n^{(i)}[t] & \text{if } i \neq z_n \end{cases}$$

➢ **Optimization:** implementation based on C3D-v1.0, stochastic gradient descent, learning rate 0.00001 for all layers except 0.0001 for CDC8 layer since CDC8 is randomly initialized, momentum 0.9, weight decay 0.005, converges after 4 training epochs (within half day) on THUMOS'14.

➢ **Testing:**

- **Per-Frame Labeling:** Given a test window, directly feed into the CDC network to output predictions for every frame.
- **Temporal Localization:** use CDC per-frame predictions to refine temporal boundaries of segment proposal and predict segment-level category. Given a segment proposal,
  – Extend the segment slightly to search a wider interval.
  – Set segment-level category to the class with the maximum average confidence score over all frames in the segment.
  – Perform Gaussian kernel density estimation.
  – Trim the proposal segment from both sides until we reach a frame with the confidence score not lower than the score threshold (mean minus standard deviation).

Contact: zheng.shou@columbia.edu
Code: https://bitbucket.org/columbiadvmm/cdc
Project website: http://www.ee.columbia.edu/ln/dvmm/researchProjects/cdc/

5. Experiments on THUMOS'14:

➢ **Dataset**:

– 20 categories. On average 15 instances per video.
– Training data: 2,755 trimmed videos from UCF101 + 200 untrimmed YouTube videos of 3,007 instances with temporal boundary annotations.
– Test data: 213 untrimmed videos of 3,358 action instances

➢ **Evaluation metrics:**

- **Per-Frame Labeling:** For each action class, rank all frames in the test set by their confidence scores for that class and evaluate Average Precision (AP). We average over all action classes to obtain mean AP (mAP).
- **Temporal Localization:** We output a rank list of predicted segments and evaluate mAP. A prediction is correct when it has the correct category and its temporal overlap IoU with GT instance is larger than the threshold.

methods	mAP
Single-frame CNN [ICLR'15]	34.7
Two-stream CNN [NIPS'14]	36.2
LSTM [CVPR'15]	39.3
MultiLSTM [Arxiv'15]	41.3
C3D + LinearInterp	37.0
Conv & De-conv	41.7
CDC (fix 3D ConvNets)	37.4
CDC	**44.4**

Table 1. Per-frame labeling mAP

IoU threshold	0.3	0.4	0.5	0.6	0.7
Karaman [THUMOS'14]	0.5	0.3	0.2	0.2	0.1
Wang [THUMOS'14]	14.6	12.1	8.5	4.7	1.5
Heilbron [CVPR'16]	-	-	13.5	-	-
Escorcia [ECCV'16]	-	-	13.9	-	-
Oneata [THUMOS'14]	28.8	21.8	15.0	8.5	3.2
Richard and Gall [CVPR'16]	30.0	23.2	15.2	-	-
Yeung [CVPR'16]	36.0	26.4	17.1	-	-
Yuan [CVPR'16]	33.6	26.1	18.8	-	-
S-CNN [CVPR'16]	36.3	28.7	19.0	10.3	5.3
C3D + LinearInterp	36.0	26.4	19.6	11.1	6.6
Conv & De-conv	38.6	28.2	22.4	12.0	7.5
CDC (fix 3D ConvNets)	36.9	26.2	20.4	11.3	6.8
CDC	**40.1**	**29.4**	**23.3**	**13.1**	**7.9**

Table 2. Temporal action localization mAP

➢ **Efficiency:**

- Speed: Titan X with 12G memory. 500 PFS. Process 20s video within 1s.
- Storage: around 1 GB. do not need to cache intermediate features.

6. Experiments on ActivityNet:

➢ **Dataset**: includes 200 activities. 10K training videos (15K instances), 5K validation videos (7.6K instances), and 5K test videos (held-out GT).

➢ **Temporal Localization**: comparisons with top results in challenge 2016.

IoU threshold	0.5	0.75	0.95	Ave-mAP
Singh and Cuzzolin	22.7	10.8	0.3	11.3
Singh	26.0	15.2	2.6	14.6
Wang and Tao	45.1	4.1	0.0	16.4
CDC	45.3	26.0	0.2	23.8

Table 3. Results on 2016 validation set

IoU threshold	0.5	0.75	0.95	Ave-mAP
Singh and Cuzzolin	36.4	11.1	0.1	17.8
Singh	28.7	17.8	2.9	17.7
Wang and Tao	42.5	2.9	0.1	14.6
CDC (train)	43.1	25.6	0.2	22.9
CDC (train+val)	43.0	25.7	0.2	22.9

Table 4. Results on 2016 test set