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

Z. Shou[1], J. Chan[1], A. Zareian[1], K. Miyazawa[2], and S.-F. Chang[1]

zheng.shou@columbia.edu

[1] Columbia University & [2] Mitsubishi Electric

Problem Definition

Image: Classification

People
Dog

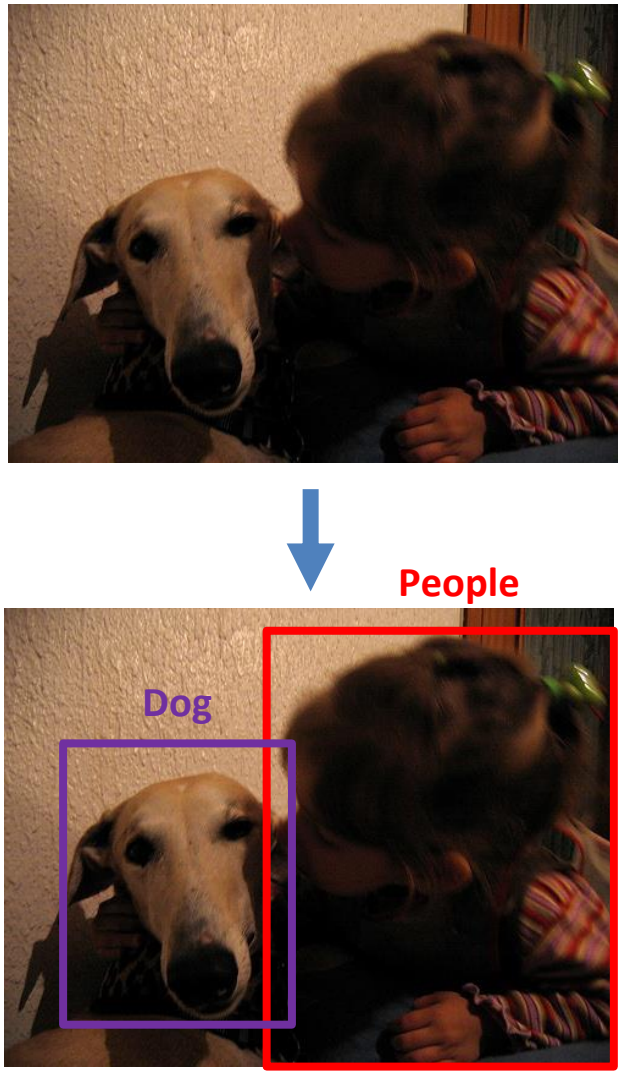
Video: Classification

Which action?

Cricket Bowling

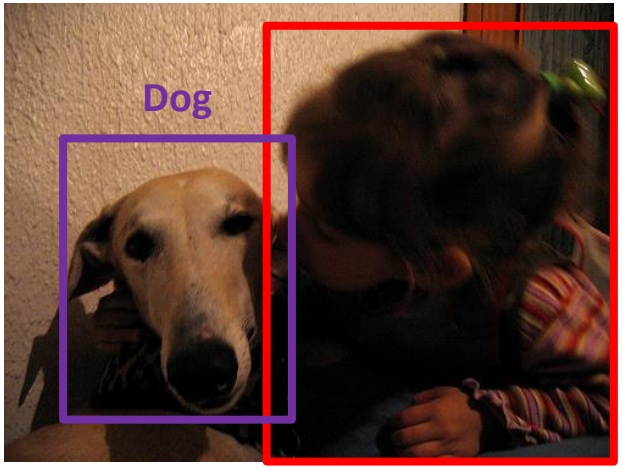
Problem Definition

Image: Object Detection



Problem Definition

Image: Object Detection



Video: Temporal Localization

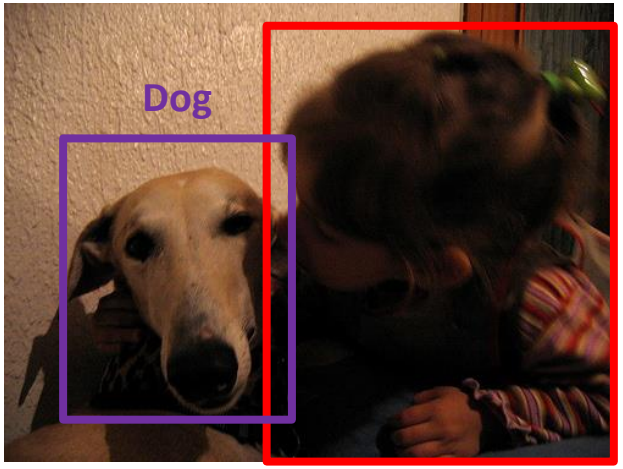
Background | Background | Background

1. Which action?
2. When does each action start/end?

Cricket Bowling | Cricket Bowling

Problem Definition

Image: Object Detection



Video: Temporal Localization

Background | Background | Background

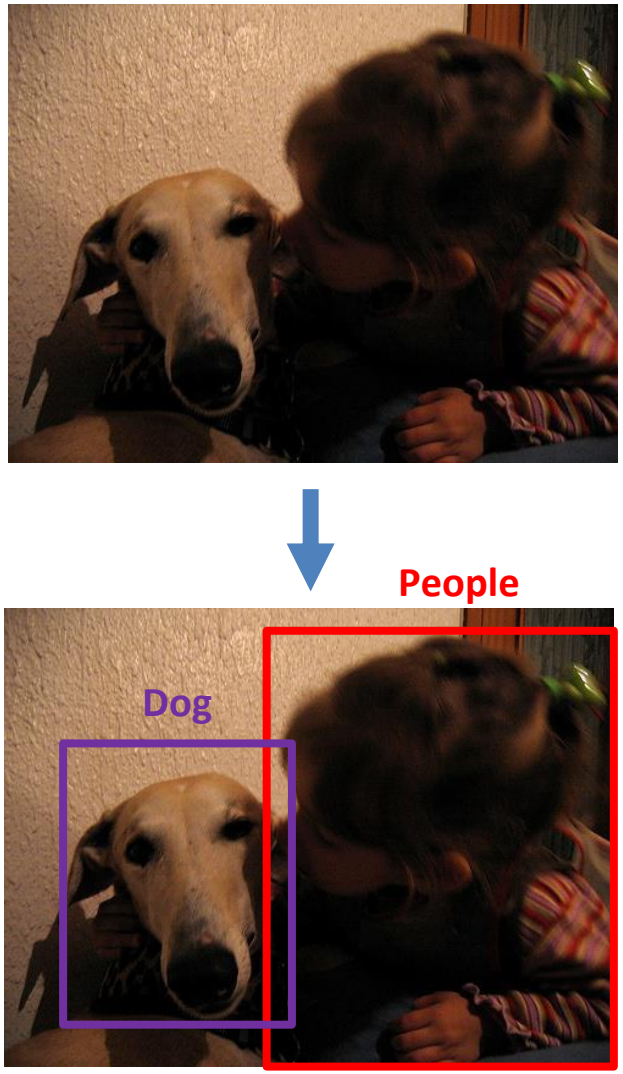
1. Which action?
2. When does each action start/end?

Cricket Bowling

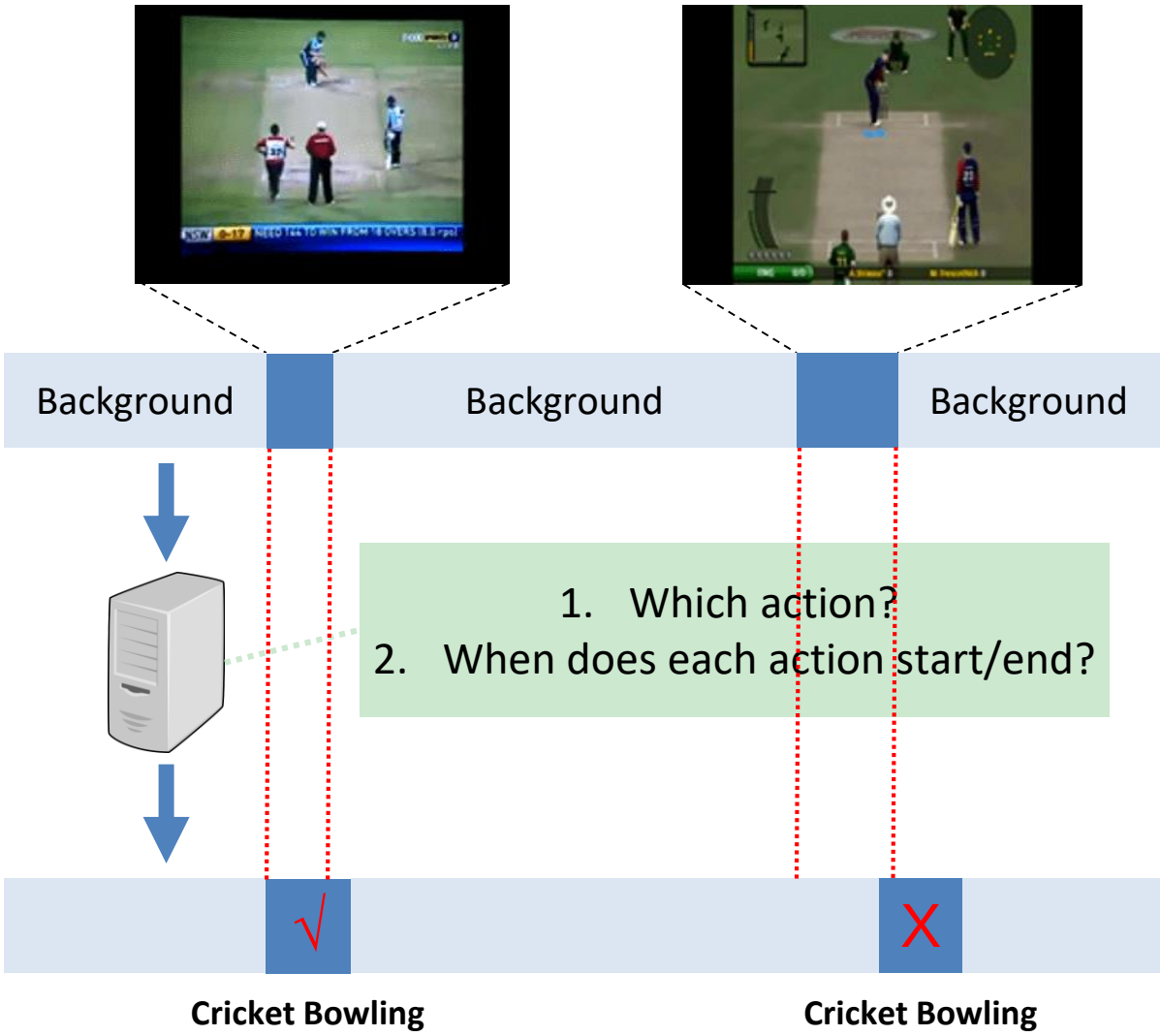
Cricket Bowling

Problem Definition

Image: Object Detection

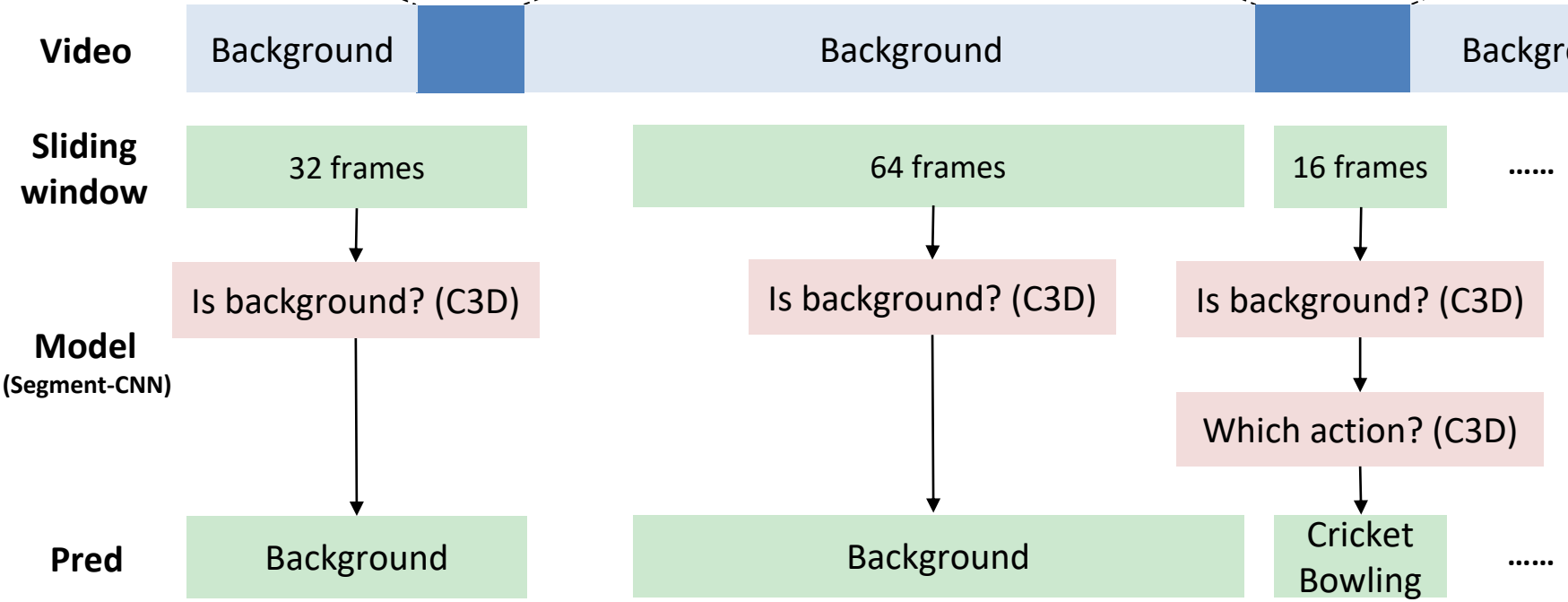


Video: Temporal Localization



Related work

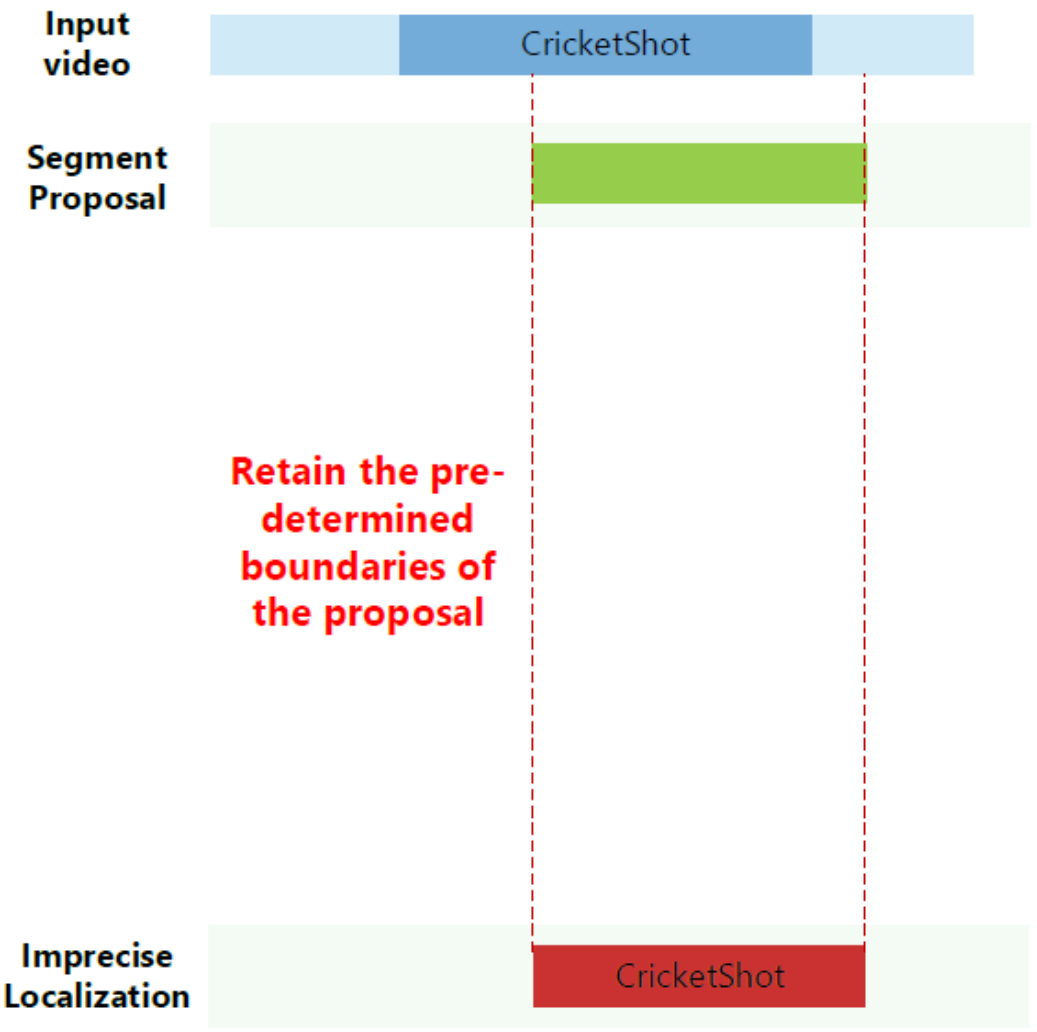
- **How do conventional methods address temporal localization?**



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

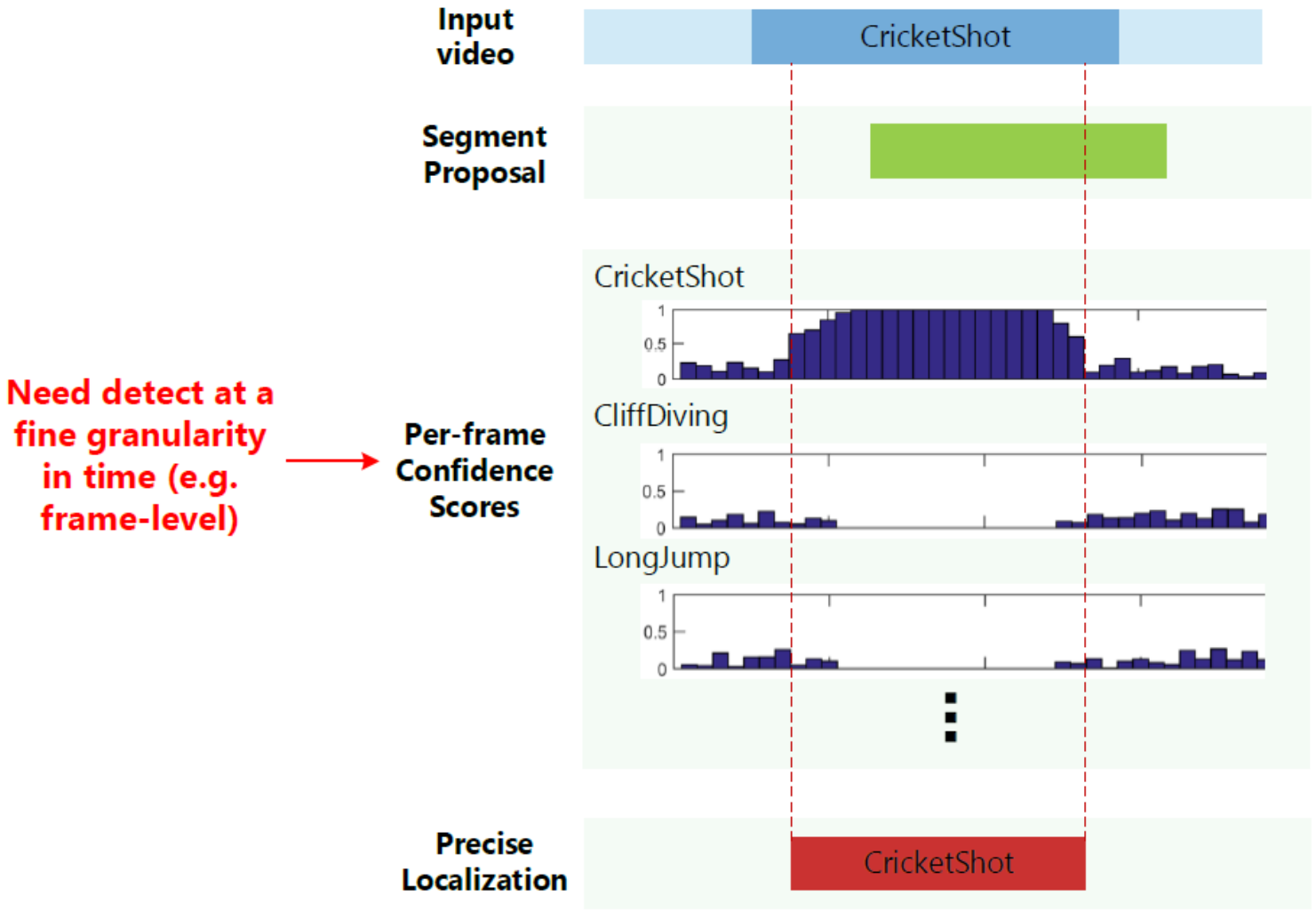
Motivation

- **Can we achieve more precise localization?**



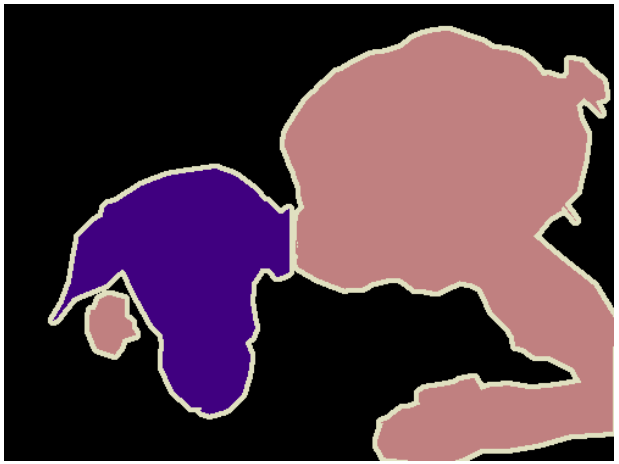
Motivation

- **How can we achieve more precise localization?**



Problem Definition

Image: Segmentation



Background People Dog

Video: Per-frame Labeling

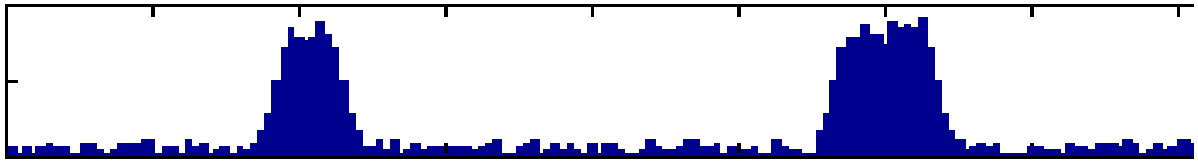
Cricket Bowling

Cricket Bowling

Background | Background | Background

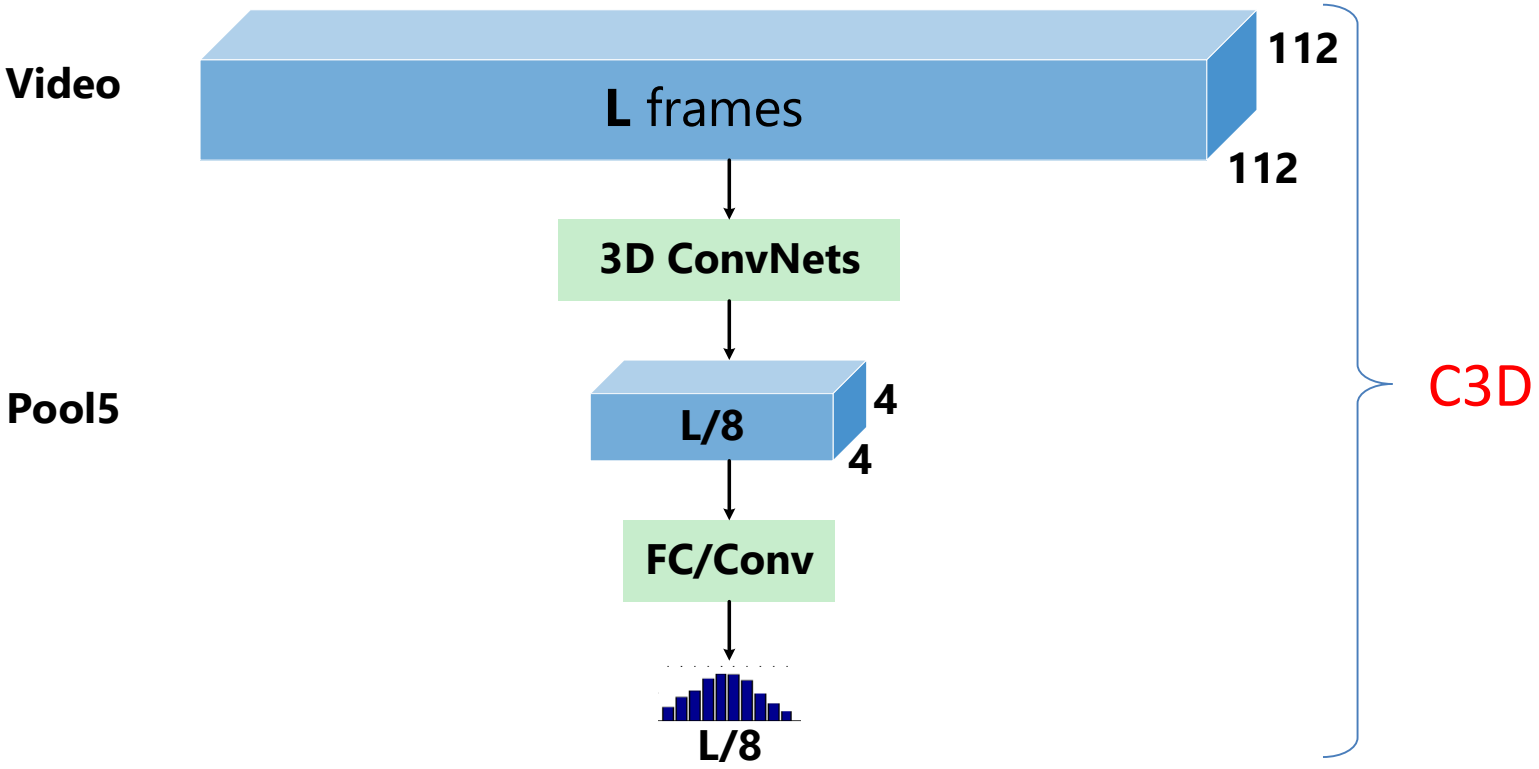
Which action contained in every frame?

Confidence score of detecting Cricket Bowling at every frame



Solution

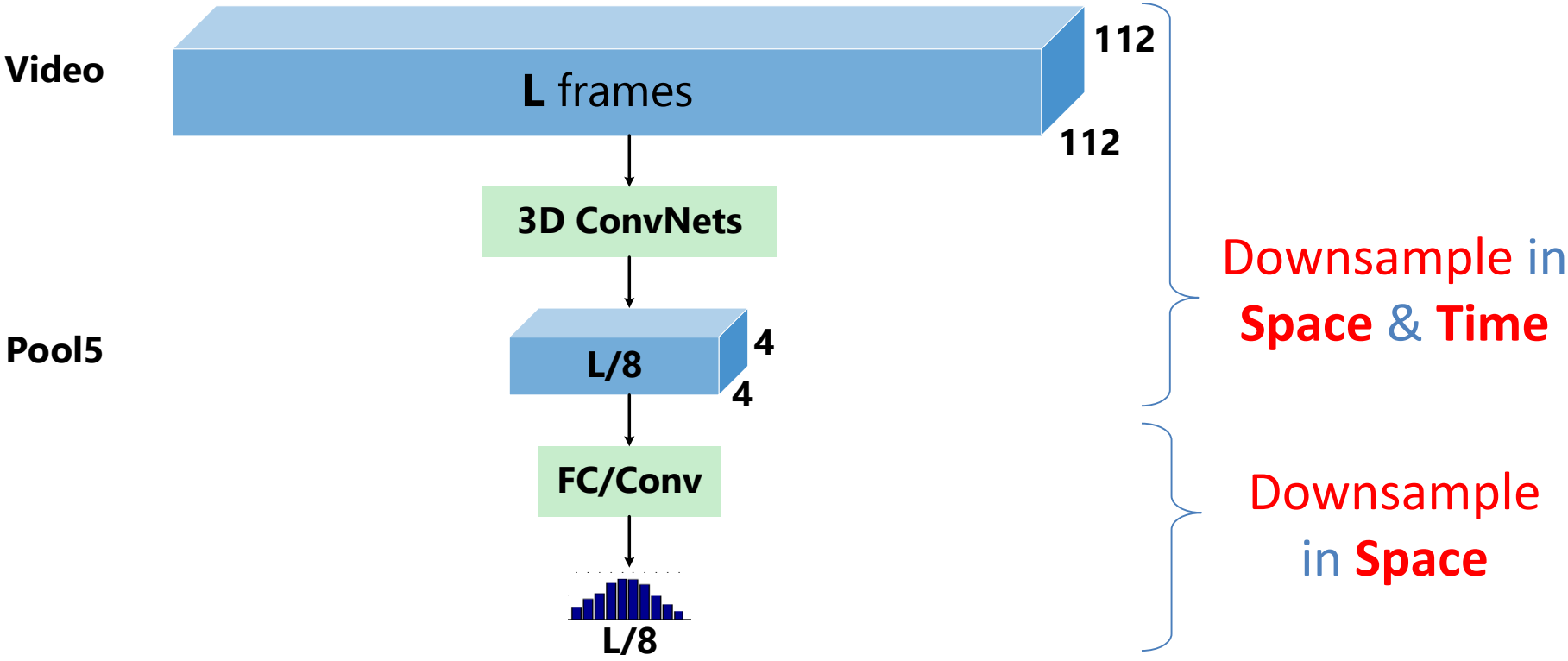
- How can we use C3D to perform per-frame labeling?



Learning Spatiotemporal Features with 3D Convolutional Networks
Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. In ICCV'15.

Solution

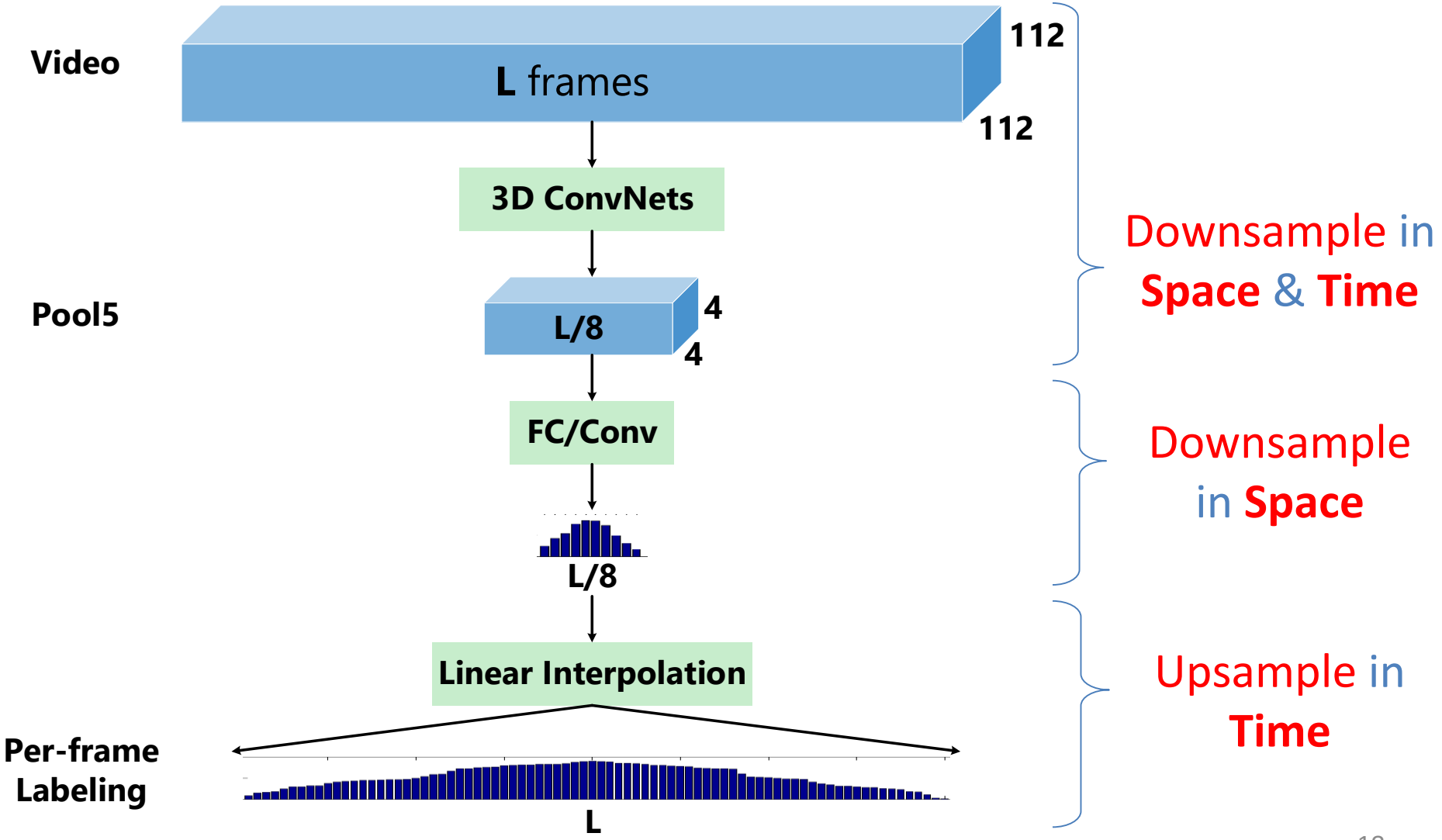
- **How can we use C3D to perform per-frame labeling?**



Learning Spatiotemporal Features with 3D Convolutional Networks
Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. In ICCV'15.

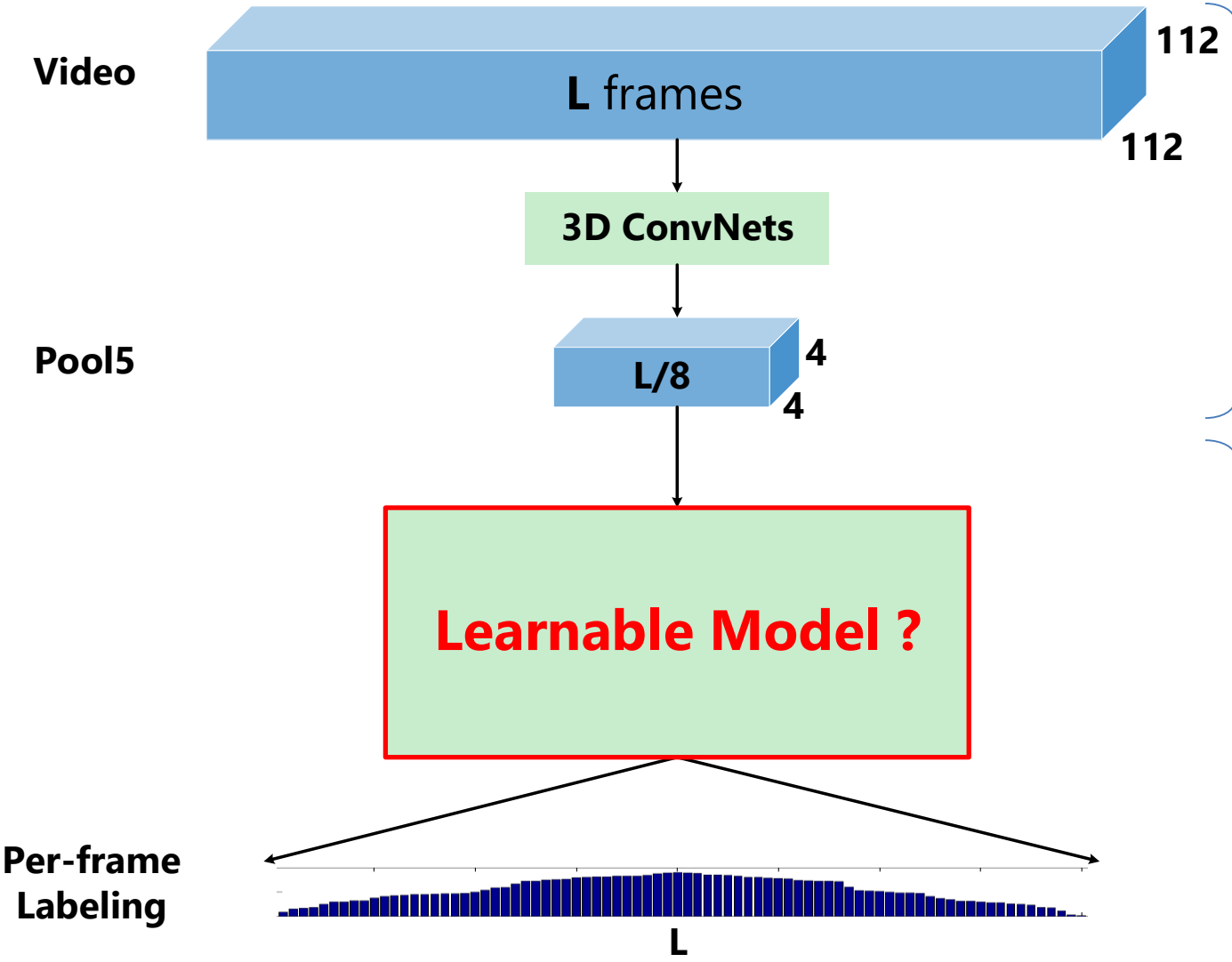
Solution

- How can we use C3D to perform per-frame labeling?



Solution

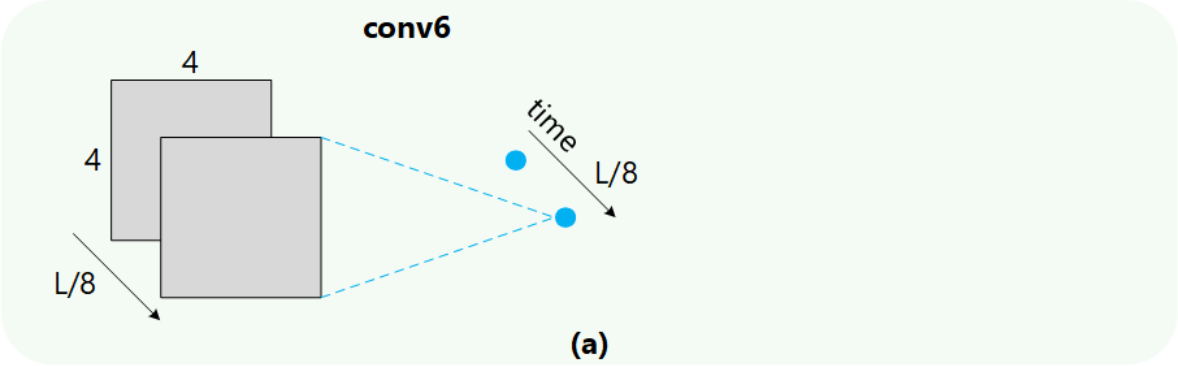
- How can we use C3D to perform per-frame labeling?



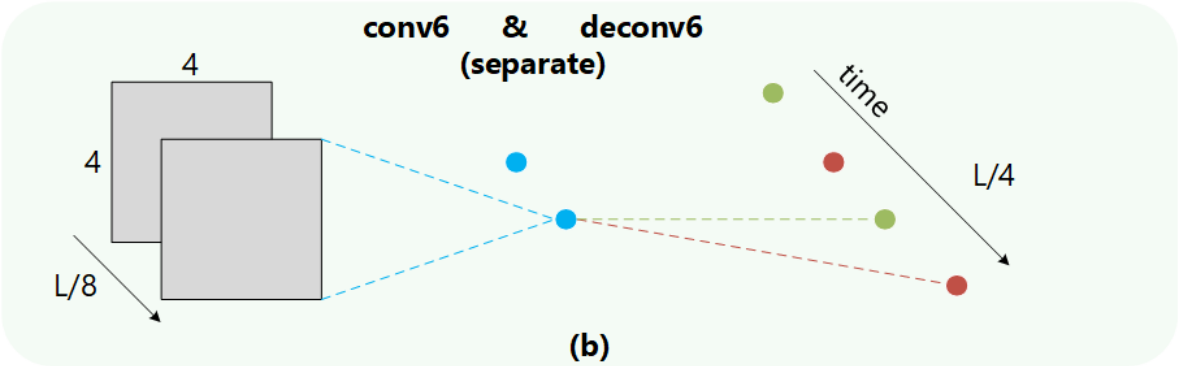
Solution

- **How can we Downsample in S and Upsample in T?**

C3D



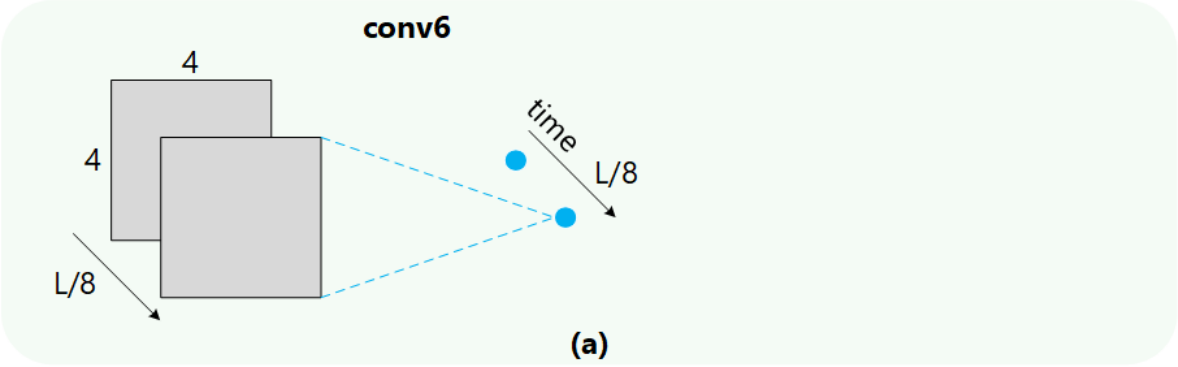
De-conv



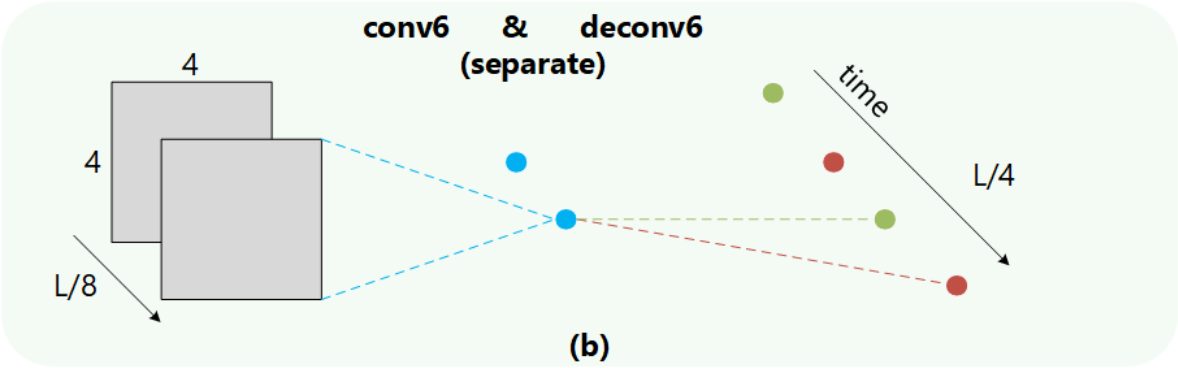
Solution

- **How can we Downsample in S and Upsample in T?**

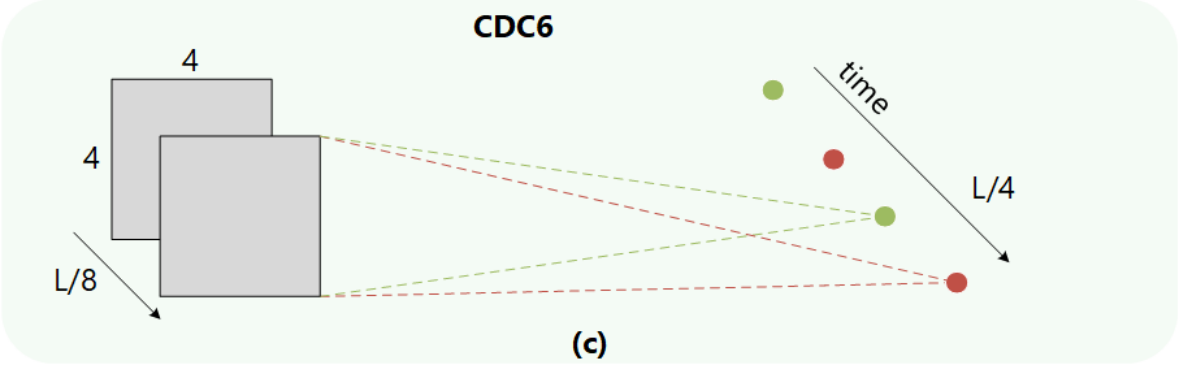
C3D



De-conv

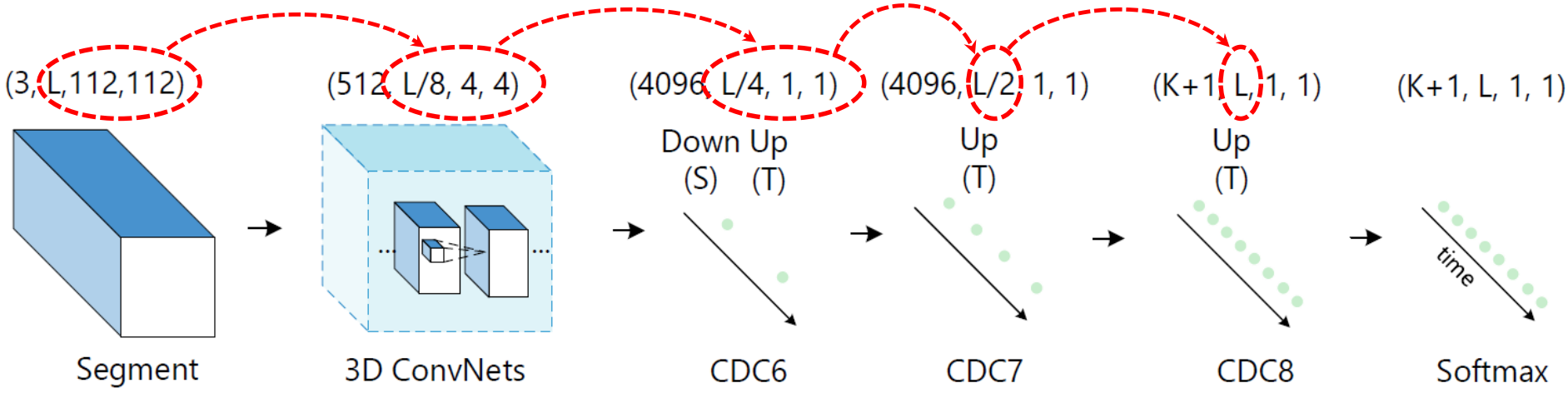


CDC



Network Architecture

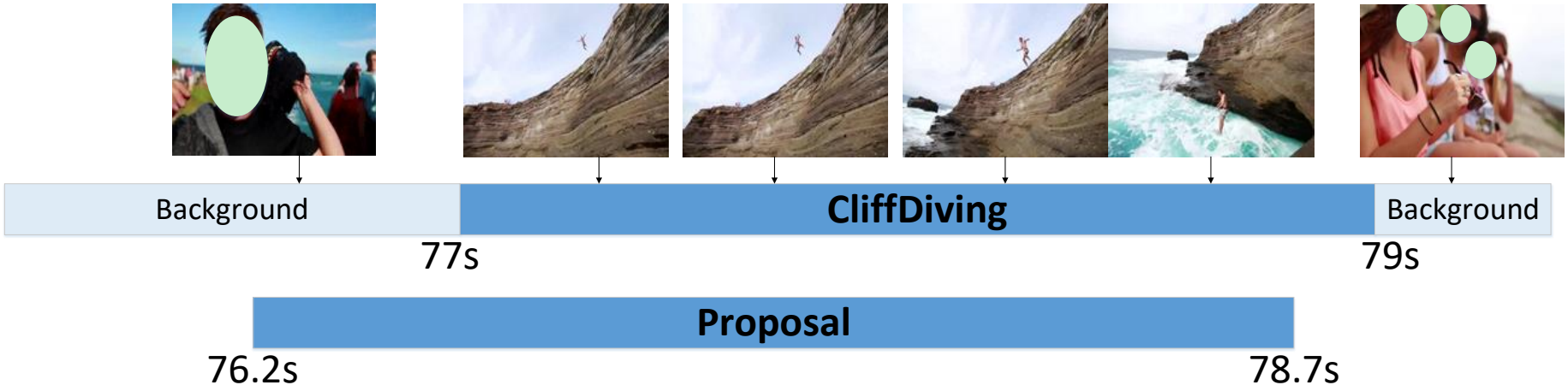
- Data dim (#channels, temporal length, height, width)
- Input video of length L. K action classes + background class



- Loss Function: Frame-wise Softmax Loss

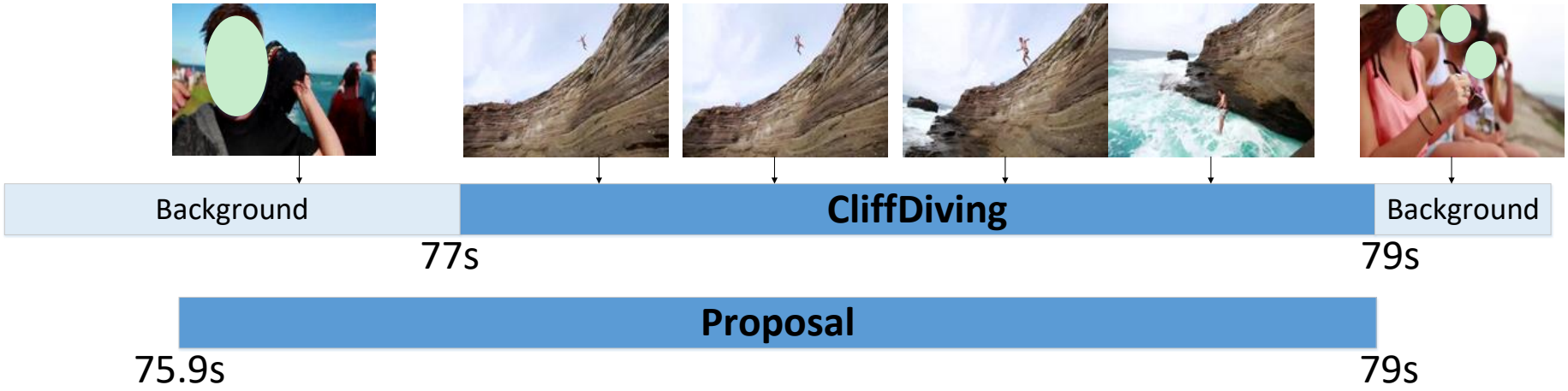
Visualization Example

- **How to refine temporal boundary?**



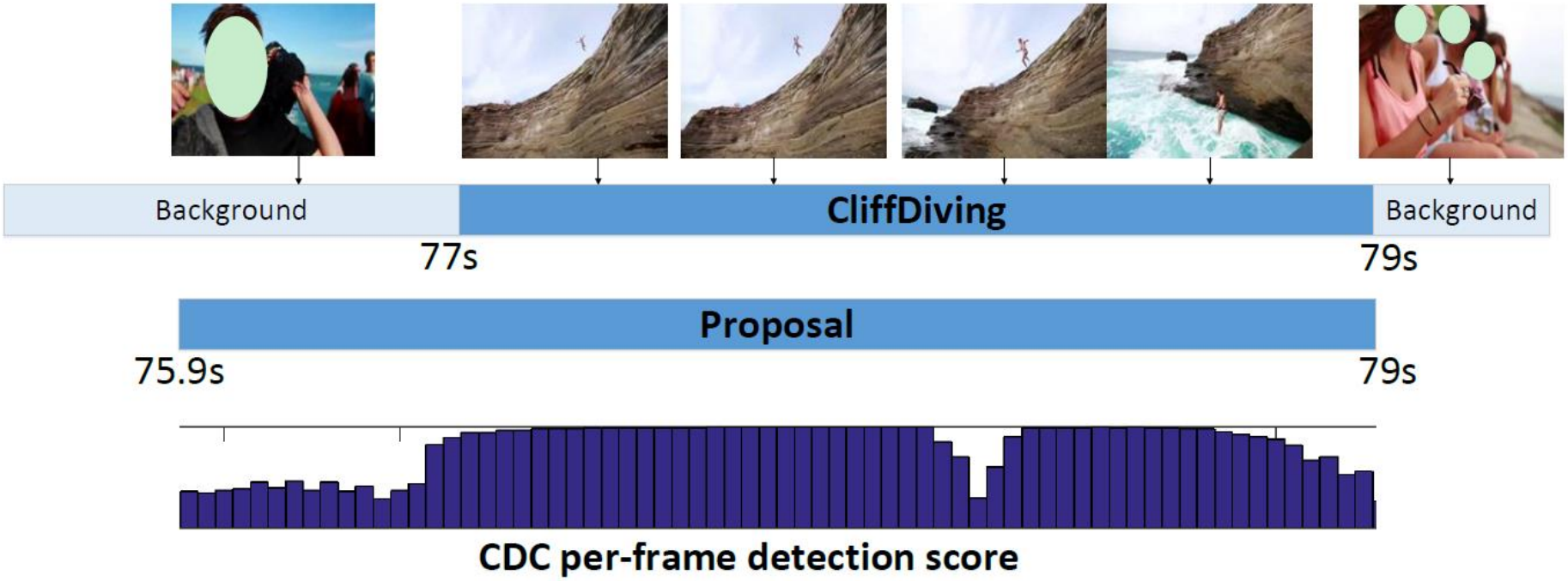
Visualization Example

- **How to refine temporal boundary?**



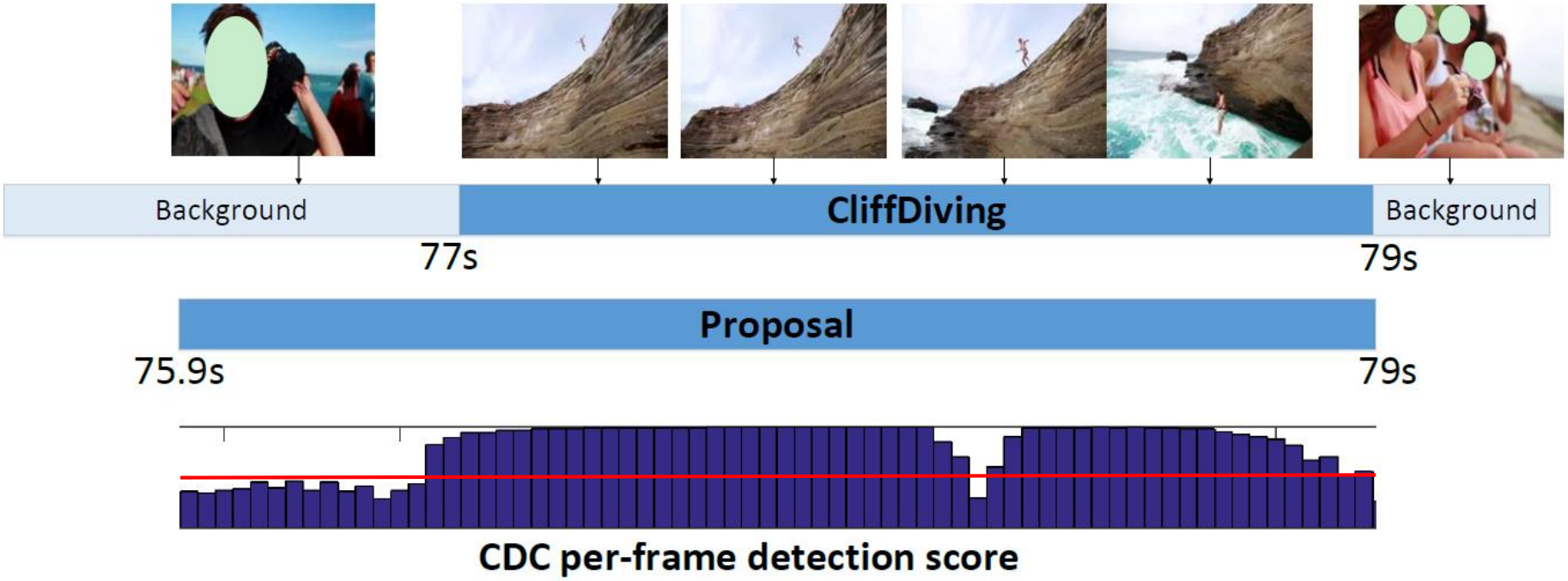
Visualization Example

- **How to refine temporal boundary?**



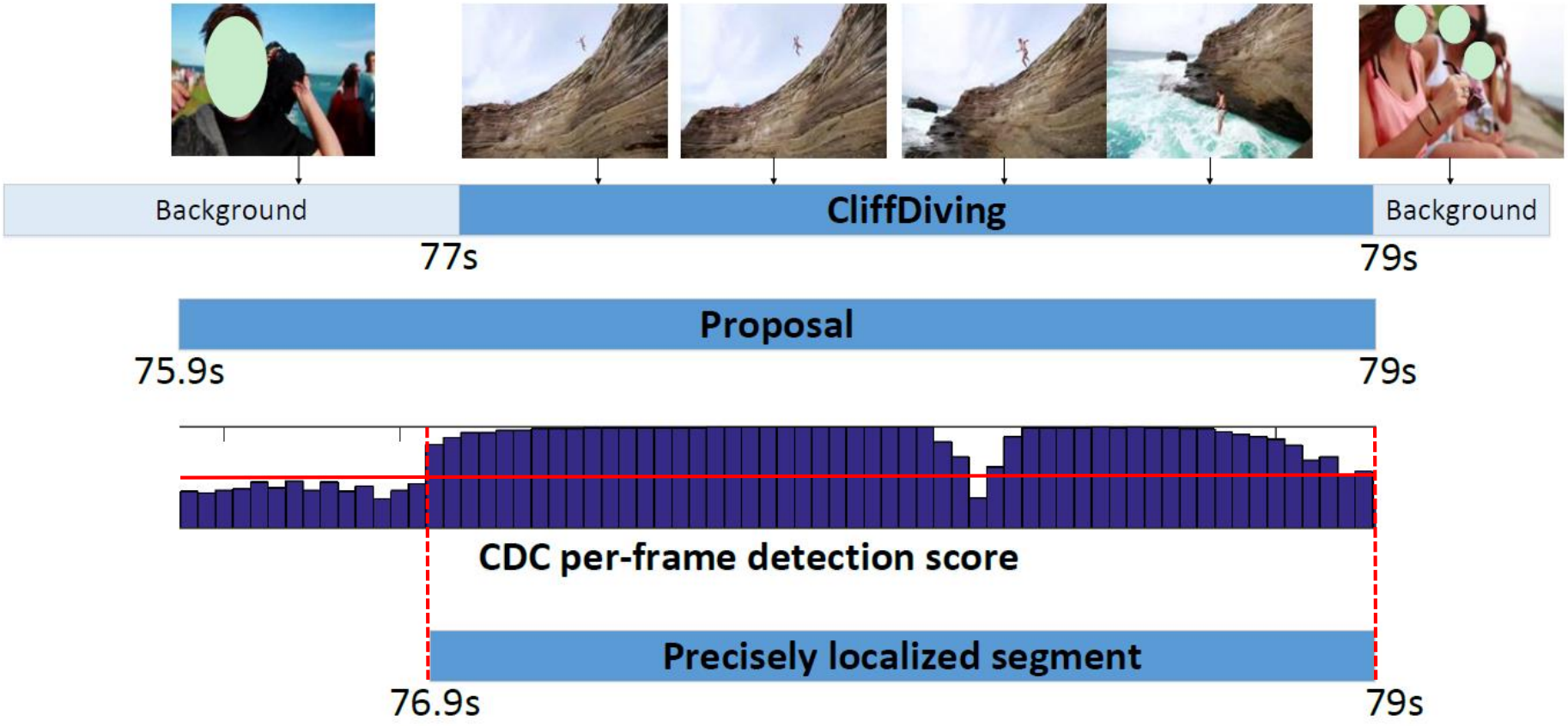
Visualization Example

- **How to refine temporal boundary?**



Visualization Example

- **How to refine temporal boundary?**



Experiments

- **Quantitative evaluation** on THUMOS'14:
  - Training data:
    - 3K trimmed short videos from UCF101
    - 200 untrimmed long videos from YouTube. 3K action instances
  - Test data: 200 untrimmed long videos. 3K action instances
  - Statistics:
    - Around 15 instances per video
    - Time duration of instances are diverse (from <1s to >20s)

THUMOS challenge: Action recognition with a large number of classes. http://crcv.ucf.edu/THUMOS14/.
Y.-G. Jiang, J. Liu, A. R. Zamir, G. Toderici, I. Laptev, M. Shah, and R. Sukthankar.

Q1: Per-frame Labeling

- Task: Predict label for every frame

- Evaluation: for each class, compute AP over all frames. Then compute mAP over 20 action classes.

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 on THUMOS'14 .

Q2: Temporal Action Localization

- Task: Predict a set of segments with label and start/end time

- Evaluation: mean Average Precision over 20 actions on THUMOS'14:

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**

Q2: Temporal Action Localization

- Task: Predict a set of segments with label and start/end time

- mean Average Precision over 200 activities on AcitivityNet challenge 2016:

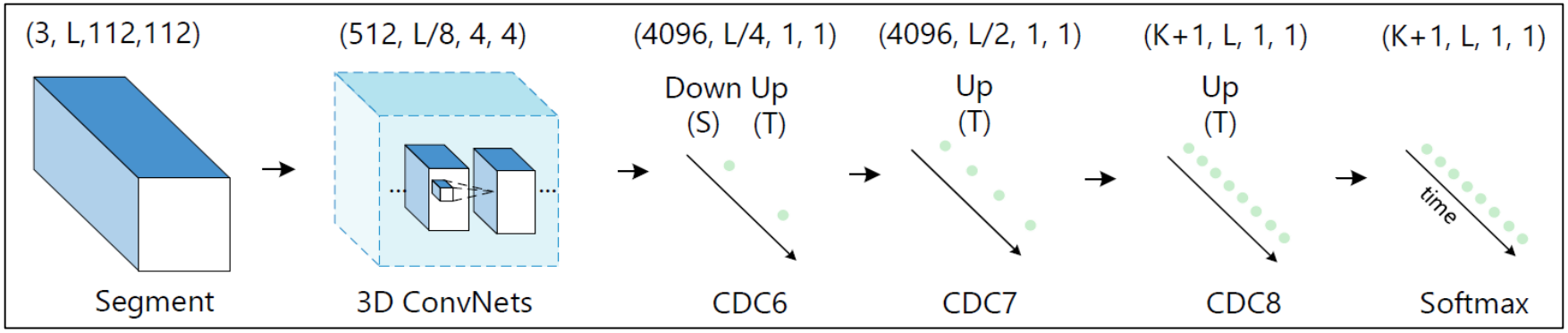
IoU threshold	0.5	0.75	0.95	Ave-mAP
Singh and Cuzzolin	22.7	10.8	0.3	11.3
Singh	26.0	15.2	2.6	14.6
Wang and Tao	45.1	4.1	0.0	16.4
CDC	45.3	26.0	0.2	23.8

Results on the 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

Results on the test set

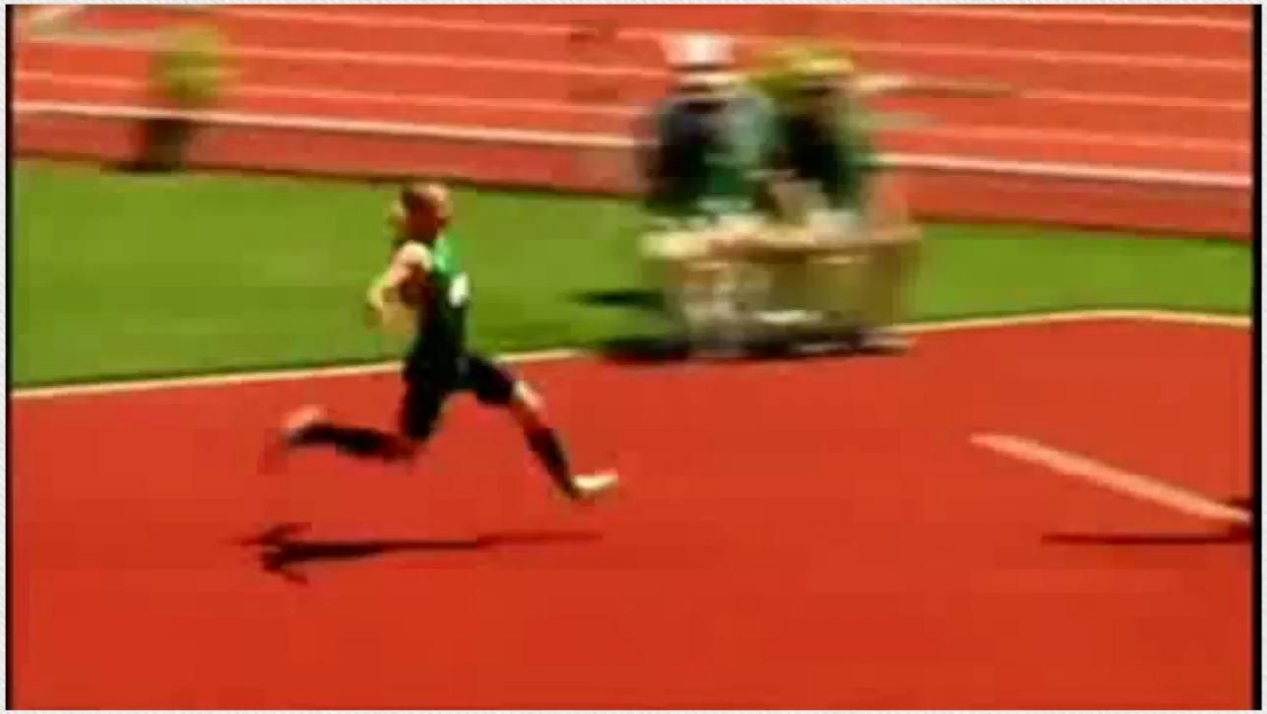
Q3: Efficiency



- Storage:
  - CDC is end-to-end. No need to cache intermediate features
  - A typical CDC network requires around 1GB storage

- Speed:
  - On Titan X GPU of 12GB memory, CDC runs at around 500 Frames Per Second

Visualization examples



[Visualization examples credit to Y. Wang]

Conclusions

- **Frame-level detection:**
  - Precise localization in untrimmed video

- **Conv-De-Conv: (simultaneously)**
  - Down-sample in Space (semantic abstraction)
  - Up-sample in Time (precise localization)

- **Extensive experiments:**
  - Per-frame labeling
  - Temporal localization
  - Efficiency (1GB storage and 500FPS speed)

Thank you! Please come to our poster at #36

Paper: https://arxiv.org/abs/1703.01515

Project: http://www.ee.columbia.edu/ln/dvmm/researchProjects/cdc

Code: https://bitbucket.org/columbiadvmm/cdc

References I

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

- D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In ICCV, 2015.

- J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In CVPR, 2015.

- Y.-G. Jiang, J. Liu, A. R. Zamir, G. Toderici, I. Laptev, M. Shah, and R. Sukthankar. THUMOS challenge: Action recognition with a large number of classes. In ECCV Workshop, 2014.

- F. C. Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In CVPR, 2015.

References II

- R. Wang and D. Tao. Uts at activitynet 2016. In CVPR ActivityNet Workshop, 2016.
- S. Karaman, L. Seidenari, and A. D. Bimbo. Fast saliency based pooling of fisher encoded dense trajectories. In ECCV THUMOS Workshop, 2014.
- L. Wang, Y. Qiao, and X. Tang. Action recognition and detection by combining motion and appearance features. In ECCV THUMOS Workshop, 2014.
- F. C. Heilbron, J. C. Niebles, and B. Ghanem. Fast temporal activity proposals for efficient detection of human actions in untrimmed videos. In CVPR, 2016.
- V. Escorcia, F. C. Heilbron, J. C. Niebles, and B. Ghanem. Daps: Deep action proposals for action understanding. In ECCV, 2016.
- D. Oneata, J. Verbeek, and C. Schmid. The lear submission at thumos 2014. In ECCV THUMOS Workshop, 2014.

References III

- A. Richard and J. Gall. Temporal action detection using a statistical language model. In CVPR, 2016.
- S. Yeung, O. Russakovsky, G. Mori, and L. Fei-Fei. End-to-end learning of action detection from frame glimpses in videos. In CVPR, 2016.
- J. Yuan, B. Ni, X. Yang, and A. Kassim. Temporal action localization with pyramid of score distribution features. In CVPR, 2016.
- A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In CVPR, 2014.
- M. Zeiler, D. Krishnan, G.W. Taylor, and R. Fergus. Deconvolutional networks. In CVPR, 2010.
- R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In CVPR, 2014