Active Perception

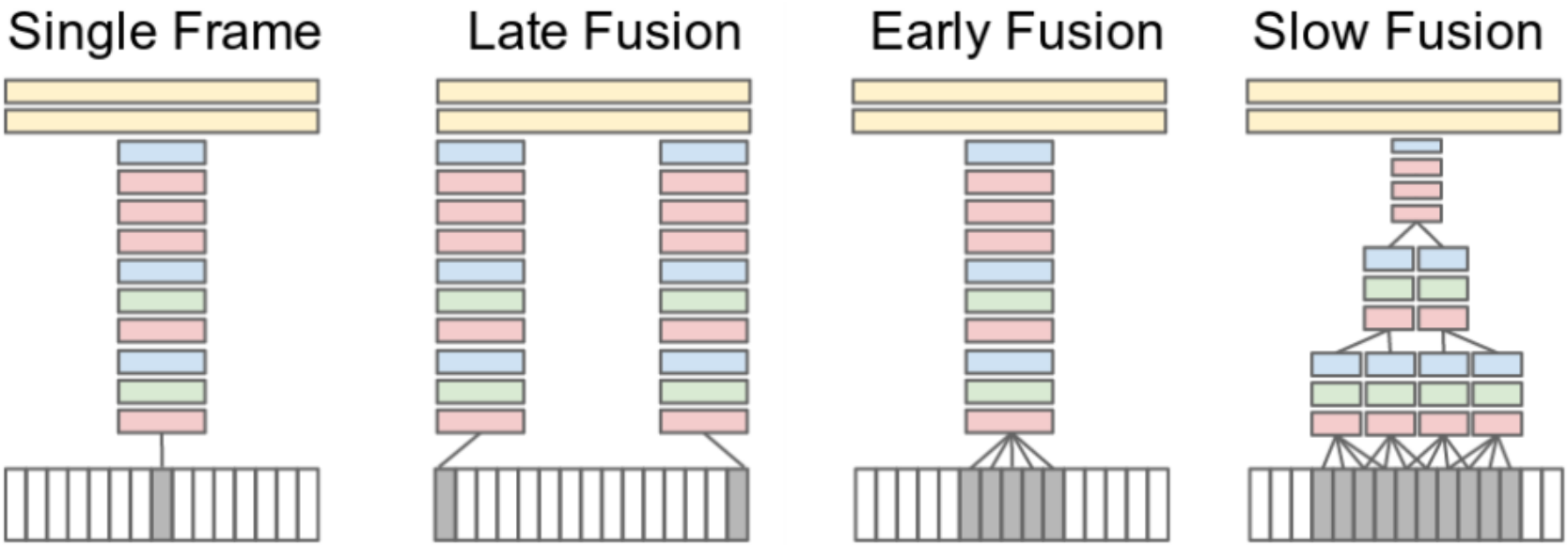
Jia-Bin Huang

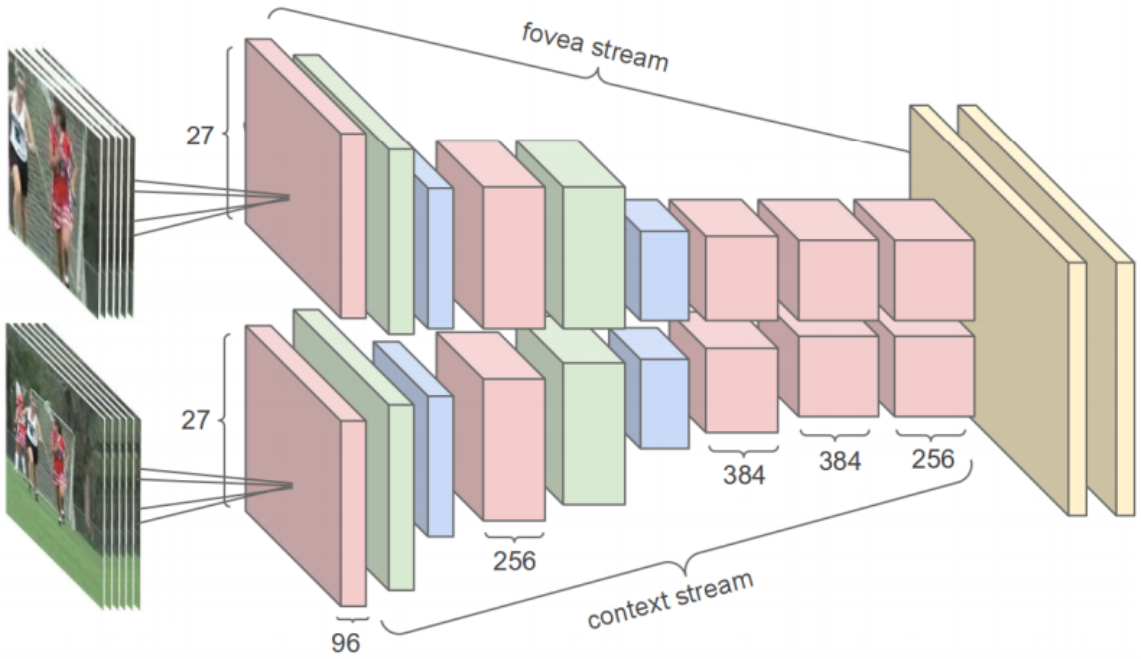
Virginia Tech

ECE 6554 Advanced Computer Vision

Today's class

- Review action recognition
- Topic presentation by Shruti
- Paper discussion
  - Kevin ("for" discussion lead)
  - Ashish ("against" discussion lead)
- Next lecture: Group of objects by Jia-Bin

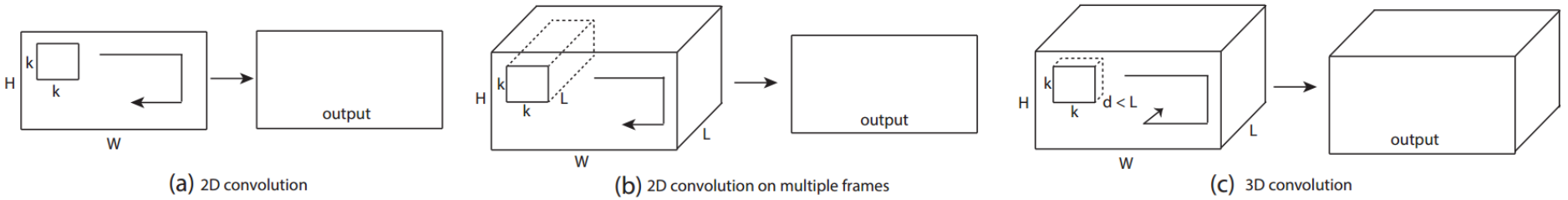


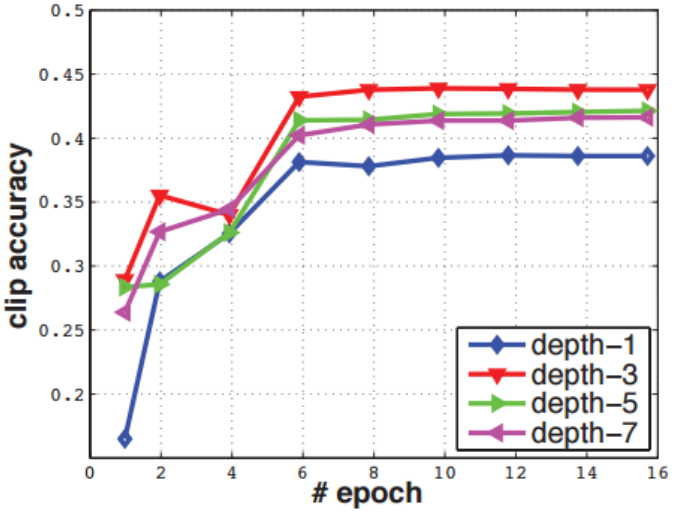


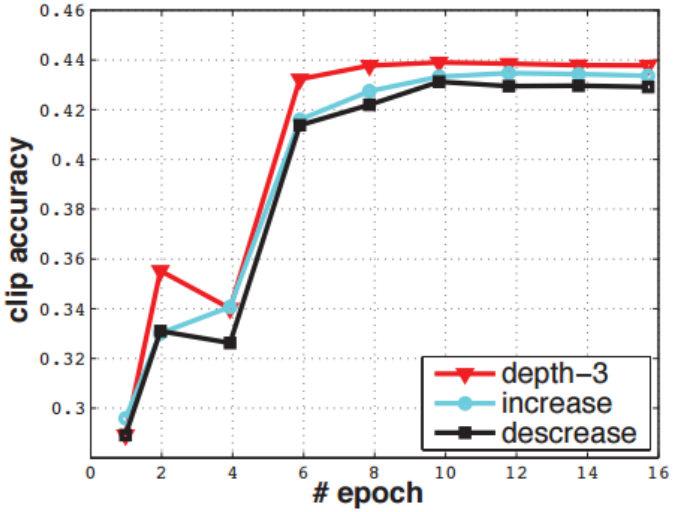
Large-scale Video Classification with Convolutional Neural Networks, CVPR 2014

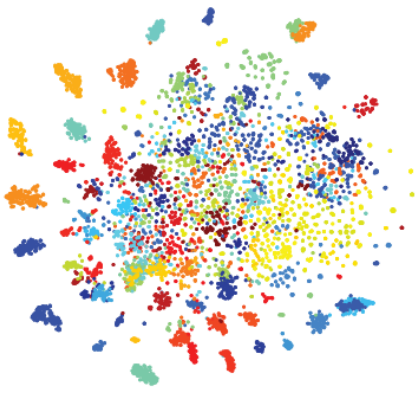
Model	Clip Hit@1	Video Hit@1	Video Hit@5
Feature Histograms + Neural Net	-	55.3	-
Single-Frame	41.1	59.3	77.7
Single-Frame + Multires	**42.4**	**60.0**	**78.5**
Single-Frame Fovea Only	30.0	49.9	72.8
Single-Frame Context Only	38.1	56.0	77.2
Early Fusion	38.9	57.7	76.8
Late Fusion	40.7	59.3	78.7
Slow Fusion	**41.9**	**60.9**	**80.2**
CNN Average (Single+Early+Late+Slow)	41.4	63.9	82.4

Large-scale Video Classification with Convolutional Neural Networks, CVPR 2014

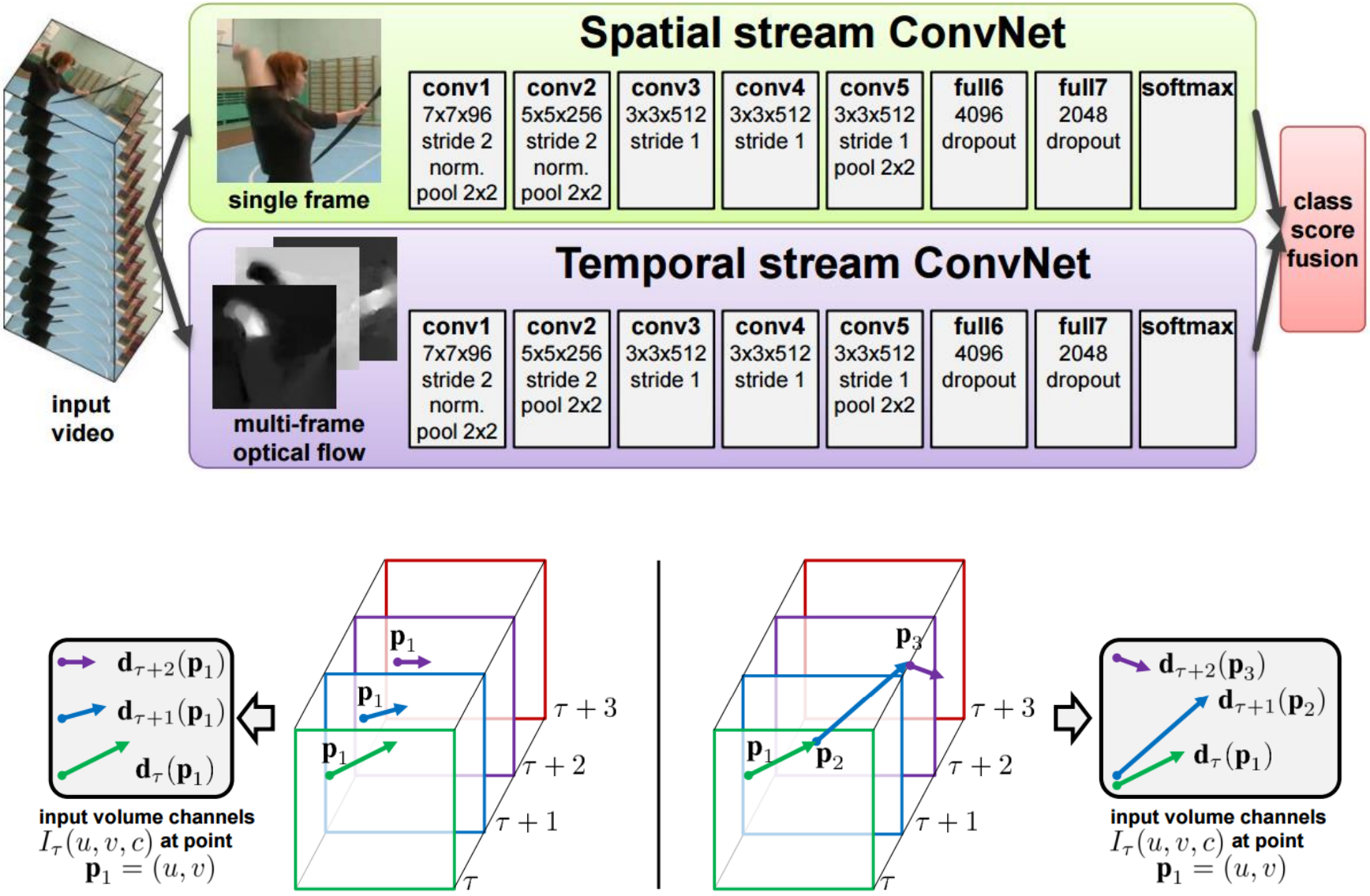








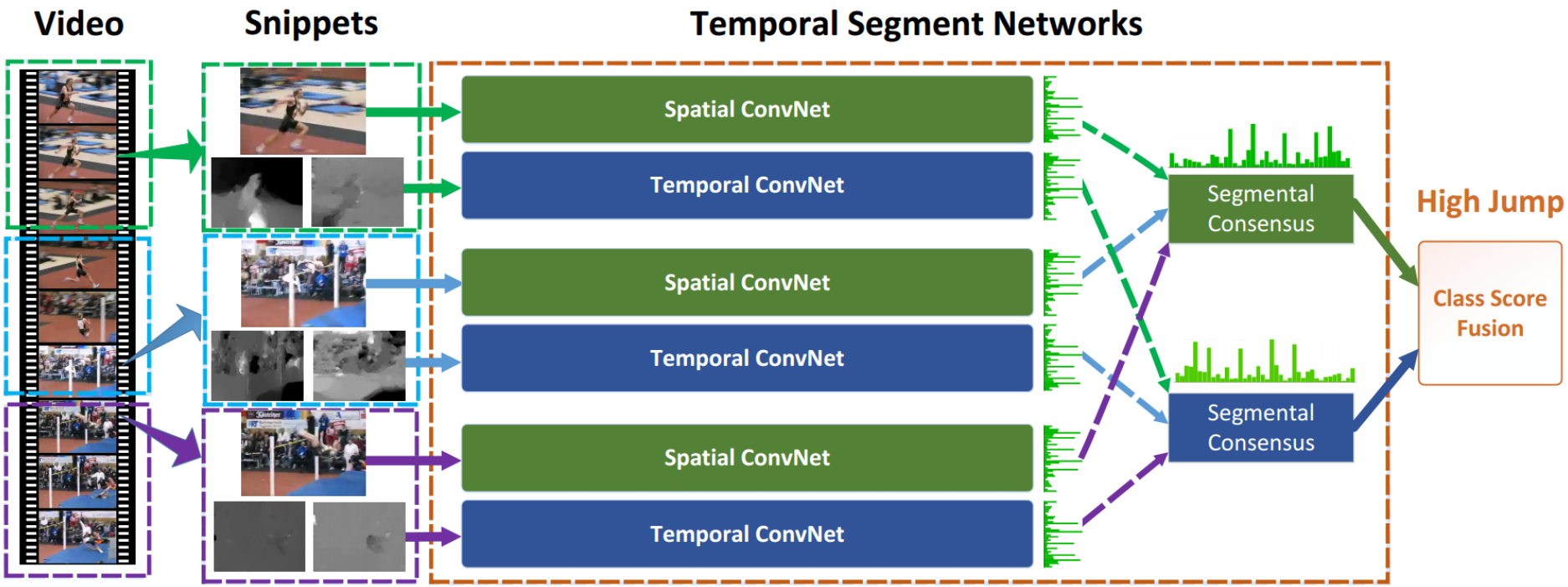
Learning Spatiotemporal Features with 3D Convolutional Networks, ICCV 2015



Two-Stream Convolutional Networks for Action Recognition in Videos, NIPS 2014

Method	UCF-101	HMDB-51
Improved dense trajectories (IDT) [26, 27]	85.9%	57.2%
IDT with higher-dimensional encodings [20]	**87.9%**	61.1%
IDT with stacked Fisher encoding [21] (based on Deep Fisher Net [23])	-	**66.8%**
Spatio-temporal HMAX network [11, 16]	-	22.8%
"Slow fusion" spatio-temporal ConvNet [14]	65.4%	-
Spatial stream ConvNet	73.0%	40.5%
Temporal stream ConvNet	83.7%	54.6%
Two-stream model (fusion by averaging)	86.9%	58.0%
Two-stream model (fusion by SVM)	**88.0%**	**59.4%**

Two-Stream Convolutional Networks for Action Recognition in Videos, NIPS 2014

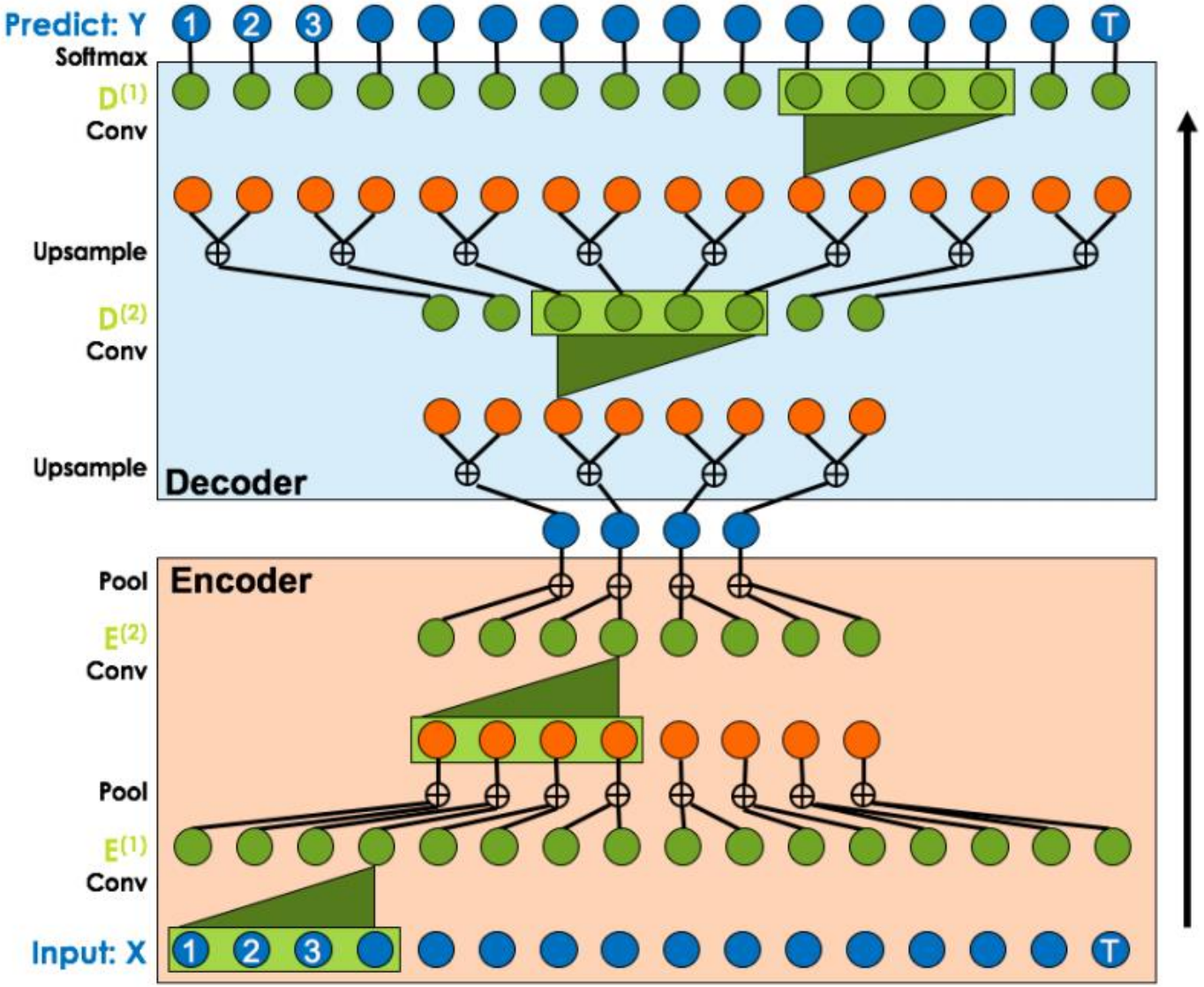


Temporal Segment Networks: Towards Good Practices for Deep Action Recognition, ECCV 2016

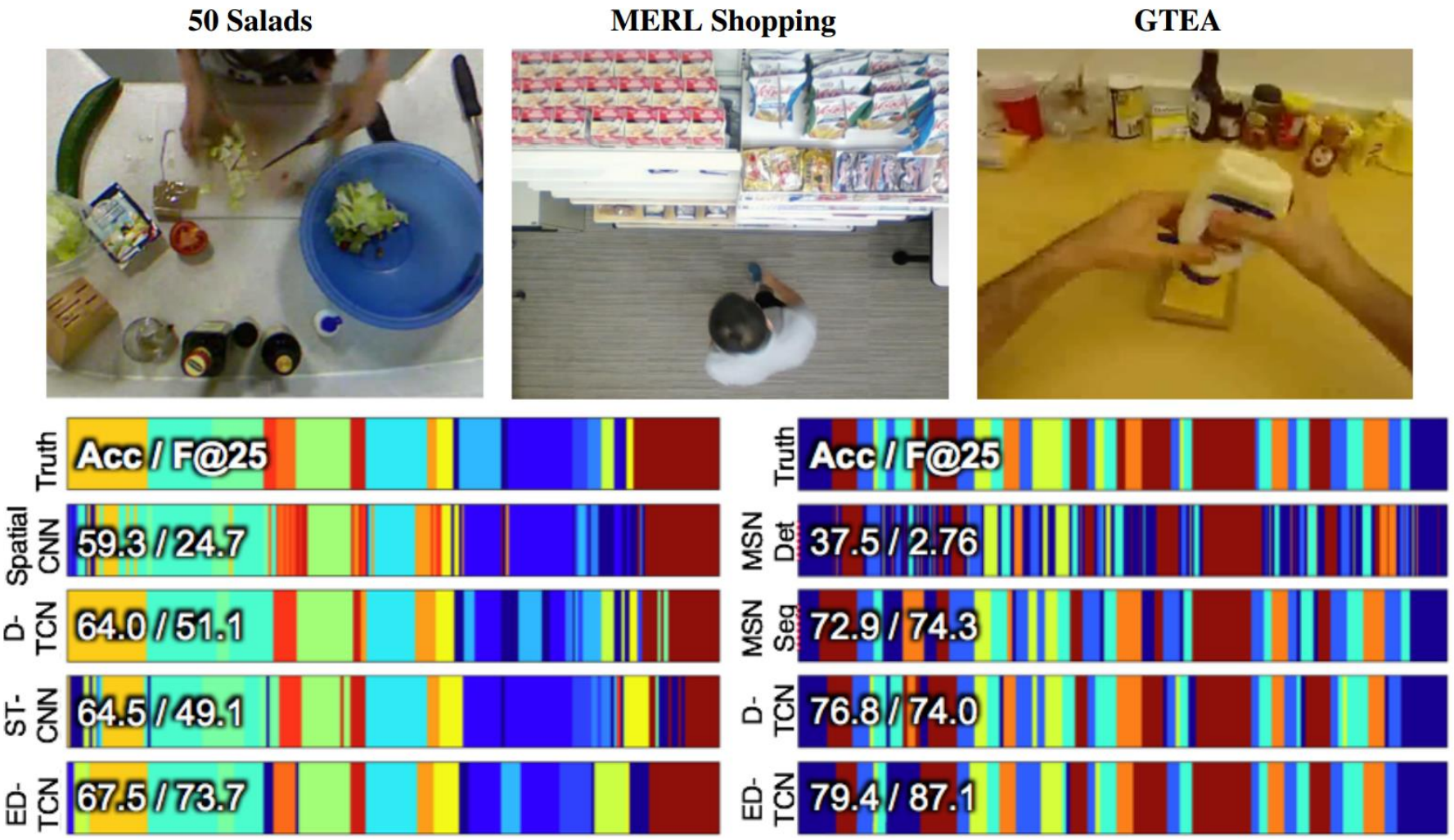
Training setting	Spatial ConvNets	Temporal ConvNets	Two-Stream
Baseline [1]	72.7%	81.0%	87.0%
From Scratch	48.7%	81.7%	82.9%
Pre-train Spatial(same as [1])	84.1%	81.7%	90.0%
+ Cross modality pre-training	84.1%	86.6%	91.5%
+ Partial BN with dropout	84.5%	87.2%	92.0%

Modality	Performance
RGB Image	84.5%
RGB Difference	83.8%
RGB Image + RGB Difference	87.3%
Optical Flow	87.2%
Warped Flow	86.9%
Optical Flow + Warped Flow	87.8%
Optical Flow + Warped Flow + RGB	**92.3%**
All Modalities	91.7%

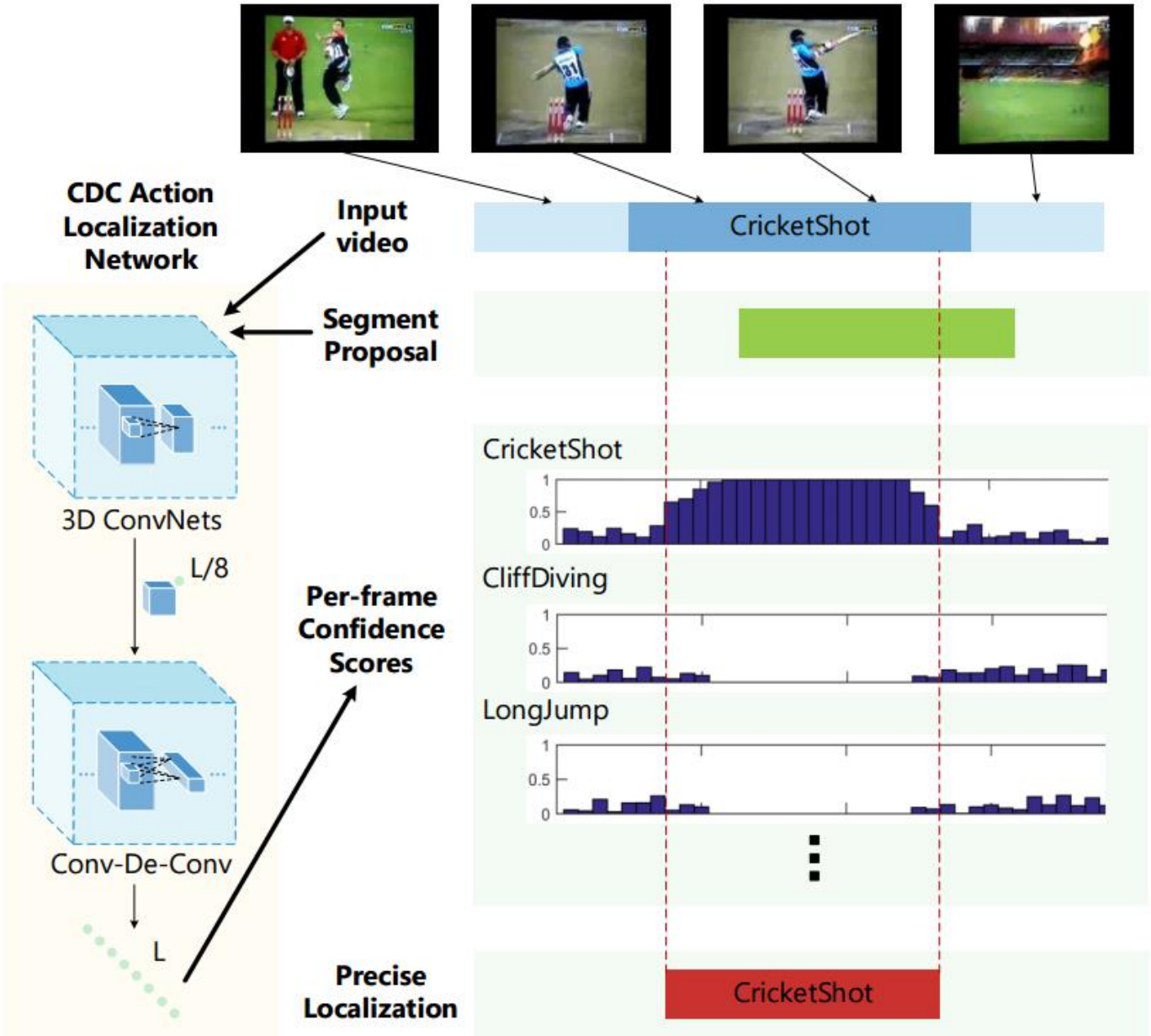
Temporal Segment Networks: Towards Good Practices for Deep Action Recognition, ECCV 2016



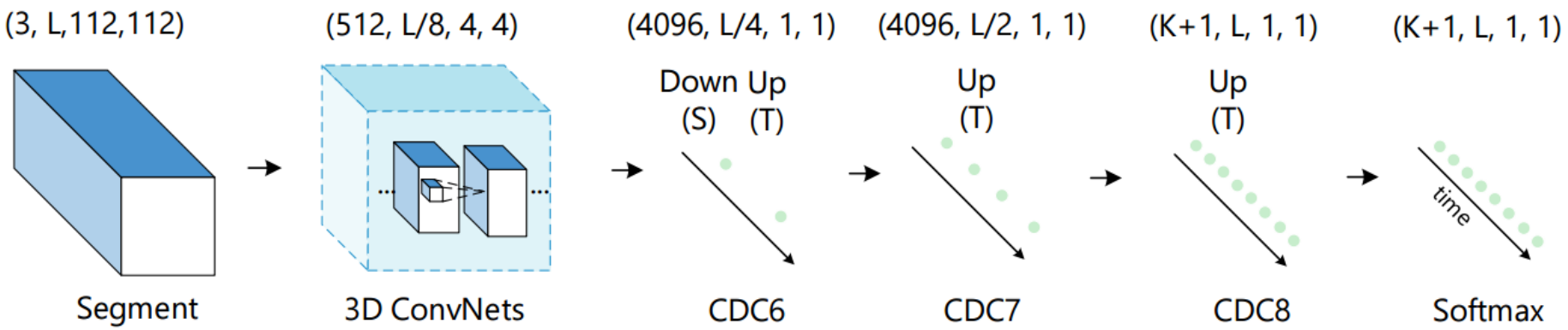
Temporal Convolutional Networks for Action Segmentation and Detection, CVPR 2017

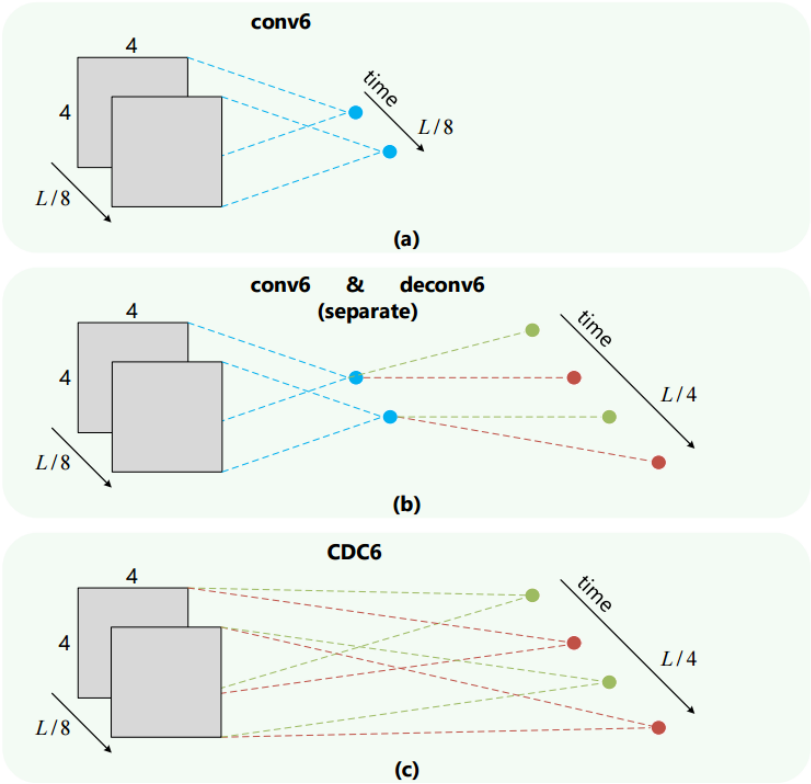


Temporal Convolutional Networks for Action Segmentation and Detection, CVPR 2017

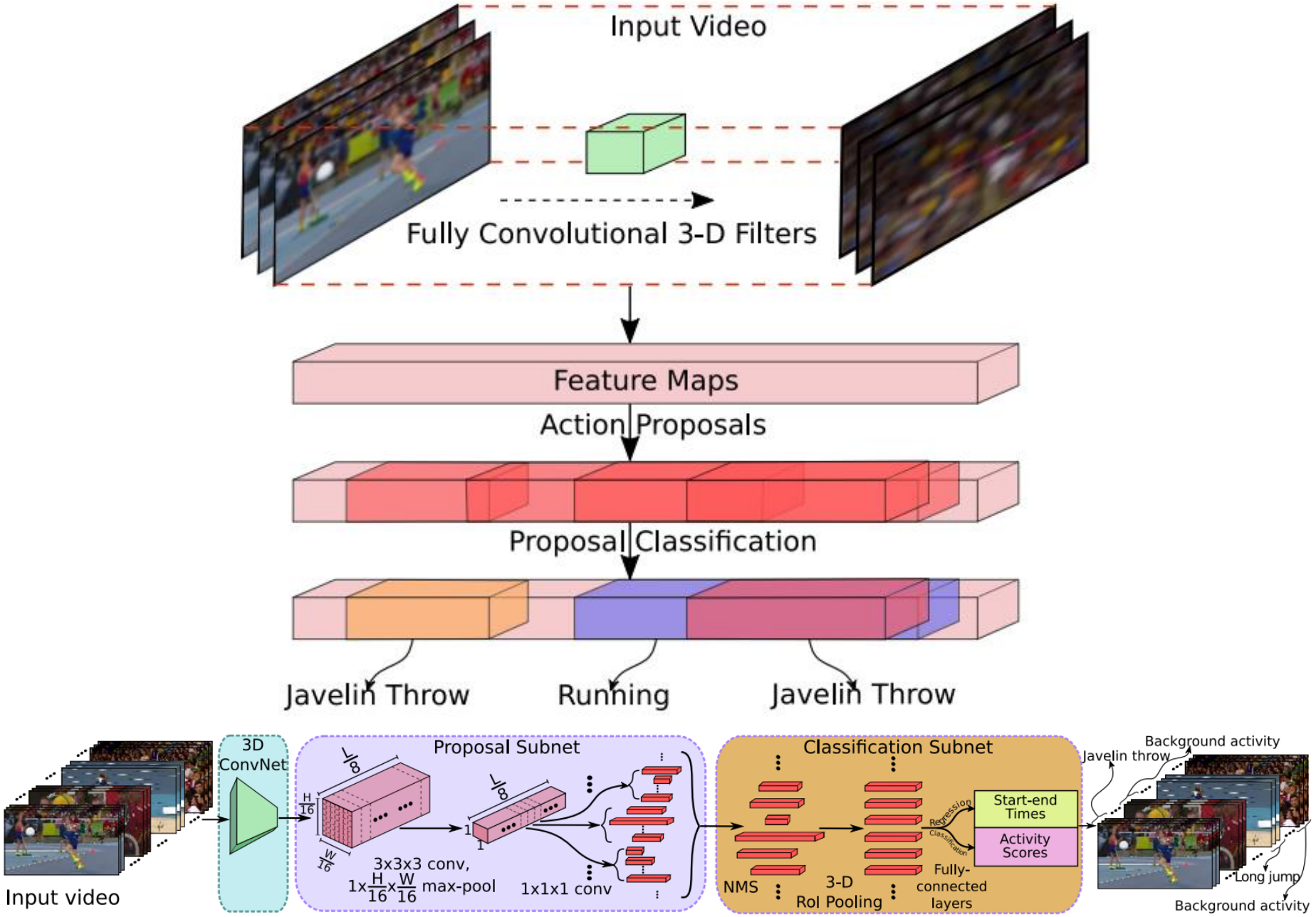


CDC: Convolutional-De-Convolutional Networks for Precise Temporal Action Localization in Untrimmed Videos, CVPR 2017

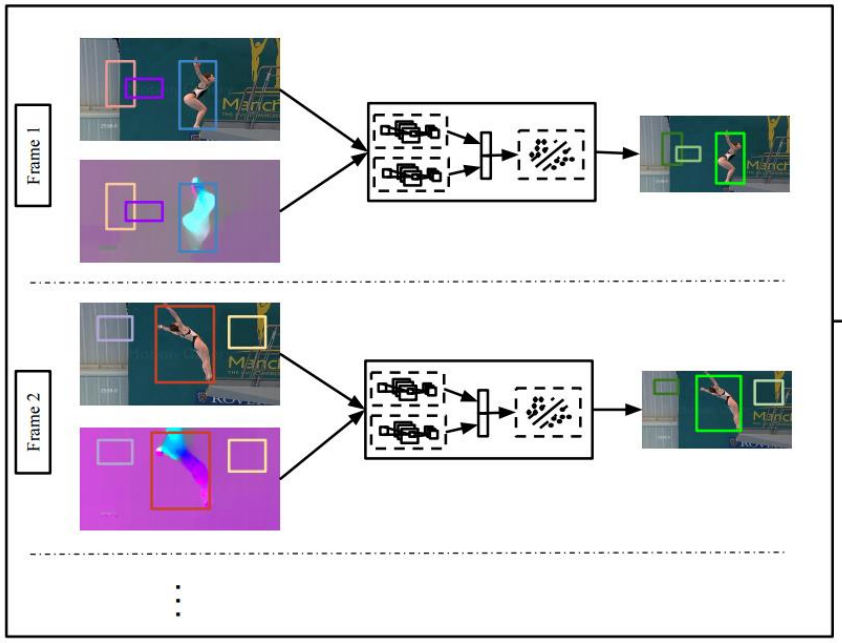


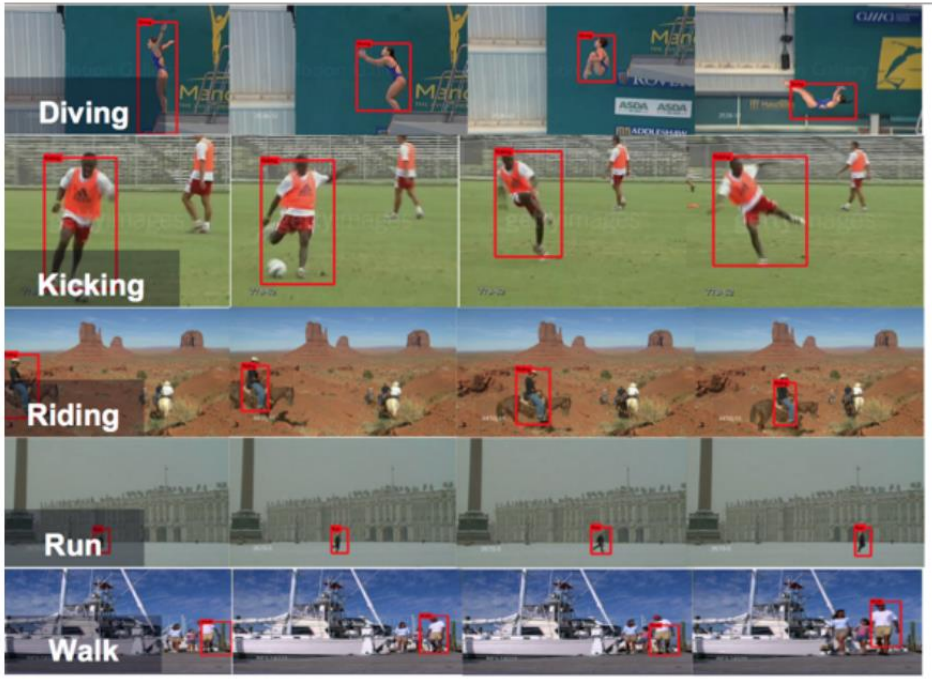


CDC: Convolutional-De-Convolutional Networks for Precise Temporal Action Localization in Untrimmed Videos, CVPR 2017

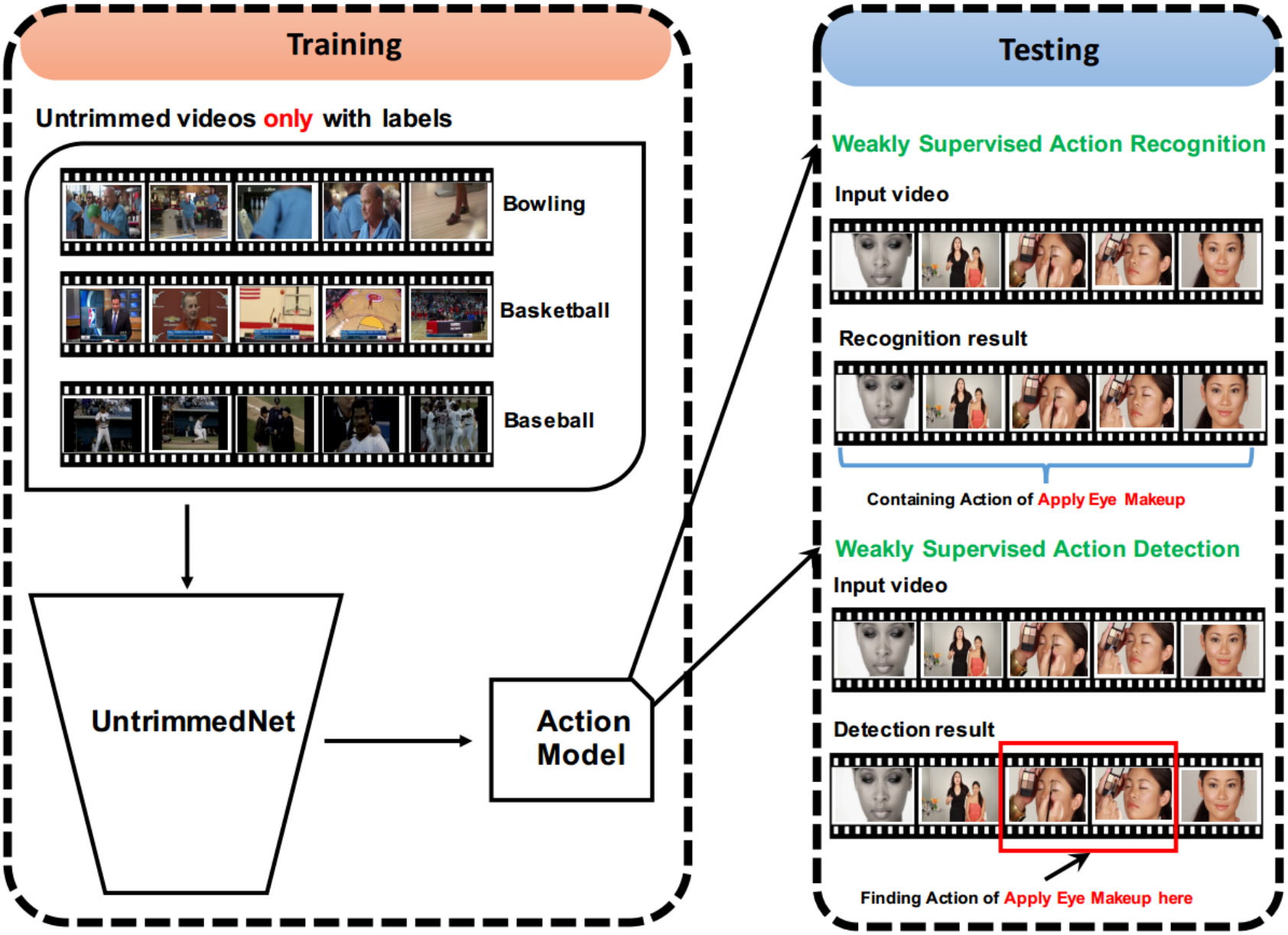


R-C3D: Region Convolutional 3D Network for Temporal Activity Detection, Arxiv 2017





Finding Action Tubes , ICCV 2015



UntrimmedNets for Weakly Supervised Action Recognition and Detection , CVPR 2017