S^3D: Single Shot multi-Span Detector via Fully 3D Convolutional Network

Da Zhang[1], Xiyang Dai[2], Xin Wang[1], and Yuan-Fang Wang[1]

dazhang@cs.ucsb.edu

[1]UC Santa Barbara & [2]University of Maryland

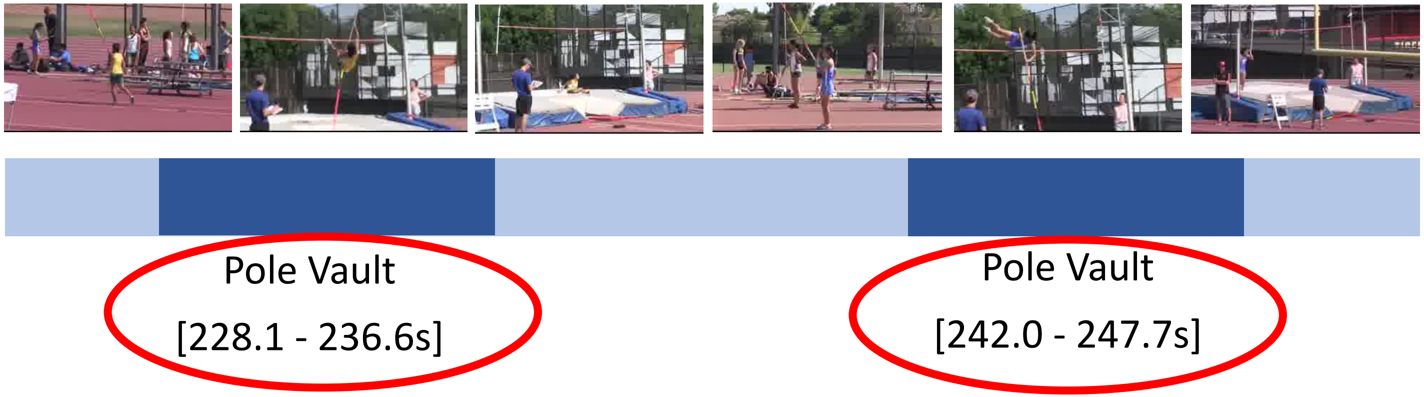
BMVC 2018
Newcastle upon Tyne

Task: Temporal Activity Detection

Input: untrimmed videos

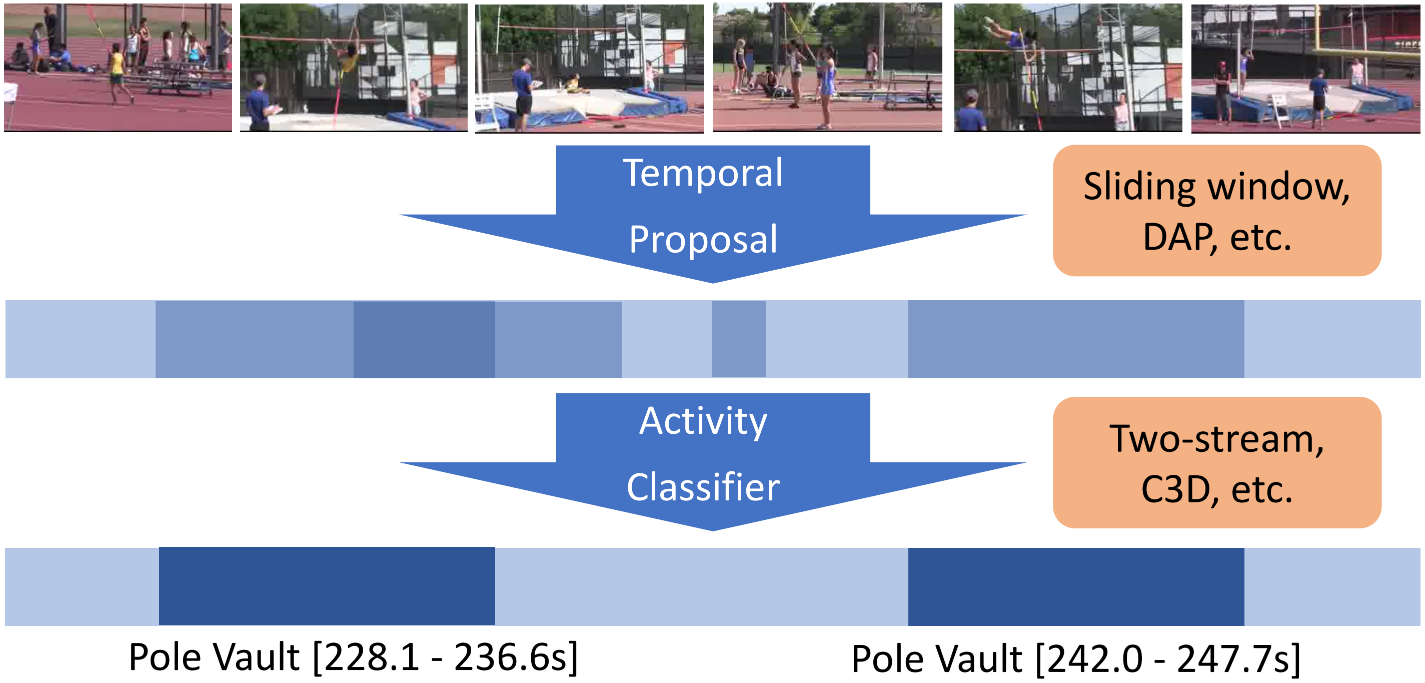
1. *Localization*: when do activities start/end?
2. *Classification*: what are the activities?

Detection Results



Related Works

Conventional two-stage approach: Proposal + Classification



Temporal Proposal

Sliding window, DAP, etc.

Activity Classifier

Two-stream, C3D, etc.

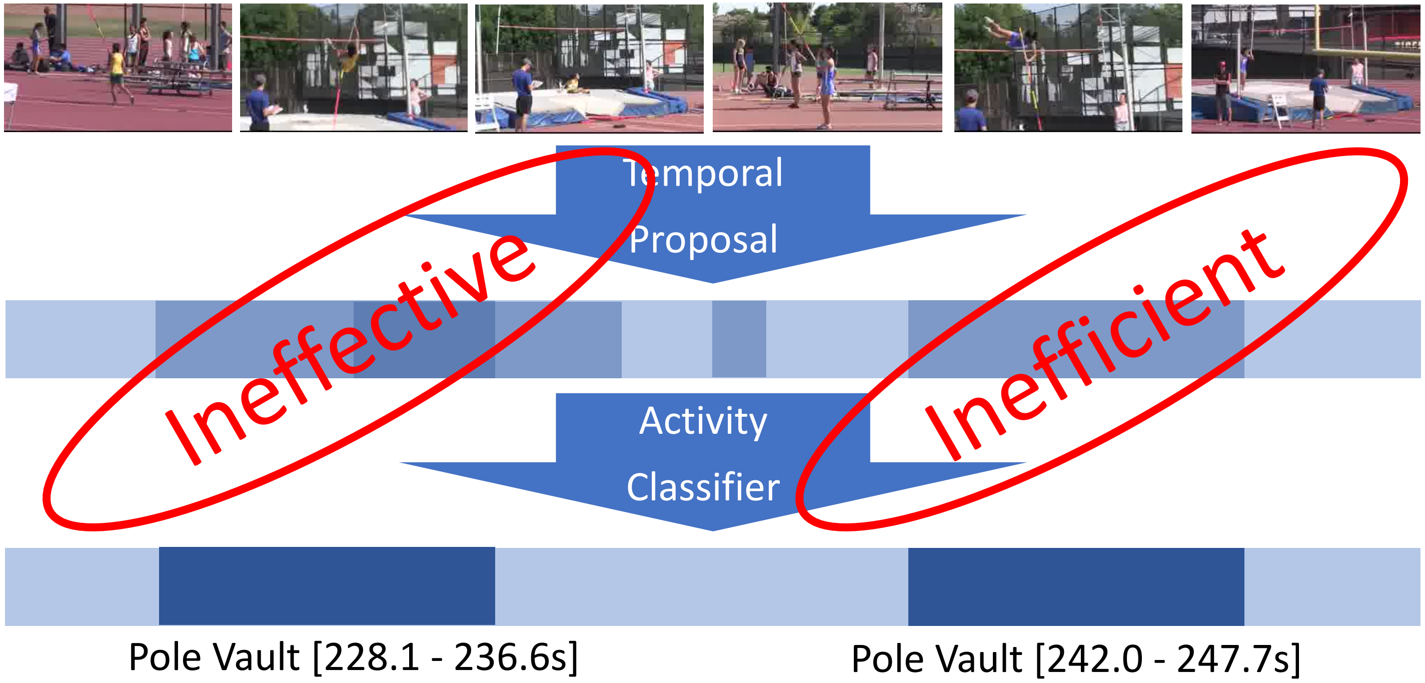
Pole Vault [228.1 - 236.6s]

Pole Vault [242.0 - 247.7s]

S-CNN (CVPR 2016), CDC (CVPR 2017), TSN (ICCV 2017), R-C3D (ICCV 2017), SSN (ICCV 2017)

Related Works

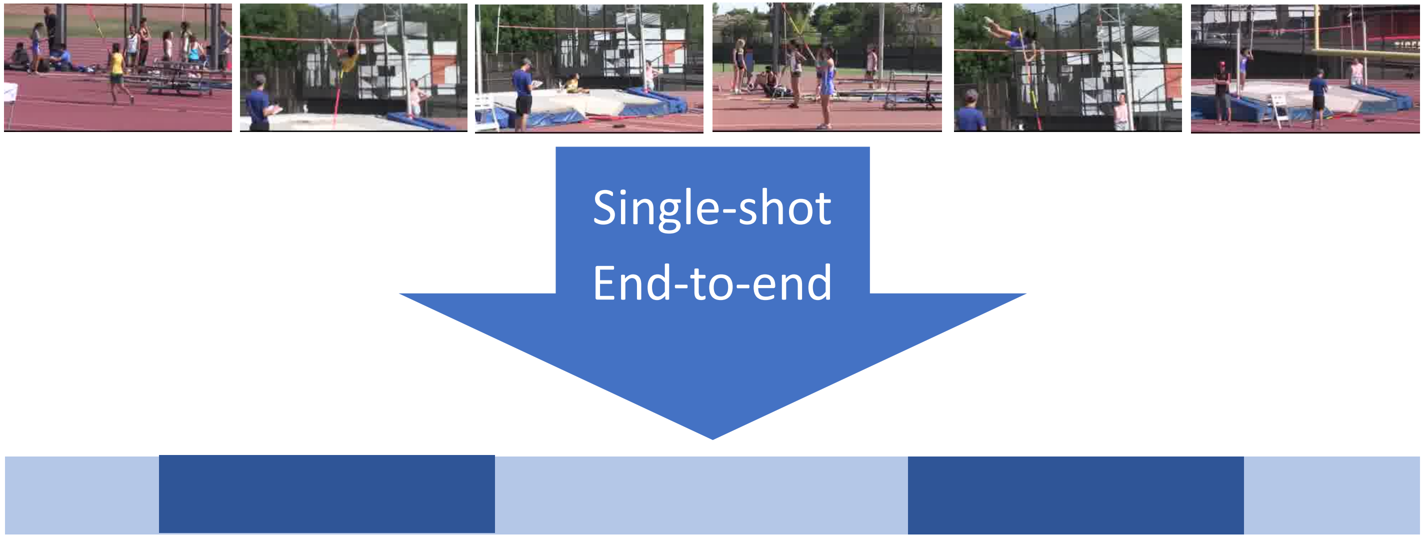
Current limitations:



S-CNN (CVPR 2016), CDC (CVPR 2017), TSN (ICCV 2017), R-C3D (ICCV 2017), SSN (ICCV 2017)

Motivation

Can we do better?



Single-shot

End-to-end

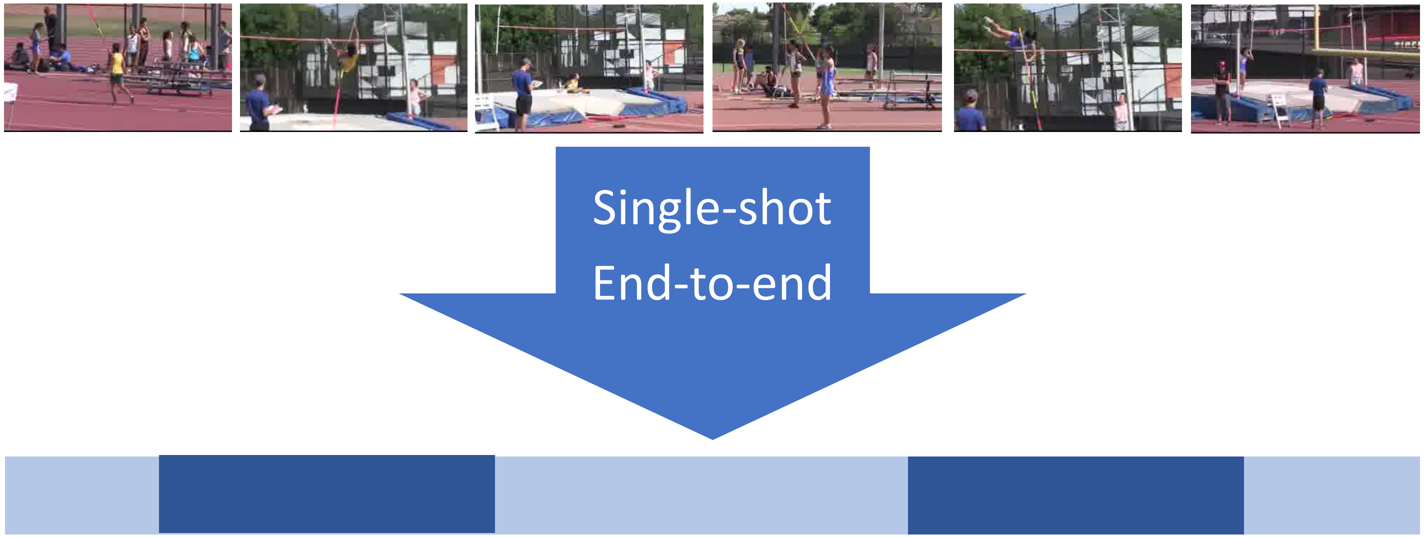
Pole Vault [228.1 - 236.6s]

Pole Vault [242.0 - 247.7s]

Introducing a novel **S**ingle **S**hot multi-**S**pan **D**etector (S^{3}D)

Motivation

Quick Summary



Single-shot

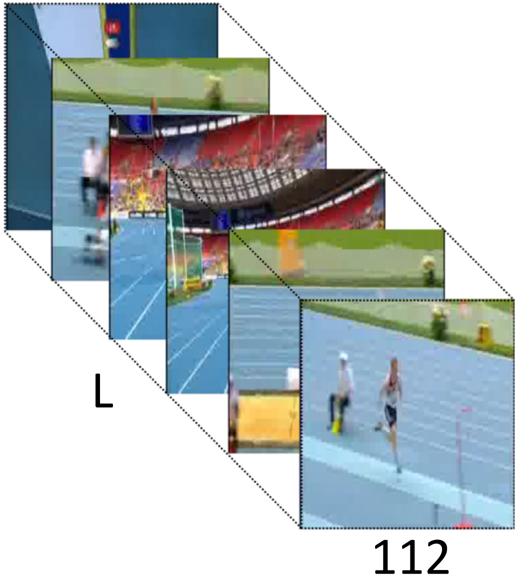
End-to-end

Pole Vault [228.1 - 236.6s]

Pole Vault [242.0 - 247.7s]

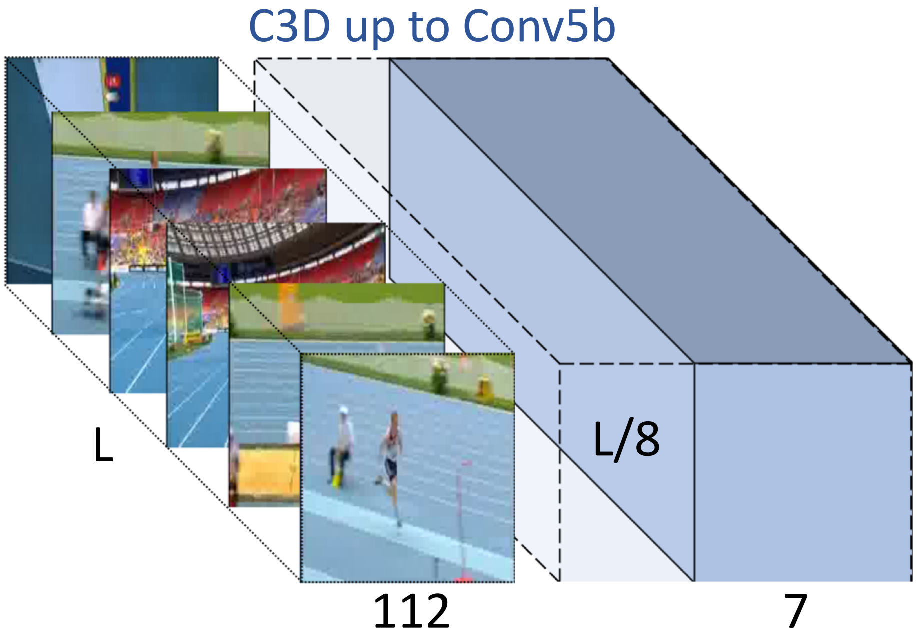
- □ Directly encode entire input video with Conv3D kernels
- □ Multi-scale default spans associated to temporal feature maps
- □ End-to-end trainable and single forward-pass inference

S^3D: Input Video



Our model takes the whole video stream as input (L frames)

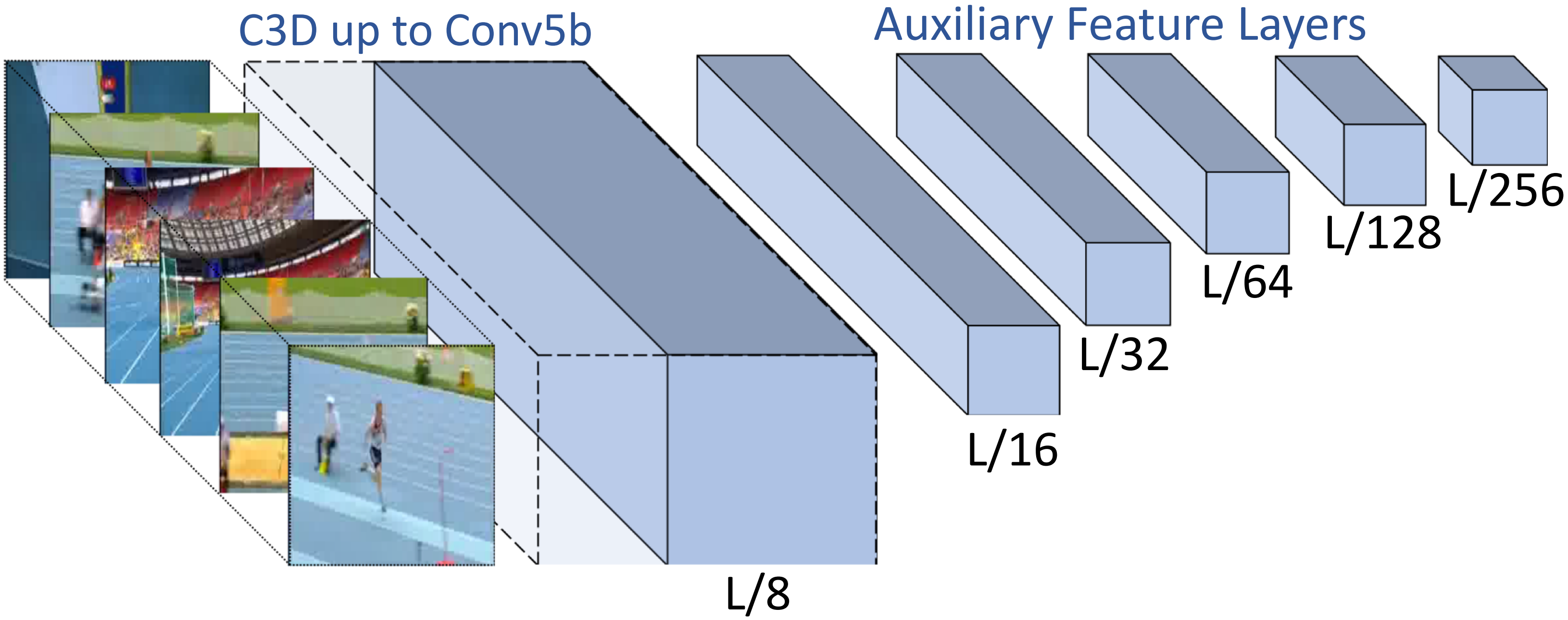
S^{3}D: Base Feature Layers



We apply the standard C3D network to extract spatial-temporal features.

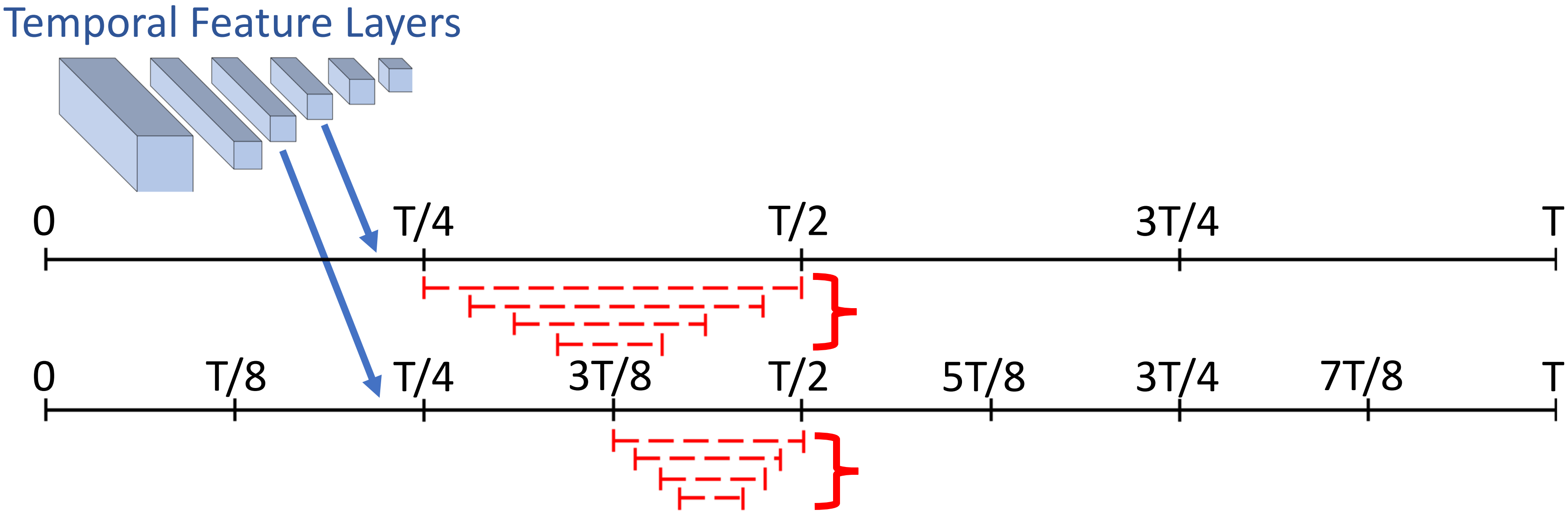
D. Tran, L. Bourdev, R. Fergus, L. Torresani and M. Paluri. Learning spatiotemporal features with 3D convolutional networks. In CVPR, 2015.

S^3D: Auxiliary Feature Layers



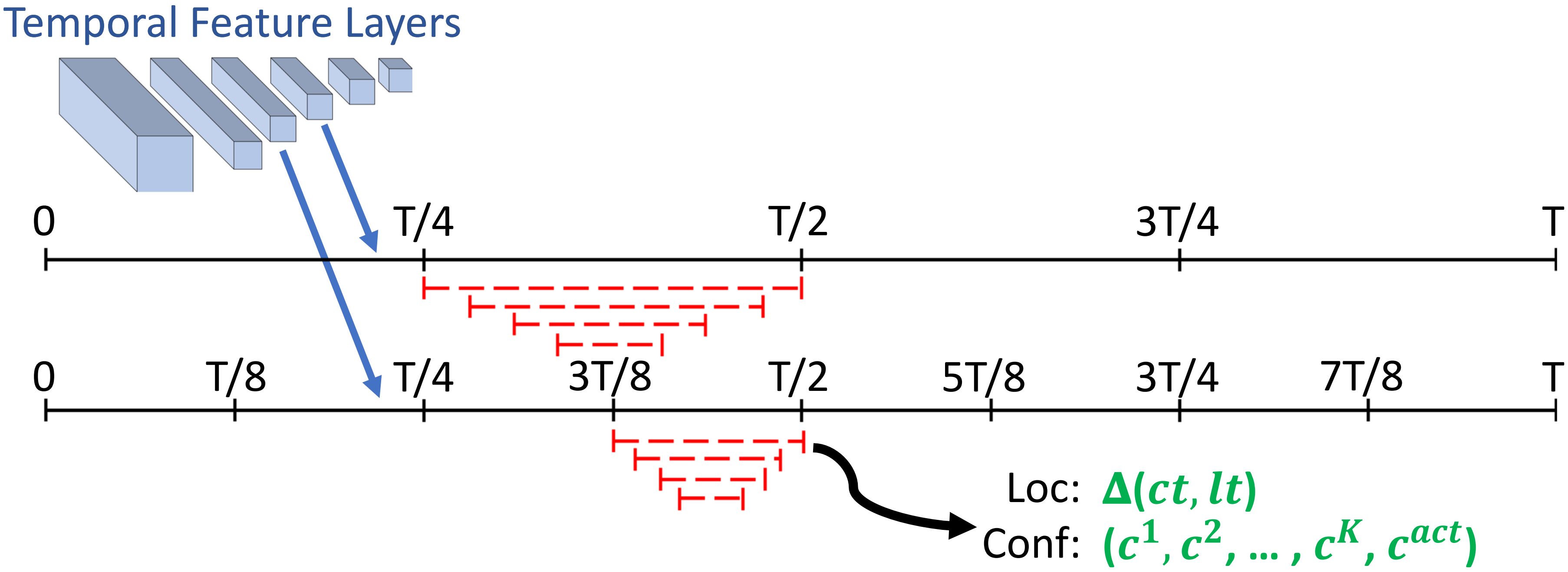
We produce a sequence of feature maps that progressively decrease in temporal dimension.

S^3D: Multi-scale Default Spans



Multi-scale default spans are associated to each temporal feature map

S³D: Multi-scale Default Spans

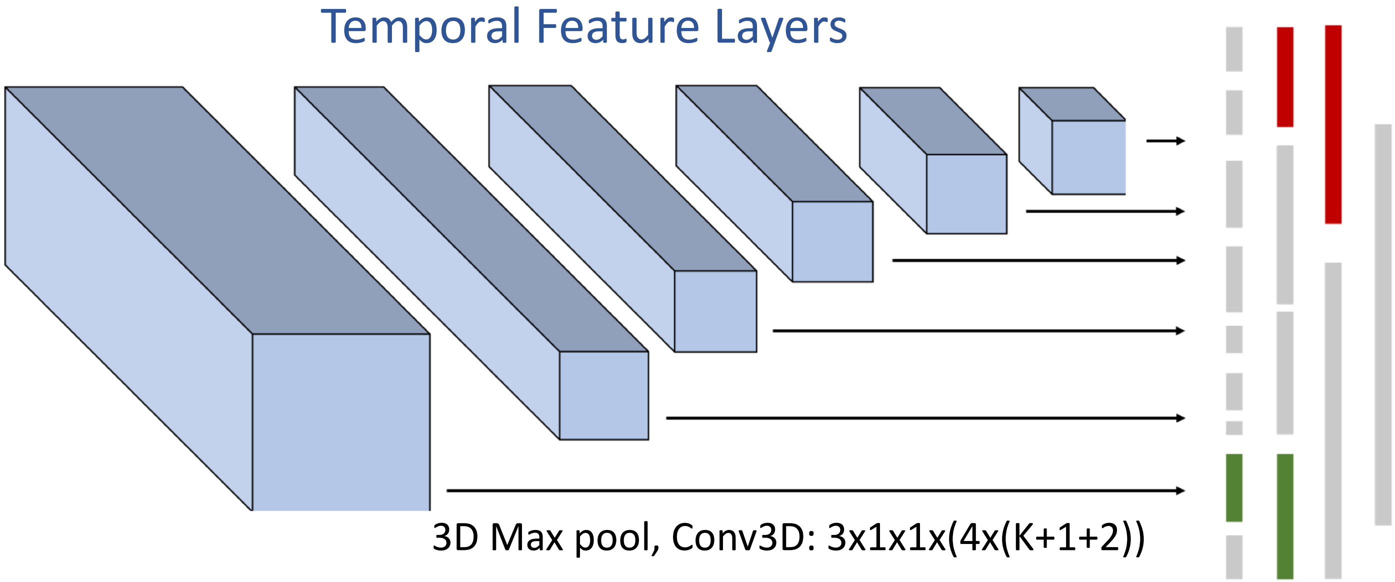


Loc: $\Delta(ct, lt)$

Conf: $(c^1, c^2, \ldots, c^K, c^{act})$

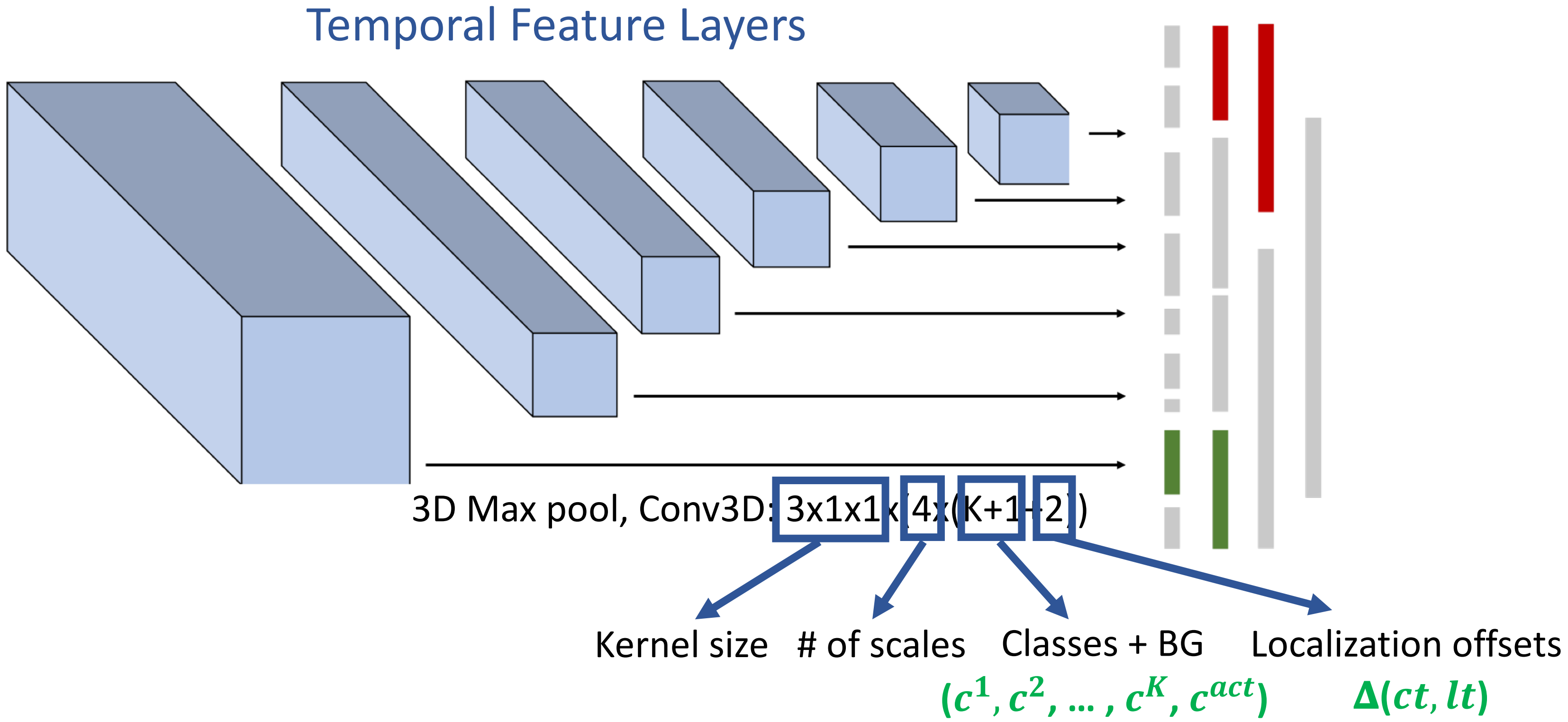
Localization and classification results are predicted at each default span.

S^3D: Convolutional Predictors



We apply on top of each feature map a Conv3D filter to produce the results.

S^{3}D: Convolutional Predictors



Temporal Feature Layers

3D Max pool, Conv3D: $\boxed{3x1x1}x\boxed{4}x(\boxed{K+1}+\boxed{2})$

Kernel size

# of scales

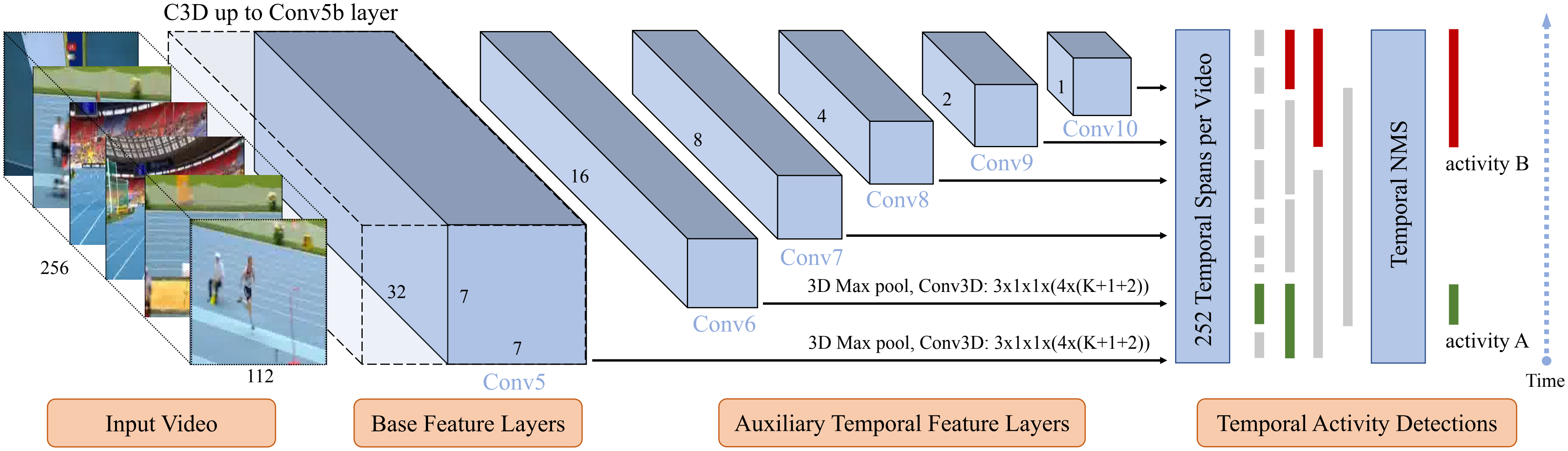
Classes + BG

$(c^1, c^2, \ldots, c^K, c^{act})$

Localization offsets

$\Delta(ct, lt)$

Single Shot multi-Span Detector



C3D up to Conv5b layer
256
112
32
7
7
16
8
4
2
1
Conv5
Conv6
Conv7
Conv8
Conv9
Conv10
3D Max pool, Conv3D: 3x1x1x(4x(K+1+2))
3D Max pool, Conv3D: 3x1x1x(4x(K+1+2))
252 Temporal Spans per Video
Temporal NMS
activity B
activity A
Time
Input Video
Base Feature Layers
Auxiliary Temporal Feature Layers
Temporal Activity Detections

Training of S^{3}D:

$$Loss = \boxed{L_{loc}(x,t,g)} + \alpha \boxed{L_{conf}(x,c)} + \beta \boxed{L_{act}(s,c)}$$

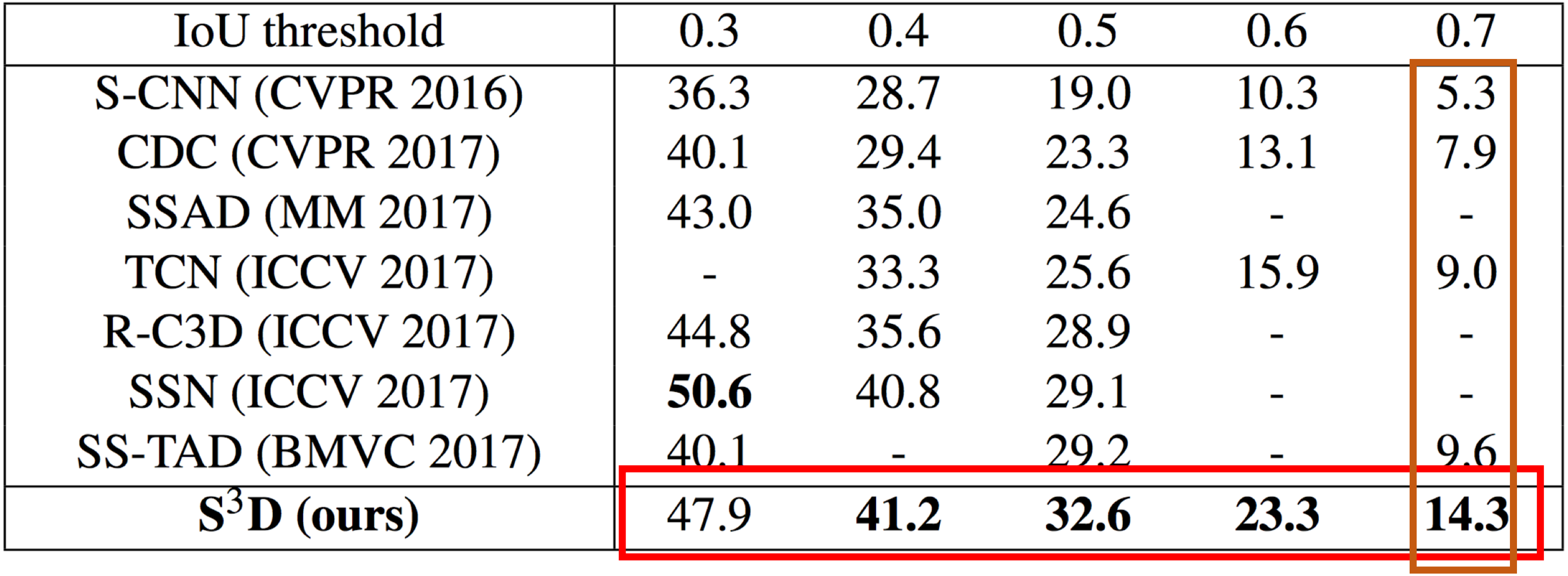
Smooth L1

Softmax Cross Entropy

Sigmoid Cross Entropy

Quantitative Results

Evaluation: mean Average Precision over 20 activities on THUMOS’14

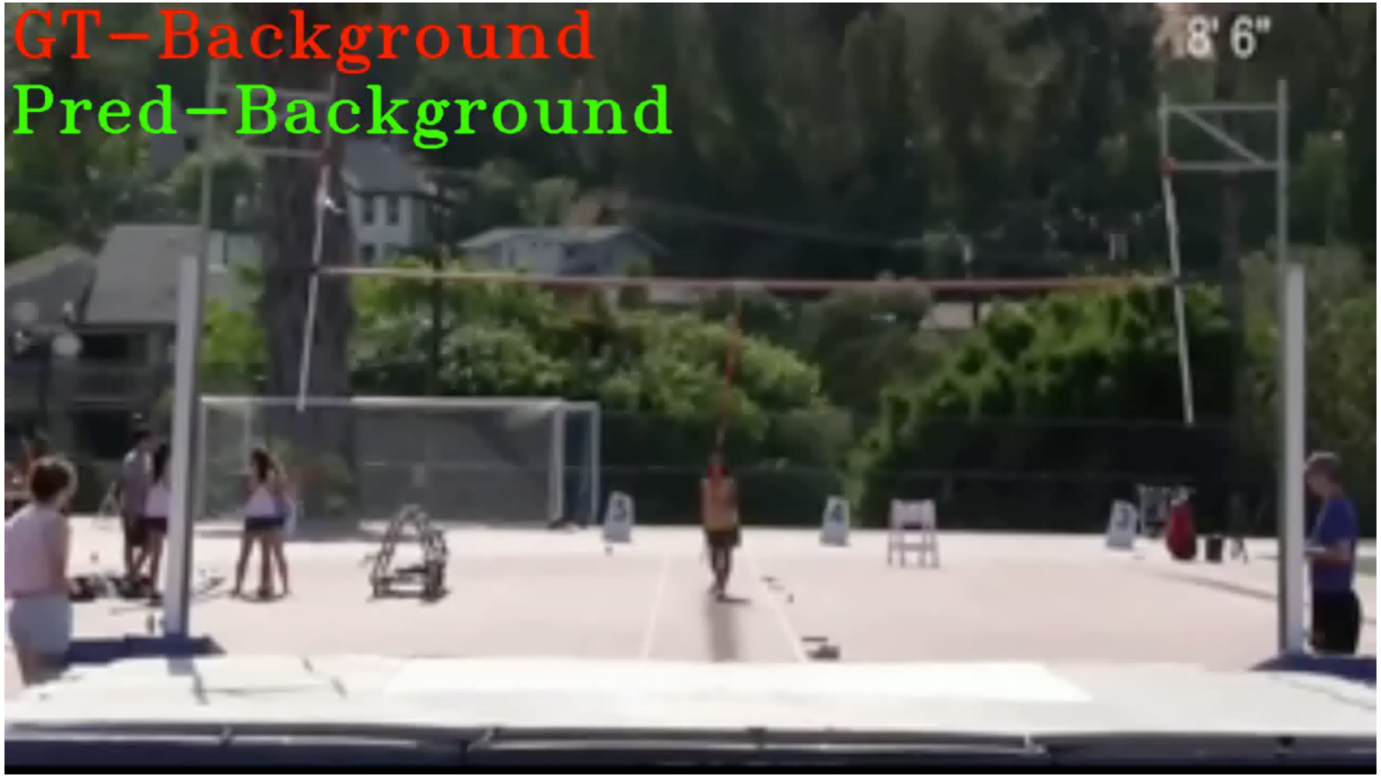


IoU threshold	0.3	0.4	0.5	0.6	0.7
S-CNN (CVPR 2016)	36.3	28.7	19.0	10.3	5.3
CDC (CVPR 2017)	40.1	29.4	23.3	13.1	7.9
SSAD (MM 2017)	43.0	35.0	24.6	-	-
TCN (ICCV 2017)	-	33.3	25.6	15.9	9.0
R-C3D (ICCV 2017)	44.8	35.6	28.9	-	-
SSN (ICCV 2017)	**50.6**	40.8	29.1	-	-
SS-TAD (BMVC 2017)	40.1	-	29.2	-	9.6
S^3D (ours)	47.9	**41.2**	**32.6**	**23.3**	**14.3**

1271 FPS on a single GTX 1080 Ti GPU

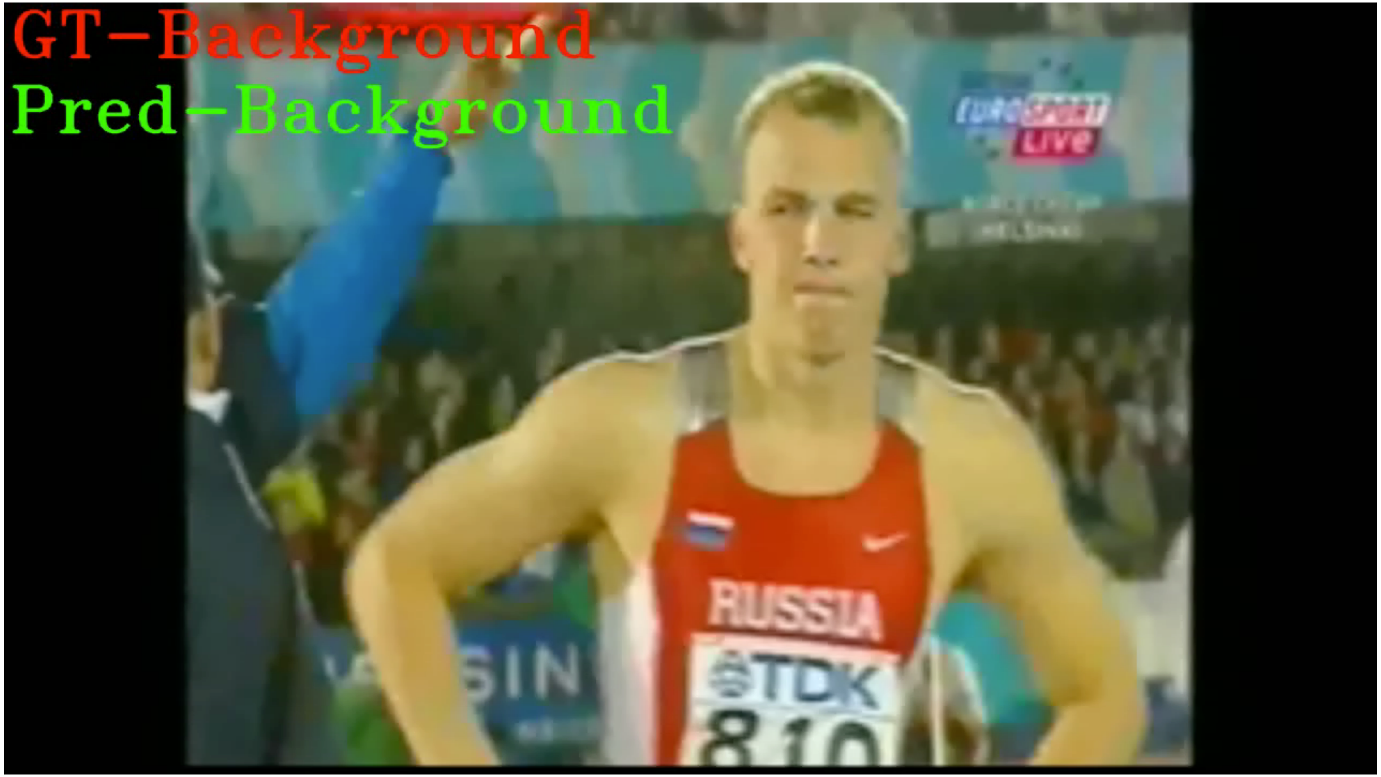
Qualitative Results

THUMOS’14 segment: Pole Vault



Qualitative Results

THUMOS'14 segment: Javelin Throw

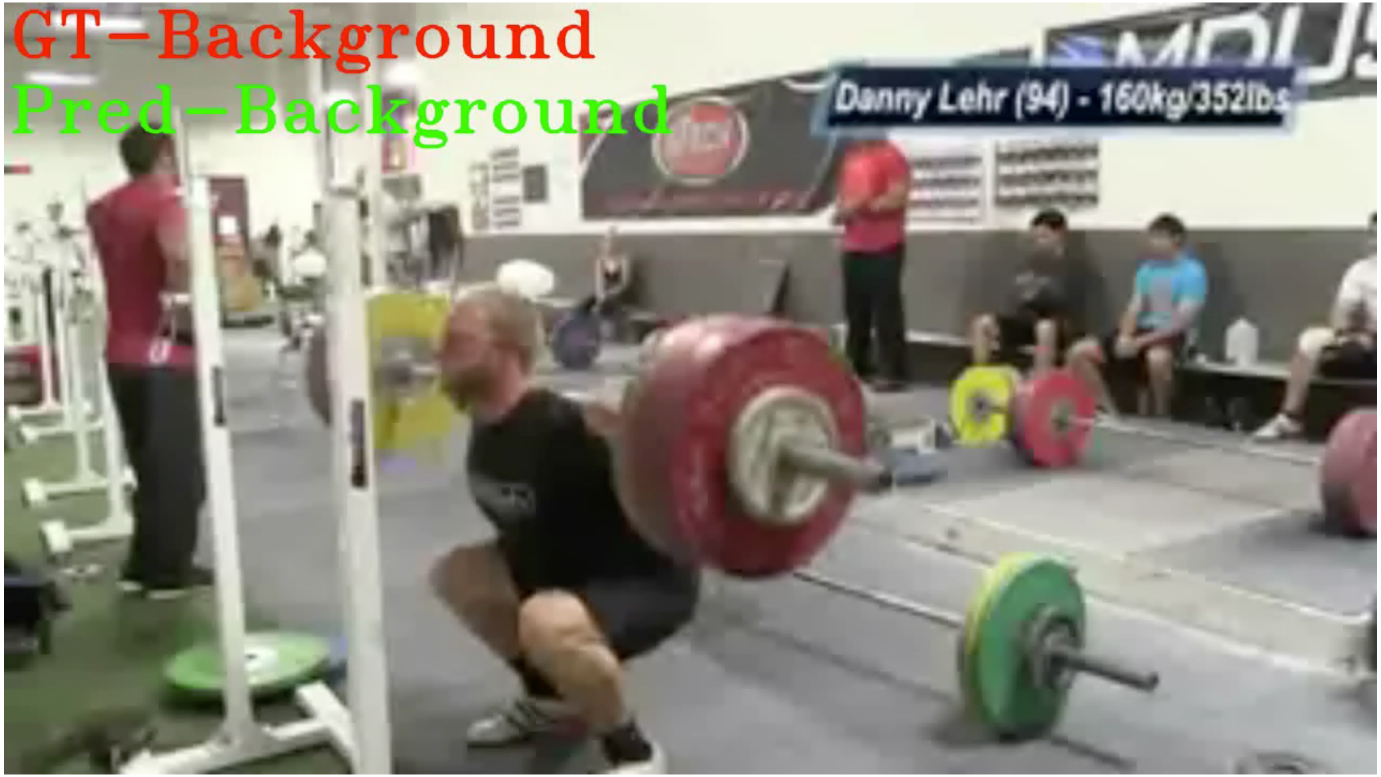


Qualitative Results

THUMOS'14 segment: Shotput

Qualitative Results

THUMOS'14 segment: Clean and Jerk



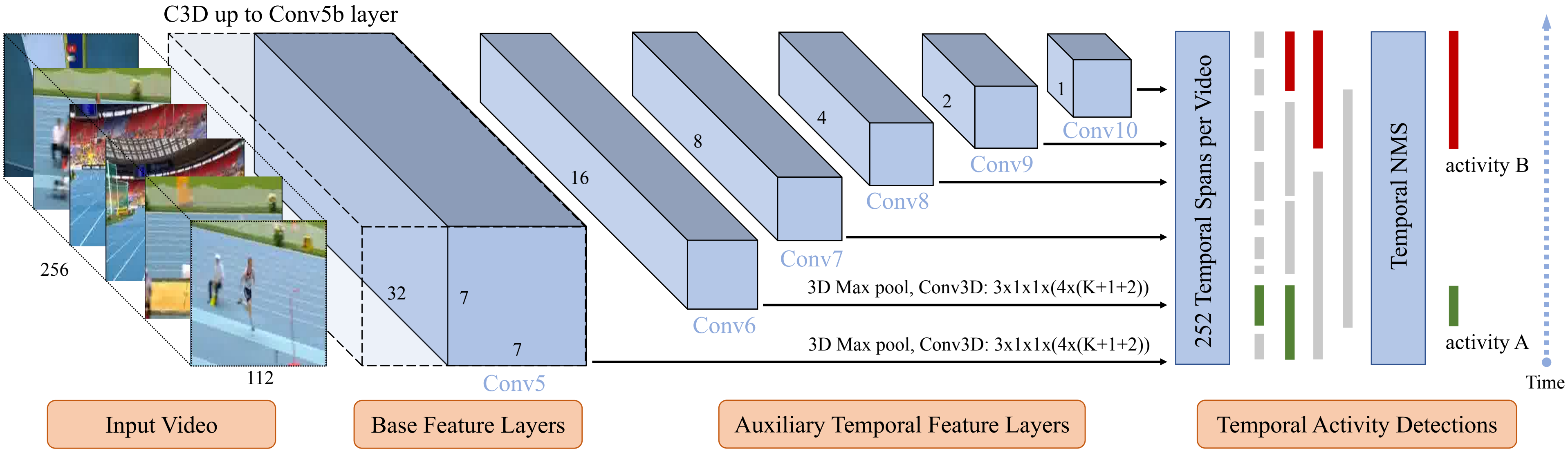
Conclusions

Introduced **S^{3}D**:

- ❑ A novel single-shot end-to-end model for Temporal Activity Detection.
- ❑ *Simple*: completely based on Conv3D kernels.
- ❑ *Strong*: state-of-the-art performance on THUMOS’14 benchmark.
- ❑ *Speed*: operates at 1271 FPS on a single GeForce GTX 1080 Ti GPU.

TensorFlow code coming soon at https://github.com/dazhang-cv/S3D

Thank you!


C3D up to Conv5b layer
256
112
32
7
7
Conv5
16
Conv6
8
Conv7
4
Conv8
2
Conv9
1
Conv10
3D Max pool, Conv3D: 3x1x1x(4x(K+1+2))
3D Max pool, Conv3D: 3x1x1x(4x(K+1+2))
252 Temporal Spans per Video
Temporal NMS
activity B
activity A
Time
Input Video
Base Feature Layers
Auxiliary Temporal Feature Layers
Temporal Activity Detections