Rethinking the Faster R-CNN Architecture for Temporal Action Localization

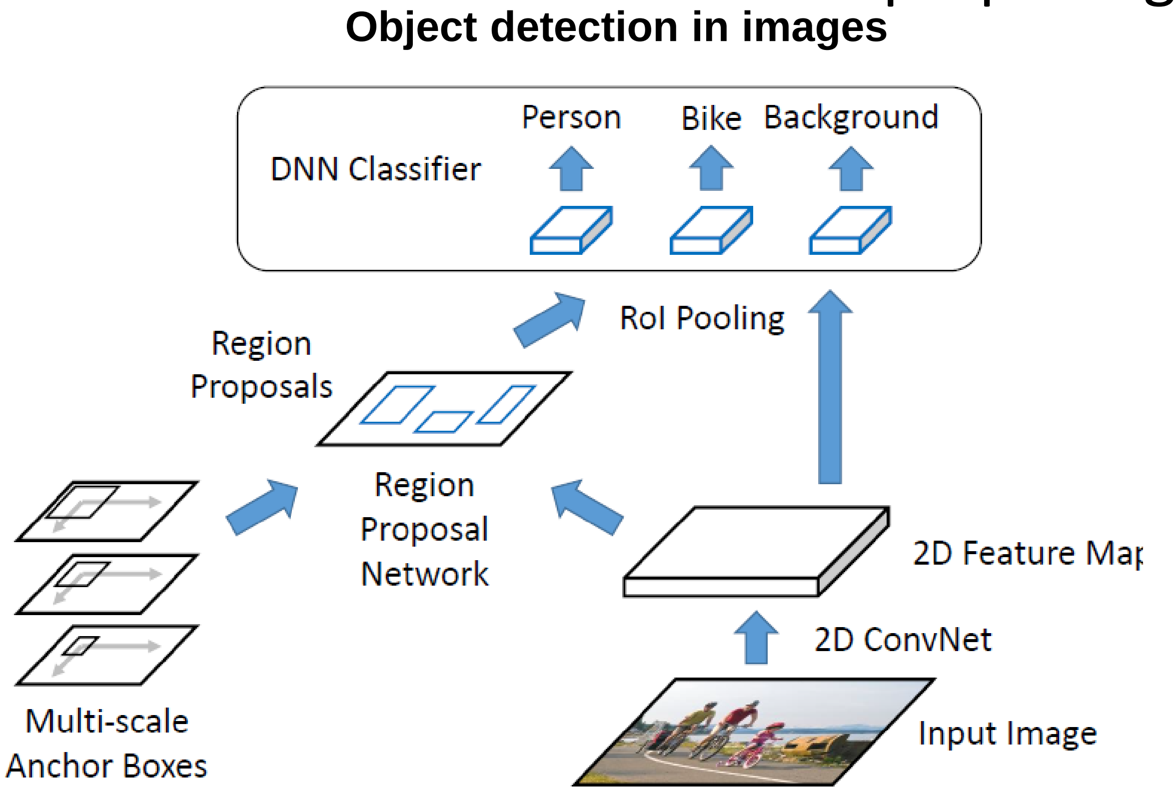
Yu-Wei Chao, Sudheendra Vijayanarasimhan, Bryan Seybold, David A. Ross, Jia Deng, Rahul Sukthankar

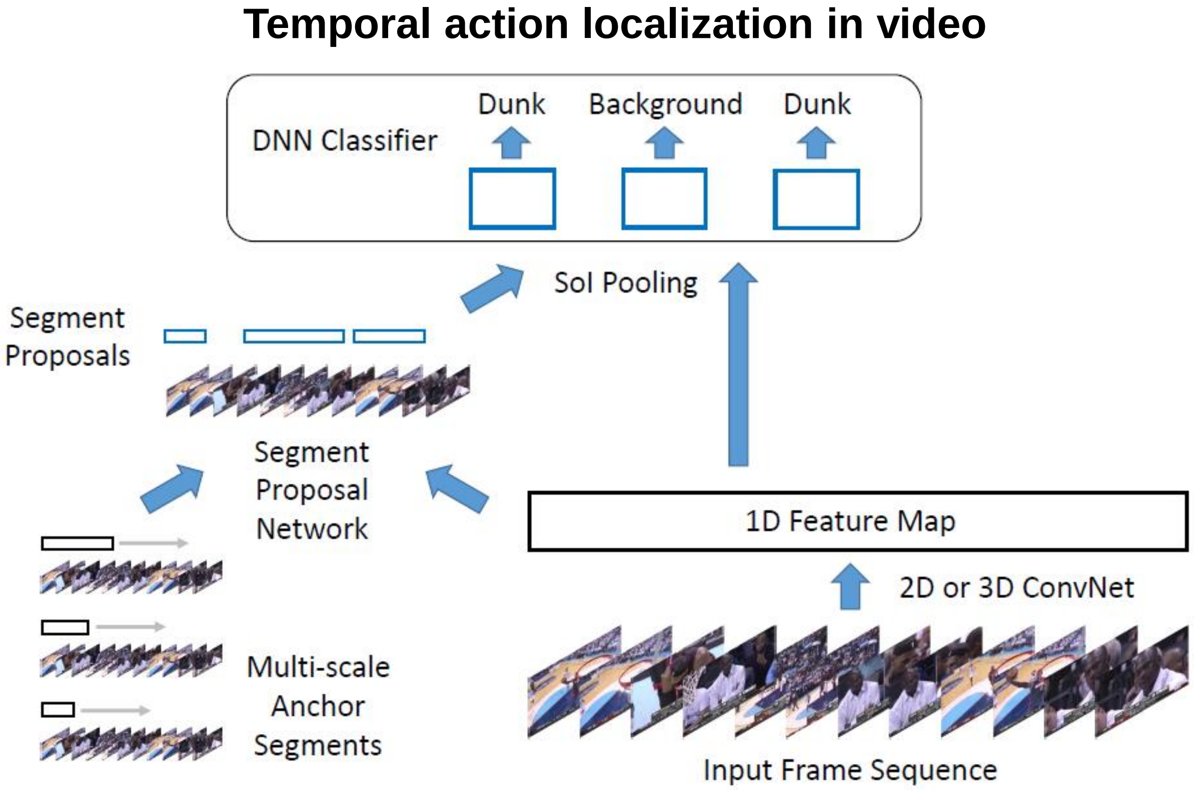
University of Michigan, Ann Arbor; Google Research

Accepted in CVPR 2018

Background

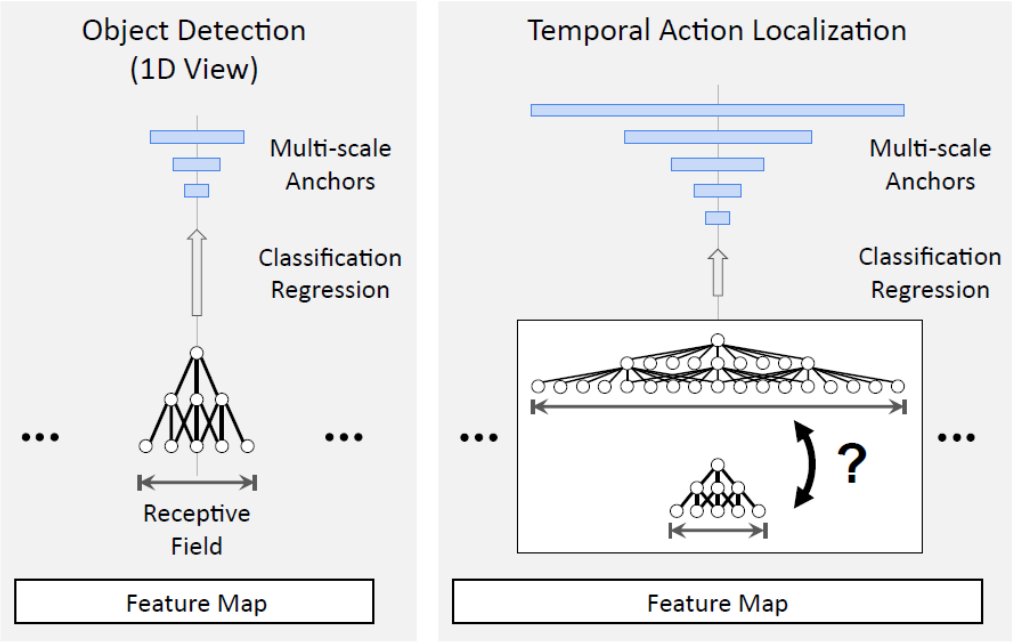
- Classify human activity through video:
  - Classification of a temporally trimmed video clip into one of several action classes.
  - untrimmed video: identify the action class + detect the start and end time of each action instance
- Faster R-CNN architecture: proposal generation and classification





https://stackoverflow.com/questions/43430056/what-is-the-purpose-of-the-roi-layer-in-a-fast-r-cnn

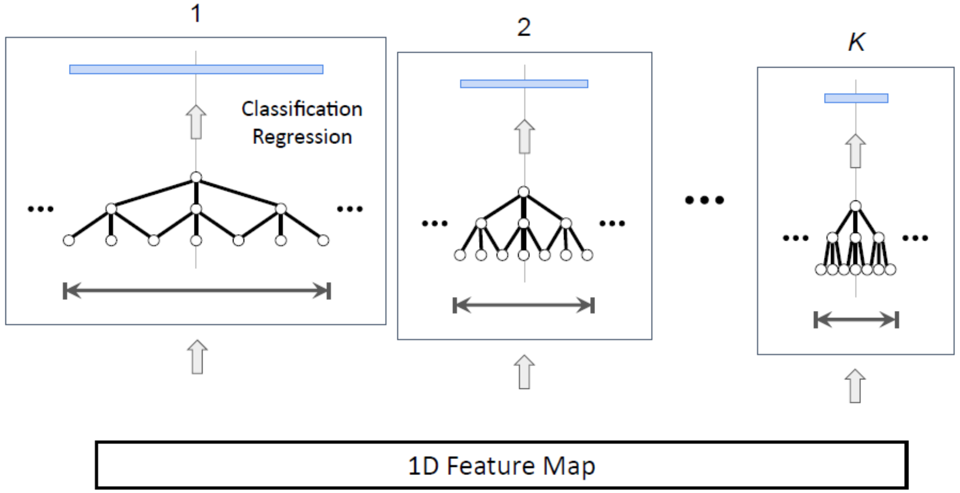
1. Receptive Field Alignment



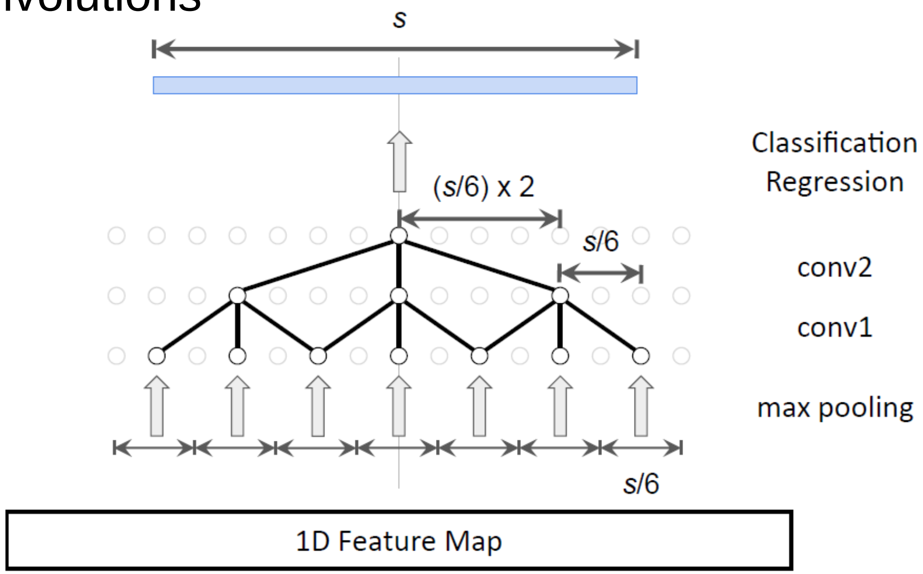
AN	10	20	50	100	200
Single	9.4	15.3	25.3	33.9	41.3
Single + TConv	12.9	20.0	30.3	37.6	44.0
Multi + TConv	13.4	20.6	31.1	38.1	43.7
Multi + Dilated	**14.0**	**21.7**	**31.9**	**38.8**	**44.7**
Single	11.0	18.0	28.9	36.8	43.6
Single + TConv	15.1	23.2	33.7	40.0	44.7
Multi + TConv	15.7	24.0	35.0	41.1	46.2
Multi + Dilated	**16.3**	**25.4**	**35.8**	**42.3**	**47.5**

Table 1: Results for receptive field alignment on proposal generation in AR (%). Top: RGB stream. Bottom: Flow stream.

Multi-tower network

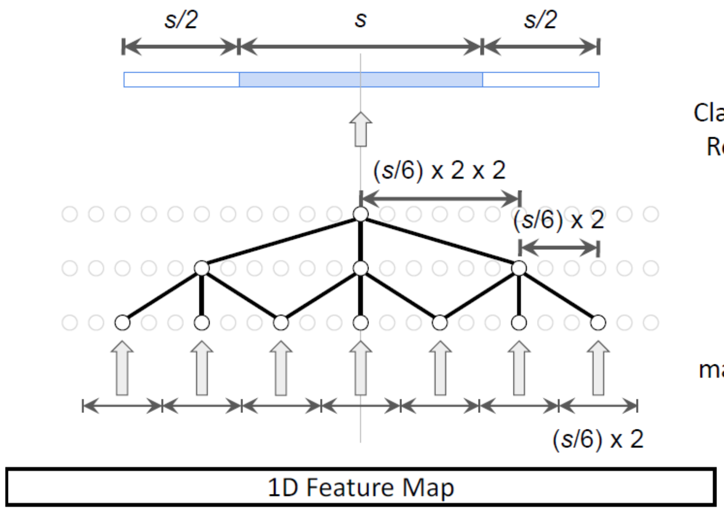


Dilated temporal convolutions



2. Context Feature Extraction

Incorporating context features in proposal generation.

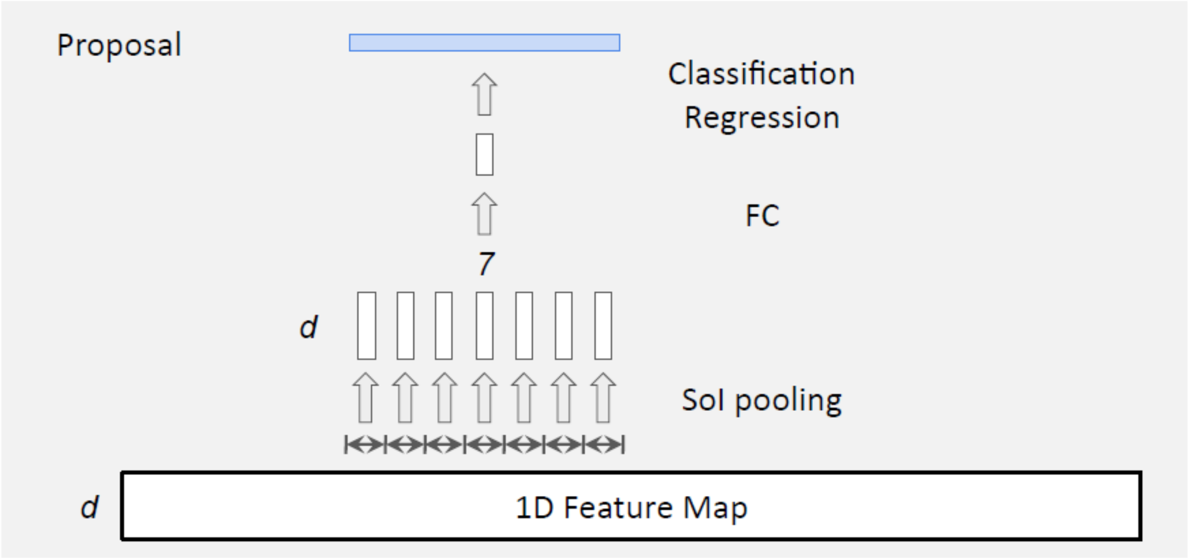


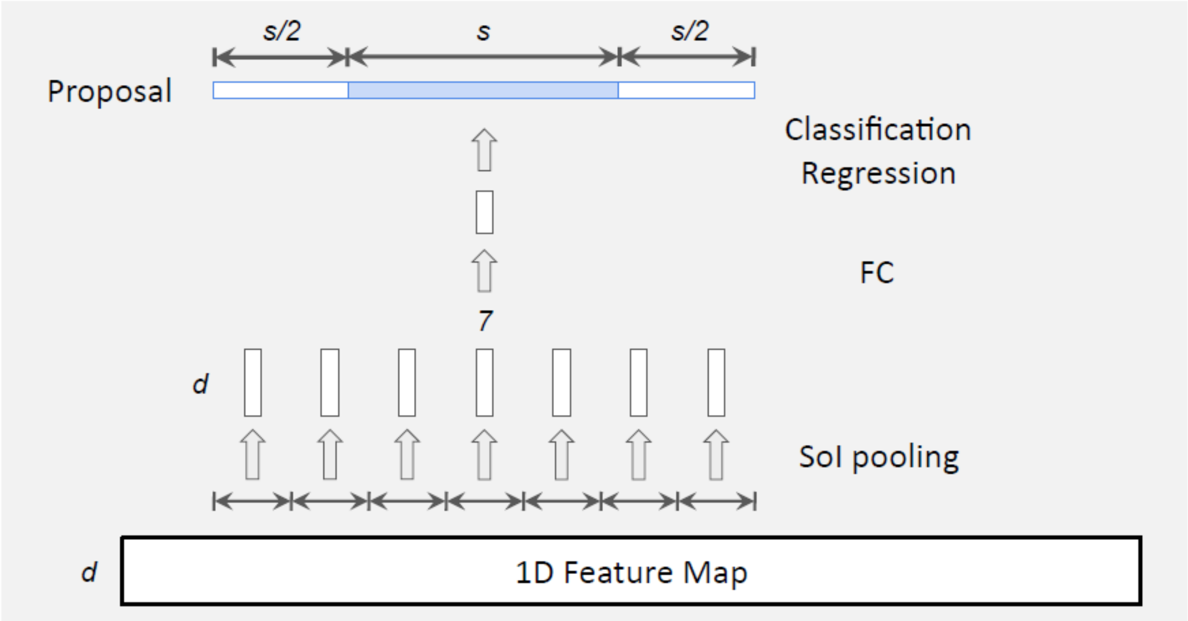
We enforce the receptive field to also cover the two segments of length s=2 immediately before and after the anchor. This can be achieved by doubling the dilation rate of the convolutional la

AN	10	20	50	100	200
Multi + Dilated	14.0	21.7	31.9	38.8	44.7
Multi + Dilated + Context	**15.1**	**22.2**	**32.3**	**39.9**	**46.8**
Multi + Dilated	16.3	25.4	35.8	42.3	47.5
Multi + Dilated + Context	**17.4**	**26.5**	**36.5**	**43.3**	**48.6**

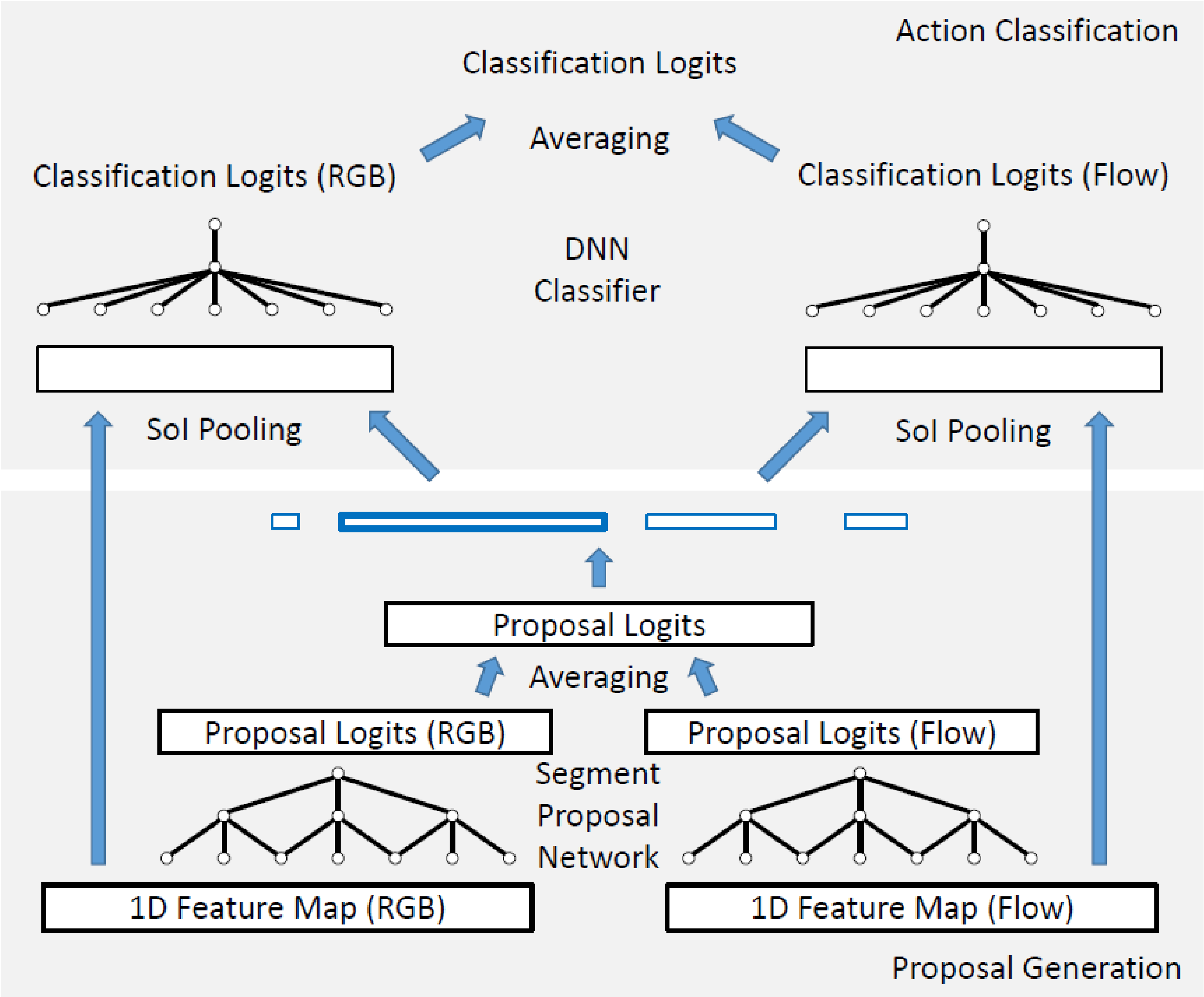
Table 2: Results for incorporating context features in proposal generation in AR (%). Top: RGB stream. Bottom: Flow stream.

SoI pooling





3. Late Feature Fusion



tIoU	0.1	0.3	0.5	0.7	0.9
RGB	49.3	42.6	31.9	14.2	0.6
Flow	54.3	48.8	38.2	18.6	**0.9**
Early Fusion	**60.5**	52.8	40.8	19.3	0.8
Late Fusion	59.8	**53.2**	**42.8**	**20.8**	**0.9**

Table 4: Results for late feature fusion in mAP (%).

Overall performance

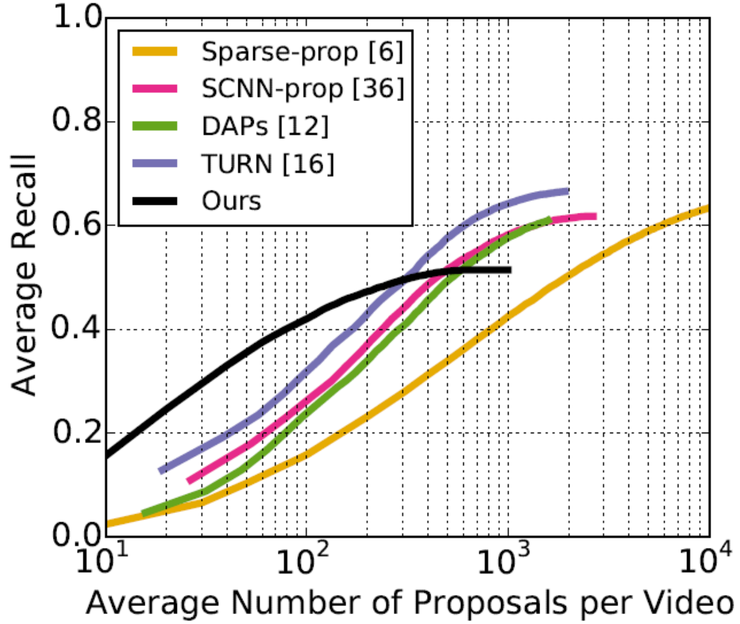


Figure 7: Our action proposal result in AR-AN (%) on THUMOS'14 comparing with other state-of-the-art methods.

tIoU	0.5	0.75	0.95	Average
Singh and Cuzzolin [39]	34.47	–	–	–
Wang and Tao [50]	43.65	–	–	–
Shou et al. [35]	**45.30**	**26.00**	0.20	**23.80**
Dai et al. [9]	36.44	21.15	**3.90**	–
Xu et al. [51]	26.80	–	–	12.70
Ours	38.23	18.30	1.30	20.22

Table 6: Action localization mAP (%) on ActivityNet v1.3 (val).

tIoU	0.1	0.2	0.3	0.4	0.5	0.6	0.7
Karaman et al. [24]	4.6	3.4	2.4	1.4	0.9	–	–
Oneata et al. [32]	36.6	33.6	27.0	20.8	14.4	–	–
Wang et al. [47]	18.2	17.0	14.0	11.7	8.3	–	–
Caba Heilbron et al. [6]	–	–	–	–	13.5	–	–
Richard and Gall [34]	39.7	35.7	30.0	23.2	15.2	–	–
Shou et al. [36]	47.7	43.5	36.3	28.7	19.0	10.3	5.3
Yeung et al. [52]	48.9	44.0	36.0	26.4	17.1	–	–
Yuan et al. [54]	51.4	42.6	33.6	26.1	18.8	–	–
Escorcia et al. [12]	–	–	–	–	13.9	–	–
Buch et al. [3]	–	–	37.8	–	23.0	–	–
Shou et al. [35]	–	–	40.1	29.4	23.3	13.1	7.9
Yuan et al. [55]	51.0	45.2	36.5	27.8	17.8	–	–
Buch et al. [2]	–	–	45.7	–	29.2	–	9.6
Gao et al. [15]	60.1	56.7	50.1	41.3	31.0	19.1	9.9
Hou et al. [20]	51.3	–	43.7	–	22.0	–	–
Dai et al. [9]	–	–	–	33.3	25.6	15.9	9.0
Gao et al. [16]	54.0	50.9	44.1	34.9	25.6	–	–
Xu et al. [51]	54.5	51.5	44.8	35.6	28.9	–	–
Zhao et al. [56]	**66.0**	**59.4**	51.9	41.0	29.8	–	–
Ours	59.8	57.1	**53.2**	**48.5**	**42.8**	**33.8**	**20.8**

Table 5: Action localization mAP (%) on THUMOS'14.

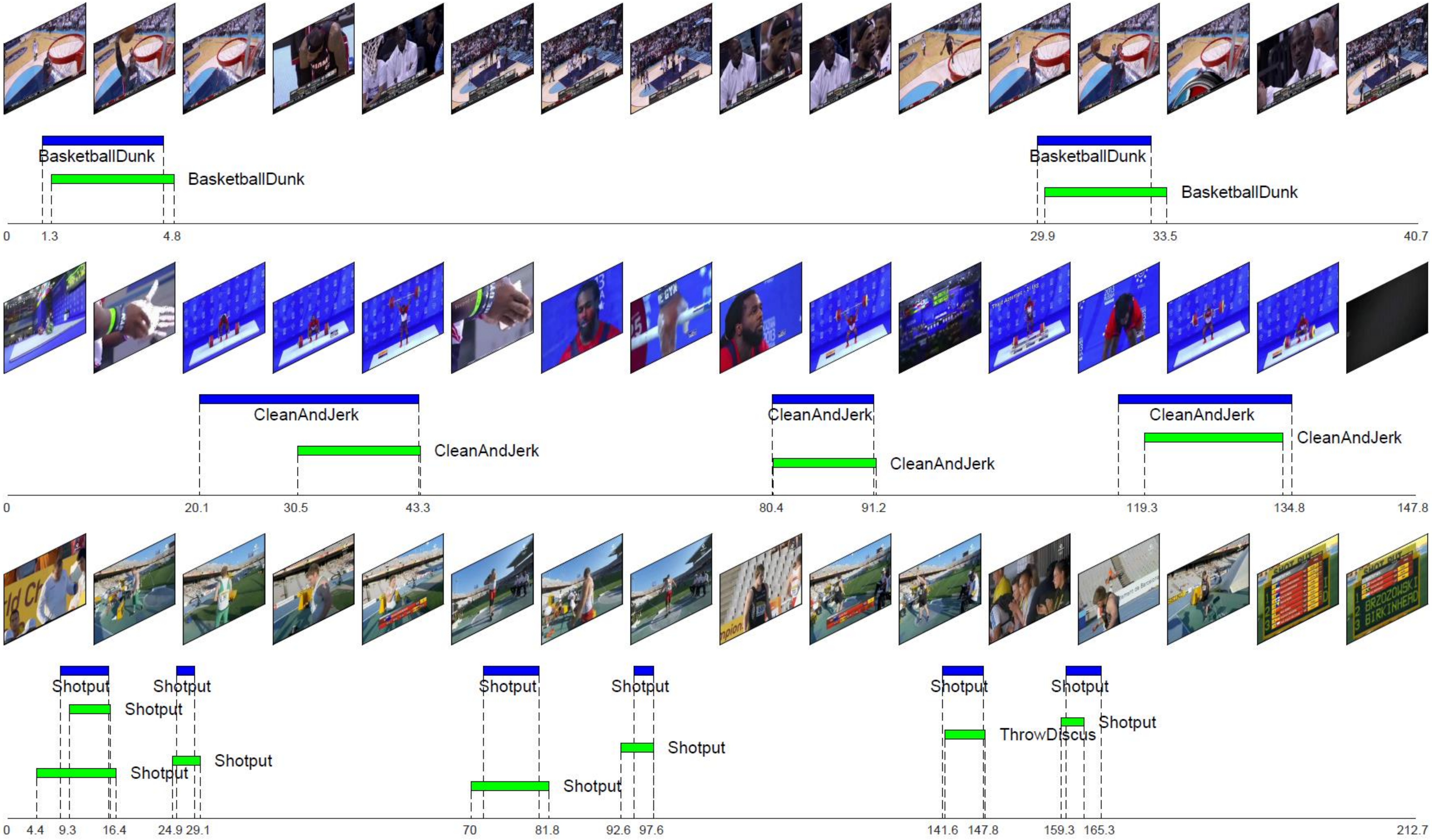


Figure 8: Qualitative examples of the top localized actions on THUMOS'14. Each consists of a sequence of frames sampled from a full test video, the ground-truth (blue) and predicted (green) action segments and class labels, and a temporal axis showing the time in seconds.