Temporal Action Detection with Structured Segment Networks Supplementary Materials

Yue Zhao[1], Yuanjun Xiong[1], Limin Wang[2], Zhirong Wu[1], Xiaoou Tang[1], and Dahua Lin[1]

[1]Department of Information Engineering, The Chinese University of Hong Kong
[2]Computer Vision Laboratory, ETH Zurich, Switzerland

1. Implementation Details

Below, we present more details related to implementing the SSN framework.

1. To generate the inputs to the optical flow stream in T-SN, we adopt the TVL1 optical flow algorithm [3] implemented in OpenCV with CUDA.

2. The linear SVMs for the stage-wise training baselines are from the implementation provided by *scikit-learn* [2].

2. Visualization of Detection Results

We visualize some detection results obtained on the validation set of ActivityNet v1.2 dataset and the testing set of THUMOS'14 dataset in Fig. 1 and Fig. 2, respectively. Notice the accuracy of the detected temporal bounding boxes and the framework's capability of detecting actions of different durations.

3. Per-Class Detection Performance

Although we obtain superior overall detection performance, it may also be of interest for audience to see the per-class performance. Due to space limit in the text, we present the per-class average of AP values using SSN on ActivityNet v1.2 validation set in Table 1. The average AP values are measured by varying the IOU thresholds from 0.5 to 0.95 in the step of 0.05. For comparison, detection results produced by SSN with proposals generated from a sliding window (486 proposals per video, AR $= 71\%$) and TAG (100 proposals per video, AR $= 67\%$) method are listed in parallel, showing that TAG-SSN achieves a higher AP on most of the classes. The results are also visualized in Fig. 3.

4. The Performance Metrics in Table 4

In Table 4 of the text we report two versions of performance metrics ("SSN" and "SSN*"). This is due to our observation of two major differences between the evaluation toolkit of THUMOS14[1] and ActivityNet[2]. First, the THUMOS14 toolkit assigns detections with the highest temporal *overlap* with groundtruth instances as true positives. The ActivityNet evaluation toolkit follows the convention of PASCAL VOC detection challenge [1] and assigns detection outputs to ground-truth annotations in the decreasing order of *confidence*. The latter one is more appropriate as the detections are later ranked by their confidence scores. Second, the arithmetic average of precision values is reported as average precision in THUMOS14, which does not consider recall in the evaluation. The ActivityNet toolkit calculates interpolated average precision, or AUC of the precision-recall curve, likewise in PASCAL VOC. Since previous results on THUMOS14 are usually reported by its own toolkit, we report both performance metrics of SSN to make the results comparable while promoting using the more appropriate metrics for evaluation.

References

[1] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111(1):98–136, 2015. 1

[2] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. 1

[3] C. Zach, T. Pock, and H. Bischof. A duality based approach for realtime tv-L^1 optical flow. In *29th DAGM Symposium on Pattern Recognition*, pages 214–223, 2007. 1

[1] https://storage.googleapis.com/www.thumos.info/thumos15_zips/THUMOS14_evalkit_20150930.zip

[2] https://github.com/activitynet/ActivityNet/tree/master/Evaluation

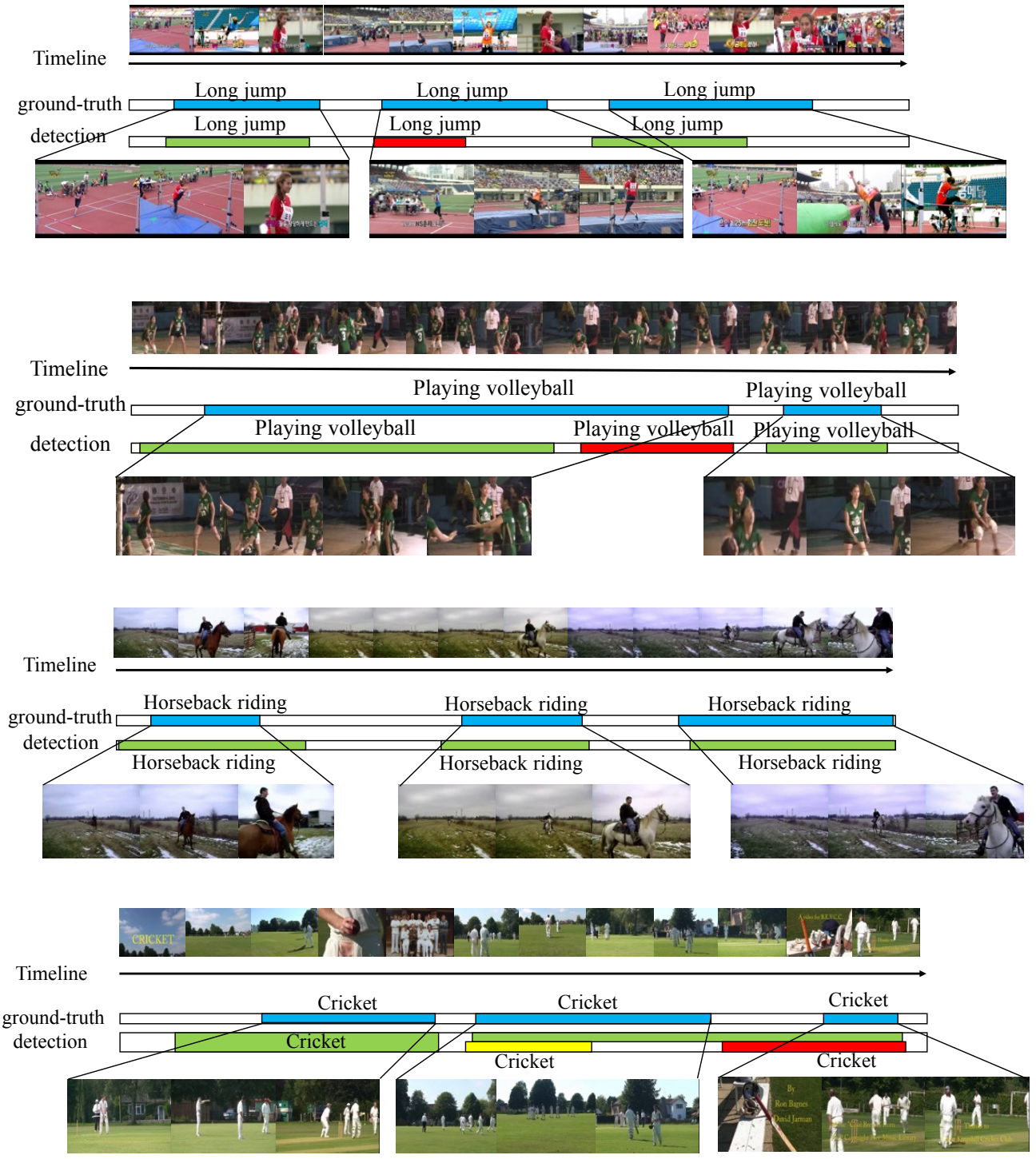

Figure 1. Examples of detection results on the ActivityNet v1.2 validation set. In each group, the video is shown as sequences of frames on top. The upper bar in each group with blue boxes denotes the annotated ground-truth instances, whose sampled frames are also illustrated at bottom. The detection results from SSN are shown in the lower bar, filled with different colors. A green box denotes a correct detection on condition that IoU ≥ 0.5. Other colors, namely red and yellow, denote the cases of bad localization (IoU < 0.5) and multiple detection, respectively.

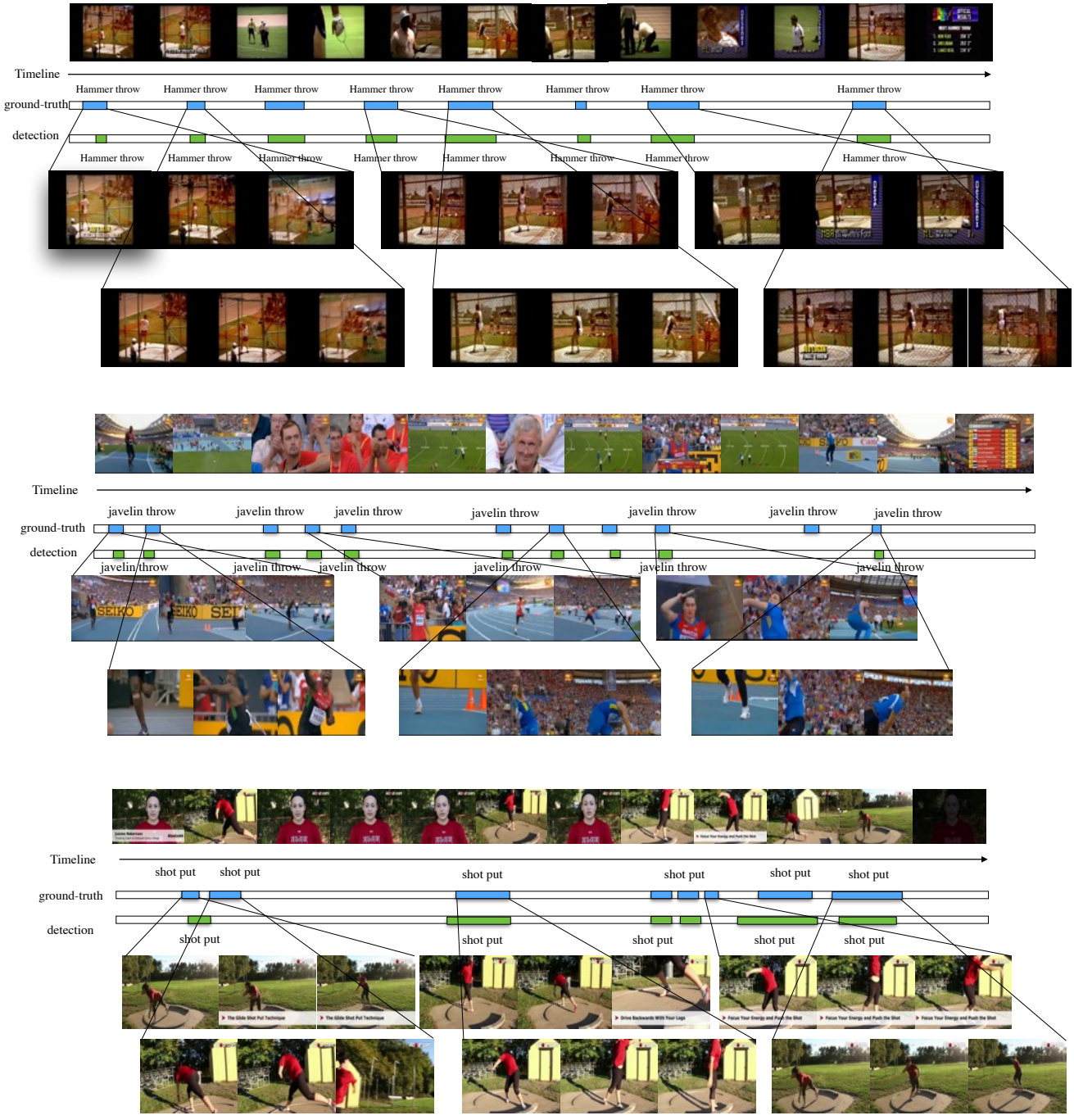

Figure 2. Examples of detection results on the THUMOS14 testing set. In each group, the video is shown as sequences of frames on top. The upper bar in each group with blue boxes denotes the annotated ground-truth instances, whose sampled frames are also illustrated at bottom. The detection results from SSN are shown in the lower bar, where a green box denotes a correct detection on condition that IoU ≥ 0.1. Note that the durations of action instances in THUMOS14 are much different from those in ActivityNet.

ActivityNet v1.2 (validation) **Per-class average AP**@{0.5:0.05:0.95}					
(Method)	Archery	Ballet	Bathing dog	Belly dance	Breakdancing
SW-SSN	8.97%	27.65%	33.60%	18.36%	19.19%
TAG-SSN	15.40%	42.90%	36.10%	55.35%	29.65%
Brushing hair	Brushing teeth	Bungee jumping	Cheerleading	Chopping wood	Clean and jerk
13.27%	18.26%	6.53%	17.05%	21.32%	20.74%
15.79%	22.33%	7.71%	30.58%	27.29%	28.27%
Cleaning shoes	Cleaning windows	Cricket	Cumbia	Discus throw	Dodgeball
7.83%	10.71%	7.71%	36.19%	9.45%	33.49%
20.16%	9.61%	11.79%	50.08%	12.91%	39.72%
Doing karate	Doing kickboxing	Doing motocross	Doing nails	Doing step aerobics	Drinking beer
13.13%	55.31%	29.61%	7.87%	30.34%	7.89%
21.42%	53.72%	42.05%	10.54%	41.07%	8.66%
Drinking coffee	Fixing bicycle	Getting a haircut	Getting a piercing	Getting a tattoo	Grooming horse
3.32%	20.59%	17.32%	14.86%	12.90%	26.83%
0.30%	30.07%	12.63%	26.09%	23.52%	33.44%
Hammer throw	Hand washing clothes	High jump	Hopscotch	Horseback riding	Ironing clothes
11.76%	9.83%	17.86%	15.70%	16.66%	9.28%
15.77%	9.98%	21.96%	19.46%	22.51%	15.67%
Javelin throw	Kayaking	Layup drill in basketball	Long jump	Making a sandwich	Mixing drinks
8.73%	15.60%	5.98%	0.73%	17.40%	27.74%
20.23%	31.39%	14.45%	3.08%	19.44%	36.19%
Mowing the lawn	Paintball	Painting	Ping-pong	Plataform diving	Playing accordion
22.01%	11.99%	12.06%	15.27%	6.48%	27.27%
24.65%	21.17%	17.66%	21.18%	10.20%	30.70%
Playing badminton	Playing bagpipes	Playing field hockey	Playing flauta	Playing guitarra	Playing harmonica
27.13%	41.38%	33.51%	22.59%	24.53%	10.45%
31.47%	53.29%	42.35%	26.28%	38.45%	13.01%
Playing kickball	Playing lacrosse	Playing piano	Playing polo	Playing racquetball	Playing saxophone
22.26%	26.42%	24.45%	6.44%	30.88%	21.10%
37.68%	33.96%	32.40%	17.08%	54.06%	26.45%
Playing squash	Playing violin	Playing volleyball	Playing water polo	Pole vault	Polishing forniture
22.38%	9.50%	44.66%	33.81%	5.80%	10.06%
40.12%	17.24%	49.35%	40.87%	17.08%	23.40%
Polishing shoes	Preparing pasta	Preparing salad	Putting on makeup	Removing curlers	Rock climbing
4.44%	19.69%	19.56%	9.90%	4.44%	9.93%
3.89%	33.09%	28.81%	13.87%	9.60%	14.54%
Sailing	Shaving	Shaving legs	Shot put	Shoveling snow	Skateboarding
21.41%	9.67%	8.31%	0.64%	6.75%	8.93%
32.07%	12.32%	8.77%	2.41%	18.98%	13.70%
Smoking a cigarette	Smoking hookah	Snatch	Spinning	Springboard diving	Starting a campfire
5.27%	15.68%	7.56%	6.16%	16.22%	24.89%
5.44%	17.09%	10.09%	12.63%	23.79%	26.95%
Tai chi	Tango	Tennis serve with ball bouncing	Triple jump	Tumbling	Using parallel bars
17.63%	47.61%	22.41%	5.59%	5.46%	34.02%
31.93%	57.56%	26.76%	7.45%	14.27%	38.47%
Using the balance beam	Using the pommel horse	Using uneven bars	Vacuuming floor	Walking the dog	Washing dishes
46.24%	36.17%	38.00%	11.95%	30.24%	14.37%
54.50%	54.98%	57.84%	14.48%	36.61%	13.77%
Washing face	Washing hands	Windsurfing	Wrapping presents	Zumba	(mean)
3.94%	4.88%	43.69%	10.15%	27.38%	18.19%
11.37%	10.20%	69.18%	18.99%	78.91%	25.95%

Table 1. Per-class average of AP values on ActivityNet v1.2 validation set. Detection results are produced by SSN with proposals generated from a sliding window (486 proposals per video, AR $= 71\%$, SW-SSN) or TAG (60 proposals per video, AR $= 67\%$, TAG-SSN) method. There are 100 classes in total, listed in the alphabetical order.

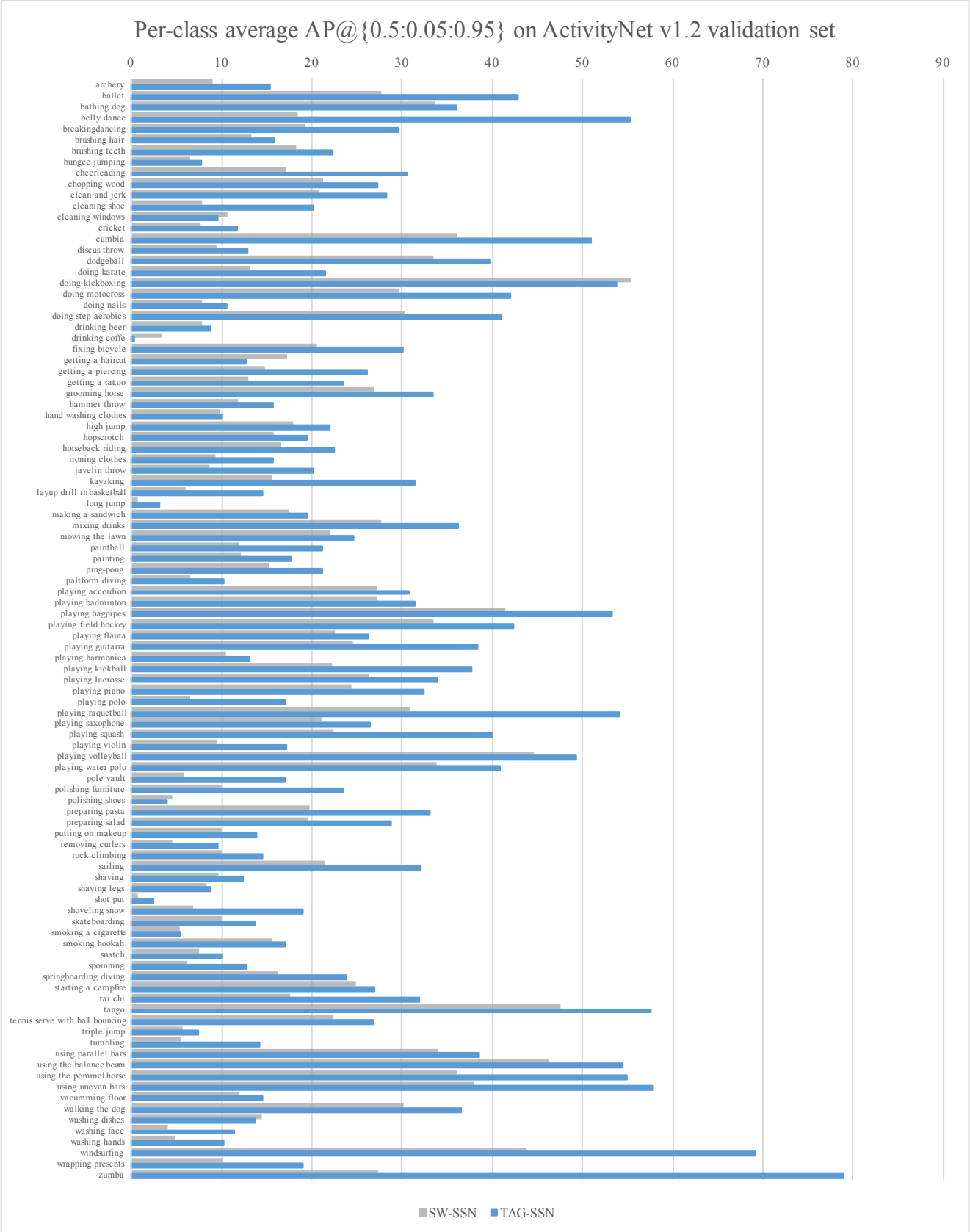

Figure 3. Illustration of per-class average of AP values on ActivityNet v1.2 validation set. There are 100 classes in total, listed in the alphabetical order.