Learning to Localize Actions from Moments — ECCV 2020 Supplementary Material

Anonymous ECCV submission

Paper ID 6101

The supplementary material contains: 1) the backbone structure of 1D convolutional networks in AherNet; 2) the structure of context generators for action moments in AherNet; 3) exemplars of temporal action localization on ActivityNet v1.3 [2] and Kinetics-600 [1] datasets.

1 The Structure of 1D ConvNet Backbone

Table 1 summarizes the structure of 1D convolutional backbone in AherNet, which mainly consists of temporal reduction layers and temporal anchor layers. The temporal reduction layers include two 1D convolutional layers ("conv1" and "conv2") and one max-pooling layer ("pool1"). We feed the input feature map into such temporal reduction layers to increase the temporal size of receptive fields. Then, eight 1D temporal convolutional layers are cascaded as temporal anchor layers to generate feature maps on different temporal scales for temporal action localization or action moment classification.

2 The Structure of Context Generators

Table 2 further details the structure of context generators for start/end contextual feature generation. Taking the moment feature as prior input, G_1 and G_2 are exploited to synthesize the context of the start and end part for action moments, respectively. The action moment features augmented with synthetic contextual features are employed for learning temporal action localization.

3 Localization Examples of AherNet

Finally, we illustrate six examples of temporal action localization on ActivityNet v1.3 and Kinetics-600, which correspond to the setting of TH14→ANet-FG and ANet→K600, in Figure 1 and Figure 2, respectively.

References

1. Ghanem, B., Niebles, J.C., Snoek, C., Heilbron, F.C., Alwassel, H., Escorcia, V., Krishna, R., Buch, S., Dao, C.D.: The ActivityNet Large-Scale Activity Recognition Challenge 2018 Summary. arXiv preprint arXiv:1808.03766 (2018)
2. Heilbron, F.C., Escorcia, V., Ghanem, B., Niebles, J.C.: ActivityNet: A Large-Scale Video Benchmark for Human Activity Understanding. In: CVPR (2015)

Table 1: The details of 1D temporal convolutional networks. RF represents the size of receptive fields.

id	layer	kernel size	#channels	#stride	RF
Temporal Reduction layers					
1	conv1	3	2048	1	3
2	conv2	3	1024	1	5
3	pool1	3	1024	2	7
Temporal Anchor layers					
4	conv_a1	3	256	2	11
5	conv_a2	3	256	2	19
6	conv_a3	3	512	2	35
7	conv_a4	3	512	2	67
8	conv_a5	3	1024	2	131
9	conv_a6	3	1024	2	259
10	conv_a7	3	2048	2	515
11	conv_a8	3	2048	2	1027

Table 2: The details of context generators for action moments.

id	layer	kernel size	#channels	#stride
Start Context Generator (G_1)				
1	gconv_s1	3	1024	1
2	gconv_s2	3	2048	2
End Context Generator (G_2)				
1	gconv_e1	3	1024	1
2	gconv_e2	3	2048	2

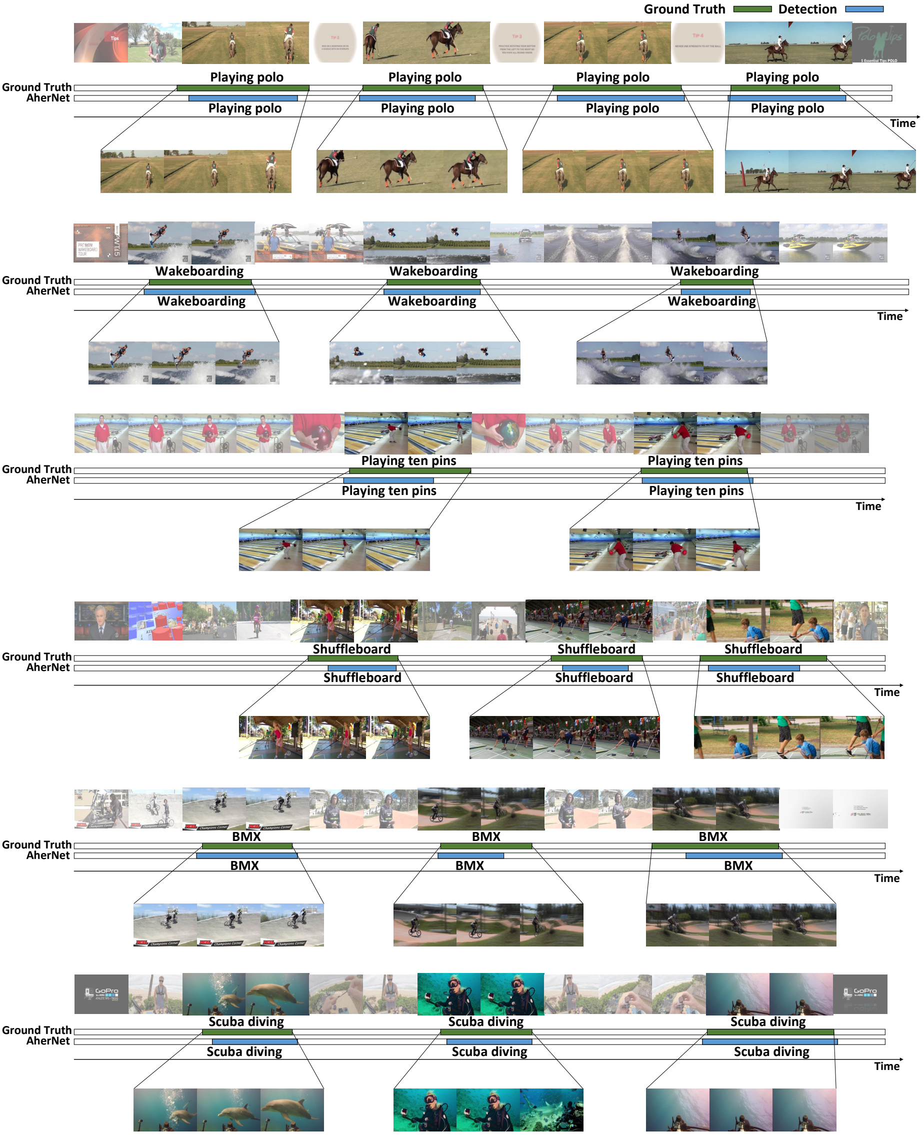

Fig. 1: Examples of six temporal action localization results on the setting of TH14→ANet-FG. In each example, a video is shown as a sequence of frames at the top. The green boxes in the upper bar denote the ground truth proposals, whose sampled frames are illustrated at the bottom. The localization results are shown in the lower bar, where a blue box denotes a predicted proposal from AherNet on the condition of IoU$\geq$0.7.

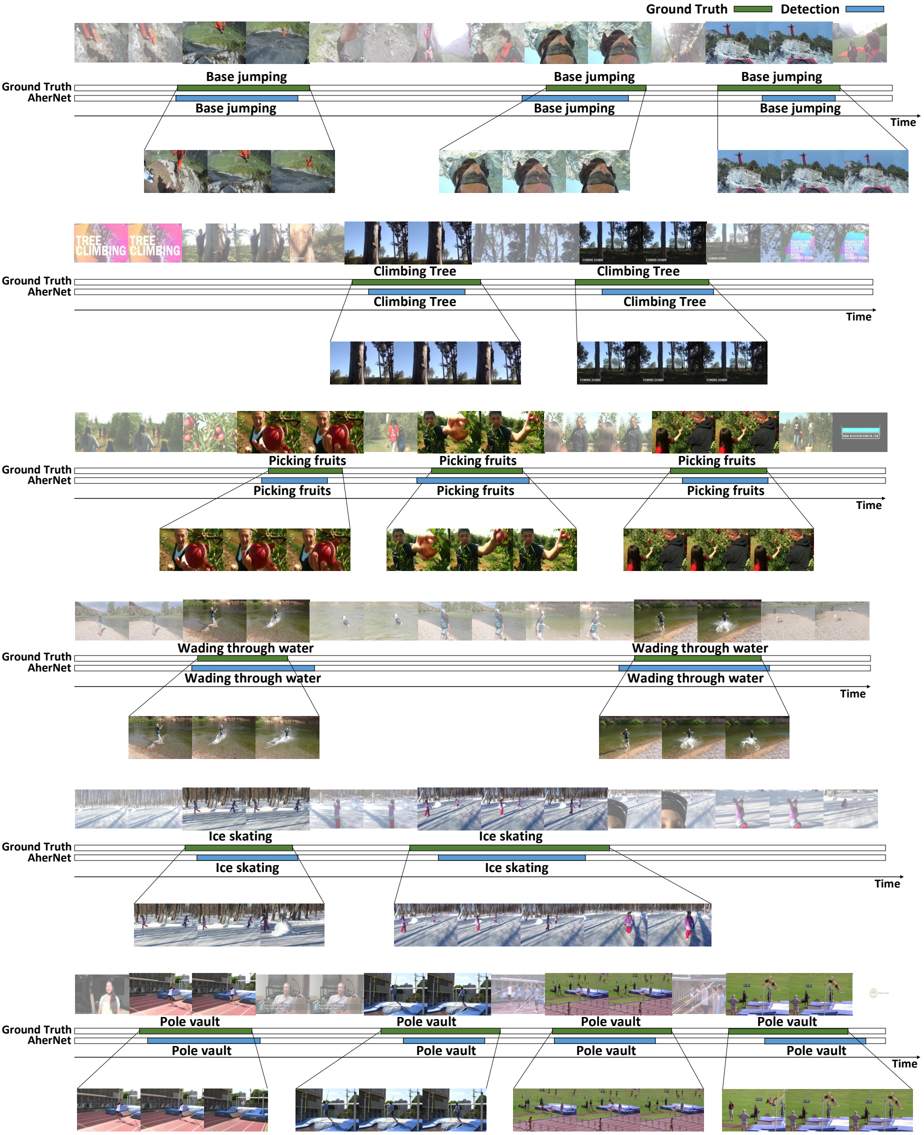

Fig. 2: Examples of six temporal action localization results on the setting of ANet→K600. In each example, a video is shown as a sequence of frames at the top. The green boxes in the upper bar denote the ground truth proposals, whose sampled frames are illustrated at the bottom. The localization results are shown in the lower bar, where a blue box denotes a predicted proposal from AherNet on the condition of IoU≥0.7.