Supplementary material for: Actor and Action Video Segmentation from a Sentence

Kirill Gavrilyuk, Amir Ghodrati, Zhenyang Li, Cees G. M. Snoek
QUVA Lab, University of Amsterdam
{kgavrilyuk, a.ghodrati, zhenyangli, cgmsnoek}@uva.nl

In this supplementary material, we first report annotation statistics on both the A2D Sentences and J-HMDB Sentences datasets in Section 1. In Section 2, we show more segmentation results of our proposed model followed by a qualitative comparison of our video-based model with the image-based models of Hu *et al*. [1] and Li *et al*. [2] in Section 3.

1. Dataset statistics

We show some statistics of the annotated sentences on A2D and J-HMDB datasets. Figure 1 shows the most frequent nouns and verbs in the A2D Sentences dataset. Segmentation from a sentence allows us to distinguish between the fine-grained actors in the same super-category. For example while in the normal A2D dataset [3] there is a general 'adult' category, we annotate fine-grained human actors like {man, woman, guy, person, girl, boy, ...} in A2D Sentences. Furthermore, natural language sentences enable us to make use of a richer set of verbs to describe the same type of action, *e.g*. {jumping (up and down), bouncing, falling} all are representative for the action label 'jumping' in the regular A2D dataset. Likewise, {flipping, turning, rolling, rotating} are representative for the action label 'rolling', and {moving, running, chasing} are representative for 'running'. Figure 2 shows the most frequent nouns and verbs in the J-HMDB Sentences dataset.

2. Segmentation results on A2D Sentences

In this section, we visualize more results of the sentence-guided segmentation using our model. Figure 3 illustrates videos with only one type of actor performing the same action. Our model segments both deformable (*e.g*., the 'woman' in the second video) and non-deformable (*e.g*., the 'ball' in the first video) objects. Also, it can handle reflecting surfaces, indicated by the 'ball' example. The third video demonstrates the ability of our model to distinguish instances among the same actor and action type by language cues like the spatial location provided in the sentence descriptions. Figure 4 illustrates videos showing human actions. While the first two videos prove the ability of our model to recognize different human actions, the last video shows a failure case of our model. The model is asked to segment 'man' and 'woman' separately, while it segments both.

3. Baseline comparison on A2D Sentences

In this section, we show a qualitative comparison of our model with two image-based baselines by Hu *et al*. [1] and Li *et al*. [2] in Figure 5. The first two rows verify that our model is able to segment relatively small actors, while both baselines struggle. The next two rows demonstrate the better segmentation accuracy of our model in comparison to the baseline models. For example, in the fourth row our model segments the car as a whole, while both baselines segment parts of the car only. In the last row, we illustrate the ability of our model to better distinguish between different types of actors.

References

[1] R. Hu, M. Rohrbach, and T. Darrell. Segmentation from natural language expressions. In *ECCV*, 2016. 1, 5

[2] Z. Li, R. Tao, E. Gavves, C. G. M. Snoek, and A. W. M. Smeulders. Tracking by natural language specification. In *CVPR*, 2017. 1, 5

[3] C. Xu, S.-H. Hsieh, C. Xiong, and J. J. Corso. Can humans fly? action understanding with multiple classes of actors. In *CVPR*, 2015. 1

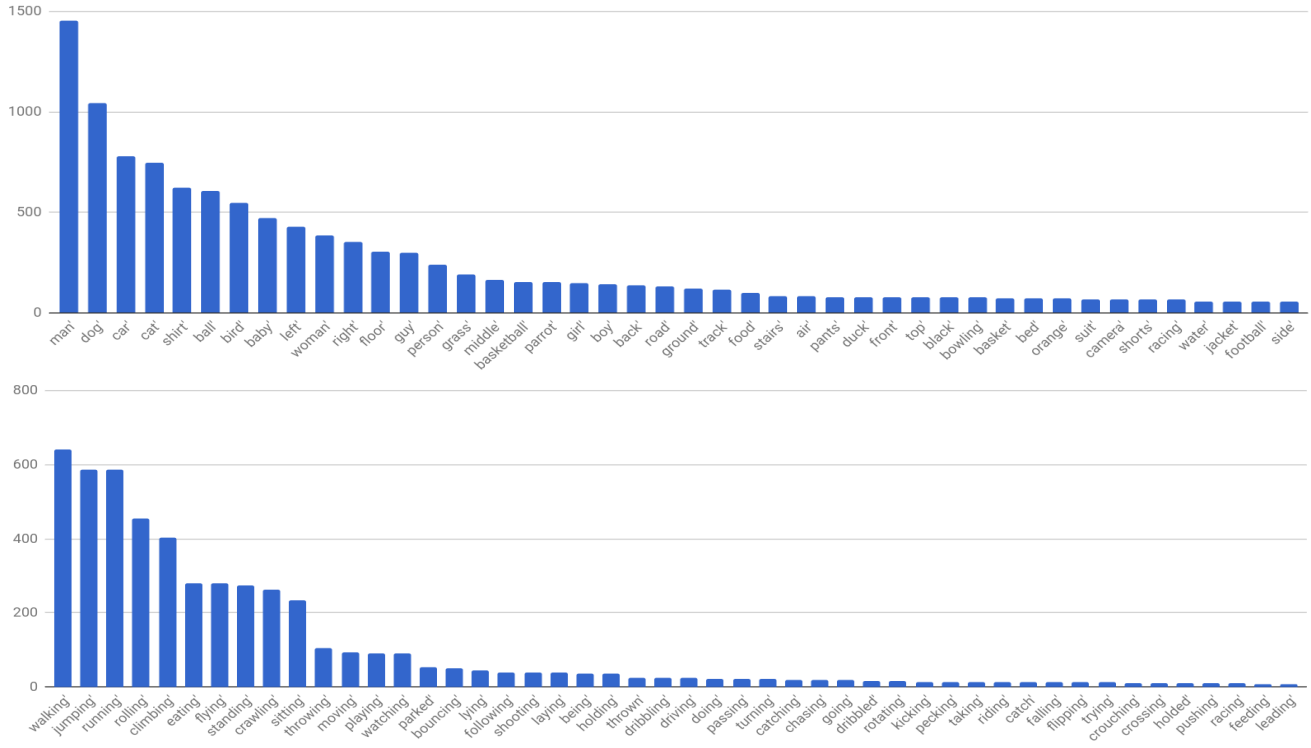

Figure 1: Most frequent nouns (top) and verbs (bottom) in the A2D Sentences dataset.

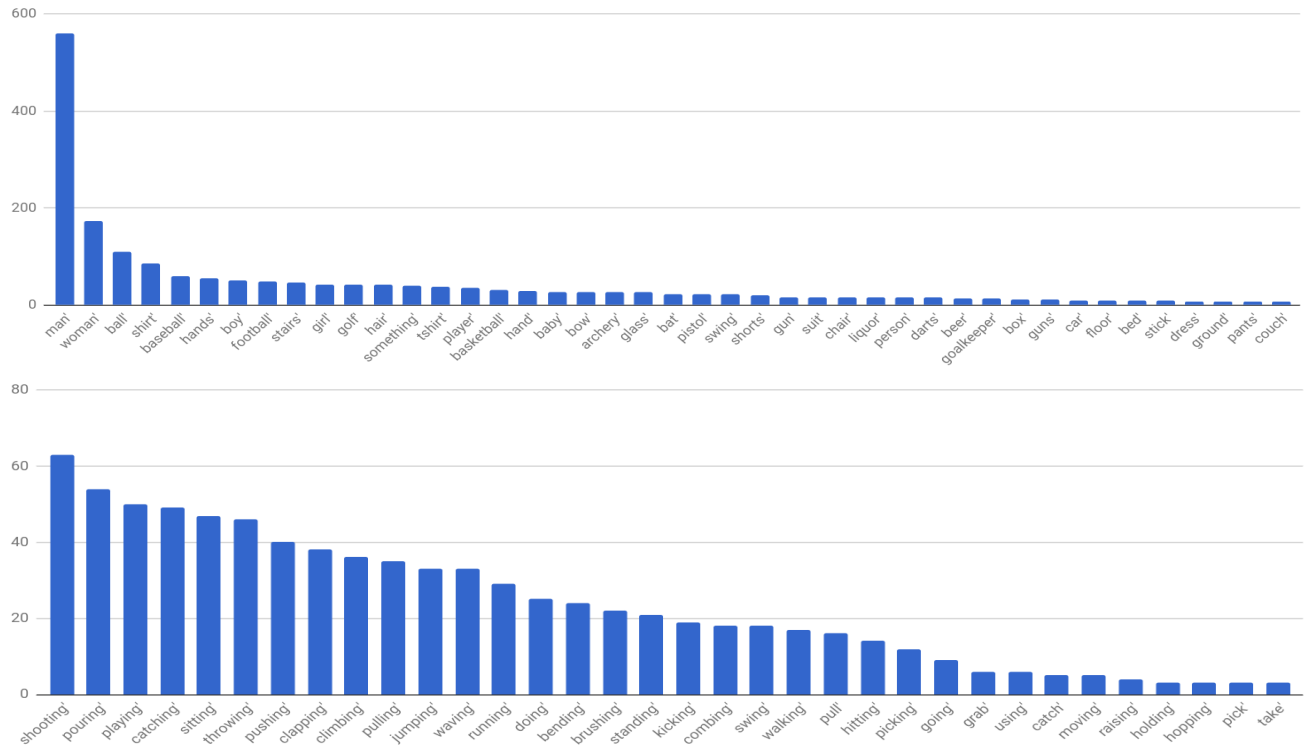

Figure 2: Most frequent nouns (top) and verbs (bottom) in the J-HMDB Sentences dataset.

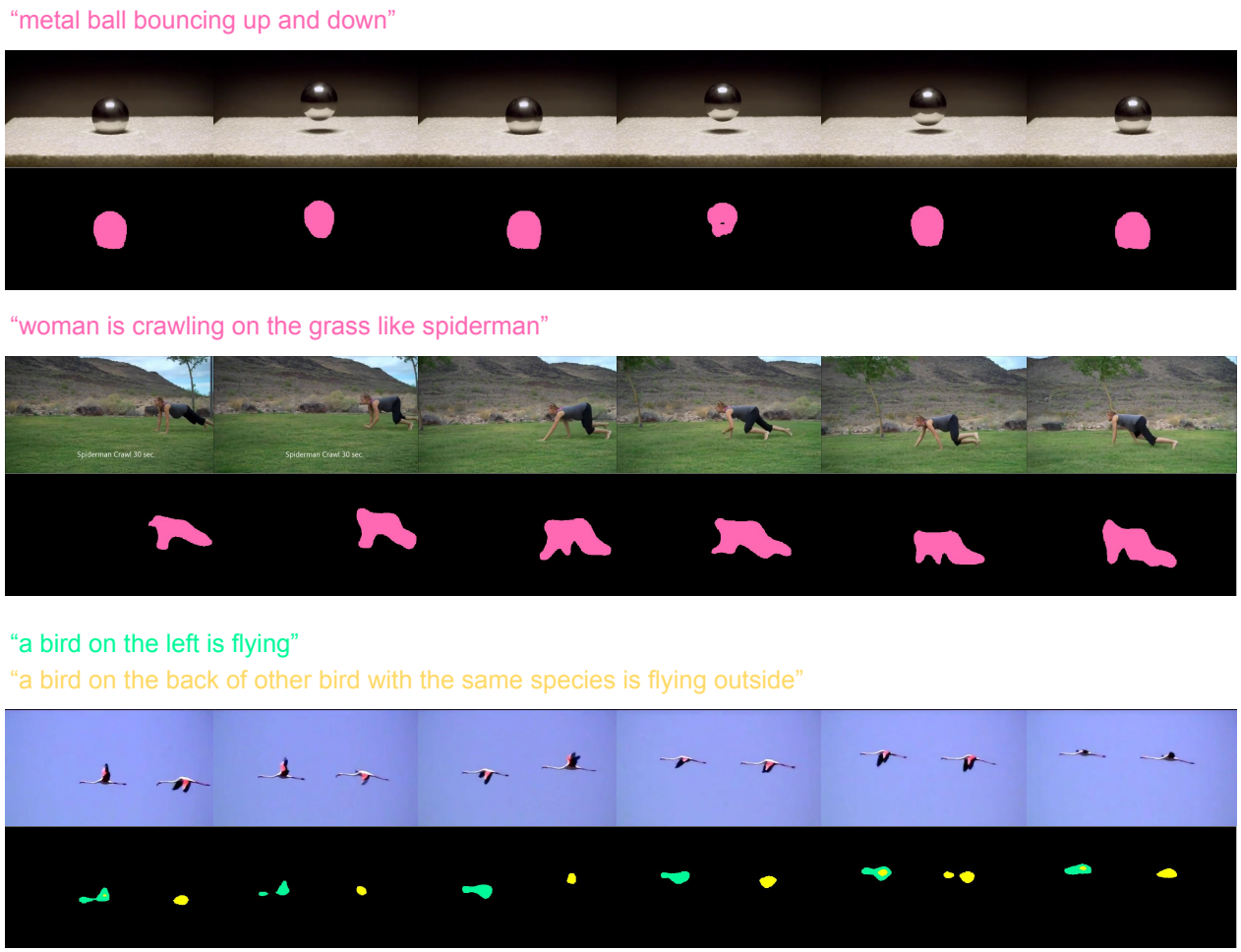

Figure 3: Visualized segmentation results from our model on A2D Sentences. In all rows we show examples with only one type of actor performing the same action. The first two videos illustrate examples with one single instance while the last video contains two instances. The colored segmentation masks are generated from the sentence with the same color above each video.

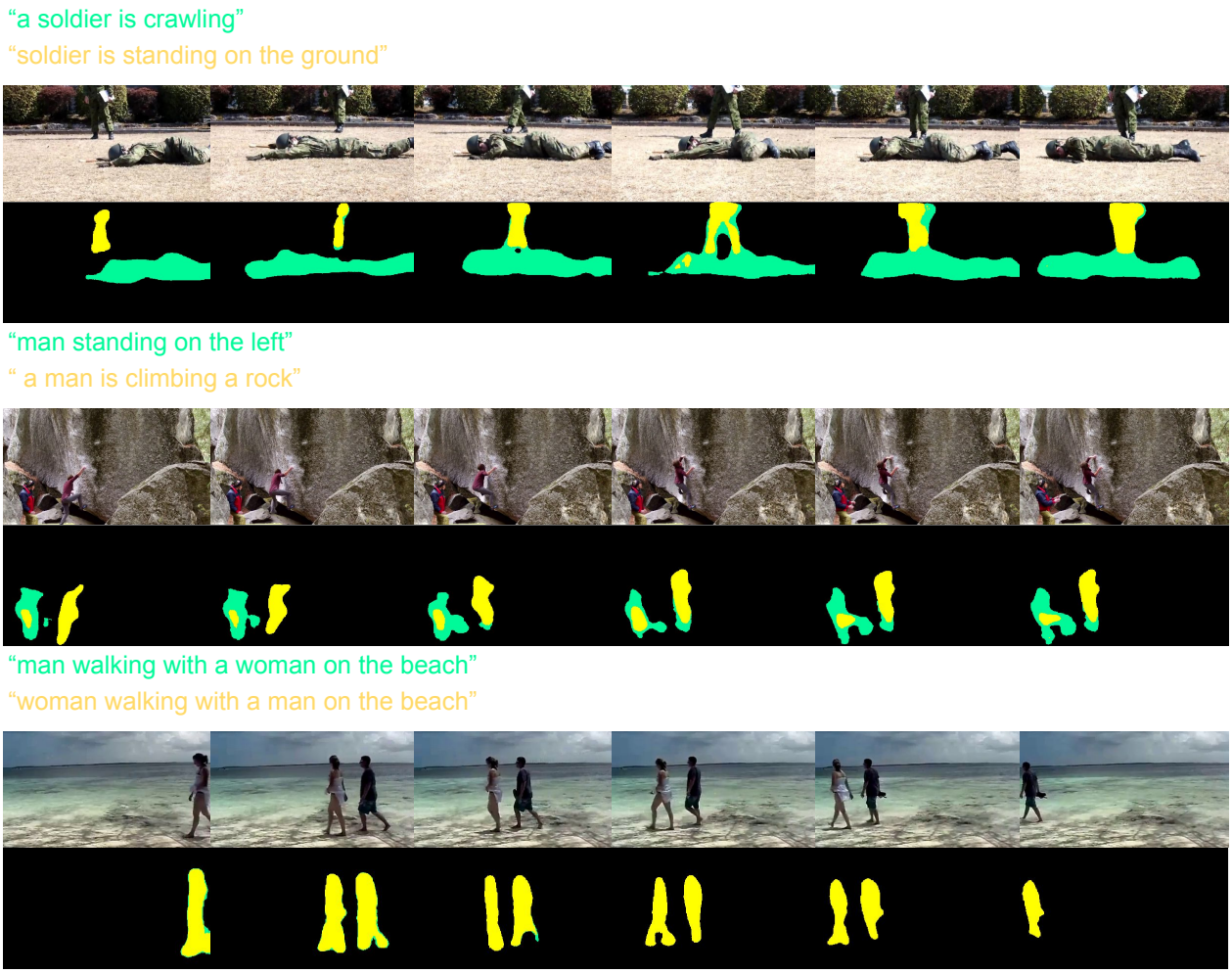

Figure 4: Visualized segmentation results from our model on A2D Sentences. In the first two rows we show examples with one type of actor performing different actions. The last row illustrates a failure case of our model. The colored segmentation masks are generated from the sentence with the same color above each video.

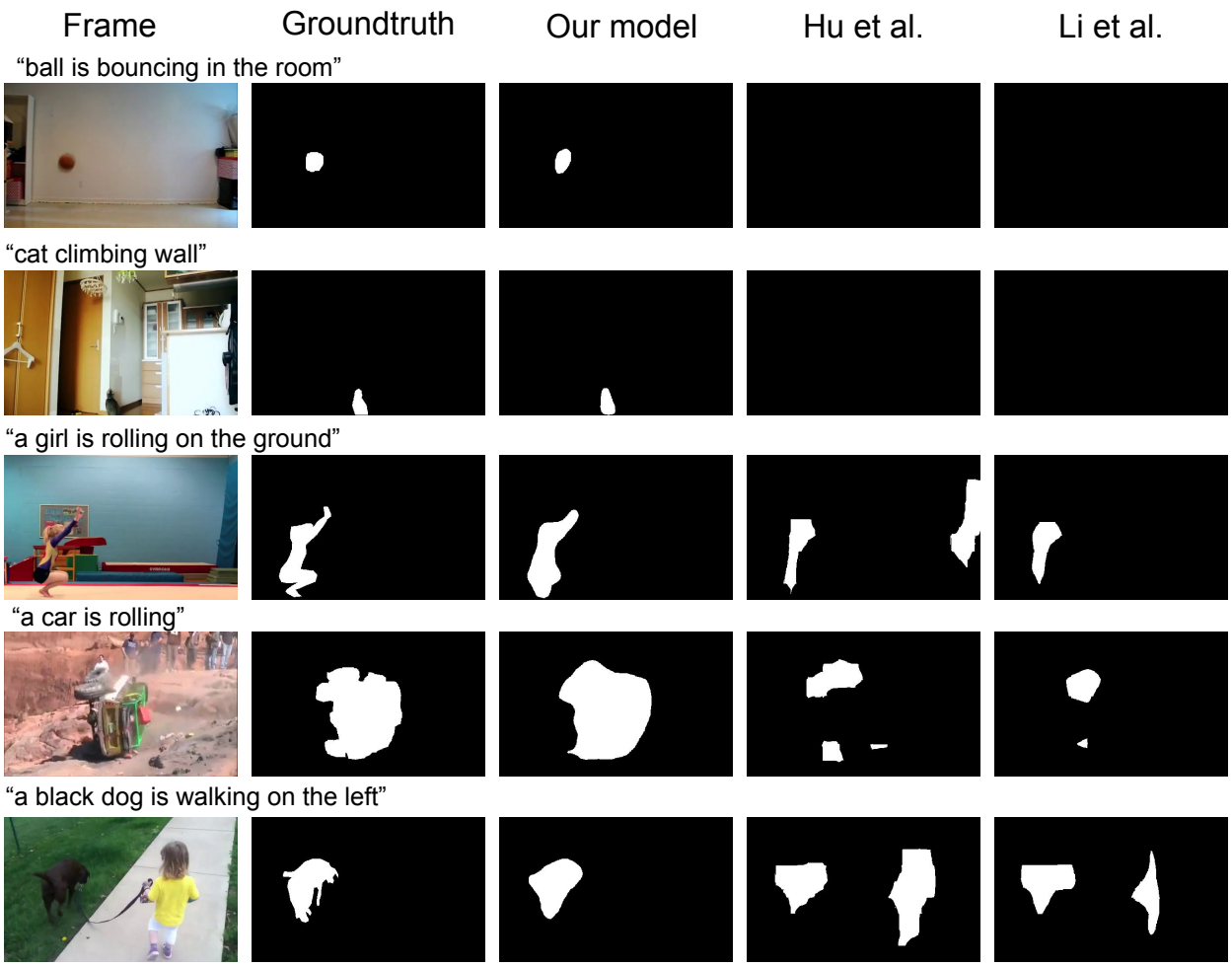

Figure 5: Qualitative results on A2D Sentences. Columns from left to right are frame to segment, groundtruth segmentation, our model output, output of Hu *et al*. [1] and output of Li *et al*. [2]. Above each example there is a sentence used as input for all methods describing what to segment in the frame.