Supplementary Material for "Action Segmentation with Mixed Temporal Domain Adaptation"

Min-Hung Chen[1*] Baopu Li[2] Yingze Bao[2] Ghassan AlRegib[1]
[1]Georgia Institute of Technology [2]Baidu USA
cmhungsteve@gatech.edu {baopuli, baoyingze}@baidu.com alregib@gatech.edu

1. More Qualitative results

1.1. GTEA dataset

Here we show more examples to compare our approaches with the baseline model MS-TCN [1] and the ground truth, as shown in Figure 1.

For the example *Hotdog* (Figure 1a), MS-TCN fails to predict the *put* action before the *take* action in the early part of the video, and the predicted *fold* action in the end of the video has much shorter time duration than it should be, as shown in the "Source only" row. With local and global temporal DA, our approach can detect all the actions with proper time duration, as shown in the row "DA (L + G)", and the domain attention mechanism further helps refine the time duration for each predicted action. For another example *Pealate* (Figure 1b), our final approach "DA (L + G + A)" also produces the best action segmentation result, which is the closest to the ground truth, compared with the baseline and other methods.

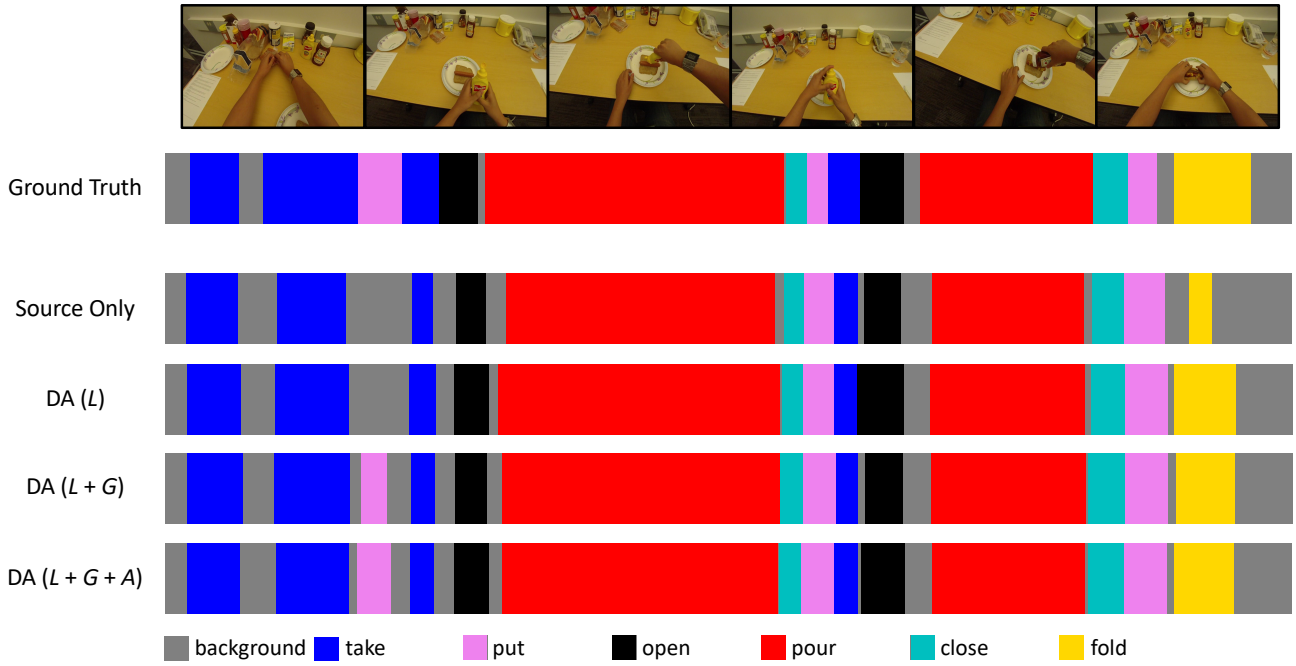
1.2. 50Salads and Breakfast datasets

In addition to the GTEA dataset, we also evaluate the qualitative performance on another two challenger datasets: *50Salads* and *Breakfast*, as demonstrated in Figure 2. The 50Salads dataset is challenging since each video is long and contains around 20 fine-grained action classes. While MS-TCN confuses with some similar classes like *"cut tomato"* and *"place tomato into bowl"*, our approach can produce smooth temporal segmentation, as shown in Figure 2a. The Breakfast dataset is challenging because of high spatio-temporal variations among videos since the videos are recorded in 18 different kitchens with 52 subjects. The total number of action classes is also much larger than the other two datasets. Figure 2b shows that MS-TCN falsely classifies *"take cup"* as *"take bowl"* and unable to detect the *"add teabag"* action for a long time. However, our proposed MTDA can continuously detect actions in the video without gaps along the temporal direction.

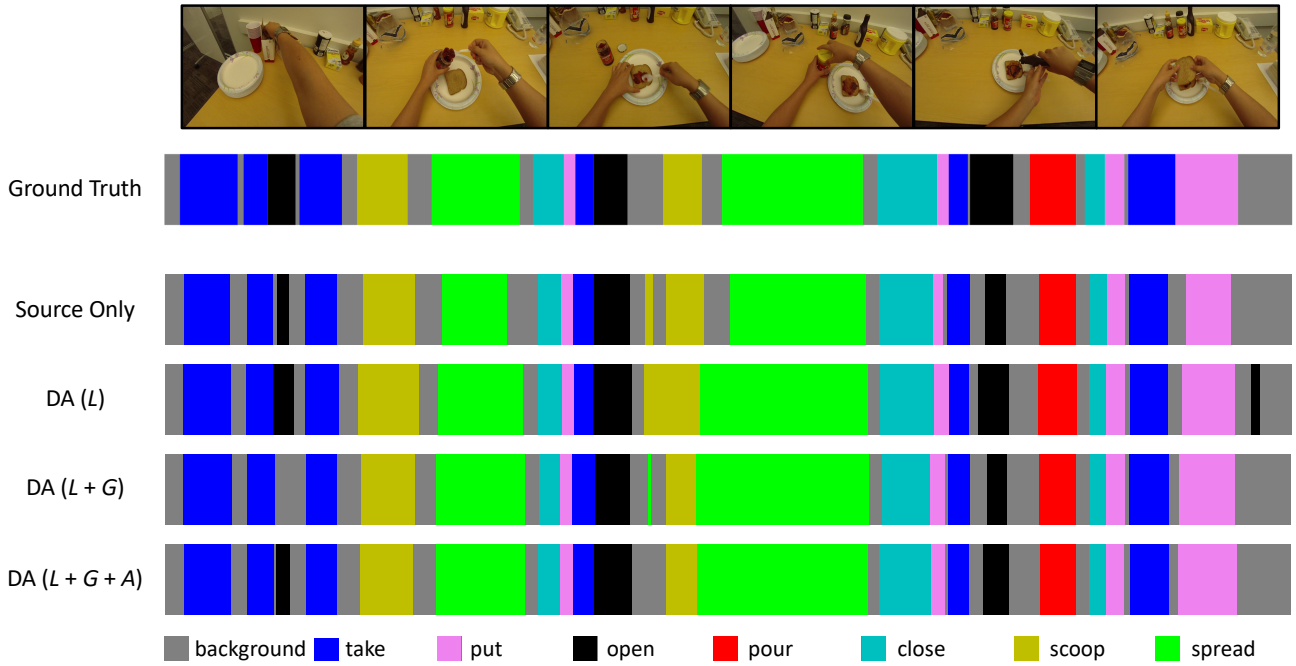
*Work done during an internship at Baidu USA

References

[1] Y. A. Farha and J. Gall. Ms-tcn: Multi-stage temporal convolutional network for action segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 2, 3

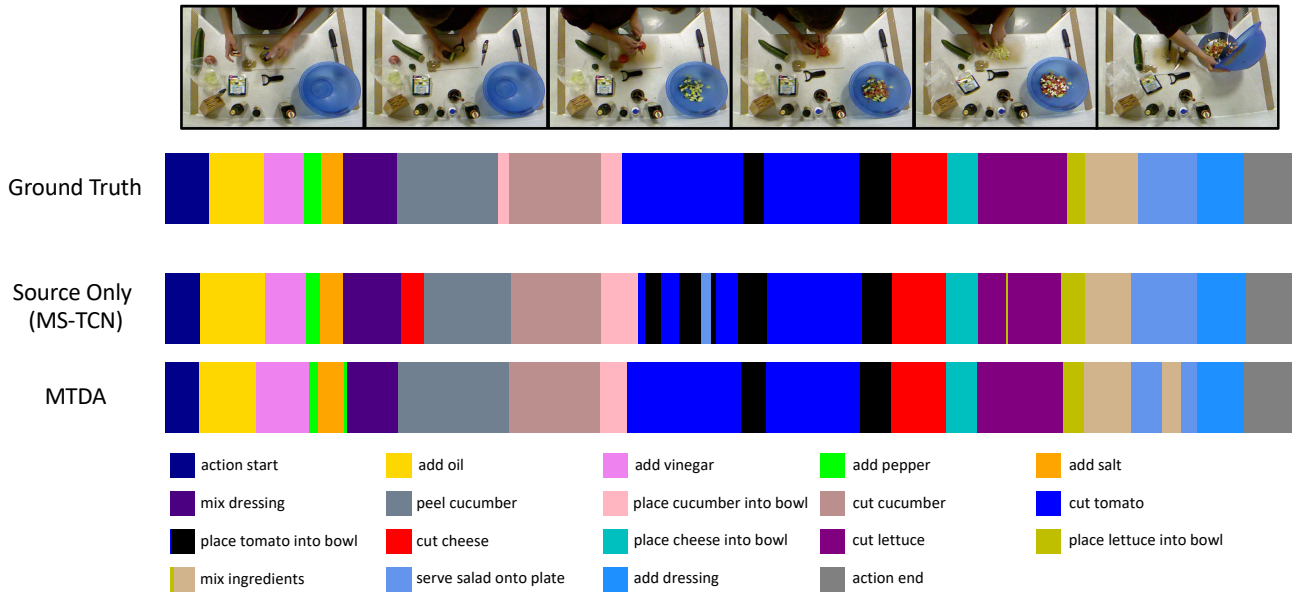


(a) *Hotdog*

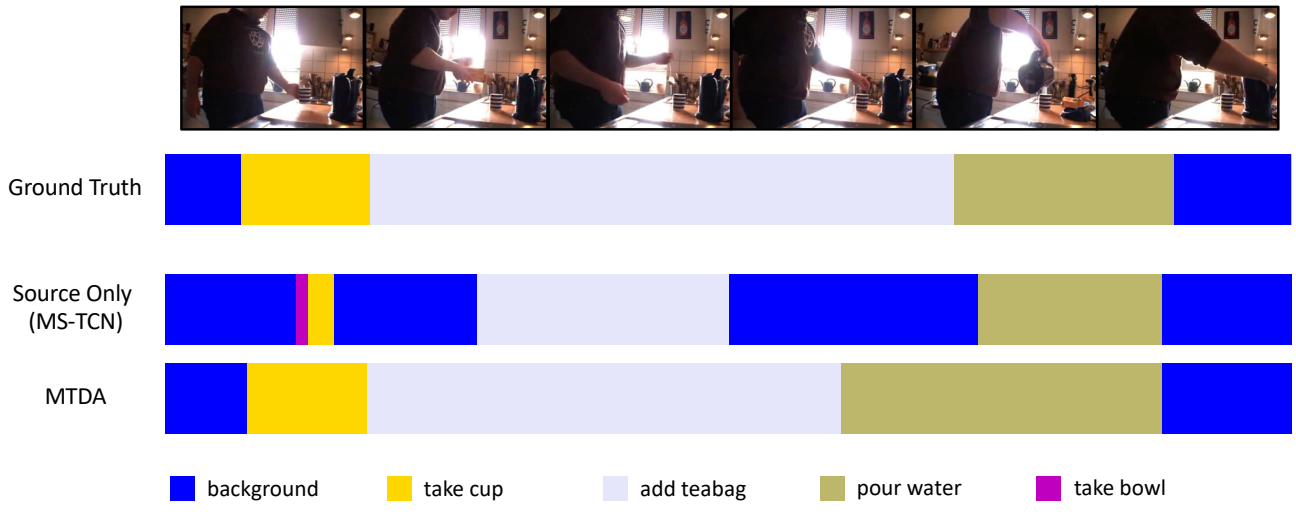


(b) *Pealate*

Figure 1: The qualitative results of temporal action segmentation on GTEA for the activity (a) *Hotdog* and (b) *Pealate*. The video snapshots are shown in the first row in a temporal order (from left to right). "Source only" refers to the baseline model MS-TCN [1].



(a) *50Salads*



(b) *Breakfast*

Figure 2: The qualitative results of temporal action segmentation on (a) *50salads* and (b) *Breakfast*. The video snapshots are shown in the first row in a temporal order (from left to right). "Source only" refers to the baseline model MS-TCN [1].