Supplementary Material: Attention is All We Need: Nailing Down Object-centric Attention for Egocentric Activity Recognition

Swathikiran Sudhakaran[1,2]
sudhakaran@fbk.eu

Oswald Lanz[1]
lanz@fbk.eu

[1]University of Trento
Trento, Italy

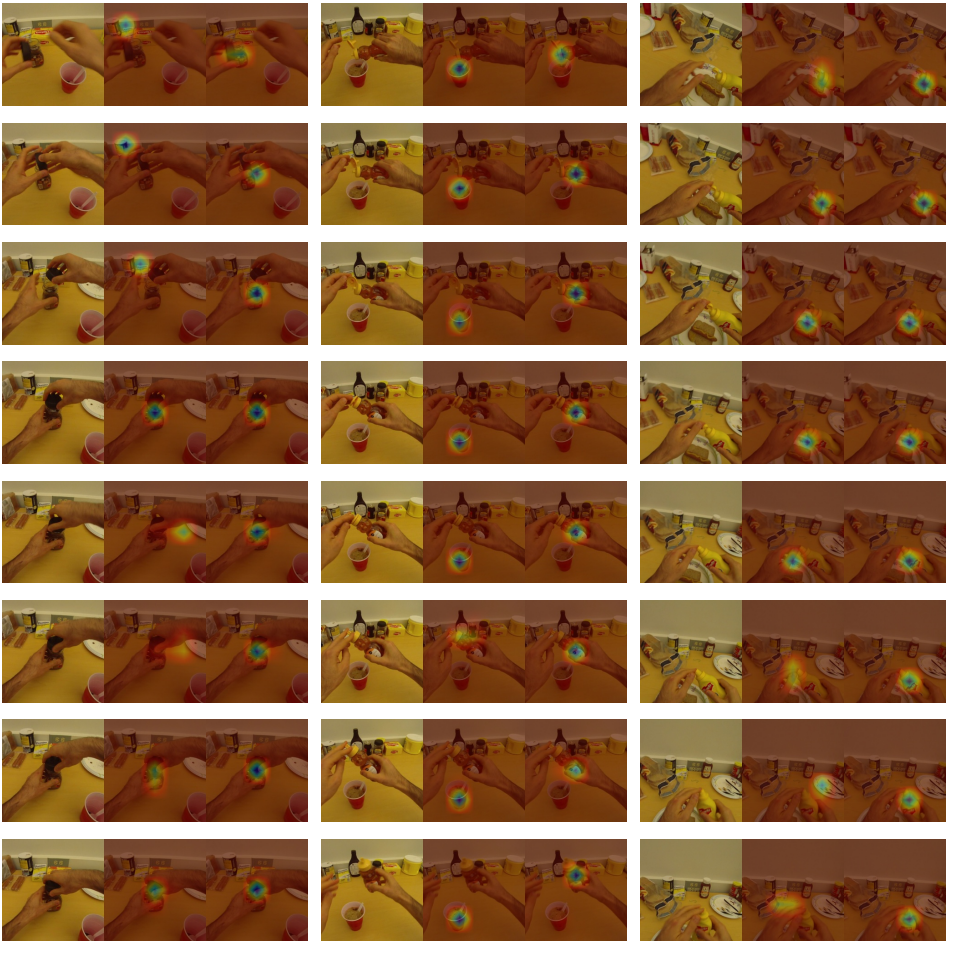
[2]Fondazione Bruno Kessler
Trento, Italy

This supplementary material of our BMVC submission shows:

- More examples of the proposed spatial attention map generation technique
- The confusion matrix obtained for subjects P1 and P3 of the GTEA Gaze+ dataset

1 GTEA 61 attention maps

Here, we show additional images with the attention map generated by the network as discussed in section 4.3 of the paper. The videos are from the GTEA 61 dataset mentioned in our paper. In the figures, each column consists of the frames extracted from the same video. In each image, the first one represents the input image, second one, the attention map obtained from the imagenet pre-trained ResNet-34 network and the third image shows the attention map obtained after the proposed fine-tuning technique(after stage 2 training of our network). From the figures, it can be seen that the network learns to attend to the relevant objects that characterize each activity and gets improved after the fine-tuning step.

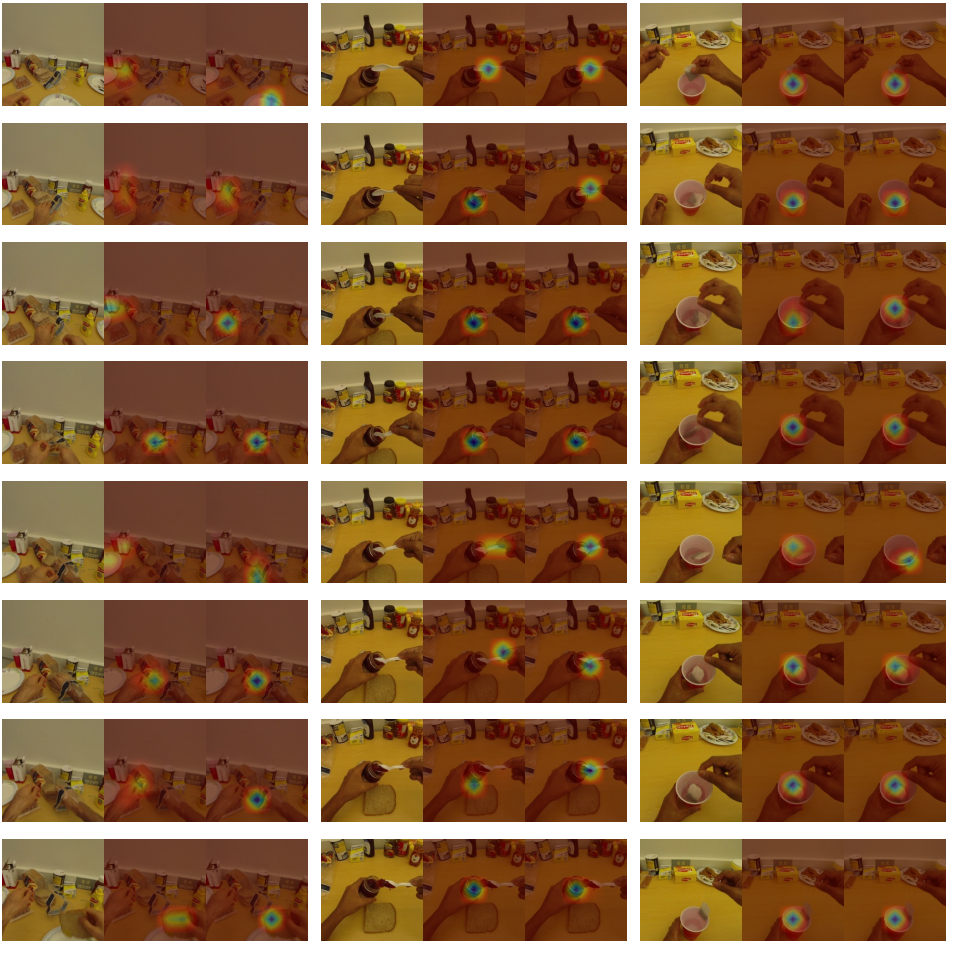


(a) Close coffee (b) Close honey (c) Close mustard

Figure 1: Spatial attention maps obtained for frames from GTEA(61) dataset. The clip identifiers are: (a) S1_Coffee_C1 (b) S1_CofHoney_C1 (c) S1_Hotdog_C1

(a) Pour coffee (b) Pour mustard (c) Pour mayonnaise

Figure 2: Spatial attention maps obtained for frames from GTEA(61) dataset. The clip identifiers are: (a) S1_Coffee_C1 (b) S1_Hotdog_C1 (c) S1_Cheese_C1



(a) Take bread (b) Scoop jam (c) Shake tea

Figure 3: Spatial attention maps obtained for frames from GTEA(61) dataset. The clip identifiers are: (a) S1_Hotdog_C1 (b) S2_Pealate_C1 (c) S2_Tea_C1

2 Confusion matrix of GTEA Gaze+ dataset

As discussed in section 4.3 of the paper, we show the confusion matrices obtained when P1 and P3 are used as the test split of the GTEA Gaze+ dataset. These two splits resulted in the least recognition accuracy out of all the splits. P1 gave an accuracy of 50.2% while P3 resulted in 48.84%. We can see that the accuracy of classes containing the meta-object labels ('spoonForkKnife' and 'cupPlateBowl') is low compared to the other classes. This verifies our hypothesis explained in section 4.3 regarding the reason for the lower performance of the proposed method on GTEA Gaze+ dataset compared to the methods that use strong supervision for training.

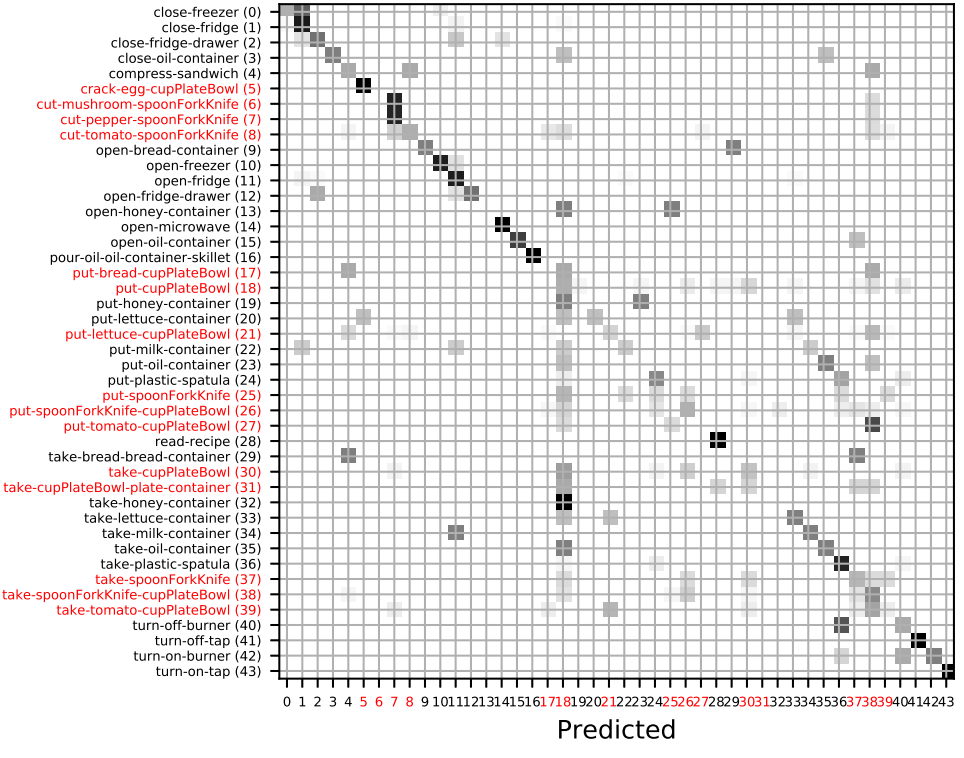


Figure 4: Confusion matrix of split P1 from GTEA Gaze+ dataset

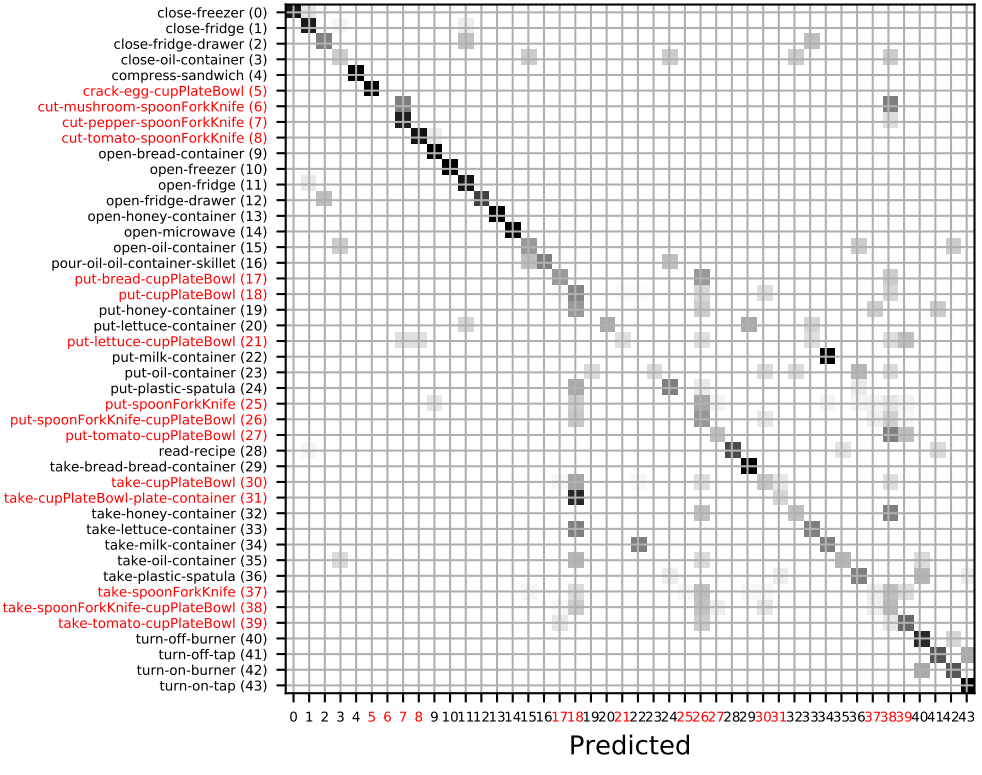


Figure 5: Confusion matrix of split P3 from GTEA Gaze+ dataset