Supplementary Materials

Han Wei[1]
whan003@e.ntu.edu.sg
Huang Hantao[2]
hantao.huang@mediatek.com
Yu Xiaoxi[2]
xiaoxi.yu@mediatek.com

[1] Nanyang Technological University
Singapore

[2] MediaTek, Inc.
Singapore

1 Part-wise visualization of attention maps and predicted locations

In Figure 1, We visualize the attention maps (As introduced in Section 3.2 of the paper. The attention for visualization is calculated by a Softmax function instead of the Gumbel-Softmax) and predicted part centers of 9 selected target parts from the template image in each of the two examples. The center location of the 9 selected parts are indicated with the points of 9 different colors in the top left image. The 9 images in row 2, 3, and 4 are the attention maps on the search region corresponding to each selected target part (with the same spatial ordering). The predicted part location of each target part is also plotted in each corresponding attention map using point with the same color. The top right image shows an ensemble of attention maps from all target parts (includes but not only includes the 9 selected points). The same observations are made in the two examples: 1. The predicted location coincide well with the attention map peak due to the proposed attention loss. 2. The predicted target part location in the search region does not always match with the corresponding target part. For example, in the right example, the green point in the template is located at the neck of the person, while in the attention map, the predicated location is on the head of the person. Since our attention-guided supervision is flexible, the network may learn to localize the most discriminative pattern inside the target part's local receptive field during training.

2 Visualization of accumulated attention maps

In Figure 2, we visualize the accumulated attention maps from more examples. Each column corresponds to one example. The images in the top, middle and bottom rows show the resized template image and ground truth bounding box, the search region image and the accumulated attention map respectively. It can be concluded that our network learns to attend to the foreground locations under large appearance changes.

1 Partial work is done during working at Mediatek Singapore.

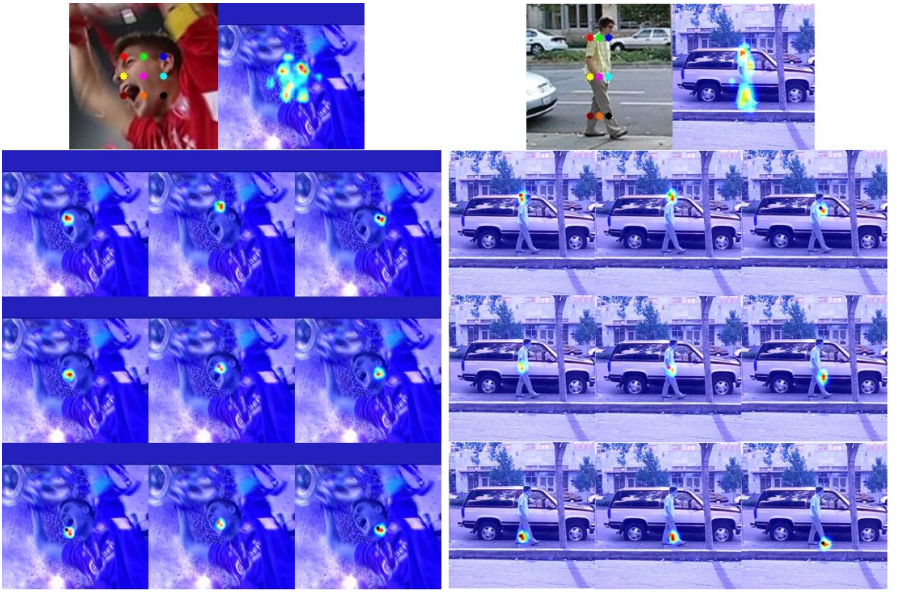

Figure 1: Part-wise visualization of attention maps and predicted locations. The center location of the 9 selected parts are indicated with the points of 9 different colors in the top left image. The 9 images in row 2, 3, and 4 are the attention maps on the search region corresponding to each selected target part (with the same spatial ordering). The predicted part location of each target part is also plotted in each corresponding attention map using point with the same color. The top right image shows an ensemble of attention maps from all target parts

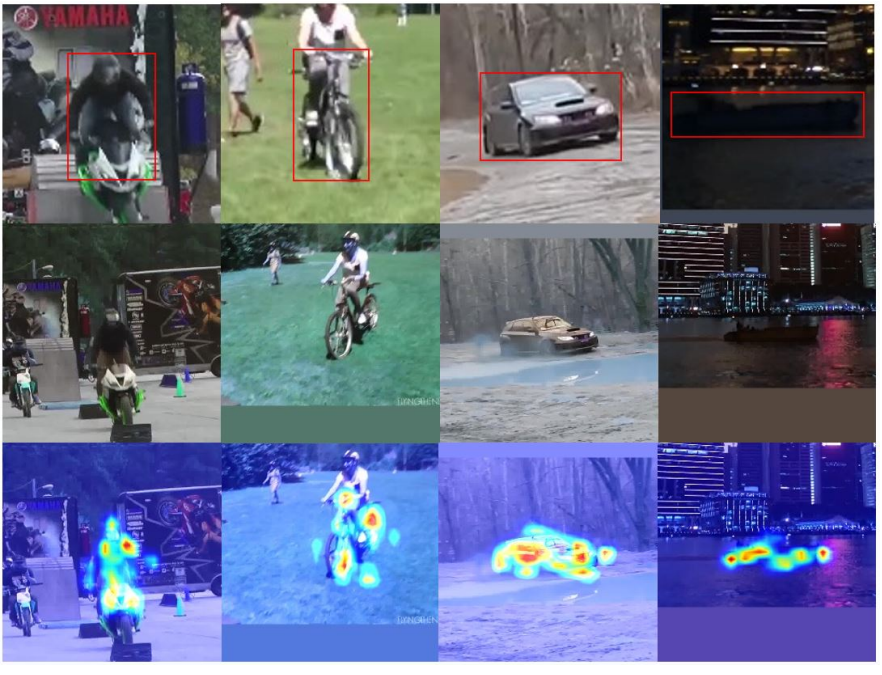

Figure 2: Visualization of accumulated attention maps. The top, middle and bottom rows show the resized template image and ground truth bounding box, the search region image and the accumulated attention map respectively.