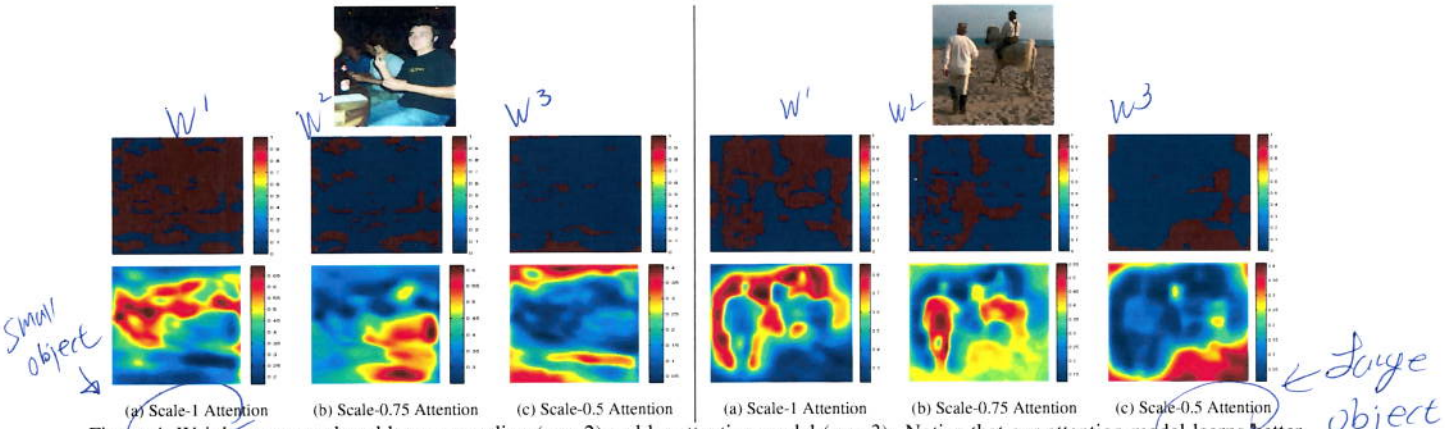


Figure 4. Weight maps produced by max-pooling (row 2) and by attention model (row 3). Notice that our attention model learns better interpretable weight maps for different scales. (a) Scale-1 attention (*i.e.*, weight map for scale $s = 1$) captures small-scale objects, (b) Scale-0.75 attention usually focuses on middle-scale objects, and (c) Scale-0.5 attention emphasizes on background contextual information.

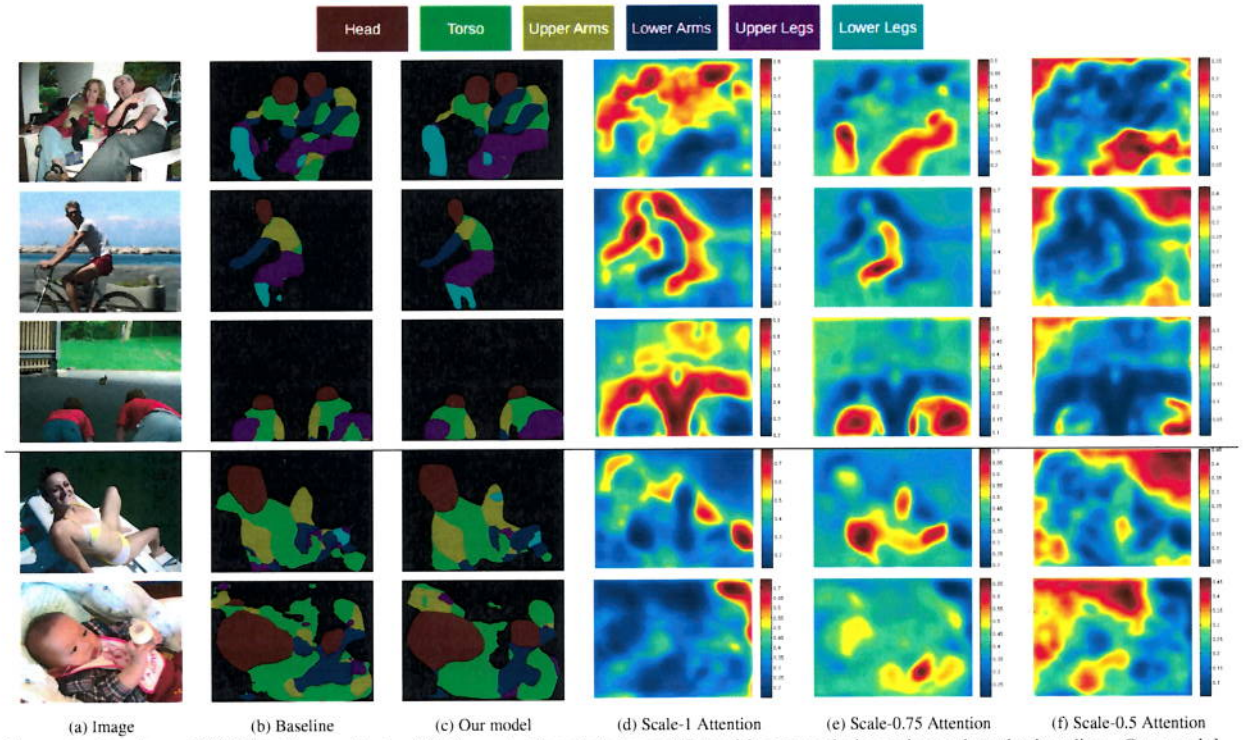


Figure 5. Results on PASCAL-Person-Part *validation* set. DeepLab-LargeFOV with one scale input is used as the baseline. Our model employs three scale inputs, attention model and extra supervision. Scale-1 attention captures small-scale parts, scale-0.75 attention catches middle-scale torsos and legs, while scale-0.5 attention focuses on large-scale legs and background. Bottom two rows show failure examples.

set, respectively. They are similar to those (62.25% and 64.21%) reported in [11]. We report results of the proposed methods on the validation set in Tab. 3. We observe similar experimental results as PASCAL-Person-Part dataset: (1) Using two input scales is better than single input scale. (2) Adding extra supervision is necessary to achieve better performance for merging three input scales, especially for average-pooling and the proposed attention model. (3) The best performance (6.8% improvement over the DeepLab-LargeFOV baseline) is obtained with three input scales, attention model, and extra supervision, and its performance is 4.69% better than DeepLab-MSc-LargeFOV (64.39%).