A. Appendix

A.1. Visual Results

Figure 1 presents examples of predicted masks of PuzzleCAM and our CLIMS for the class of train. It can be observed in the second column that, PuzzleCAM [2] usually falsely discriminates the railroad as train, which leads to an overestimation of the target region of train. By contrast, our CLIMS successfully segment the correct regions of train, without the involvement of railroad.

We also present examples of masks predicted by PuzzleCAM and our CLIMS for the class of boat and person in Figure 2. It can be obviously seen that, compared with PuzzleCAM, less false positives of water and more complete regions of person are predicted in the semantic masks of our CLIMS.

A.2. Semi-supervised Semantic Segmentation

The generated pseudo ground-truth masks can be further used in the semi-supervised semantic setting. We try to use the generated masks to improve the performance of the semi-supervised method CCT [5]. Table 1 presents the performance comparison with those recent semi-supervised methods. It can be observed that with our masks, CCT can obtain a significant improvement of 3.7% mIoU on the PASCAL VOC2012 val set.

A.3. Limitation

CLIMS leverages the power of CLIP [7] to recognize those diverse backgrounds. However, as the text descriptions of objects in the segmentation dataset, i.e., PASCAL VOC2012, may be different from that used to train CLIP, many objects like person cannot be identified well using CLIP. In this work, we must finetune the CLIP model using the text descriptions of PASCAL VOC to mitigate the above issue.
Figure 2. Examples of predicted semantic masks for the class of boat and person on PASCAL VOC2012 val set. Best viewed in color.
References


