Leveraging Class Distributions in CLIP for Weakly Supervised Semantic Segmentation

Supplementary Material

This supplementary material provides more details and results that are not included in the experiment part of the main paper. The contents are organized as follows:

- Experimental Settings
- Experimental Results
- Ablation Studies

1. Experiments

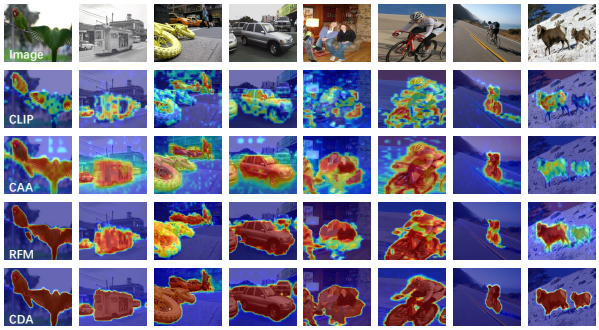

Figure 1. More visualization comparison of CAMs between other modules and our CDA method. The experimental dataset is PASCAL VOC 2012 train split.

1.1. Experimental Settings

Implementation Details We trained our CD-CLIP on the PASCAL VOC 2012 dataset using a single NVIDIA RTX 3090 GPU. Moreover, for MS COCO 2014, it was trained on four NVIDIA RTX 3090 GPUs. The designed super-class set D_s is {animal, vehicle, indoor items, outdoor items, furniture, person, kitchenware, accessory, appliance, electronic items, foods, sports}.

1.2. Experimental Results

Visualizations We provide additional comparative visualizations of Class Activation Maps (CAMs) refined by different affinity modules in Figure 1. As demonstrated in the figure, the initial CAMs generated by the CLIP model exhibit the common limitation of activating only the most discriminative regions of target objects. While the CAA module from CLIP-ES [1] expands the overall activation regions, the resulting CAMs still suffer from class confusion and background noise. In contrast, the RFM module from WeCLIP [2] effectively mitigates these issues by leveraging feature similarity as a noise filter for relationship modeling. However, CAMs refined by RFM tend to contain underactivated regions that are nevertheless accurately activated by our proposed CDA module, demonstrating its superior capability in generating comprehensive and precise CAMs.

1.3. Ablation Studies

Table 1. Ablation study of distance measurement selection. 'M' denotes the mIoU (%) of CAM performance.

Distance	KL	WS	JS
M	79.5	79.1	80.8

Analysis of Jensen-Shannon (JS) Divergence In this paper, we leverage a class-distribution similarity map based on the Jensen–Shannon (JS) divergence to rectify the initial affinity. Compared with the other two measurements, the Kullback–Leibler (KL) divergence is asymmetric, making it incompatible with the symmetric form of the initial affinity. In addition, the Wasserstein distance (WS) is difficult to apply to class-wise comparison in semantic segmentation and is generally insensitive to small distributional variations. For these reasons, JS divergence serves as the most suitable choice in our framework. As further validated in Table 1, using JS divergence within our CDA module yields the best CAM performance.

Analysis of SBE module Figure 2 shows scenarios with adjacent target classes (*e.g.*, 'person' and 'motorbike'). As observed in the "CDA only" setting, although the CDA successfully activates the object regions, it tends to overactivate at the boundaries between adjacent objects, leading to blurred separation. In contrast, the "CDA +SBE" setting demonstrates that introducing the SBE module effectively segments these boundaries. This visual improvement also corresponds to the quantitative results in Table 3 of the main paper. Comparing Setting #3 and Setting #6, SBE achieves an improvement of 1.4% in segmentation, specifically by refining these boundary regions.

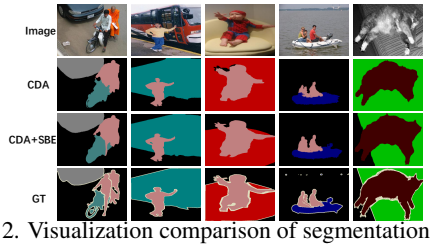

Figure 2. Visualization comparison of segmentation results.

References

[1] Yuqi Lin, Minghao Chen, Wenxiao Wang, Boxi Wu, Ke Li, Binbin Lin, Haifeng Liu, and Xiaofei He. Clip is also an efficient segmenter: A text-driven approach for weakly supervised semantic segmentation. In *CVPR*, 2023. 1

[2] Bingfeng Zhang, Siyue Yu, Yunchao Wei, Yao Zhao, and Jimin Xiao. Frozen clip: A strong backbone for weakly supervised semantic segmentation. In *CVPR*, 2024. 1