Learn to Rectify the Bias of CLIP for Unsupervised Semantic Segmentation

Supplementary Material

Table 8. **Effect of Fine-tuning CLIP Image Encoder.** Results show that fine-tuning CLIP on small datasets may destroy the image-text alignment of CLIP and result in worse performance.

Setting	PASCAL VOC	PASCAL Context
Fine-tune	54.3	21.7
Frozen	**58.5**	**25.8**

6. Experimental Details

6.1. Training Details

We show the number of rectification epochs, the number of distillation epochs, data augmentations and batch size during training in Table 12.

6.2. Details about Fig. 1(a)

In an image, we calculate the centroid of each object with the help of ground-truth annotations. We further calculate the distance between the centroids of the object and the image. We split the distance interval into several bins (*e.g.,* $[0, 40]$, $[40, 80]$, *etc.*), and calculate the average IoU for each bin with the objects whose distances to the corresponding image centroids falling into this bin. The mIoU in this context means the IoU averaged across the objects falling into a certain bin, which is slightly different from the metric used for evaluating the segmentation performance of a model.

7. More Ablations

7.1. Fix or Fine-tune Image Encoder of CLIP

We jointly train the image encoder of CLIP on PASCAL VOC and PASCAL Context in Table 8 and find a consistent performance decrease (more than 4%). It is because fine-tuning CLIP on downstream small datasets may destroy the image-text alignment of pre-trained CLIP.

7.2. Sensitivity to t

We study the sensitivity of our method to the threshold t discussed in Sec. 3.4 and show the results in Table 9. It shows that the mIoU of our method firstly increases and then decreases as t gets larger, exhibiting a typical regularization effect on the training.

Table 9. **Sensitivity to threshold t.** The experiments are conducted on PASCAL VOC. According to the results, we set t to 0.23 in our rectification and distillation stage.

t	0.21	0.22	**0.23**	0.24	0.25
mIoU	57.6	57.9	**58.5**	57.6	57.2

7.3. Comparison to zero-shot/open-vocabulary semantic segmentation (OVSS) methods

In this paper, we focus on unsupervised semantic segmentation (USS). The assumption and training protocol of OVSS are quite different from those of USS. As methods for OVSS require additional large-scale annotations to train, *e.g.,* class-agnostic masks or image captions, it is unfair to directly compare those works with ours. We attempt to compare to OVSS methods under *the same unsupervised setting*. For fairness, we re-run the code of previous works with the same training data as ours. Results in Table 10 show our method outperforms previous OVSS methods remarkably.

Table 10. **Comparison with Open-Vocabulary methods.** For a fair comparison, we re-run the codes for previous OVSS methods under the USS setting.

Method	CLIP-based	Pre-training	VOC	Context
GroupViT [57]	✗	✗	very low	very low
GroupViT [57]	✗	CC12M+YFCC	very low	very low
TCL [5]	✓	✗	32.6	12.0
Ours (w/o. distill.)	✓	✗	**58.5**	**25.8**

7.4. Comparison after distillation

We apply distillation to previous works, *i.e.,* distilling the knowledge of previous models into DeepLab V2 on PASCAL VOC. As the code of CLIP-S4 is not released, we cannot conduct a comparison with it. Results in Table 11 (+ means distillation) further prove the effectiveness of our method.

Table 11. **Comparison after Distillation on PASCAL VOC.** We apply distillation to previous works with DeepLab V2 as the student on PASCAL VOC.

Model	CLIPpy+	ReCo+	MaskCLIP+	Ours
VOC	59.4	69.7	70.0	**75.4**

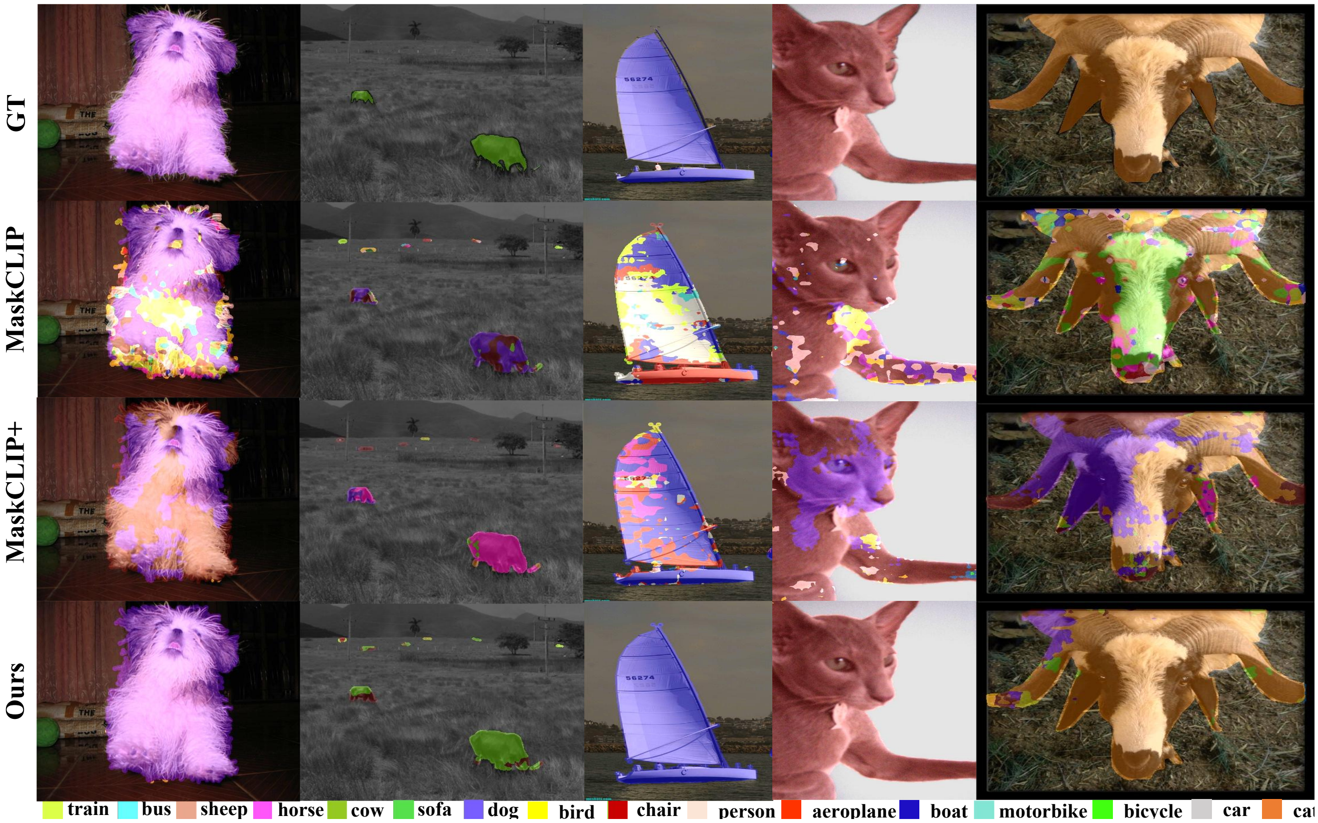

Figure 4. **Visualization of Class-Preference Bias Rectification.** These groups of images show more evidence of class-preference bias existing in CLIP (row 2) / MaskCLIP+ (row 3), and prove the effectiveness of our rectification (row 4).

Table 12. **Training Details for Each Dataset**

Dataset	VOC	Context	ADE20K
Rectification epoch	200	400	300
Distillation epoch	100	50	20
Resized Scale	2048×512	520×520	2048×512
Crop Size	520×520	480×480	512×512
Batch Size	12	12	8

8. More Qualitative Results

8.1. Rectification of class-preference bias.

We show more evidence of class-preference bias existing in CLIP and our rectification results in Fig. 4. The examples are drawn from PASCAL VOC [18]. By comparing the masks generated by MaskCLIP (row 2) / MaskCLIP+ (row 3) and Ground Trurh (row 1), we observe that without rectification, CLIP prefers to assign an incorrect but relevant label to a pixel in quite a few cases. For example, in the second column, the ground trurh (row 1) is "cow" (green mask), but MaskCLIP (row 2) and MaskCLIP+ (row 3) tend to mistakenly classify the pixels of "cow" as "dog" (purple mask) and "horse" (pink mask) respectively. In contrast, the segmentation results of our method (row 4) best match the ground truth, which verifies the rectification effect of our method.

8.2. Rectification of space-preference bias.

We show more evidence of space-preference bias existing in CLIP and our rectification results in Fig. 5. The examples are drawn from PASCAL VOC [18]. By comparing the masks generated by MaskCLIP (row 2) / MaskCLIP+ (row 3) and Ground Trurh (row 1), we observe that the space-preference bias commonly exists in the CLIP. Results of ours (row 4) verify the effectiveness of our rectification. For example, in the first column, the segmentation results of "chair" in MaskCLIP (row 2) / MaskCLIP+ (row 3) are poor as most pixels in the chair area (red mask in ground truth) are misclassified as a group of irrelevant classes, *e.g.,* "boat" (blue mask), "bus" (light blue mask), *etc.* In contrast, our model yields pretty clean and good segmentation results (row 4).

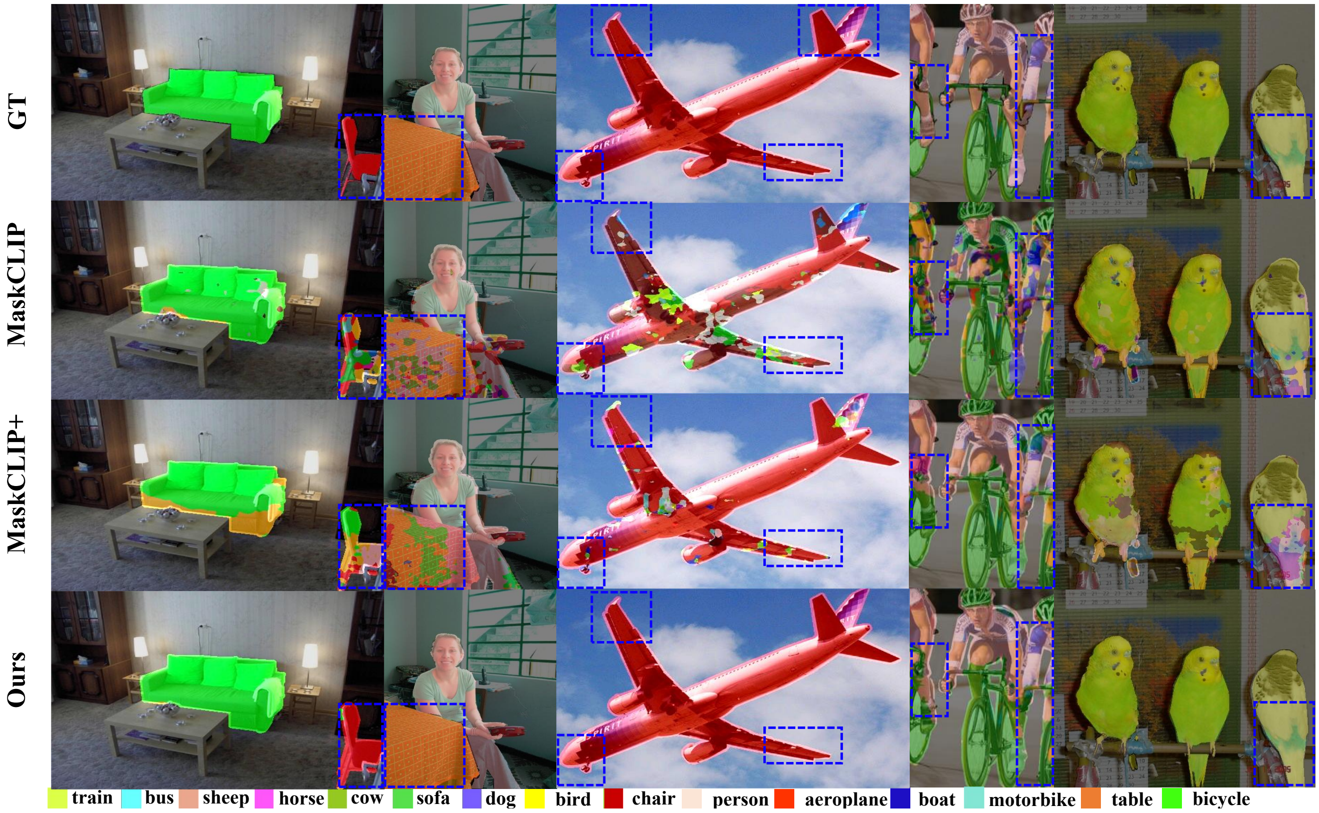

Figure 5. **Visualization of Space-Preference Bias Rectification.** These groups of images show more evidence of space-preference bias existing in CLIP (row 2) / MaskCLIP+ (row 3), and prove the effectiveness of our rectification (row 4). Please refer to the areas within the blue dashed boxes to check the differences.