Supplementary Material: Side Adapter Network for Open-Vocabulary Semantic Segmentation

1. More Design Choices

1.1. Attention bias

In Tab. 11, we add more ablation on the design of he attention bias. Using different biases for each attention head achieves better performance, but further introducing layer-wise attention bias bring no gains.

1.2. [SLS] tokens.

As discussed in the paper, we use the copies of `[CLS]` token as our `[SLS]` tokens. The `[SLS]` tokens are conceptually similar to the Mask Class Tokens (MCT) [10] but differ in their implementation details, such as how they are updated and where they are introduced in the CLIP model. Furthermore, while `[SLS]` tokens are initialized from the `[CLS]` token by default, Tab. 12 shows that using the learned embedding is only marginally worse.

1.3. Prompt engineering

Prompt engineering has been proven useful for open-vocabulary semantic segmentation. Following the common practice [22, 33], we use multiple templates to decorate the class names and average their text embeddings as the final used text embedding for each class in inference. The templates are listed in Tab. 13, and the effects of prompt engineering are shown in Tab. 14, which can improve 1.2 mIoU, 0.7 mIoU on ADE-150 and ADE-847, respectively.

2. Discussion on the parameter efficiency

Tab. 15 examines how the capacity of the side adapter network affects the performance. We find that small models can already achieve good performance, while larger model capacity does not provide significant gains. We speculate that this is because our approach has taken advantage of the features of CLIP, thus relieving the need for a large model capacity.

3. Comparison with fine-tuned models

We study the effects of fine-tuning the pre-trained CLIP model in ADE-847 dataset based on the side adapter network. For a fair comparison, we applied different initial learning rates to the CLIP model while using the same initial learning rate of the side adapter network (*i.e*. 1e-4) and keeping other hyper-parameters unchanged. The results are shown in Fig. 7. The learning rate of 0 indicates *frozen* CLIP model. As the learning rate applied in CLIP mode increases, models perform worse in ADE-847 (*but perform better on COCO Stuff dataset*), indicating that the open-vocabulary capability of the CLIP model is disrupted.

Per Head?	Per Layer?	mIoU
no.	no.	26.2
yes.	no.	**27.8**
no.	yes.	26.3
yes.	yes.	27.4

Table 11. Ablation on the design of attention bias.

Init. method	mIoU
Learned embedding	27.2
`[CLS]`	**27.8**

Table 12. Initialization method of `[SLS]` token.

“a photo of a {}.”,
“This is a photo of a {}”,
“There is a {} in the scene”,
“There is the {} in the scene”,
“a photo of a {} in the scene”,
“a photo of a small {}.”,
“a photo of a medium {}.”,
“a photo of a large {}.”,
“This is a photo of a small {}.”,
“This is a photo of a medium {}.”,
“This is a photo of a large {}.”,
“There is a small {} in the scene.”,
“There is a medium {} in the scene.”,
“There is a large {} in the scene.”,

Table 13. Prompt templates used in our method.

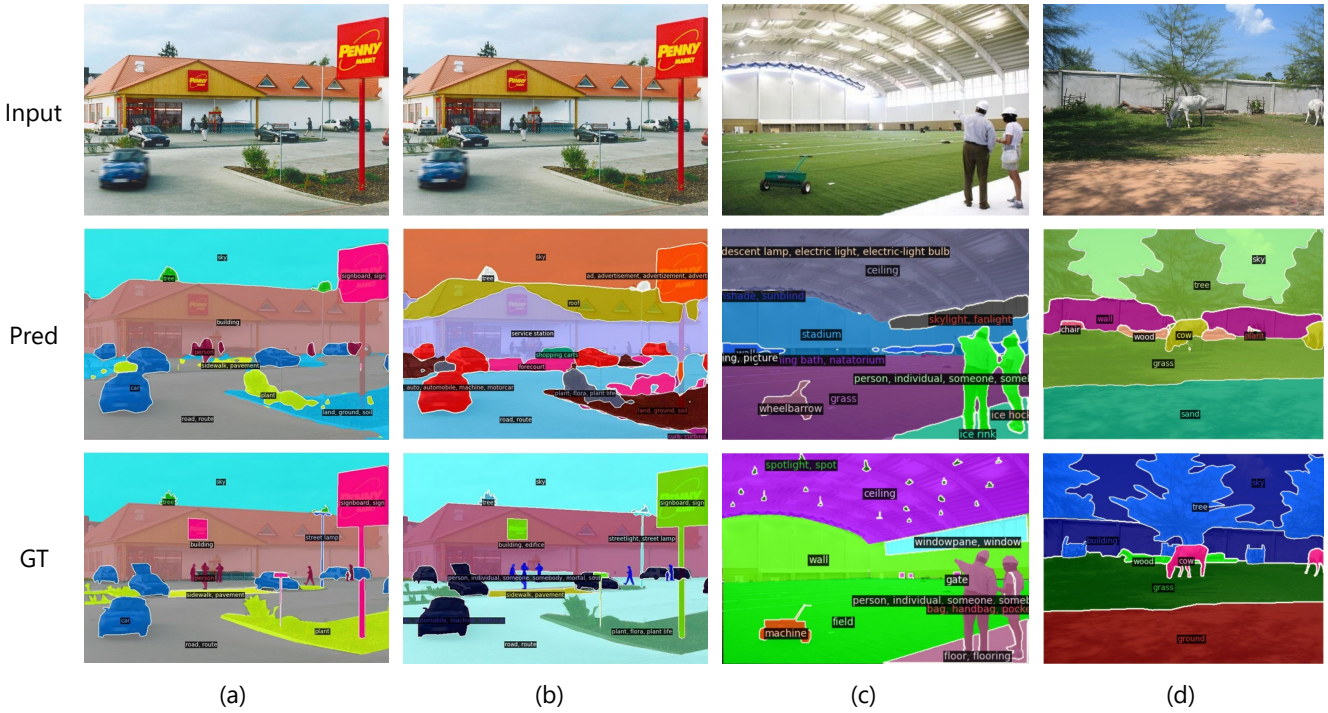

Figure 6. Qualitative results of our method. (a) and (b) are the results of the same input images with different vocabularies (ADE-150 and ADE-847 respectively).

method	w/ prompt.	ADE-150	ADE-847
SimSeg [33]	yes.	21.1	6.9
OvSeg [22]	yes.	24.8	7.1
SAN	no.	26.6	9.5
SAN	yes.	**27.8**	**10.2**

Table 14. Effects of prompt engineering. The single template *"a photo of* {}." is used for models without using prompt engineering.

Width of SAN	Param. (M)	GFLOPs	mIoU
144	4.2	53.6	26.7
192	6.1	58.6	27.4
240	8.4	64.3	**27.8**
288	11.1	70.9	27.3

Table 15. The influence of capacity of SAN. *Param.* stands for the total number of trainable parameters in the model in millions.

4. Visualization

Fig. 6 (a) and (b) are of the same input image but with different vocabularies (from ADE-150, ADE-847 respectively), and e.g., the *signboard, sign* is correctly classified in ADE-150 but is mis-classified as *ad, advertisement* in ADE-847. And we assume that the model is not good at verifying the difference between *advertisement* and *signboard*.

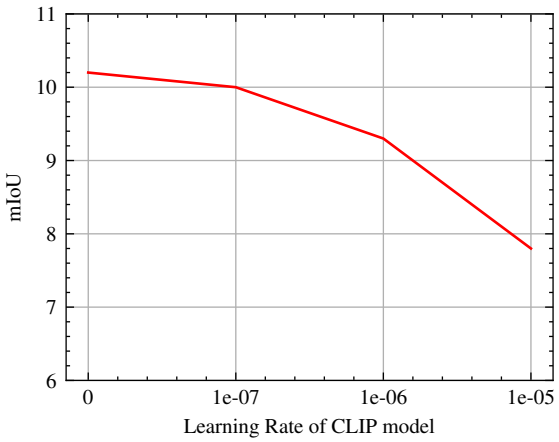

Figure 7. Performance on ADE-847 when fine-tuning the CLIP model with different learning rates. The learning rate of 0 is the *frozen* CLIP model.