A Simple Baseline for Open-Vocabulary Semantic Segmentation with Pre-trained Vision-language Model Supplementary Material

1 Definition of hIoU

Following previous works [5, 3], harmonic mean IoU (hIoU) is defined among the *seen* classes and *unseen* classes as:

$$hIoU = \frac{2 * mIoU_{seen} * mIoU_{unseen}}{mIoU_{seen} + mIoU_{unseen}}. \tag{1}$$

2 Sliding Window Testing in Fully Convolutional Network

We study the different inference methods in this section for Fully Convolutional Network(FCN). For a fair comparison, we use ResNet-101 in FCN and ViT-B/16 in CLIP, same as our two-stage framework. Table. 11 shows the results. The FCN without sliding window test achieved 11.7 hIoU and 10.4 mIoU-unseen. In comparison, employing the window test improved the performance by +9.2 on hIoU and +5.6 on mIoU-unseen. This significant difference in performance is caused by the inconsistent image size between pre-training and testing of the CLIP model. Although the sliding window test can strengthen the FCN approach, it is still worse than our two-stage approach by -16.8 on hIoU and -20.3 on mIoU unseen, indicating that our two-stage framework is more suitable for the CLIP model.

Table 11. Performance of FCN approach on COCO Stuff dataset under the *zero-shot* setting. SW: Sliding Window Testing, each image is splited into several 224×224 patches.

Method	hIoU	pACC	mIoU-unseen
FCN [2]	11.7	54.9	10.4
FCN + SW [2]	20.9	50.8	16.0
Ours	**37.7**	**60.3**	**36.3**

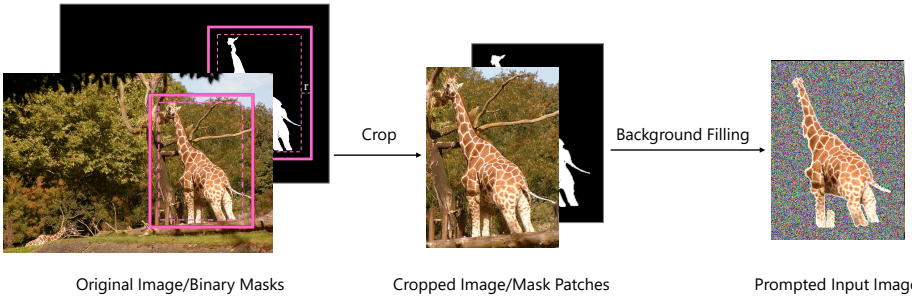

Fig. 3. Pipeline of image prompt engineering. The pink dotted box is the minimum bounding box of the foreground region. The pink solid box is the expanded bounding box of the dotted box by the ratio r. The prompted image is the input of the CLIP image encoder.

3 Prompt Engineering for Image and Text

3.1 Prompt Engineering for Image

As the CLIP model is trained with low-resolution realistic images, given a mask proposal $\mathcal{M}^p$, and the input image I, it is a problem how to extract the visual representation of the proposal with the CLIP model through the proper way, which we call image prompt engineering. The whole process is shown in Figure. 3. We crop the image with the bounding boxes of $\mathcal{M}^p$ and expand the bounding boxes by a ratio r to involve more context information. And then, we fill the background pixels with 0 values in the proposal with some patterns. We studied four choices for such patterns: a) Keep the background pixels unchanged; b) Fill the background pixels with manually designed values; c) Fill the background pixels with learnable values; d) Fill the background patches with mask token, presented in Figure. 4. The results are shown in Table. 12. The value filled in the background area can greatly affect the segmentation performance. Though our exploration to learn proper image prompts failed to achieve improvement like text [1, 6], it is still an interesting problem for future research.

Table 12. Study the effects of background filling. Filling background with learnable pixel prompts or mask token is very unstable and usually leads to negative impact.

Prompt	hIoU	mIoU	
		seen	unseen
Preserving	9.3	8.9	9.5
Zero	17.2	16.3	18.2
Mean Values	18.3	17.3	19.5
Pixel Prompts	Failed	-	-
Mask Token	Failed	-	-

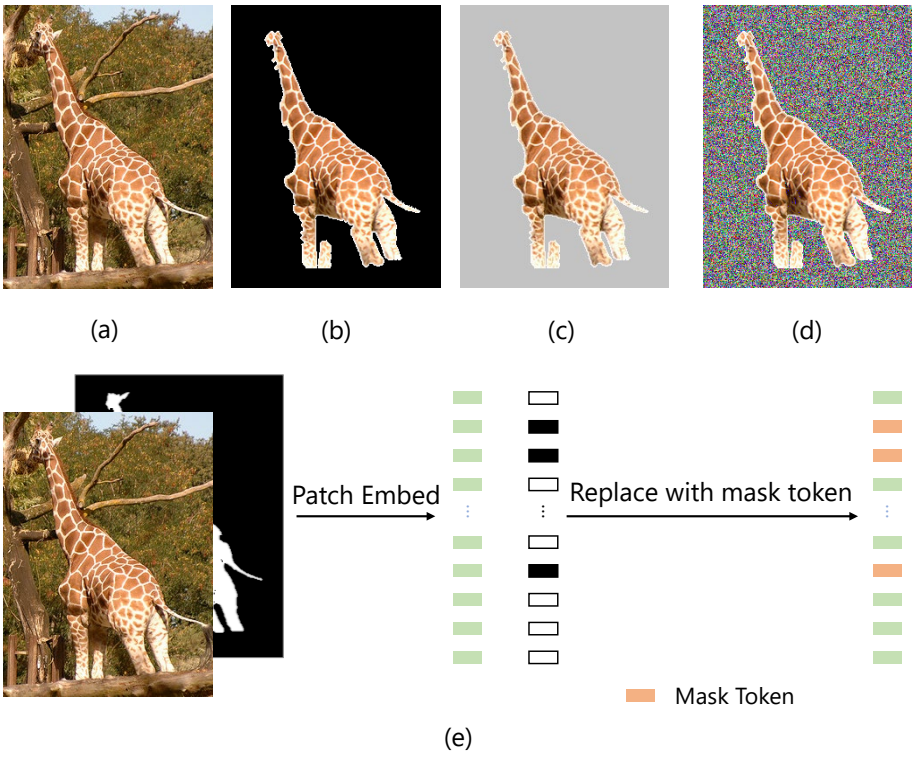

Fig. 4. Choices for background filling. a) Preserving the context pixels; b) Filling the background pixels with zero; c) Filling the background pixels with the mean values of the dataset [4]; d) Filling the background pixels with learnable pixel prompts. The prompts are tuned on *seen* classes; e) Filling the background patches with mask token. The mask token is tuned on *seen* classes.

3.2 Prompt Engineering for Text

We compare two prompt tuning methods described in Sec. 4.2. The results are shown in Table. 14. The learnable prompt outperforms the manually searched prompt by +9.9 hIoU, clearly showing the power of the learnable prompt. In addition, although the learnable prompt is only trained on *seen* classes, we notice that it achieves similar improvement on *seen* classes and *unseen* classes (+9.6 mIoU-seen and +10.2 mIoU-unseen), indicating the learnable prompt has a strong generalization ability to the *unseen* class.

We further study how prompt length and training data size affect the performance of learnable prompts by training on *seen* classes and testing on *unseen* classes. Table. 13 shows that using 32 samples for each category reaches the best performance, in either prompt length of 16 or 32, and more training samples will degrade the performance. We speculate that more samples may lead to the overfitting issue, which is also reported by other prompt learning attempts [6].

Table 13. Study the effect of prompt length and sample number of each category for prompt learning.

Prompt Len	#Sample	Unseen Acc
16	16	28.9
	32	30.0
	64	29.5
	all	25.5
32	16	32.1
	32	**32.8**
	64	31.0
	all	27.6

Table 14. Manually designed prompt v.s. Learnable prompt.

Prompt	hIoU	mIoU	
		seen	unseen
Manual	18.3	17.3	19.5
Learnable	**28.2**	**26.8**	**29.7**

4 Detailed Study on MaskFormer and CLIP

The ablations have been studied in Table. 3 and Table. 4. We re-organize the results as shown in Table. 15. We can conclude: 1) MaskFormer outperforms FCN by +26.8 hIoU (2-th row vs 5-th row); 2) CLIP pre-training outperforms ImageNet pre-training by +24.7 hIoU (3-rd row vs 4-th row); 3) Our method outperforms SPNet by +24.4 hIoU with the same pre-training data (1-st row vs 3-rd row).

Table 15. Study the effects of MaskFormer and CLIP.

Method	Image Encoder	Pre-train Data	hIoU	mIoU	
				Seen	Unseen
SPNet[45]	R-101	ImageNet	25.1	73.3	15.0
FCN	VIT/B-16	CLIP-VL	50.7	**85.5**	36.0
MaskFormer	R-101	ImageNet	49.5	71.1	38.0
	R-101	CLIP-VL	74.2	84.6	66.1
	VIT/B-16	CLIP-VL	**77.5**	83.5	**72.5**

5 The Randomness of the Data split

In the experiments under the zero-shot setting, we use the official unseen/seen split as [5] for a fair comparison. The thing/stuff ratio is 0.88 for seen and 0.87

Table 16. Results of different seen/unseen splits on COCO Stuff. 0* denotes the split used in previous work [5].

Split	Thing/Stuff Ratio		hIoU	$\text{mIoU}_{\text{unseen}}$			
	Seen	Unseen		All	Thing	Stuff	Δ
0*	0.88	0.87	37.8	36.3	44.3	29.5	14.9
1	0.95	0.36	31.9	25.6	44.8	18.6	26.2
2	0.93	0.50	30.9	24.3	41.5	15.6	25.9
3	0.86	1.14	36.6	32.9	38.8	26.2	12.6

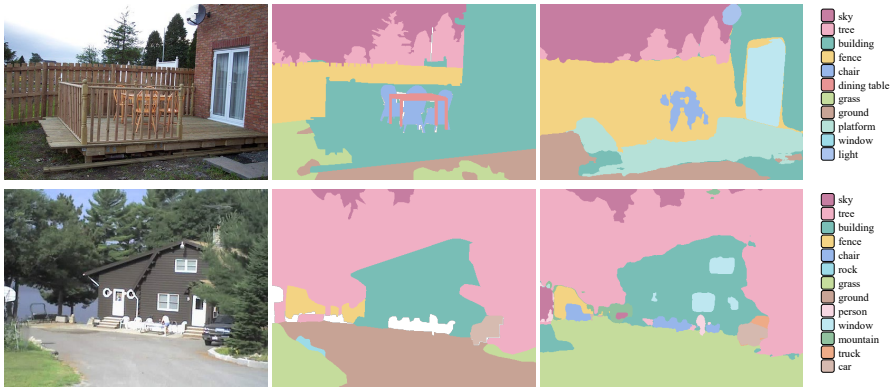

Fig. 5. Qualitative results on Pascal Context dataset under *cross-dataset* setting. From left to right are the original input images, the ground truth semantic segmentation maps and the predictions. The white areas in the ground truth maps is ignored during annotating.

for unseen classes. To study the impacts of different splits, we conduct studies on randomly generated seen/unseen splits in Table. 16. We find a more balanced unseen thing/stuff ratio yields higher hIoU.

6 Visualization of Results under Cross-dataset Setting

We illustrate more qualitative results in Figure. 5, 6 under the cross-dataset setting.

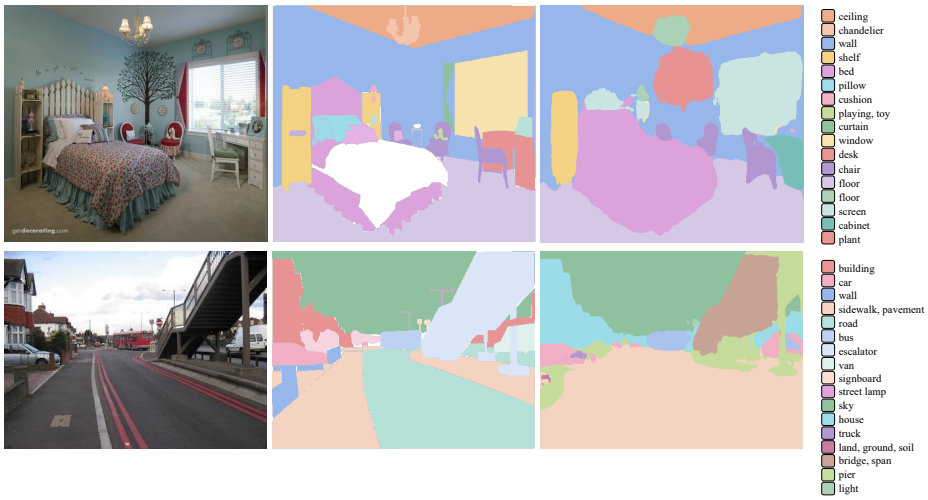

Fig. 6. Qualitative results on ADE20k dataset under *cross-dataset* setting. The original input images, the ground truth semantic segmentation maps, and the predictions are left to right. The white areas in the ground truth maps is ignored during annotating.

References

1. Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., Neubig, G.: Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. arXiv preprint arXiv:2107.13586 (2021)
2. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3431–3440 (2015)
3. Pastore, G., Cermelli, F., Xian, Y., Mancini, M., Akata, Z., Caputo, B.: A closer look at self-training for zero-label semantic segmentation (2021)
4. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. arXiv preprint arXiv:2103.00020 (2021)
5. Xian, Y., Choudhury, S., He, Y., Schiele, B., Akata, Z.: Semantic projection network for zero-and few-label semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8256–8265 (2019)
6. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. arXiv preprint arXiv:2109.01134 (2021)