Supplementary – ProxyCLIP: Proxy Attention Improves CLIP for Open-Vocabulary Segmentation

Mengcheng Lan[1], Chaofeng Chen[1], Yiping Ke[2], Xinjiang Wang[3], Litong Feng[3], and Wayne Zhang[3,4⋆]

[1] S-Lab, Nanyang Technological University
[2] CCDS, Nanyang Technological University [3] SenseTime Research
[4] Guangdong Provincial Key Laboratory of Digital Grid Technology
lanm0002@e.ntu.edu.sg {chaofeng.chen, ypke}@ntu.edu.sg
{wangxinjiang, fenglitong, wayne.zhang}@sensetime.com

Appendix

A Implementation for Stable Diffusion.

Several recent studies [5,10,13] have demonstrated the effectiveness of large-scale text-image diffusion models in open-vocabulary semantic segmentation tasks. ODISE [13] noted that the internal representations of stable diffusion models exhibit strong semantic coherence. Given the flexibility of our framework, our ProxyCLIP can also utilize stable diffusion models as VFM to extract dense visual representations for images. Specifically, we employ the stable diffusion [8] model pre-trained on a subset of the LAION dataset as our VFM. We set the time step for the diffusion process to $t = 0$ and extract feature maps from the 9-th block of UNet. The input image is directly resized from 336×336 to 672×672. Consequently, we obtain a feature map with dimensions $22\times22\times1280$, a downsampling factor of 15.3.

B Hyperparameters.

We further conduct experiments to investigate the effects of varying the shifting and scaling factors in the normalization step. As depicted in Fig. 1, we examined the segmentation performance achieved across four datasets using different values for the shifting factor β and the scaling factor γ. Notably, we observe that ProxyCLIP consistently achieves good results when β is localized within the range of 1.0 to 1.6 and γ falls within the range of 2.0 to 5.0. We set these two parameters to $\beta = 1.2$ and $\gamma = 3.0$ for all datasets by default. These results further underscore the robustness of our normalization strategy within the proxy attention mechanism, affirming its efficacy in enhancing segmentation performance across diverse datasets.

⋆ Corresponding author.

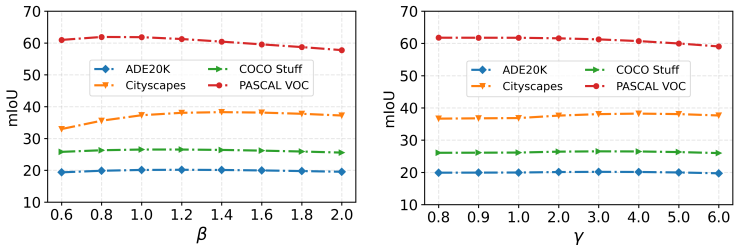

Fig. 1: Open-vocabulary semantic segmentation results mIoU w.r.t. (left) the shifting factor β, and (right) the scaling factor γ.

C Efficiency comparison.

We conduct an efficiency comparison among various training-free models during inference. These experiments are performed on an RTX 3090 GPU, using a batch size of 1 and an input image resolution of 336×336. Inference is conducted using half precision (fp16) for all models. The results in Tab. 1 indicate that ProxyCLIP, which utilizes an additional VFM (e.g., DINO-B/8), consumes more computational resources compared to models that use only the CLIP architecture. However, adopting VFM with a smaller ViT architecture or a larger patch size may balance the accuracy and efficiency.

Table 1: Efficiency comparison of training-free methods based on CLIP-ViT-B/16 architecture. IPS: Image Per Second.

Models	Params ↓ (M)	Speed ↑ (IPS)	GPU ↓ (MiB)	FLOPs ↓ (G)
CLIP	142.7	72.5	3540	41.7
MaskCLIP	142.7	74.1	3540	41.6
GEM	142.7	55.9	3716	51.5
ProxyCLIP (DINO-B/16)	224.5	52.9	3640	81.1
ProxyCLIP (DINO-B/8)	224.5	26.9	3926	253.3

D Hard masking.

To further verify the effectiveness of our adaptive normalization and masking strategy, we conduct an experiment focusing solely on the masking strategy. Implementing attention masking necessitates determining a suitable threshold, which proves challenging given the variability in median attention scores across different VFMs. To illustrate this, we perform experiments on the COCOStuff

dataset using the masking strategy alone, with the threshold α varied within the range $[0.0, \ldots, 0.8]$. The masking function is defined as follows:

$$\mathcal{M}_{ij} = \begin{cases} 0, & A_{ij} \geq \alpha \\ -\infty, & A_{ij} < \alpha \end{cases} \tag{1}$$

Results in Tab. 2 indicate that different VFMs may require different thresholds to achieve good results. In contrast, our adaptive normalization and masking strategy consistently delivers promising results across different VFMs.

Table 2: Comparison of using different thresholds on COCOStuff.

α	0.0	0.2	0.4	0.6	0.8	Adaptive
MAE	15.2	16.9	19.8	23.0	23.3	23.1
SAM	12.6	12.6	14.0	21.4	25.2	25.0
DINOv2	15.5	22.4	25.2	25.1	23.7	25.4
DINO	15.5	22.2	25.8	24.4	22.0	26.5

E Adopting attention embeddings of CLIP in PAM.

To further evaluate our method, we conducte experiments using different attention embeddings from CLIP in the Proxy Attention Module (PAM). Specifically, we compare the use of *query-key*, *query-query*, and *key-key* attention embeddings from CLIP with the use of x-x features from Vision-and-Language Models (VFMs) in PAM (referred to as the Proxy). The results, based on the CLIP-ViT-B/16 architecture, are summarized in Tab. 3. Notably, adopting *query-key* embeddings of CLIP in PAM significantly enhances vanilla CLIP, improving from 11.7 mIoU to 26.1 mIoU. The performance is further improved to 38.2 when *query-query* or *key-key* embeddings are used. Despite these enhancements, the performance of using CLIP embeddings in PAM is still inferior compared to the proposed proxy attention using VLM embeddings.

Table 3: Results of using CLIP's q and k embedings in PAM.

Attn	VOC	Context	Object	VOC20	Context59	Stuff	City	ADE	Avg.
q-k	34.0	18.2	20.6	75.3	20.8	13.6	15.7	10.3	26.1
q-q	55.2	31.3	33.7	77.7	35.1	23.5	31.6	17.8	38.2
k-k	55.7	30.6	33.9	77.8	34.6	23.3	31.6	17.9	38.2
Proxy	61.3	35.3	37.5	80.3	39.1	26.5	38.1	20.2	42.3

Table 4: Comparison of open-vocabulary semantic segmentation performance under different models and architectures.

Method		Annotation	ADE847	PC459	Avg.
Fully-supervised					
ODISE [13]	CVPR2023	✓	11.1	14.5	12.8
SAN [14]	CVPR2023	✓	12.4	15.7	14.1
X-Decoder [18]	CVPR2023	✓	9.2	16.1	12.7
OVSeg [6]	CVPR2023	✓	9.0	12.4	10.7
MaskCLIP [2]	ICML2023	✓	8.2	10.0	9.1
DeOP [3]	ICCV2023	✓	7.1	9.4	8.3
MasQCLIP [15]	ICCV2023	✓	10.7	18.2	14.5
GKC [4]	ICCV2023	✓	3.5	7.1	5.3
MAFT [4]	NeurIPS2023	✓	12.1	15.7	13.9
HIPIE [11]	NeurIPS2023	✓	9.7	14.4	12.1
Weakly-supervised					
TCL [1]	CVPR2023	✗	4.9	5.3	5.1
CLIP-DINOiser [12]	Arxiv2023	✗	7.1	8.4	7.8
Training-free					
CLIP	ICML2021	✗	0.8	1.3	1.1
MaskCLIP [17]	ECCV2022	✗	3.6	4.6	4.1
SCLIP [9]	Arxiv2023	✗	4.9	6.3	5.6
ProxyCLIP		✗	11.1	9.9	10.5

F Additional quantitative results.

We further present a performance comparison on the ADE847 [16] and Pascal Context459 [7] (PC459) datasets, two challenging benchmarks containing 847 and 459 classes, respectively. These datasets are widely utilized for evaluating the performance of fully-supervised open-vocabulary semantic segmentation methods. For fully-supervised open-vocabulary semantic segmentation methods, we directly cite the best results from their respective papers. For weakly-supervised methods, *i.e.*, TCL and CLIP-DINOiser, we obtain results using the checkpoints provided by the authors. For training-free methods, we report results based on our implementation. To our knowledge, we are among the first to report results for weakly-supervised and training-free methods on these challenging datasets.

The experimental results are summarized in Tab. 4. It's worth noting that fully-supervised methods typically achieve better results, as they undergo in-domain training using the COCOStuff training set with fully annotated labels. Remarkably, ProxyCLIP, as a training-free method, achieves performance on par with fully-supervised methods. For instance, ProxyCLIP achieves 11.1 mIoU on the ADE847 dataset, surpassing the performance of most fully-supervised methods. Additionally, ProxyCLIP significantly outperforms all weakly-supervised and training-free methods, with an average mIoU of 10.5, compared to the 7.8 mIoU of CLIP-DINOiser. We attribute this superiority to its proxy attention

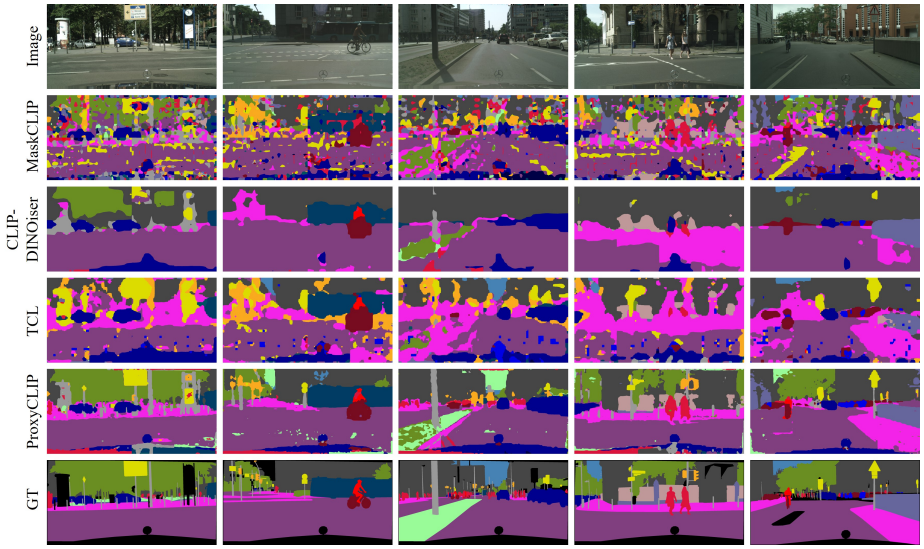

Fig. 2: Additional qualitative comparison on the Cityscapes dataset.

mechanism, which naturally inherits the robust local consistency of VFMs while maintaining CLIP's exceptional zero-shot recognition capacity.

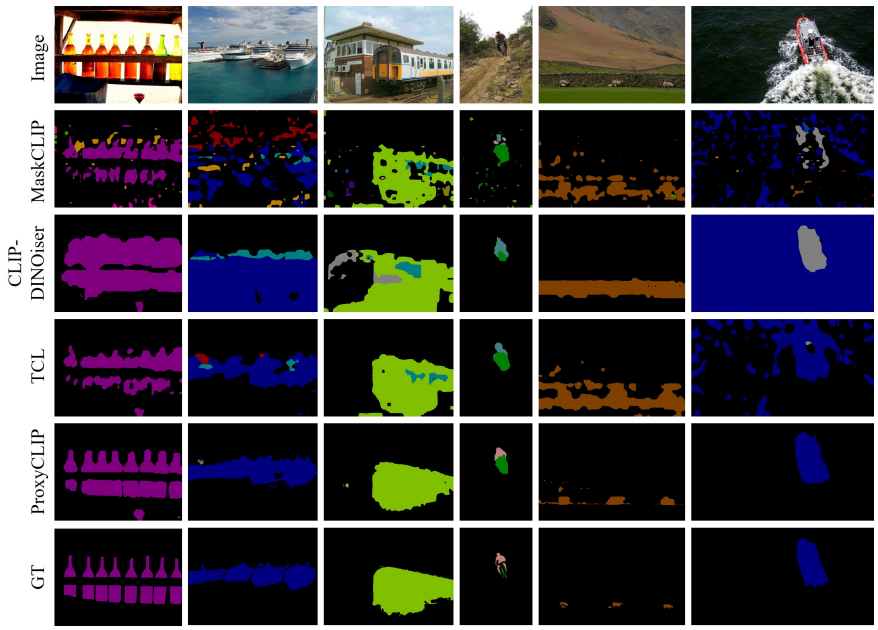

Fig. 3: Additional qualitative comparison on the Pascal VOC dataset.

G Additional qualitative results.

We present additional qualitative results on three datasets: Cityscapes, Pascal VOC, and Pascal Context59, illustrated in Figs. 2 to 4, respectively. In Fig. 2, our ProxyCLIP demonstrates its ability to effectively segment regions belonging to different categories, including small objects. Figures 3 and 4 showcase ProxyCLIP's capability to produce more accurate and high-quality segmentation maps with clearer boundaries for various objects.

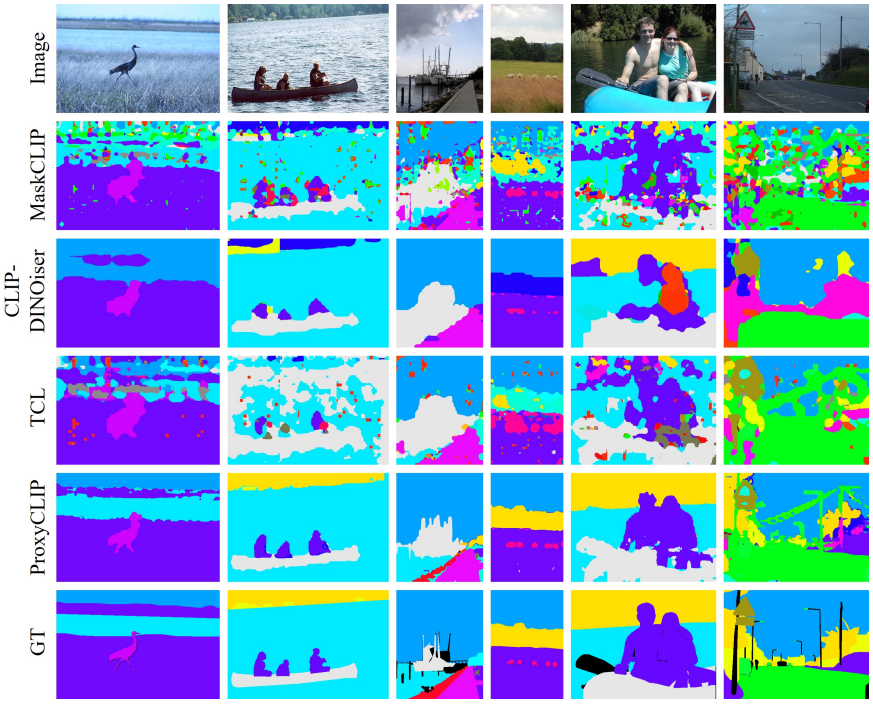

Fig. 4: Additional qualitative comparison on the Pascal Context59 dataset.

References

1. Cha, J., Mun, J., Roh, B.: Learning to generate text-grounded mask for open-world semantic segmentation from only image-text pairs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11165–11174 (2023)
2. Ding, Z., Wang, J., Tu, Z.: Open-vocabulary universal image segmentation with maskclip. arXiv preprint arXiv:2208.08984 (2022)
3. Han, C., Zhong, Y., Li, D., Han, K., Ma, L.: Open-vocabulary semantic segmentation with decoupled one-pass network. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1086–1096 (2023)
4. Han, K., Liu, Y., Liew, J.H., Ding, H., Liu, J., Wang, Y., Tang, Y., Yang, Y., Feng, J., Zhao, Y., et al.: Global knowledge calibration for fast open-vocabulary segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 797–807 (2023)
5. Li, Z., Zhou, Q., Zhang, X., Zhang, Y., Wang, Y., Xie, W.: Open-vocabulary object segmentation with diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7667–7676 (2023)
6. Liang, F., Wu, B., Dai, X., Li, K., Zhao, Y., Zhang, H., Zhang, P., Vajda, P., Marculescu, D.: Open-vocabulary semantic segmentation with mask-adapted clip. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7061–7070 (2023)
7. Mottaghi, R., Chen, X., Liu, X., Cho, N.G., Lee, S.W., Fidler, S., Urtasun, R., Yuille, A.: The role of context for object detection and semantic segmentation in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 891–898 (2014)
8. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
9. Wang, F., Mei, J., Yuille, A.: Sclip: Rethinking self-attention for dense vision-language inference. arXiv preprint arXiv:2312.01597 (2023)
10. Wang, J., Li, X., Zhang, J., Xu, Q., Zhou, Q., Yu, Q., Sheng, L., Xu, D.: Diffusion model is secretly a training-free open vocabulary semantic segmenter. arXiv preprint arXiv:2309.02773 (2023)
11. Wang, X., Li, S., Kallidromitis, K., Kato, Y., Kozuka, K., Darrell, T.: Hierarchical open-vocabulary universal image segmentation. Advances in Neural Information Processing Systems **36** (2024)
12. Wysoczańska, M., Siméoni, O., Ramamonjisoa, M., Bursuc, A., Trzciński, T., Pérez, P.: Clip-dinoiser: Teaching clip a few dino tricks. arXiv preprint arXiv:2312.12359 (2023)
13. Xu, J., Liu, S., Vahdat, A., Byeon, W., Wang, X., De Mello, S.: Open-vocabulary panoptic segmentation with text-to-image diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2955–2966 (2023)
14. Xu, M., Zhang, Z., Wei, F., Hu, H., Bai, X.: Side adapter network for open-vocabulary semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2945–2954 (2023)
15. Xu, X., Xiong, T., Ding, Z., Tu, Z.: Masqclip for open-vocabulary universal image segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 887–898 (2023)

16. Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., Torralba, A.: Semantic understanding of scenes through the ade20k dataset. International Journal of Computer Vision **127**, 302–321 (2019)
17. Zhou, C., Loy, C.C., Dai, B.: Extract free dense labels from clip. In: European Conference on Computer Vision. pp. 696–712. Springer (2022)
18. Zou, X., Dou, Z.Y., Yang, J., Gan, Z., Li, L., Li, C., Dai, X., Behl, H., Wang, J., Yuan, L., et al.: Generalized decoding for pixel, image, and language. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15116–15127 (2023)