Effective SAM Combination for Open-Vocabulary Semantic Segmentation (Supplementary Material)

Minhyeok Lee[1] Suhwan Cho[1] Jungho Lee[1] Sunghun Yang[1]
Heeseung Choi[2] Ig-Jae Kim[2] Sangyoun Lee[1]
[1]Yonsei University
[2]Korea Institute of Science and Technology (KIST)
{hydragon516, chosuhwan, 2015142131, sunghun98, syleee}@yonsei.ac.kr
{hschoi, drjay}@kist.re.kr

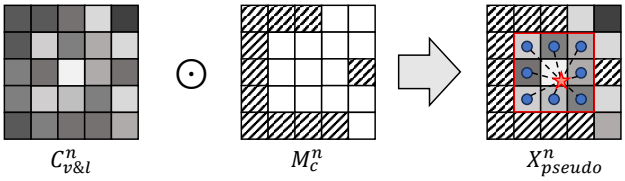

Figure 1. A detailed explanation of the pseudo point determination process. The proposed method introduces diversity in the pseudo point coordinates at the low-resolution feature level.

1. Details of Pseudo Point Generation

In this section, we provide additional details about the pseudo point generation process in the proposed PPG module. In the traditional prompt encoder of SAM [6], integer coordinate values corresponding to the actual image size are used as input. However, the proposed pseudo points are measured at the level of a low-resolution feature map, which reduces the diversity of possible pseudo point coordinates. To address this issue, we employ a linear sampling method that utilizes neighboring pixel values to generate pseudo point coordinates, as illustrated in Figure 1.

First, let the vision-language correlation map for class n be denoted as $C^n_{v\&l} \in \mathbb{R}^{1\times H\times W}$, and the clustering mask as $M^n_c \in \{0,1\}^{1\times H\times W}$. The position (i^*, j^*) with the highest probability within the mask region can be expressed as follows:

$$(i^*, j^*) = \underset{(i,j)\in\{(i,j)|(M^n_c)_{ij}=1\}}{\operatorname{argmax}} (C^n_{v\&l})_{ij}\,. \tag{1}$$

Next, as shown by the red box in Figure 1 , a 3×3 region around (i^*, j^*) is defined. Also, the coordinates (i^*, j^*) are normalized to the range $[0, 1]$, resulting in the normalized coordinates (x, y). In other words, if the original image size is $W \times H$, then $x = j^*/W$ and $y = i^*/H$. Finally, linear sampling is performed within the 3×3 region centered at the position (x, y). This process can be expressed as follows:

$$S(x,y) = \sum_{i=i^*-1}^{i^*+1} \sum_{j=j^*-1}^{j^*+1} w_{ij} \cdot (C^n_{v\&l})_{ij}\,, \tag{2}$$

where w_{ij} represents the linear interpolation weight. The weight w_{ij} is calculated based on the Euclidean distance and can be expressed as follows:

$$w_{ij} = \max\left(0, 1 - \sqrt{(x - j/W)^2 + (y - i/H)^2}\right). \tag{3}$$

Therefore, for class n, the final sampled pseudo prompt point X_p can be expressed as $X^n_{pseudo} = (\lfloor y\cdot H \rfloor, \lfloor x\cdot W \rfloor)$.

Index	CLIP	A-847	PC-459	A-150	PC-59	PAS-20	PAS-20
(a)	Freeze	11.2	17.3	30.1	55.8	90.9	75.4
(b)	Fine-tune	**18.1**	**27.0**	**41.8**	**65.6**	**98.3**	**86.3**

Table 1. Quantitative performance comparison with and without CLIP fine-tuning.

Index	Methods	A-847	PC-459	A-150	PC-59	PAS-20	PAS-20
(a)	OVSeg + CLIP	9.0	12.4	29.6	55.7	94.5	-
(b)	OVSeg + Alpha-CLIP	13.4	21.3	32.3	59.1	95.7	79.7
(c)	ESC-Net (Ours)	**18.1**	**27.0**	**41.8**	**65.6**	**98.3**	**86.3**

Table 2. Comparison between the two-stage method using Alpha-CLIP and ESC-Net.

2. Implementation Detail for CLIP Fine-tuning

In traditional open-vocabulary segmentation methods [2, 4, 10], the image-text relationship modeling of pre-trained CLIP is typically preserved by freezing its encoders. However, recent studies have demonstrated that fine-tuning CLIP's image and text encoders can achieve higher performance in open-vocabulary segmentation tasks. This suggests that CLIP's linguistic discrimination capability can

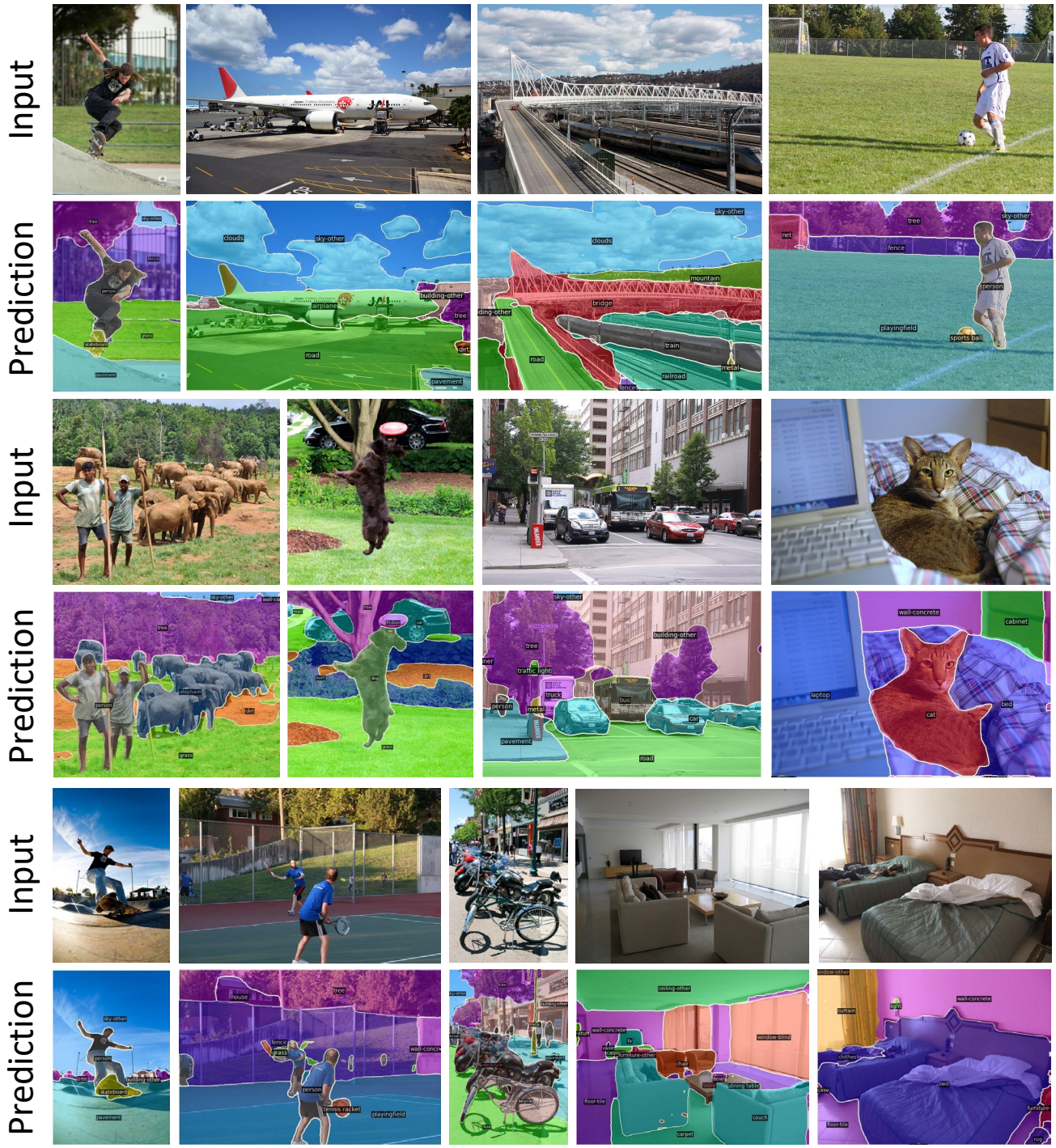

Figure 2. Visualizing ESC-Net's prediction mask results on various datasets

be retained even with training datasets containing a limited number of classes. Following this insight, the proposed ESC-Net adopts the approach of CAT-Seg [1] to fine-tune the attention layers of CLIP's image and text encoders. Additionally, as shown in Table 1, fine-tuned CLIP encoders achieve superior performance compared to frozen CLIP encoders.

3. Comparison with Alpha-CLIP

Traditional two-stage methods [3, 5, 7, 9, 10] demonstrate lower performance because, while the CLIP model can understand the overall content of an image, it lacks the ability to focus on specific regions of interest. However, a recently proposed enhanced CLIP model, Alpha-CLIP [8], addresses this limitation by incorporating an additional alpha channel to focus on specific areas. In other words, Alpha-CLIP introduces an alpha channel alongside the RGB channels to define regions of interest, enabling the model to be trained to concentrate on designated areas. This capability to emphasize regions of interest provides greater accuracy and control in tasks such as image classification, caption generation, and visual question answering.

In this section, we compare the proposed ESC-Net with a modified OVSeg [7] model, which is one of the most representative two-stage methods. The comparison involves applying Alpha-CLIP to replace the CLIP backbone in OVSeg. Table 2 presents the results of comparing the original OVSeg model, the OVSeg model with Alpha-CLIP, and our proposed model. As shown in the table, Alpha-CLIP significantly improves the performance of the traditional two-stage method but still achieves lower performance compared to our approach. This is because Alpha-CLIP is trained on relatively large and clear object masks, making it less robust in scenarios with small and numerous objects or complex scenes.

4. More Qualitative Results

In this section, we provide additional qualitative results of the proposed ESC-Net. As shown in Figure 2, the proposed ESC-Net demonstrates robust predictive performance across a variety of challenging scenes.

References

[1] Seokju Cho, Heeseong Shin, Sunghwan Hong, Anurag Arnab, Paul Hongsuck Seo, and Seungryong Kim. Cat-seg: Cost aggregation for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4113–4123, 2024. 2

[2] Jian Ding, Nan Xue, Gui-Song Xia, and Dengxin Dai. Decoupling zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11583–11592, 2022. 1

[3] Zheng Ding, Jieke Wang, and Zhuowen Tu. Open-vocabulary universal image segmentation with maskclip. *arXiv preprint arXiv:2208.08984*, 2022. 3

[4] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *European Conference on Computer Vision*, pages 540–557. Springer, 2022. 1

[5] Dat Huynh, Jason Kuen, Zhe Lin, Jiuxiang Gu, and Ehsan Elhamifar. Open-vocabulary instance segmentation via robust cross-modal pseudo-labeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7020–7031, 2022. 3

[6] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 1

[7] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7061–7070, 2023. 3

[8] Zeyi Sun, Ye Fang, Tong Wu, Pan Zhang, Yuhang Zang, Shu Kong, Yuanjun Xiong, Dahua Lin, and Jiaqi Wang. Alpha-clip: A clip model focusing on wherever you want. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13019–13029, 2024. 3

[9] Xiaoqi Wang, Wenbin He, Xiwei Xuan, Clint Sebastian, Jorge Piazentin Ono, Xin Li, Sima Behpour, Thang Doan, Liang Gou, Han-Wei Shen, et al. Use: Universal segment embeddings for open-vocabulary image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4187–4196, 2024. 3

[10] Mengde Xu, Zheng Zhang, Fangyun Wei, Yutong Lin, Yue Cao, Han Hu, and Xiang Bai. A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model. In *European Conference on Computer Vision*, pages 736–753. Springer, 2022. 1, 3