Supplementary Materials for Mask-Adapter: The Devil is in the Masks for Open-Vocabulary Segmentation

Yongkang Li, Tianheng Cheng, Bin Feng, Wenyu Liu, Xinggang Wang
School of EIC, Huazhong University of Science & Technology
{liyk, thch, fengbin, liuwy, xgwang}@hust.edu.cn

A. Additional Details

A.1. Geometric Ensemble Strategy

Following [3, 9, 11], we employ a geometric ensemble strategy to fuse the class probabilities predicted by Mask2Former, denoted as $\hat{y}_{\text{in}}$ [11], and those predicted by Mask-Adapter, denoted as $\hat{y}_{\text{out}}$. The geometric ensemble is defined as:

$$y(c) = \begin{cases} (\hat{y}_{\text{in}}(c))^{1-\alpha} \cdot \hat{y}_{\text{out}}(c)^{\alpha}, & \text{if } c \in C_{\text{seen}} \\ (\hat{y}_{\text{in}}(c))^{1-\beta} \cdot \hat{y}_{\text{out}}(c)^{\beta}, & \text{if } c \in C_{\text{unseen}} \end{cases}$$

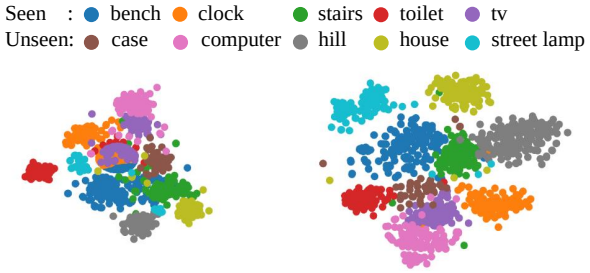
Here, α and β are hyperparameters that control the relative contributions of predictions for seen (C_{seen}) and unseen (C_{unseen}) categories. For FC-CLIP, we set $\alpha = 0.7$ and $\beta = 0.9$. For MAFTP-Base, $\alpha = 0.7$ and $\beta = 1.0$, while for MAFTP-Large, $\alpha = 0.8$ and $\beta = 1.0$. This geometric ensemble effectively balances the strengths of both predictions, enhancing the model's recognition capability.

A.2. Datasets

COCO-Stuff [1] extends COCO with fine-grained annotations for 80 thing and 91 stuff classes, covering 118k training images. It is commonly used for training open-vocabulary segmentation models.
COCO-Panoptic [6] is derived from COCO and designed for panoptic segmentation, combining instance and semantic segmentation. It includes 118k training images with 133 categories: 80 thing and 53 stuff classes, a subset of COCO-Stuff's 91 stuff classes. Due to the smaller number of stuff classes, training on COCO-Panoptic yields lower performance compared to COCO-Stuff for open-vocabulary segmentation.
ADE20K-150 (A-150) [12] is a large-scale scene parsing dataset with 20,210 training and 2,000 validation images, annotated with 150 classes. A-150 serves as the primary evaluation dataset for open-vocabulary segmentation and ablation studies. We manually classify the categories in A-150 into seen and unseen groups. Specifically, if a supercategory in ADE20K has a corresponding category in COCO, such as the ADE20K supercategory field corresponding to the COCO category playing field, we classify it as seen. This approach enables precise evaluation of the segmentation performance on unseen categories.
ADE20K-847 (A-847) [12] contains the same 20,210 training and 2,000 validation images as A-150, with annotations for 847 categories.
PASCAL VOC (PAS-20) [2] includes 1,464 training and 1,449 test images annotated with 20 object classes.
PASCAL-Context (PC-59) [8] extends PASCAL VOC with 4,998 images annotated for 59 categories.
PASCAL-Context (PC-459) [8] further extends PC-59 with annotations for 459 categories using the same 4,998 images.

B. Additional Results

Ablation of mask consistency. Fig. 1 illustrates the effects of different mask extraction methods. The adoption of mask consistency constraint increases inter-class distances and enhances the distinctiveness of mask embeddings, thereby improving the model's ability to recognize unseen classes.



(a) Mask Pooling (b) Mask-Adapter

Figure 1. ***t*-SNE visualization of mask embeddings from different extraction methods.** Mask embeddings extracted using the Mask-Adapter demonstrate better separability compared to those obtained by mask pooling.

(a) mask (b) single map (c) multiple maps (d) mask (e) single map (f) multiple maps

Figure 2. **Visualization of different numbers of semantic activation maps.** We compare the visualization of a single semantic activation map and multiple semantic activation maps. Using multiple semantic activation maps effectively reduces excessive contextual noise, enhancing the model's recognition capability for masks.

Tab. 1 presents the effect of varying cosine loss weights in combination with the IoU-based matcher (threshold 0.7). With a cosine loss weight of 5, we observe a 0.4 improvement in overall mIoU and a 0.5 increase in mIoU for unseen classes.

Cosine Loss Weight	ADE20K		
	mIoU^s	mIoU^u	mIoU
0.0	47.8	25.0	36.2
2.0	47.8	24.7	36.1
5.0	**48.0**	**25.5**	**36.6**
10.0	47.4	24.8	36.0

Table 1. **Ablation study of cosine loss weight in mask consistency.** We evaluate the effect of different cosine loss weights on the ADE20K dataset combining FC-CLIP with Mask-Adapter.

Warmup training with additional datasets. Fig. 3 demonstrates the model's performance for unseen categories using different datasets during ground-truth warmup training. Training on COCO-Panoptic often results in overfitting, impairing the model's ability to recognize unseen classes. In contrast, training on a combined COCO-Panoptic and LVIS dataset enhances the model's ability to recognize unseen categories and improves generalization, while maintaining the same number of training epochs. This improvement is primarily due to the richer categories in LVIS compared to COCO-Panoptic, which enhances the model's open-vocabulary recognition capability and reduces overfitting. This observation highlights a limitation in the current open-vocabulary segmentation setup, where training solely on COCO-Panoptic or COCO-Stuff restricts the model's generalization ability for unseen categories.

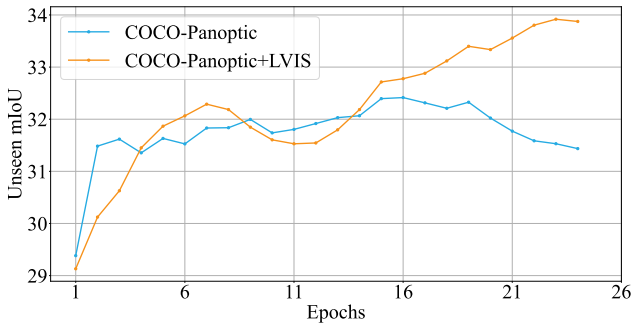


Figure 3. **Ground-truth warmup training with different datasets.** We compare the results of training on COCO-Panoptic alone and COCO-Panoptic with additional LVIS data. The unseen mIoU is evaluated every 10,000 iterations.

Ablations on the number of semantic activation maps. In our Mask-Adapter, we extract 16 semantic activation maps for each mask, aggregate their corresponding CLIP features separately, and compute the average. This design effectively mitigates the excessive contextual noise in the semantic activation maps, as shown in Fig. 2. We also evaluate the impact of using different numbers of semantic activation maps during ground-truth warmup training in Tab. 2. Compared to a single semantic activation map, multiple semantic activation maps reduce the excessive contextual noise and improve the FC-CLIP mIoU by 0.8, demonstrating the effectiveness of our method in enhancing the model's recognition capability and robustness.

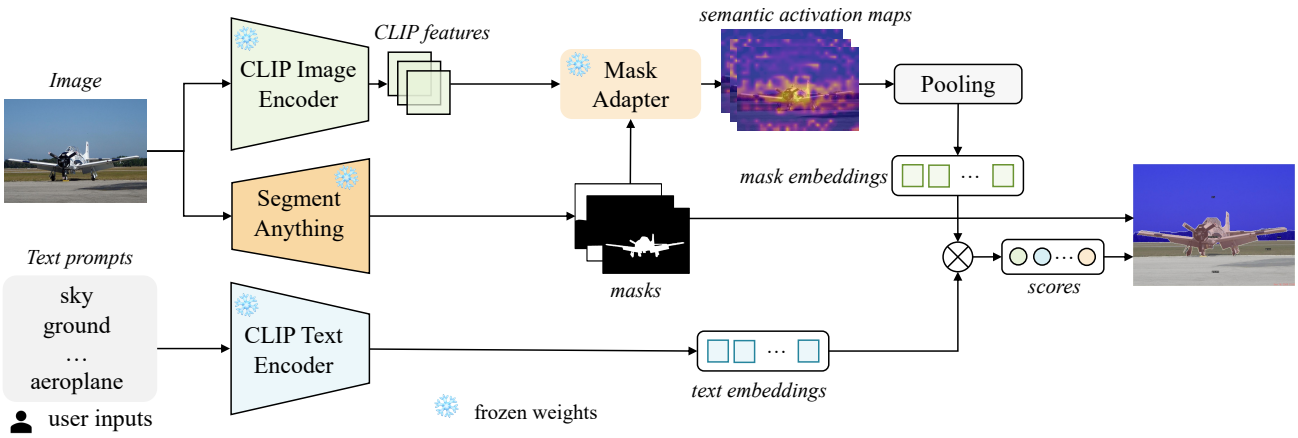


Figure 4. **Illustration of the framework for Segment Anything with Mask-Adapter.** The SAM generates class-agnostic masks, while CLIP extracts features. These are processed by the Mask-Adapter to produce semantic activation maps and obtain mask embeddings, which are matched with text embeddings for classification.

num. of maps	GT mIoU	FC-CLIP mIoU
1	41.9	34.6
16	**43.4**	**35.4**

Table 2. **Comparison with different numbers of semantic activation maps during ground-truth warmup.** GT mIoU and FC-CLIP mIoU represent results using Ground-truth and FC-CLIP predicted masks, respectively.

Ablations on the block structures. In our paper, we primarily adopt ConvNeXt blocks, which are consistent with the backbone. We also provide the ablations about different blocks in Tab. 3. The results demonstrate that the ConvNeXt structure yields superior performance compared to other block configurations, and CNN-based blocks are generally more suitable for dense prediction tasks.

Block	GT mIoU	FC-CLIP mIoU
ResNet Block	43.0	35.1
ConvNeXt Block	**43.4**	**35.4**
Transformer Block	**43.4**	35.0
Swin Transformer Block	42.6	34.9

Table 3. **Ablation on different blocks structures.**

Experimental analysis on PASCAL VOC. As mentioned in [10], PASCAL VOC and PASCAL-Context(PC-59) exhibit a Hausdorff similarity of approximately 0.9 with COCO-Stuff. similarity, all categories in PASCAL VOC overlap with those in COCO-Stuff, which limits its ability to evaluate the model's performance on unseen classes. Mask-Adapter is trained to extract semantic activation maps, aiming to retain the original generalization capabilities of CLIP while improving mask classification accuracy. To demonstrate the effectiveness of our Mask-Adapter over original Mask-Pooling, we evaluate PASCAL VOC (PAS-20) using the predicted probabilities from Mask2Former, mask pooling, and Mask-Adapter independently. From the results in

Classification	PASCAL VOC mIoU
In-vocab.	94.6
Out-vocab.	92.1
In-vocab. & Out-vocab.	95.4
Mask-Adapter	95.4
In-vocab. & Mask-Adapter	95.5

Table 4. **Comparison of class branches on PASCAL VOC.** In-vocab. represents the results obtained using the predictions from Mask2Former, while Out-vocab. refers to the results using the predictions from Mask-Pooling.

Tab. 4, we observe that Mask-Adapter improves upon the In-vocab. and Out-vocab. branches by 0.8 and 3.3, respectively. This indicates that our model demonstrates substantial improvements for both seen and unseen categories. However, due to the limited number of categories in Pascal VOC, *i.e.*, 20 common categories, the overall improvement brought by Mask-Adapter is somewhat minor.

C. Extending to Segment Anything

Our Mask-Adapter can be seamlessly integrated into SAM [5] in a plug-and-play manner, recognizing class-agnostic masks predicted by SAM. As illustrated in Fig. 4, given an input image, SAM generates class-agnostic masks, while CLIP extracts image features. These outputs are fed into the Mask-Adapter to extract semantic activation maps. Subsequently, the CLIP features are aggregated to generate mask embeddings, which are matched with text embeddings

for mask classification.

In our experiments, we utilize SAM-H for mask generation and CLIP ConvNeXt-L [7] for feature extraction. We use SAM's default `AutomaticMaskGenerator`, which first samples points from the image and then uses these points as prompts to perform segmentation on the image. Without any additional training or fine-tuning, our Mask-Adapter combined with SAM achieved a remarkable mIoU of 31.4 on the A-150 dataset, surpassing the performance of ODISE [9]. This demonstrates the adaptability and effectiveness of our approach. However, one challenge of SAM lies in its exceptionally fine-grained mask outputs, which can negatively impact its performance on open-vocabulary semantic segmentation tasks. Addressing this limitation remains an open problem and is beyond the scope of this paper.

D. Visualizations

We provide additional visualization results comparing segmentation performance on A-150 [12] using FC-CLIP [11] and MAFTP [4], as well as the results after integrating the Mask-Adapter (Fig. 5). Comparisons on PC-459 [8] are shown in Fig. 6. Additionally, we demonstrate the performance of the combined SAM [5] and Mask-Adapter on COCO-Stuff with different vocabularies in Fig. 7.

References

[1] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1209–1218, 2018. 1

[2] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–338, 2010. 1

[3] 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

[4] Siyu Jiao, Hongguang Zhu, Jiannan Huang, Yao Zhao, Yunchao Wei, and Humphrey Shi. Collaborative vision-text representation optimizing for open-vocabulary segmentation. *arXiv preprint arXiv:2408.00744*, 2024. 4, 5

[5] 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. 3, 4

[6] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 1

[7] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022. 4

[8] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. 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*, pages 891–898, 2014. 1, 4, 6

[9] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2955–2966, 2023. 1, 4

[10] Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. San: Side adapter network for open-vocabulary semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 3

[11] Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip. *Advances in Neural Information Processing Systems*, 36:32215–32234, 2023. 1, 4, 5, 6

[12] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127:302–321, 2019. 1, 4, 5

(a) FC-CLIP (b) FC-CLIP+Ours (c) GT (d) MAFTP (e) MAFTP+Ours (f) GT

Figure 5. **Comparison of qualitative results on A-150 [12].** We compare the segmentation results of existing open-vocabulary methods [4, 11] with those after integrating the Mask-Adapter.

(a) Image (b) FC-CLIP (c) FC-CLIP+Ours (d) GT

Figure 6. **Comparison of qualitative results on PC-459 [8].** We compare the segmentation results of FC-CLIP [11] and FC-CLIP integrated with Mask-Adapter on PC-459, highlighting the improvements achieved with Mask-Adapter.

Figure 7. **Visualization results of Segment Anything with Mask-Adapter.** We present the segmentation results of Segment Anything integrated with Mask-Adapter on COCO-Stuff, using different vocabularies.