Learning Mask-aware CLIP Representations for Zero-Shot Segmentation
(*Supplementary material*)

Anonymous Author(s)
Affiliation
Address
`email`

In the supplementary material, we first introduce technical details of the "frozen CLIP" approaches in Sec. 1. Then the dataset settings are shown in Sec. 2. Moreover, we provide additional qualitative results in Sec. 3.

1 Technical details of the "frozen CLIP" approaches

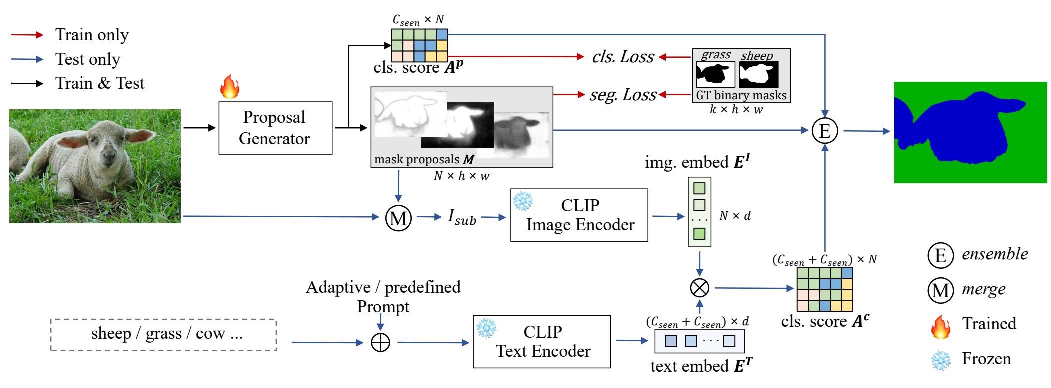

Figure 1: Overview of the "decoupling-paradigm".

Fig. 1 presents an overview of the "frozen CLIP" approach. **During training**, a standard MaskFormer or Mask2Former is used as Proposal Generator to generate N mask proposals (M, $M \in \mathbb{R}^{N\times h\times w}$) and classification score (A^p, $A^p \in \mathbb{R}^{N\times|C_{seen}|}$). **During testing**, the input image is merged with M to obtain N sub-images (I_{sub}, $I_{sub} \in \mathbb{R}^{N\times\hat{h}\times\hat{w}}$). These sub-images are fed into a frozen CLIP to get the CLIP classification score (A^c, $A^c \in \mathbb{R}^{N\times|C_{seen}\cup C_{unseen}|}$). Here C_{seen} and C_{unseen} represent a set of seen classes and unseen classes. An *ensemble* operation is used to ensemble A^p and A^c for the final prediction. The *merge* and the *ensemble* operations will be introduced in detail in following:

Merge operation. To generate appropriate sub-images based on mask proposals, [2] presents three different *merge* operations: 1) mask, 2) crop, 3) mask & crop. Through experimentation, they demonstrate that the mask & crop option yields the best results. Figure 2 provides an example of these operations. It's worth noting that all sub-images are resized to $\hat{h} \times \hat{w}$, here $\hat{h}$ and $\hat{w}$ typically take a value of 224, which is the default input size of CLIP Image Encoder. Although acceptable results can be obtained with the *merge* operation, it involves repeatedly feeding images into CLIP, which leads to significant computational redundancy.

Ensemble operation. Comparatively, A^p provides higher confidence classification scores for the seen classes and A^c provides higher confidence classification scores for the unseen classes. Therefore,

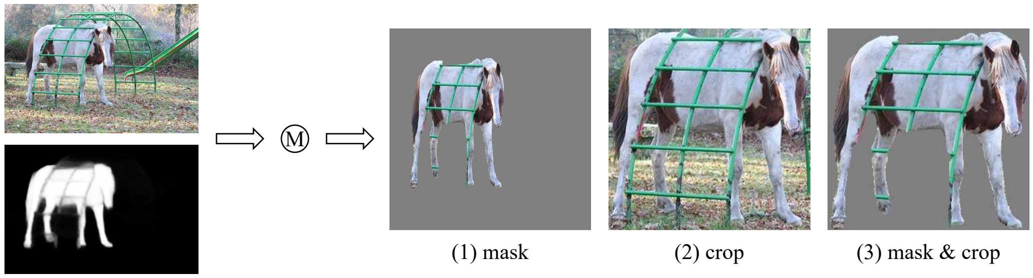

Figure 2: Comparison among three *merge* operations.

an ensemble of A^p and A^c achieves better results. The *ensemble* operation can be formulated as:

$$\hat{A}(c) = \begin{cases} A^p(c)^{\lambda} \cdot A^c(c)^{(1-\lambda)} & , \ c \in C^{seen} \\ A^c(c)^{\lambda} & , \ c \in C^{unseen} \end{cases} \quad (1)$$

here a geometry mean of A^p and A^c is calculated (dubbed as $\hat{A}$), and the contribution of both classification scores is balanced by λ. As per literature [2, 7, 6], λ usually takes values from 0.6 to 0.8. Therefore, the final output ($O, O \in \mathbb{R}^{|C_{seen} \cup C_{unseen}| \times h \times w}$) can be obtained by matrix multiplication: $O = \hat{A}^T \cdot M$. With the *ensemble* operation, the classification results of seen classes primarily depend on A^p, whereas the classification results of unseen classes mainly rely on A^c.

2 Dataset

We follow [1, 3, 5, 2, 7] to conduct experiments on three benchmarks of the popular *zero-shot* setting, Pascal-VOC, COCO-Stuff and ADE20K, to evaluate the performance of MAFT. Additionally, we evaluate MAFT on the *cross-dataset* setting [4, 7], *i.e.*, training on COCO-Stuff and testing on ADE20K, Pascal-Context, and Pascal-VOC.

- **COCO-Stuff**: COCO-Stuff is a large-scale semantic segmentation dataset that includes 171 classes. For the *zero-shot* setting [2, 7, 6], it is divided into 156 seen classes for training and 15 unseen classes for testing. For the *cross-dataset* setting, all 171 classes are used for training.
- **Pascal-VOC**: There are 10582 images for training and 1,449 images for testing. For the *zero-shot* setting, Pascal-VOC is split into 15 seen classes and 5 unseen classes. For the *cross-dataset* setting, all 20 classes are used for evaluation (dubbed as PAS-20).
- **ADE20K**: ADE20K contains 25k images for training and 2k images for validation. For the *zero-shot* setting, we follow [2] to choose 847 classes present in both training and validation sets, and split them into 572 seen and 275 unseen classes. For the *cross-dataset* setting, we use two settings of ADE20K: 150 classes (dubbed as A-150) and 847 classes (dubbed as A-847).
- **Pascal-Context** is an extensive dataset of Pascal-VOC 2010. Two versions are used for *cross-dataset* setting, one with 59 frequently used classes (dubbed as PC-59) and another with the whole 459 classes (dubbed as PC-459).

3 Visualization

We provide more qualitative results, including typical proposals and top-5 A^c (Fig. 3), as well as examples of models train on COCO-Stuff and text on A-847 (Fig. 4), A-150 (Fig. 5), PC-459 (Fig. 6), PC-59 (Fig. 7), Pascal-VOC (Fig. 8), and COCO-Stuff(Fig. 9).

Typical Proposals and Top-5 A^c. Fig. 3 shows frozen CLIP and mask-aware CLIP classifications of typical proposals. In the 2^{nd} column, we provide high-quality proposals of *thing* classes. Both the frozen CLIP and mask-aware CLIP provide high classification scores for the correct classes. In the 3^{rd} column, we provide proposals that only contain part of the objects (row 1-3), and proposals containing more than 1 class (row 4). The mask-aware CLIP provides more proper results compared to the frozen CLIP. In the 4^{th} column, we provide some high-quality background proposals. The frozen CLIP typically gives incorrect predictions, but the mask-aware CLIP assigns high scores for the correct classes.

Qualitative Analysis. Fig. 4,5,6,7,8,9 show segmentation results on Pascal-VOC, COCO-Stuff, ADE20K. In Pascal-VOC dataset (Fig. 8), which only contains 20 *thing* classes, the FreeSeg+MAFT model tends to assign background regions to the similar *thing* classes, *e.g.*, "train" in row 1, "potted-plant" in row3-4. "boat" in row 8. In A-847, A-150, PC-459, PC-59 and COCO-Stuff datasets, both seen classes and unseen classes exist in the input images, the FreeSeg+MAFT model generates better segmentation results compared to FreeSeg.

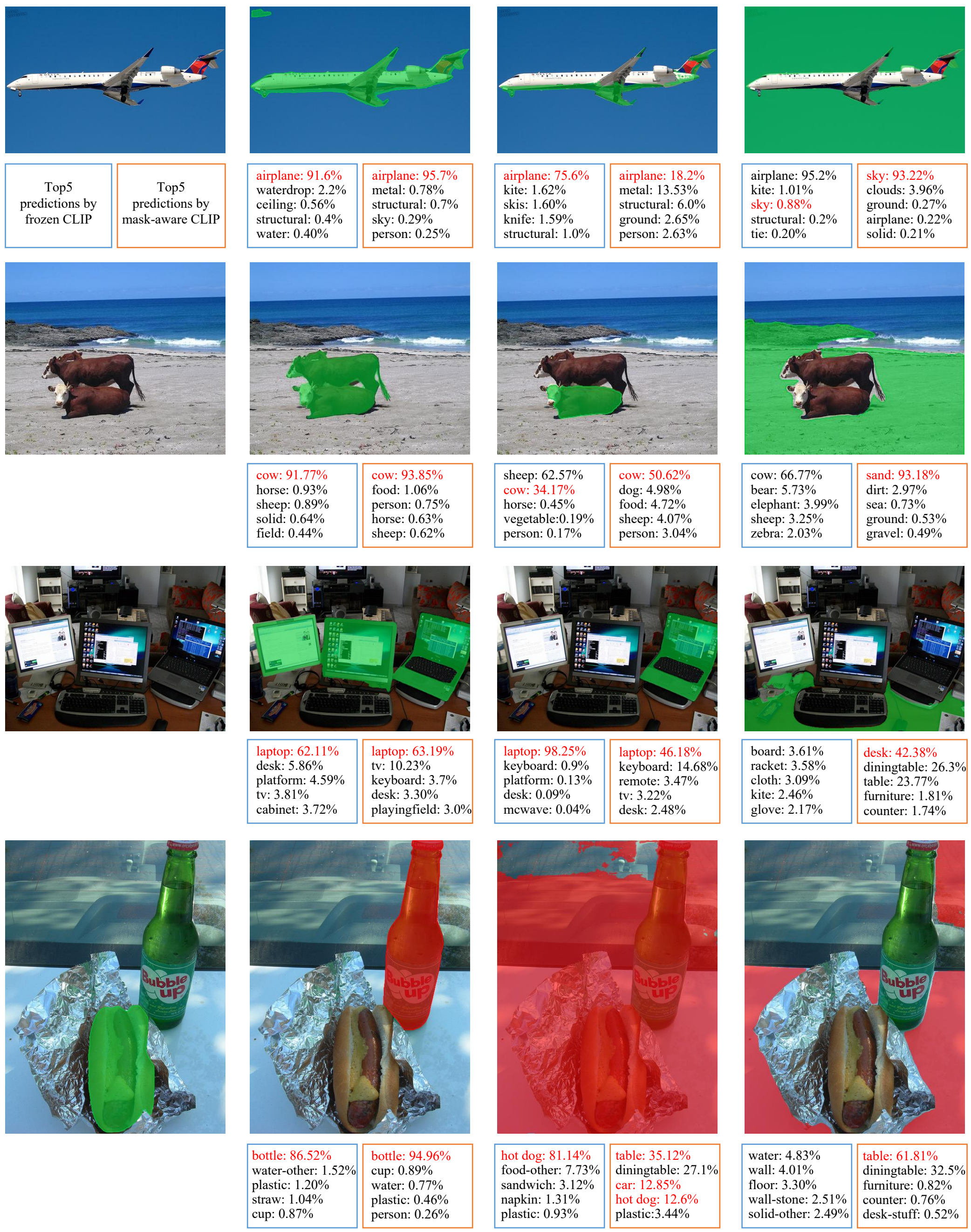

Figure 3: Visualizations of typical proposals top 5 A^c by frozen CLIP and mask-aware CLIP. The correct classes are highlighted in red.

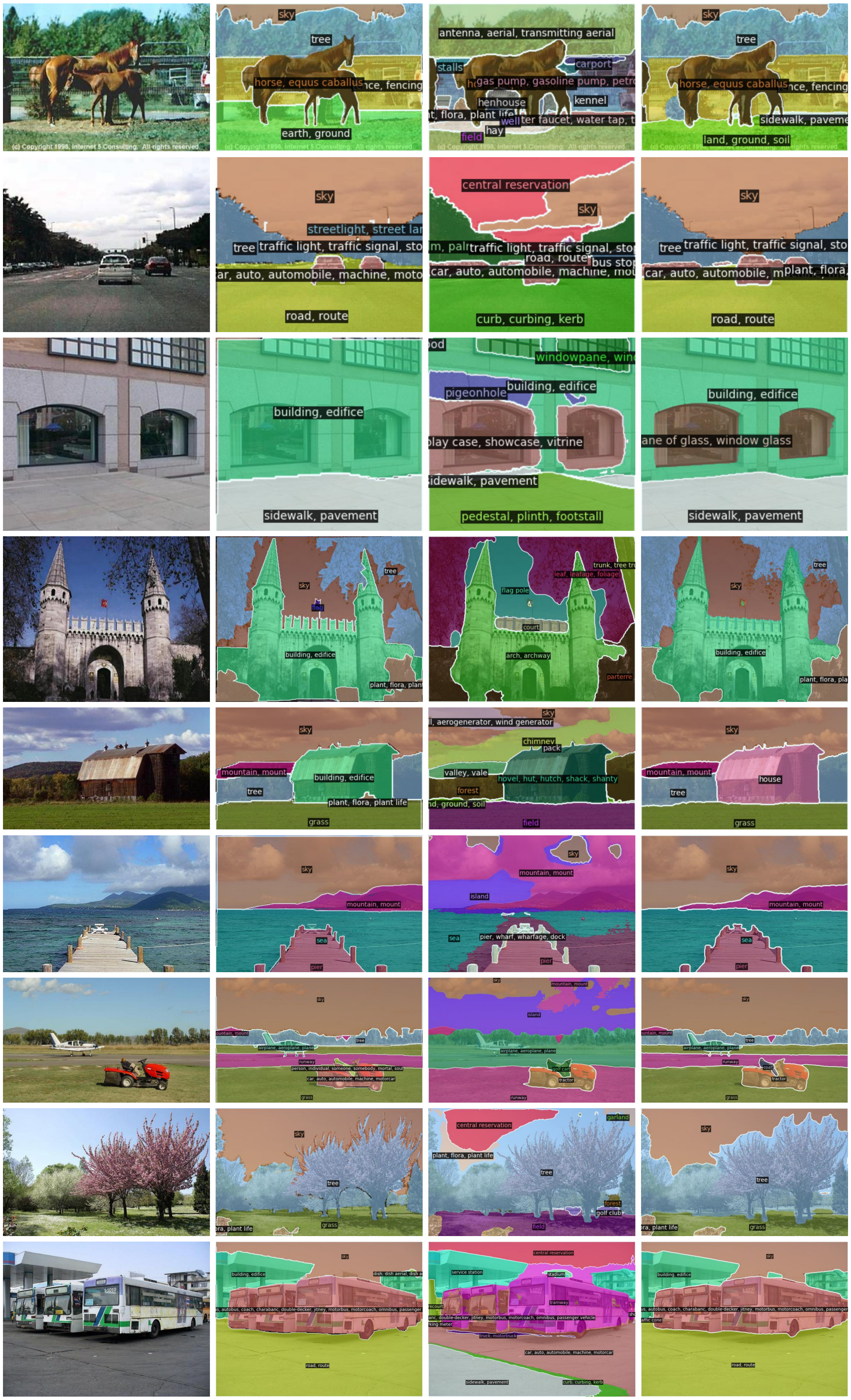

Image | GT | FreeSeg | FreeSeg+MAFT

Figure 4: Qualitative results on A-847, using 847 class names in ADE20K to generate text embeddings.

Figure 5: Qualitative results on A-150, using 150 class names in ADE20K to generate text embeddings.

Figure 6: Qualitative results on PC-459, using 459 class names in Pascal-Context to generate text embeddings.

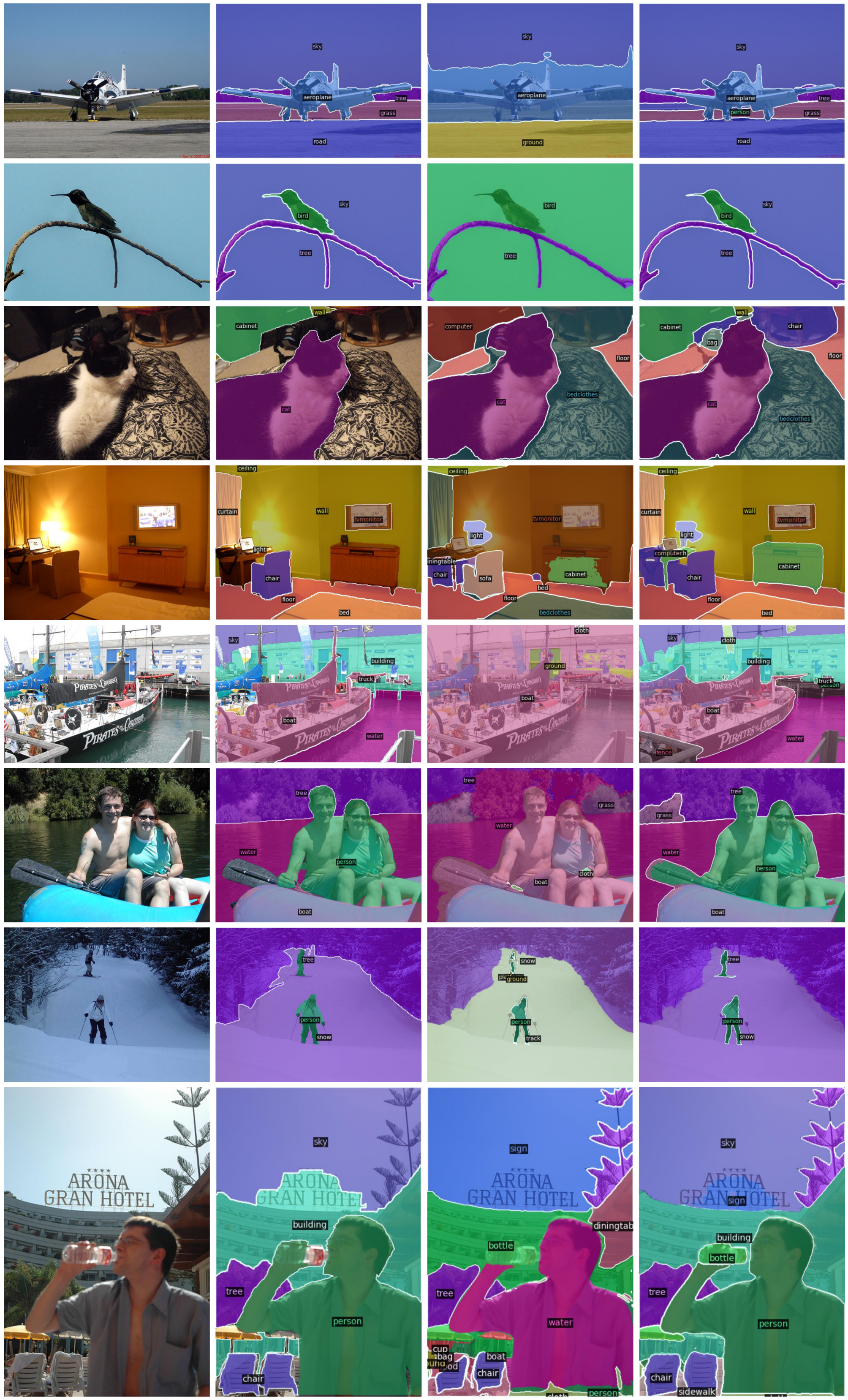

Image GT FreeSeg FreeSeg+MAFT

Figure 7: Qualitative results on PC-59, using 59 class names in Pascal-Context to generate text embeddings.

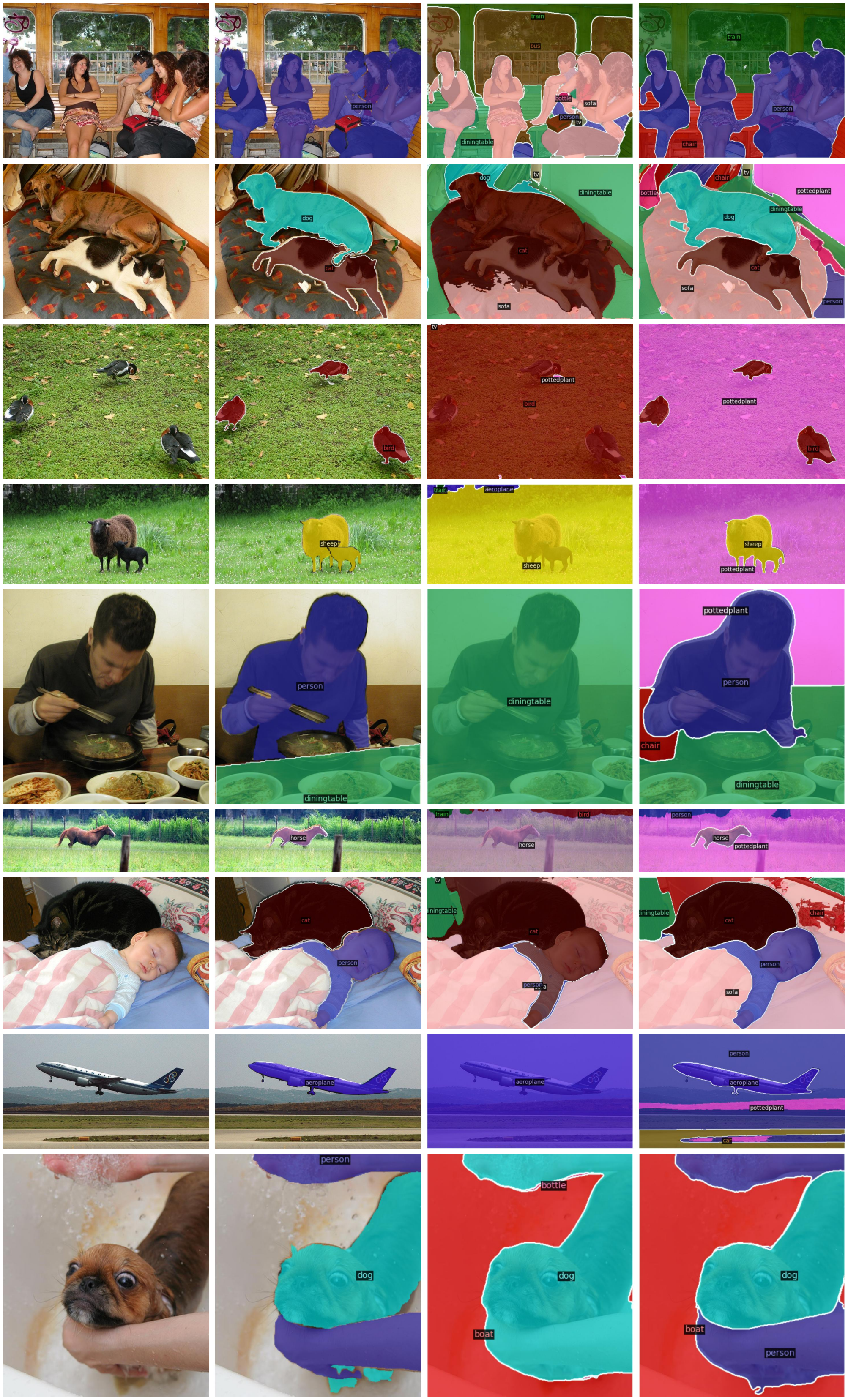

Image GT FreeSeg FreeSeg+MAFT

Figure 8: Qualitative results on Pascal-VOC, using 20 class names in Pascal-VOC to generate text embeddings.

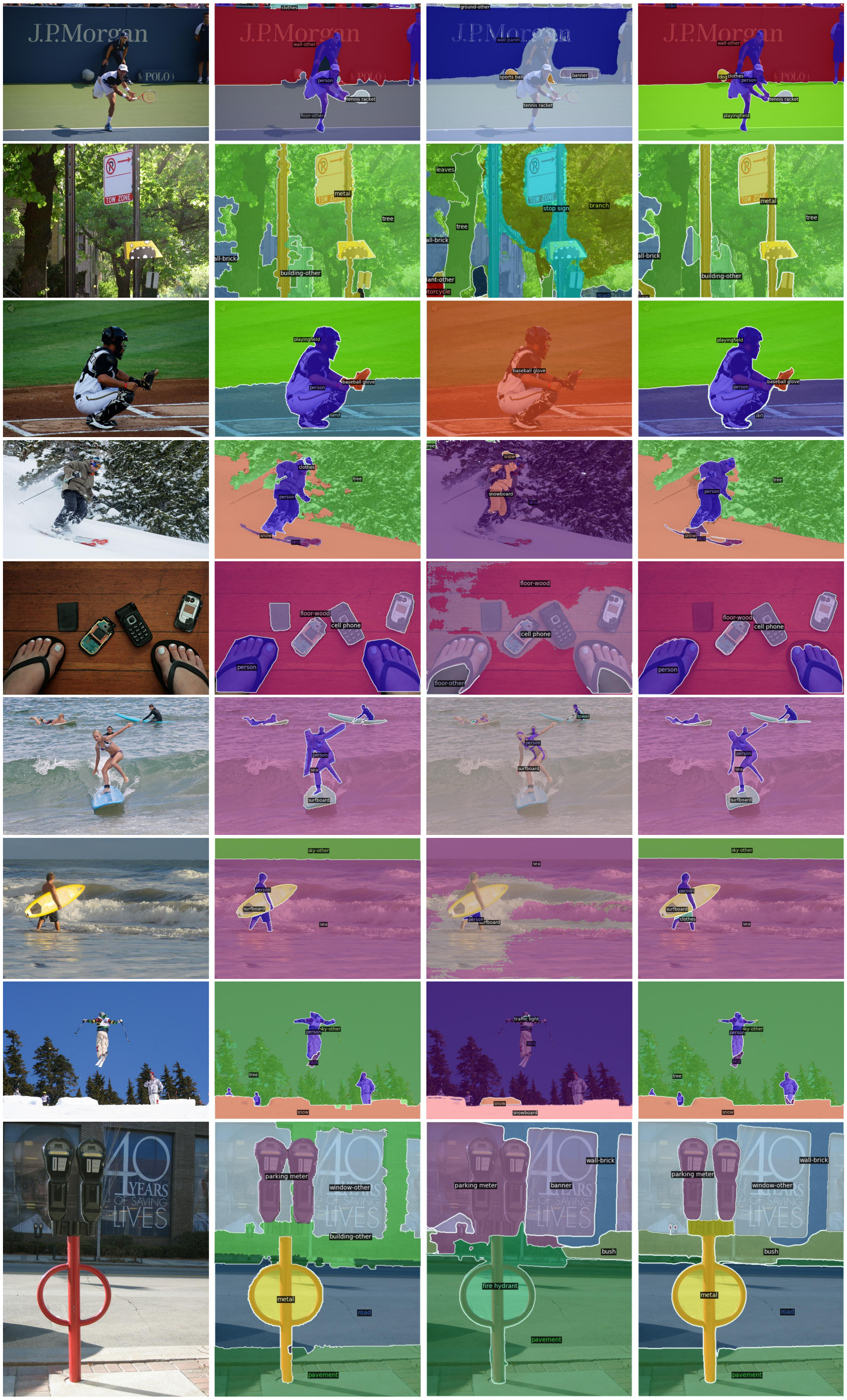

Image GT FreeSeg FreeSeg+MAFT

Figure 9: Qualitative results on COCO, using 171 class names in COCO-Stuff to generate text embeddings.

References

[1] Maxime Bucher, Tuan-Hung Vu, Matthieu Cord, and Patrick Pérez. Zero-shot semantic segmentation. *Advances in Neural Information Processing Systems*, 32, 2019.

[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.

[3] Zhangxuan Gu, Siyuan Zhou, Li Niu, Zihan Zhao, and Liqing Zhang. Context-aware feature generation for zero-shot semantic segmentation. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1921–1929, 2020.

[4] 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. *arXiv preprint arXiv:2210.04150*, 2022.

[5] Giuseppe Pastore, Fabio Cermelli, Yongqin Xian, Massimiliano Mancini, Zeynep Akata, and Barbara Caputo. A closer look at self-training for zero-label semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2693–2702, 2021.

[6] Jie Qin, Jie Wu, Pengxiang Yan, Ming Li, Ren Yuxi, Xuefeng Xiao, Yitong Wang, Rui Wang, Shilei Wen, Xin Pan, et al. Freeseg: Unified, universal and open-vocabulary image segmentation. *arXiv preprint arXiv:2303.17225*, 2023.

[7] 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 *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIX*, pages 736–753. Springer, 2022.