Supplementary Material: Learning Open-vocabulary Semantic Segmentation Models From Natural Language Supervision

Here, we start by providing additional experimental results in Sec. 1 and give more visualization results in Sec. 2.

1. Additional Experiments

1.1. Details on the Entity Set

The constructed entity set contains 100 frequently appeared entities, including: people, man, men, woman, women, girl, boy, lady, kid, child, children, baby, student, bride, groom, couple, prince, princess, car, bus, truck, motorcycle, train, bicycle, boat, aeroplane, airplane, motorbike, bike, cup, bottle, bowl, knife, spoon, glass, fork, chair, table, bench, clock, laptop, light, vase, plant, remote, microwave, toaster, oven, mouse, keyboard, sofa, monitor, desk, tv, TV, couch, flower, refrigerator, house, building, hotel, handbag, umbrella, book, backpack, phone, shirt, tie, suitcase, T-shirt, bag, box, sink, bed, toilet, cat, dog, horse, bird, cow, sheep, elephant, bear, zebra, giraffe, ball, racket, skateboard, skis, snowboard, surfboard, kite, pizza, cake, apple, banana, sandwich, orange, carrot, donut. **Note that**, we exclude the word "person" in the entity set as CC12M [2] claimed that they performed person-name substitutions to protect the privacy of the individuals in the images, specifically, all named entities of type Person (*e.g.*, the name of the artist) detected by the natural language APIs are replaced with "person".

1.2. Additional Ablation Studies

Effect of the Pre-trained Backbones. We show the effect of applying different unimodal/multimodal pre-trained weights for visual and textual encoders in Table 1, with $\mathcal{L}_{\text{contrast}}$ being adopted only. Training both encoders from scratch only achieves 28.8 mIoU on PASCAL VOC. Initialization from CLIP visual and text encoders (including the visual/textual projection heads) brings significant improvement. However, it requires 400M image-text pairs for pre-training. Besides, a potential drawback of applying CLIP pre-trained weights is that the model can easily learn the visual-text alignment while ignoring the visual grouping. In comparison, initializing the model from single-modality sources, i.e. DINO and BERT, yields better performance. This design choice requires no manual annotation as both DINO and BERT use self-supervised training.

Pre-training scheme		PASCAL VOC	PASCAL Context
Visual Enc.	Text Enc.		
✗	✗	28.8	12.1
CLIP-V	CLIP-T	38.4	**15.3**
DINO	BERT	**40.5**	15.1

Table 1. Comparison of different unimodal/multimodal pre-training schemes for visual encoder and text encoder.

On the Choice of Mask Threshold. Here, we study the influence of different mask thresholds δ as mentioned in Sec.3.2 in the manuscript. As observed in Table 2, our model reaches a decent mIoU of 53.6 on PASCAL VOC when δ is 0.6, while smaller thresholds lead to false-positive pixels of the objects.

Effect of the Momentum Model. Our proposed OVSegmentor adopts a momentum model for encoding the cross-image, which is updated by the exponential-moving-average (EMA) of the online model. Table 3 reveals that applying the momentum model brings about 2% mIoU gain on PASCAL VOC and COCO Object. We attribute this to the improved quality of the

δ	0.3	0.4	0.5	0.6	0.7	0.8	0.9
mIoU	51.2	51.6	51.8	**53.6**	53.1	53.2	53.4

Table 2. Comparison of different mask thresholds δ on PASCAL VOC.

Method	PASCAL VOC	PASCAL Context	COCO Object
w/o momentum	50.9	20.2	23.7
w/ momentum	**53.8**	**20.4**	**25.1**

Table 3. Effect of the momentum model in OVSegmentor.

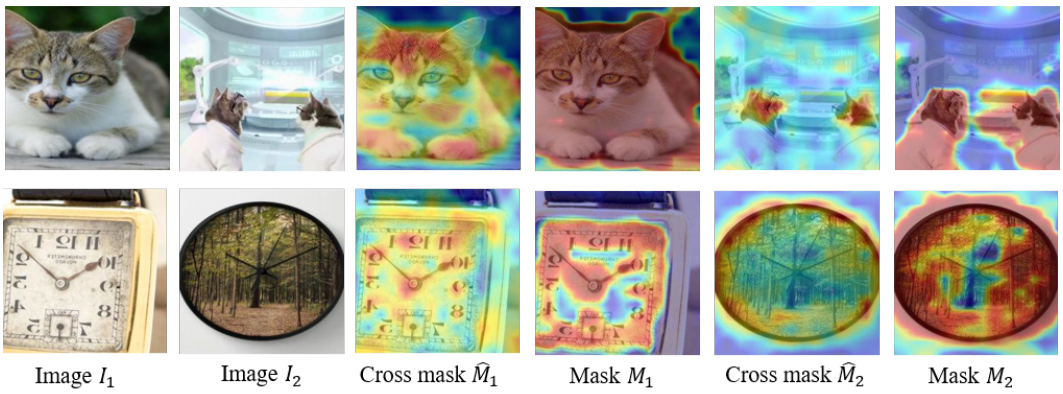

Figure 1. Qualitative results of the masks and cross-image masks in our proposed cross-image mask consistency.

pseudo targets generated by the momentum model. In Fig. 1, we also show the object masks generated by our online model $\hat{M}_1, \hat{M}_2$ and momentum model M_1, M_2 for both the input image I_1 and the sampled cross-image I_2 with the shared entity.

Mask Probing. Following DINO [1] and GroupViT [6], we evaluate the quality of the generated masks regardless of the class predictions, termed as mask probing. Mask probing directly reflects the effect of the pixel-to-group assignment in our proposed model. For ViT-based methods that adopt finetuning transfer, *i.e.*, DeiT [5], MoCo [3] and DINO [1], the self-attention maps in the last ViT block are probed.

Specifically, we denote the self-attention maps as $S \in \mathbb{R}^{nh\times(L+1)\times(L+1)}$, where nh refers to the number of heads and $L+1$ is the token numbers (L image tokens and 1 class token), the self-attention masks $M \in \mathbb{R}^{nh\times1\times L}$ are derived by taking the similarities of the class token and all image tokens. M is then binarized by keeping the highest values (*e.g.* 60%) as the foreground and the remaining regions as the background. The Jaccard similarity is computed between the attention mask for each head and the ground-truth mask, and the one with the highest similarity is taken as the mask probing result. For grouping-based methods GroupViT [6] and our proposed OVSegmentor, the pixel-to-group affinity $\mathbb{A} \in \mathbb{R}^{HW\times K}$ is considered as the attention masks. We directly choose one of the K groups that has the highest Jaccard similarity to the ground-truth mask. As demonstrated in Table 4, OVSegmentor surpasses methods using finetuning transfer. Despite comparable mask probing performance to GroupViT, our proposed OVSegmentor still outperforms GroupViT in terms of open-vocabulary semantic segmentation, indicating that OVSegmentor learns better group-text alignment with 85% less data (4M vs 27M) used during pre-training.

Per-class Segmentation Performance. We compare the mIoU over total 20 object categories in PASCAL VOC, as shown in Table 5. Our proposed OVSegmentor surpasses VIL-Seg [4] on all the categories, while significantly outperforming GroupViT [6] on categories such as aeroplane, car, motorbike. OVSegmentor achieves inferior results on the "person" class, owing to its large variation of visual appearance in web-collected images, posing additional challenges for our proposed cross-image mask consistency to learn visual invariance.

Method	Pretrain dataset	Supervision	Zero-shot transfer	Mask probing	Open-vocabulary segmentation
DeiT [5]	ImageNet-1K	class	✗	24.6	53.0
MoCo [3]	ImageNet-1K	self	✗	28.2	34.3
DINO [1]	ImageNet-1K	self	✗	45.9	39.1
DINO [1]	CC12M+YFCC15M	self	✗	41.8	37.6
GroupViT [6]	CC12M+YFCC15M	text	✓	51.8	51.2
GroupViT* [6]	CC4M	text	✓	45.2	25.8
OVSegmentor	CC4M	self+text	✓	**50.9**	**53.8**

Table 4. Comparison of mask probing results on PASCAL VOC. The results of DeiT, MoCo, DINO and GroupViT are reported in [6].

Method	Pretrain	aeroplane	bicycle	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motorbike	person	plant	sheep	sofa	train	monitor
VIL-Seg [4]	12M	40.2	21.6	41.0	17.3	35.3	52.8	10.1	59.3	15.4	42.4	21.4	49.5	56.1	49.4	11.3	21.6	41.5	18.6	54.1	13.4
GroupViT [6]	27M	38.3	31.4	50.6	31.9	**63.2**	**78.8**	65.1	**79.2**	**18.1**	**74.0**	**30.9**	**76.2**	59.3	55.0	**44.1**	40.9	66.6	**31.5**	49.6	29.8
OVSegmentor	4M	**70.8**	**32.8**	**57.5**	**40.2**	57.3	76.7	**71.7**	77.4	16.5	72.7	28.2	61.4	**60.0**	**70.4**	17.8	**43.3**	**69.7**	31.2	**58.7**	**33.2**

Table 5. Comparison of per-class mIoU results on PASCAL VOC.

1.3. Model efficiency

Table 6 shows the computation cost on a single A100 GPU, using ViT-S/16 with 224×224 input, and BERT-base text encoder. Despite extra training cost, our model outperforms GroupViT while retaining lower inference cost as the decoder is discarded after training.

Method	$\mathcal{L}_{\text{con}}$	$\mathcal{L}_{\text{con}}$+$\mathcal{L}_{\text{ent}}$	$\mathcal{L}_{\text{con}}$+$\mathcal{L}_{\text{ent}}$+$\mathcal{L}_{\text{mask}}$	Inference	mIoU
GroupViT*	145M / 0.51s	-	-	145M / 51	19.8
Ours	141M / 0.47s	148M / 0.65s	148M / 1.60s	141M / 59	44.5

Table 6. Training cost (#params / sec-per-iter), inference cost (#params / FPS) and performance (mIoU) on PASCAL VOC.

2. More Visualization Results

Additional qualitative results on PASCAL VOC, PASCAL Context, and COCO Object can be found in Fig. 2, Fig. 3, and Fig. 4, respectively. Generally, our proposed OVSegmentor successfully groups semantically related pixels together and aligns the group to the correct category. On PASCAL VOC, OVSegmentor successfully segments objects with various scales (*e.g.* small aeroplanes and distant cars in the 2$^{\text{nd}}$, 5$^{\text{th}}$ and 7$^{\text{th}}$ rows) and multiple objects of the same class (4$^{\text{th}}$, 6$^{\text{th}}$ and 8$^{\text{th}}$ rows). In terms of PASCAL Context where objects of more categories are annotated, our model manages to segment the salient objects while failing to recognize stuff classes that usually appear as the background in web-collected data (*e.g.* grass, floor, wall, etc.). On COCO Object, we observe that our model can not separate co-occurring objects from different classes into distinct groups very well (*e.g.* laptop and mouse), which we conjecture is because the captions sourced from the Internet usually lack fine-grained descriptions to cover the full image content.

References

[1] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 2, 3

[2] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, 2021. 1

[3] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *ICCV*, 2021. 2, 3

[4] Quande Liu, Youpeng Wen, Jianhua Han, Chunjing Xu, Hang Xu, and Xiaodan Liang. Open-world semantic segmentation via contrasting and clustering vision-language embedding. In *ECCV*, 2022. 2, 3

[5] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *ICML*, 2021. 2, 3

[6] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. In *CVPR*, 2022. 2, 3

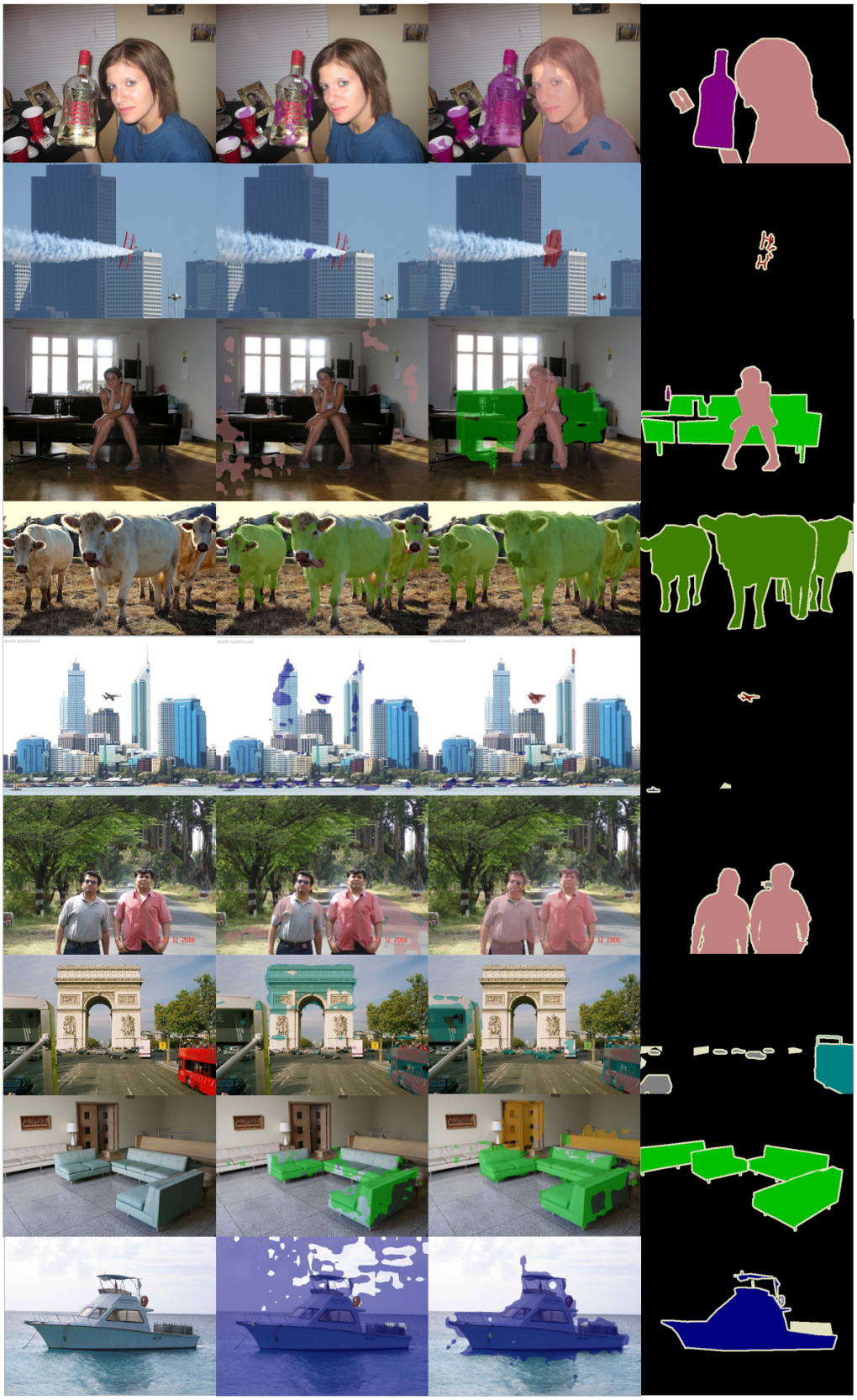

Figure 2. Qualitative results on PASCAL VOC.

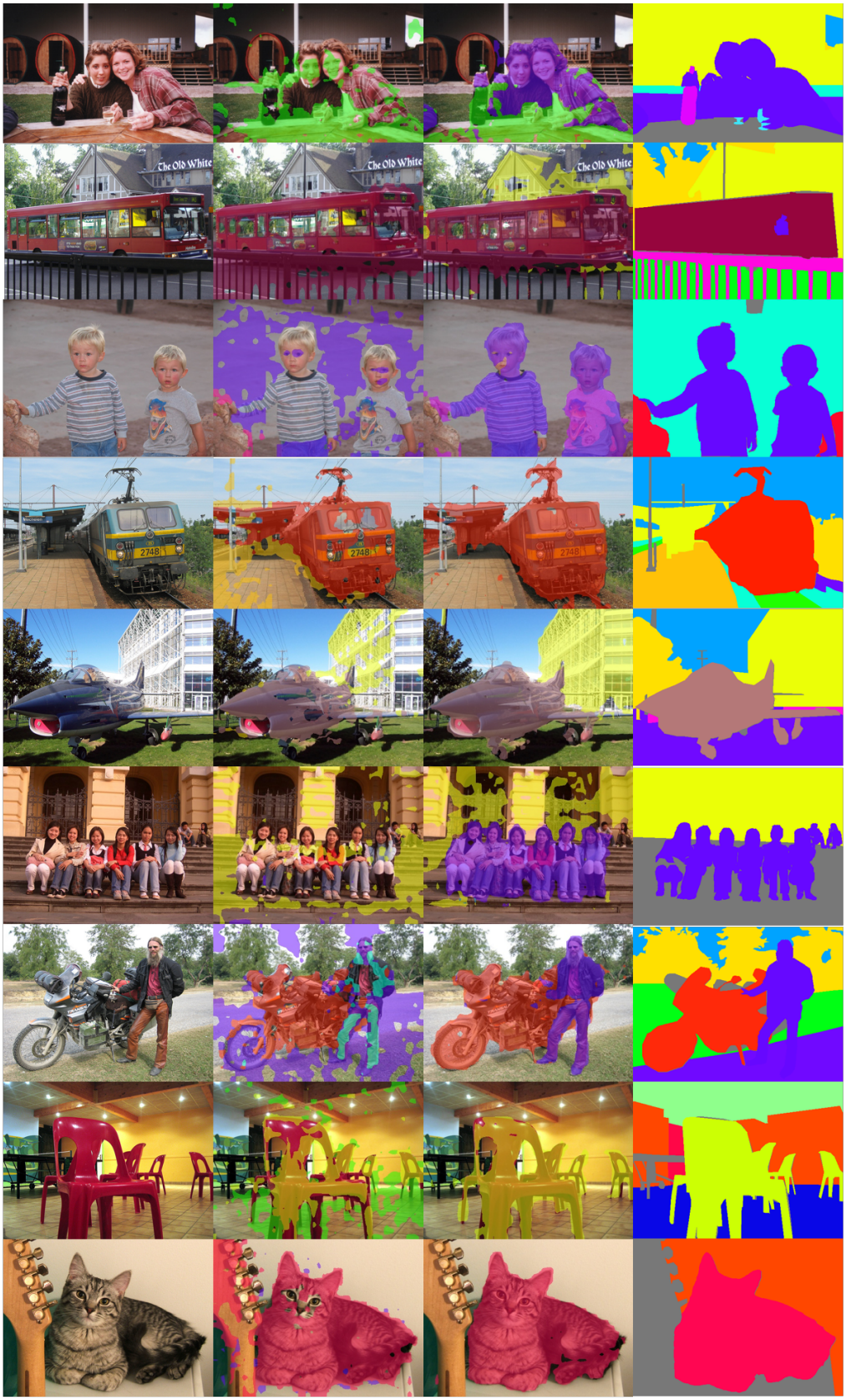

Figure 3. Qualitative results on PASCAL Context.

Figure 4. Qualitative results on COCO Object.