Exploring Simple Open-Vocabulary Semantic Segmentation (Supplementary Material)

Zihang Lai
Visual Geometry Group (VGG), University of Oxford
zlai@robots.ox.ac.uk

1. Additional implementation details

1.1. S-Seg experiments

Training. During training, our model is trained end-to-end using AdamW [13] with intial learning rate $5e^{-4}$ and batch size 4096. We use cosine learning rate decay schedule with 2 warmup epochs, and the training lasts for 30 epochs in total. We use a 0.05 weight decay. See Table 1 for our complete training hyperparameter settings. All input images are random resized and cropped to 224×224 in resolution. Following [18], we extract nouns and verbs from raw sentence because these words are more likely to describe the image.

Inference. We evaluate S-Seg on the validation set of three datasets: Pascal VOC 2012 [6], Pascal Context [15] and COCO [11]. As in [18], we combine all instances of the same class to get semantic segmentation mask for each image in COCO. Following GroupViT [18], we threshold the maximum probability to obtain background prediction. During inference, we set the input resolution to 448×448, which is consistent with [18].

1.2. S-Seg+ experiments

Self-training. For self-training experiments, we use UperNet [17] with MAE [9] pretrained ViT backbone. We utilize a pyramid-structured network to merge the features obtained from layer 4, 6, 8, and 12 of the ViT, following the implementation of BEiT [1]. We use the same model that we used to evaluate our main results to generate training data from the train set of the respective dataset. Training hyperparameters are provided in Table 2. Following [1, 9], we use a layerwise learning rate decay [4]. We do *not* use relative position embeddings in our backbone ViT model (which is used by [1, 9] at fine-tuning stage for extra improvement).

1.3. Reimplemented baselines

CLIP [16]. We utilized the CLIP ViT-B/16 model along with the official pretraining weights. The ViT model incorporates attentional pooling in its last layer, using an additional [CLS] token to aggregate other tokens. We choose to employ the *value* embedding as the representation of each token, as the query and key embedding of the final layer is not fully trained during CLIP pretraining (only the similarity between the query embedding of the [CLS] token and the key embedding of other tokens is utilized). Finally, we leverage the language model to encode all classes and classify the visual tokens, similar to CLIP's zero-shot classification approach.

config	value
optimizer	AdamW [13]
base learning rate	5e-4
weight decay	0.05
optimizer momentum	β_1, β_2=0.9, 0.999
batch size	4096
learning rate schedule	cosine decay [12]
warmup epochs [8]	2
training epochs	30

Table 1. **S-Seg setting.**

config	value
optimizer	AdamW [13]
base learning rate	1e-4
weight decay	0.05
optimizer momentum	β_1, β_2=0.9, 0.999
batch size	16
learning rate schedule	polynomial decay
warmup iters [8]	1.5k
training iters	20k (voc), 40k (ctxt), 80k (coco)
layer-wise lr decay [4]	0.7

Table 2. **S-Seg+ setting.**

MaskCLIP [3]. We use the testing code and weights provided by the authors, but re-evaluating them on the commonly-used protocol that includes the background class. To further assess the efficacy of our approach, as well as baseline methods, we employed the evaluation metric utilized by MaskCLIP, which specifically disregards background pixels.

GroupViT [18]. The GroupViT project has provided pre-trained models for two configurations. Without specific clarification, we opt to use the model with the highest average accuracy, which was trained on CC12M, CC15M, and

(a) **Scaling training data provide consistent gain**: We train our model using different size of data: 12M (CC12M), 15M (+CC3M), and 26M (+RedCaps). We note a steady improvement in the model's performance as the data size increases.

data	S-Seg			S-Seg+		
	VOC	Context	COCO	VOC	Context	COCO
12M	44.9	22.9	22.5	53.1	25.5	26.2
15M	45.1(+0.2)	23.8(+0.9)	27.9(+5.4)	54.2(+1.1)	29.2(+3.7)	28.0(+1.8)
26M	**53.2**(+8.3)	**27.9**(+5.0)	**30.3**(+7.8)	**62.0**(+8.9)	**30.2**(+4.7)	**35.7**(+9.5)

(b) **Self-training offers constant improvement**: We observe that self-training consistently leads to significant improvement on performance across 3 datasets.

method	3-Average		
	12M	15M	26M
w/o self-train	30.1	30.8	37.1
w/ self-train	**34.9**	**37.1**	**42.6**
Δ	+4.8	+6.3	+5.5

Table 3. **Ablations on data scalability and self-training**. We report mIoU evaluated on three datasets. Higher values are better.

OV Methods	BG.	a.plane	bicycle	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	m.bike	person	plant	sheep	sofa	train	monitor	mIoU
CLIP	13.2	10.4	4.4	8.0	5.9	19.4	27.0	17.5	26.0	3.1	19.6	9.0	21.5	16.8	11.2	11.7	5.2	13.1	7.6	21.1	12.2	13.5
MaskCLIP	41.3	12.8	18.7	22.5	6.7	22.8	50.7	23.4	56.8	13.6	34.1	8.1	46.3	29.5	39.9	22.7	9.5	29.5	25.1	30.8	18.2	26.8
GroupViT	79.0	37.4	29.9	33.3	33.9	64.4	60.2	62.4	76.7	16.2	68.8	28.0	75.9	62.5	64.2	51.6	38.7	63.0	37.4	44.0	38.4	50.8
S-Seg(Ours)	81.0	47.2	40.1	38.6	30.0	63.5	74.6	67.6	75.7	**18.6**	65.3	34.4	72.2	56.3	68.0	50.7	45.7	60.2	33.6	53.1	41.0	53.2
S-Seg+(Ours)	**86.5**	**53.8**	**42.0**	**48.1**	**49.3**	**76.0**	**84.7**	**74.5**	**87.2**	17.1	**81.8**	**35.0**	**83.4**	**65.2**	**74.3**	**65.3**	**46.6**	**78.2**	**40.2**	**58.5**	**53.6**	**62.0**

Table 4. **Per-category open vocabulary semantic segmentation performance over 21 Pascal VOC classes.** Our method surpass baseline methods such as GroupViT on the Pascal VOC dataset, particularly in segmenting large objects and categories with consistent textures.

Method	OV	Sup.	LVIS (1103 classes)	ImageNet-S (919 classes)
CLIP [16]	✓	text	1.3	8.0
MaskCLIP [19]	✓	text	4.3	9.1
GroupViT [18]	✓	text	7.2	32.2
S-Seg (Ours)	✓	text	**8.5**	**34.9**
ViT-FCN [1]	✗	GT	9.6	40.4

Table 5. **Open-vocabulary semantic segmentation results on LVIS and ImageNet-S.** Our method demonstrates competitive performance on these challenging datasets with a significantly larger number of classes.

Redcaps datasets. This particular model also closely aligns with our method in terms of training data.

Fully supervised models (DeepLabV3+ [2] and MaskFormer [3]). We leverage public checkpoints when available. In cases where a checkpoint is not available, we retrain the model using the original training hyperparameters (*e.g.* optimizer, learning rate, momentum, and weight decay) along with the standard training schedule, which varies depending on the dataset (40k iterations for P. VOC, 80k for P. Context, and 160k for COCO). We show the performance of DeepLabV3+ in qualitative comparisons (*Fully Sup.*).

2. Additional results

2.1. Additional datasets

We evaluate our method on two new challenging datasets that contain significantly more classes, LVIS (1103 classes) and ImageNet-S (919 classes). The results are shown in Table 5. We observe that our model outperforms several existing open-vocabulary baseline methods and approaches supervised models, indicating its robustness in challenging scenarios.

[1]We also tried DeepLabV3+ but failed to obtain satisfactory results.

2.2. Ablation results

In Table 3a and 3b, we show numerical results corresponding to Figure 10 and 12 in the main paper. As seen from the table, scaling data and self-training provide consistent gain in performance for our model.

2.3. Per-category result

Table 4 presents the mIoU results of our models and baseline methods on the Pascal VOC dataset, where each class is evaluated separately. Our models outperform GroupViT in most classes, and S-Seg+ achieves superior performance across *all* categories. Our models are particularly effective at segmenting large objects such as aeroplanes, buses, and trains, with an average improvement of 11.1 compared to 2.5 for all classes. This improvement could suggest that our models benefit from the pseudo-mask generator, which works better for larger objects (which shows a 83.3% oracle performance compared to 77.2% for other classes). On the other hand, our self-training model performs better on categories that share consistent texture, such as cats, cows, dogs, and sheep, with an average improvement of 14.3 compared to 8.8 for all classes. This indicates that self-training can identify common features and reduce noise in the self-training labels.

2.4. Additional visualizations

Figures 3 and 4 present more detailed open-vocabulary segmentation results in higher resolution. As shown in the results, our approach can effectively segment object-centric images from [6] (fig. 3) as well as context-rich images from [11] (fig. 4) accurately. Our method can segment objects based solely on their category name, without requiring any annotations from specific target datasets during training. Figure 5 and 6 provide additional comparison with previous methods.

Method	Algorithm	VL Pretrain	Pretrain data	Anno. masks	I-T pairs	Custom model	Loss	mIoU (VOC)
OpenSeg [7]	Adapt&Refine image-level VL alignment models	Yes (ALIGN)	1800M	Yes (COCO)	-	Not required	image+pixel	77.2
ZegFormer [5]	Directly training pixel&language alignment	Yes (CLIP)	400M	Yes (COCO)	-	Not required	image+pixel	80.7
MaskCLIP [19]	Adapt&Refine image-level VL alignment models	Yes (CLIP)	400M	Not required	-	Not required	image	49.5
GroupViT [18]	Extract segments from language alignment	Not required	-	Not required	30M	Yes (GroupViT)	image	77.2
S-Seg (Ours)	Directly training pixel&language alignment	Not required	-	Not required	26M	Not required	image+pixel	81.8

Table 6. **Comparing S-Seg (Ours) with closely-related methods (OpenSeg [7], ZegFromer [5], MaskCLIP [19], and GroupViT [18]).** We conduct a comparative analysis of our method against a range of closely-related approaches, which are further detailed in Section 3.

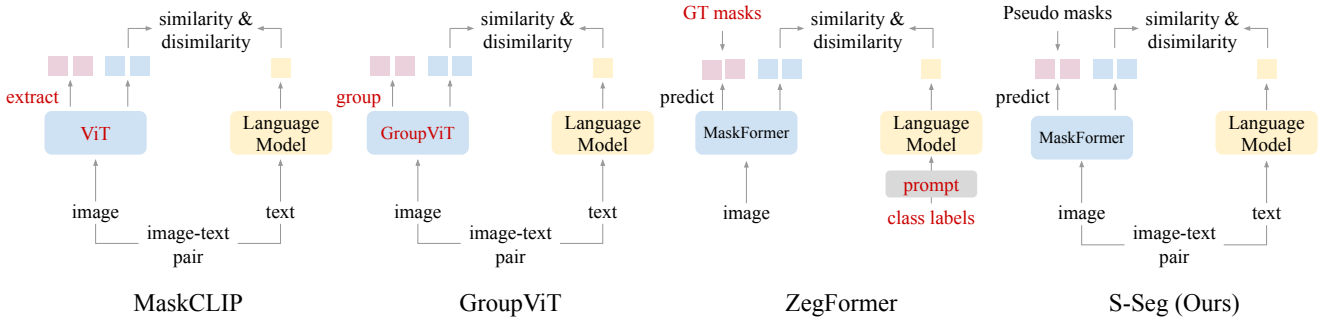


Figure 1. **Comparing S-Seg (Ours) with closely-related methods.** The components in red are those different from S-Seg.

3. Methodology Comparisons

We present a comparative analysis of our method against several closely-related exemplary approaches. Our method serves as a connection among these methodologies. The primary similarities and differences are outlined in Table 6 and Figure 1, with further discussion below.

Relation to OpenSeg. OpenSeg (and similar methods, *e.g.* LSeg [10]) refines image-level models like CLIP/ALIGN by training on annotated semantic masks. The pretrained image-level model provides language alignment and utilize ground truth mask for refining pixel-level feature. In contrast, S-Seg trains directly on pixel features from pseudo-masks and learns language alignment through text. Conceptually, S-Seg offers an end-to-end alternative to OpenSeg, with the added advantage of training exclusively on image-text pairs. Our approach removes the need for the resource-intensive VL pretraining step, streamlines the learning process, and reduces the reliance on extensive supervised data.

Relation to ZegFormer. Our method can be conceptualized as a variant of "ZegFormer trained from scratch with pseudo-masks and language," albeit with notable implementation distinctions. Training with *seen* ground truth masks benefits in-domain classes, but may not extend to *unseen* classes. Interestingly, while our method underperforms compared to ZegFormer on *seen* classes, it surpasses ZegFormer in handling *unseen* classes and demonstrates superior average performance across the dataset. This suggests that our solution offers better generalization than ZegFormer, despite not utilizing CLIP, annotated masks, or pixel-wise labels. The architectural and training similarities between the two methods suggest that their integration could lead to enhanced performance, a hypothesis we leave for future exploration.

Relation to CLIP/MaskCLIP. Our method closely parallels CLIP in the image-text contrastive training paradigm and can be seen as a "CLIP with MaskFormer as the image encoder," supplemented by an additional mask supervision branch. Despite these similarities, CLIP primarily aims to learn image-level alignment, whereas S-Seg is focused on pixel-level alignment. This is evident from the fact that even with the MaskCLIP adaptation, the segmentation performance significantly lags behind that of other compared methods. This highlights the importance of incorporating both the MaskFormer and mask supervision in S-Seg.

Relation to GroupViT. GroupViT and S-Seg share a similar problem setup, where both methods avoid CLIP pretraining and manual annotations. Methodologically, S-Seg resembles "GroupViT with MaskFormer as the grouping model." A key difference, however, is that GroupViT *extracts* segments from a trained model, while S-Seg *directly predicts* segmentation, supervised by pseudo-masks. This more explicit form of supervision allows S-Seg to leverage standard segmentation models like MaskFormer more effectively and offers a potentially simpler pathway for updates with future advancements in segmentation models.

4. Limitations

Our model has several limitations. First, it may struggle in scenes with high color contrast, where sharp transitions between adjacent regions can confuse the model and lead to inaccurate segmentation boundaries (see Fig. 2, top). Second, performance tends to degrade in visually complex scenes, such as those with dense object arrangements, intricate textures, or highly cluttered backgrounds, where distinguishing individual instances becomes more challenging (see Fig. 2, bottom). Third, the model can fail to correctly separate overlapping objects, especially when their visual features are similar, resulting in merged or incomplete segmentations. Finally, rare categories that appear infrequently in the training data are prone to misclassification or may be entirely missed, reflecting a limitation in generalization.

5. Discussions

Q: Is vision-language (VL) training still used in your approach, and how does it differ from large-scale methods like CLIP?

A: Yes, VL training is still used. However, our approach avoids relying on massive cross-modal pretraining like CLIP. Instead, we focus on making training more accessible and reproducible from scratch by using open datasets that are over 10 times smaller (26M vs. 400M). While we do use pretrained models like DINO, it's based on unlabeled ImageNet and requires far less compute, without needing curated text–image pairs.

Q: Could you provide deeper insights into the core motivation behind your approach?

A: Our central motivation is to show that **directly learning pixel-level vision–language alignment** is both viable and effective. By removing complex components to barebone, two core interests emerge: (1) exploring a simpler pipeline that is not tied to massive curated text–image datasets or adapting proprietary models, and (2) demonstrating the feasibility of an alternate route to open-vocabulary segmentation that focuses on learning from self-supervised visual features in conjunction with textual embeddings.

Q: Why did you choose a weakly supervised setting instead of fully supervised methods?

A: We opted for a weakly supervised approach for two main reasons. First, creating segmentation masks is costly, whereas image–text pairs are much easier to collect at scale, such as from internet images and alt-text. Second, weak supervision tends to improve generalization to unseen data, which is crucial for open-vocabulary segmentation. Fully supervised models, in contrast, are often limited by the specific categories they've been annotated with.

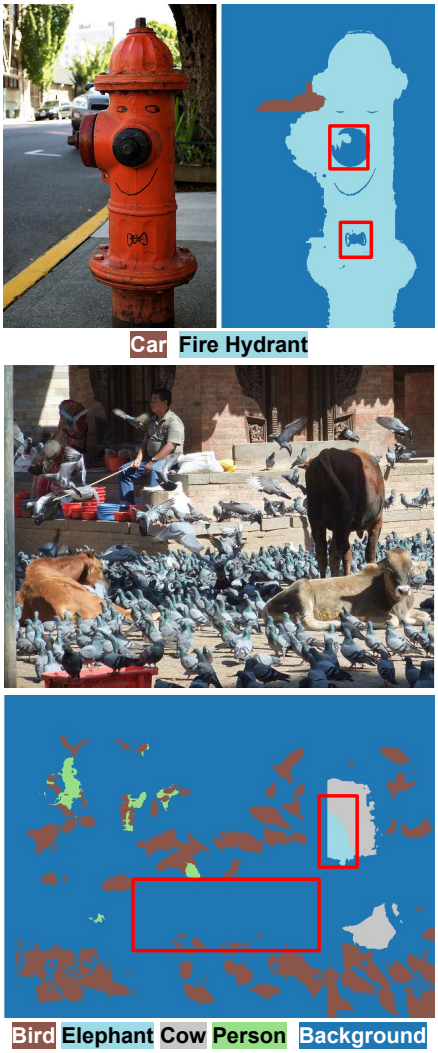
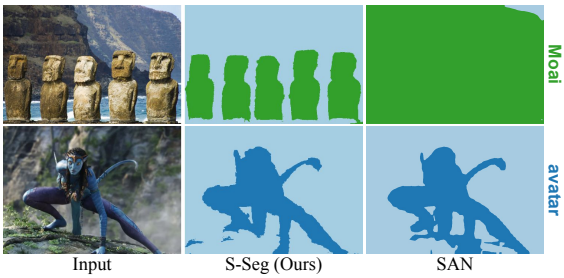


Figure 2. **Limitations.** Our model struggles with high color contrast (top) and complex scenes with overlapping objects (bottom), highlighting some key failure modes.

To illustrate, we compared S-Seg to a strong fully supervised model (SAN [14], trained on COCO) on 20 uncommon ImageNet-S classes[2]. SAN achieved 56.5 mIoU, whereas our model obtained 62.7 mIoU. Qualitative results on web images below further show SAN's limitations on new visual domains despite greater annotation use.



[2]tench, peacock, ostrich, jellyfish, albatross, magpie, indian cobra, hummingbird, snail, flamingo, tarantula, platypus, tiger shark, american lobster, quail, kite, conch, bullfrog, axolotl, koala

Figure 3. **Additional qualitative results of S-Seg in higher resolution (object-centric images).** Our method demonstrates robustness in dealing with challenging scenarios, such as objects with unconventional shapes and poses (row 1), images with unusual color and tone (row 2), objects of the same class but with differing colors (row 3), objects with the similar color but of different classes (row 4), concealed objects (row 5), and various other difficult situations.

Figure 4. **Additional qualitative results of S-Seg in higher resolution (context-rich images).** Although context-rich images pose challenges in segmentation due to the presence of an increased number of small and cluttered objects, our method can still accurately segment the objects with precision.

Figure 5. **Additional qualitative comparison with existing methods.** CLIP [16] is primarily designed for classification and does not perform well in segmentation. MaskCLIP [19] adapts CLIP for segmentation, although it produces noisy predictions and cannot handle background classes. GroupViT [18] is a strong competitor, but it could struggle in challenging scenarios.

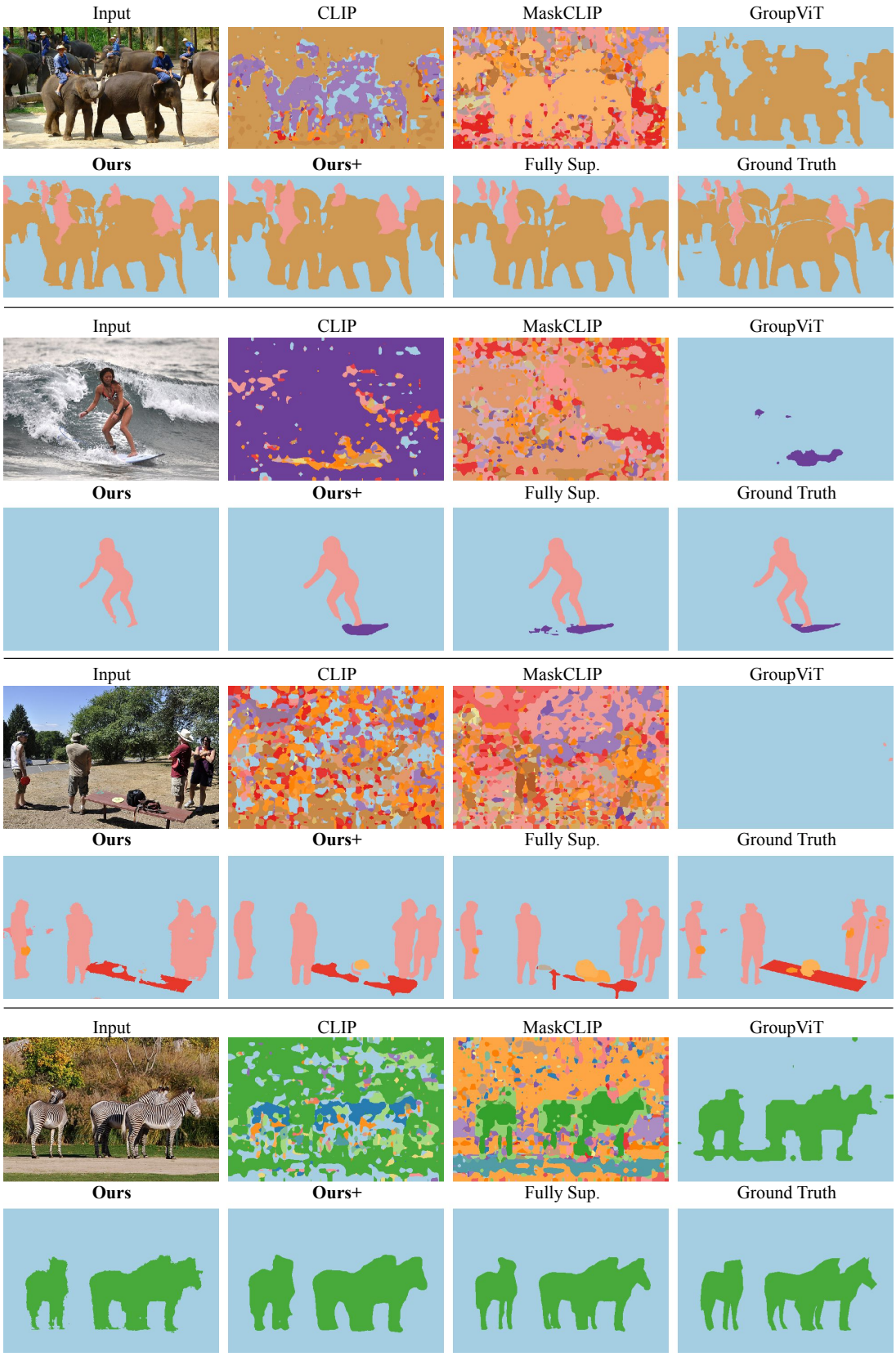


Figure 6. **Additional qualitative comparison with existing methods (continued).**

References

[1] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 1

[2] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018. 2

[3] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. 2021. 1, 2

[4] Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. Electra: Pre-training text encoders as discriminators rather than generators. *ICLR*, 2020. 1

[5] Jian Ding, Nan Xue, Gui-Song Xia, and Dengxin Dai. Decoupling zero-shot semantic segmentation. In *CVPR*, 2022. 3

[6] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 2009. 1, 2

[7] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *ECCV*, 2022. 3

[8] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017. 1

[9] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022. 1

[10] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. In *ICLR*, 2022. 3

[11] 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 *ECCV*, 2014. 1, 2

[12] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *ICLR*, 2016. 1

[13] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICCV*, 2019. 1

[14] Fangyun Wei Han Hu Xiang Bai Mengde Xu, Zheng Zhang. Side adapter network for open-vocabulary semantic segmentation. 2023. 4

[15] 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 *CVPR*, 2014. 1

[16] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. 2021. 1, 2, 7

[17] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *ECCV*, 2018. 1

[18] 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. 1, 2, 3, 7

[19] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *ECCV*, 2022. 2, 3, 7