Supplementary: Exploring Open-Vocabulary Semantic Segmentation from CLIP Vision Encoder Distillation Only

This Supplemental Material contains the following contents:

Contents

1. Hyperparameters

The training hyperparamers are displayed in Table 1.

config	value
optimizer	AdamW
base learning rate	1e-4
weight decay	0.05
optimizer momentum	β_1, β_2=0.9,0.95
batch size	4096
total epochs	80
warmup epochs	20
masked decoder layer	8
first-stage grouping layer	2
second-stage grouping layers	2
first-stage group tokens	32
second-stage group tokens	8

Table 1. **Hyperparameter setting**.

2. More visualization examples

We sample more examples from ImageNet 1k [2] and Conceptual Caption [1] val dataset and demonstrate them in the Fig. 1 and Fig. 2

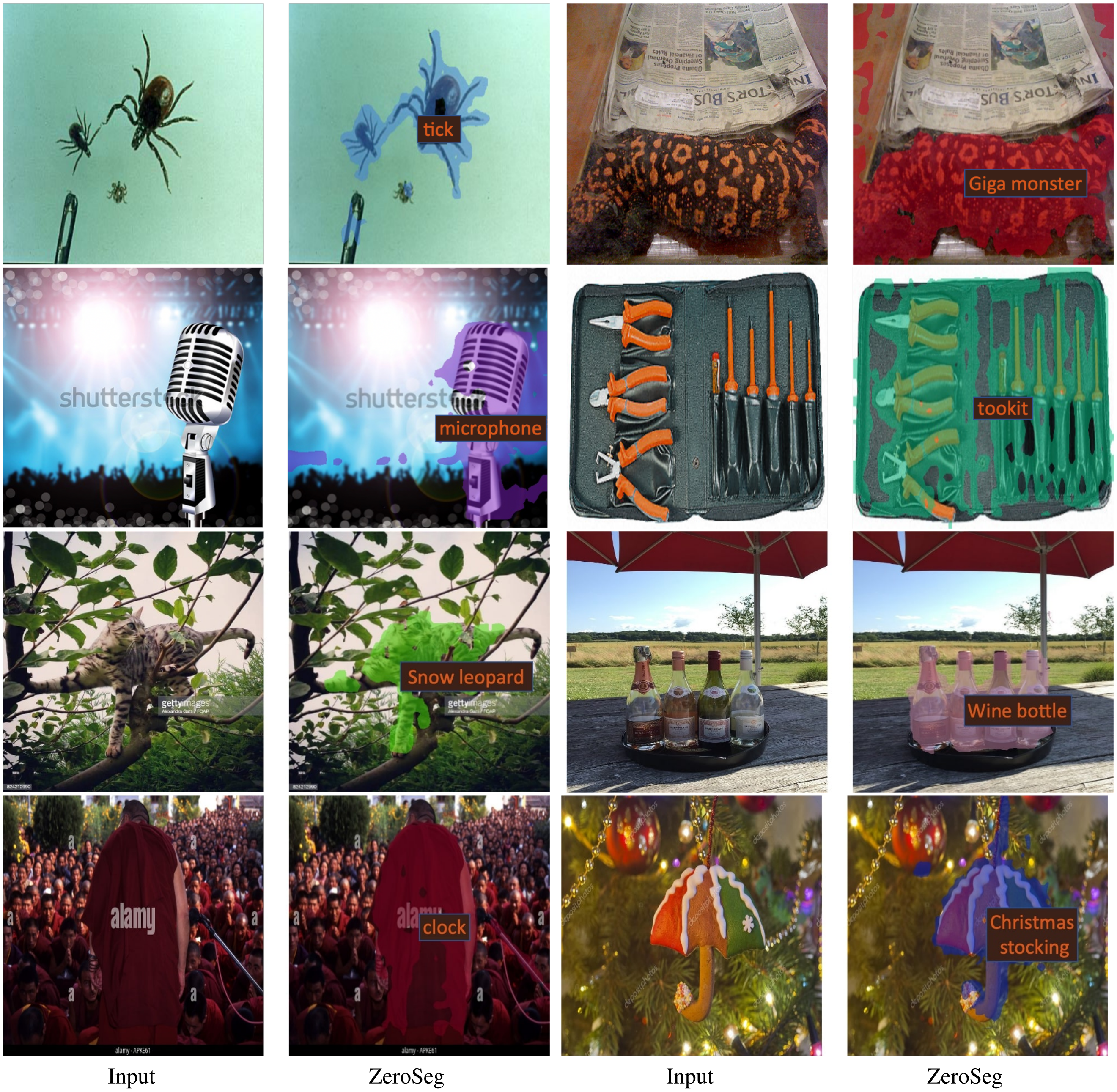

Figure 1. **More sampled example from ImageNet and Conceptual Caption val set**

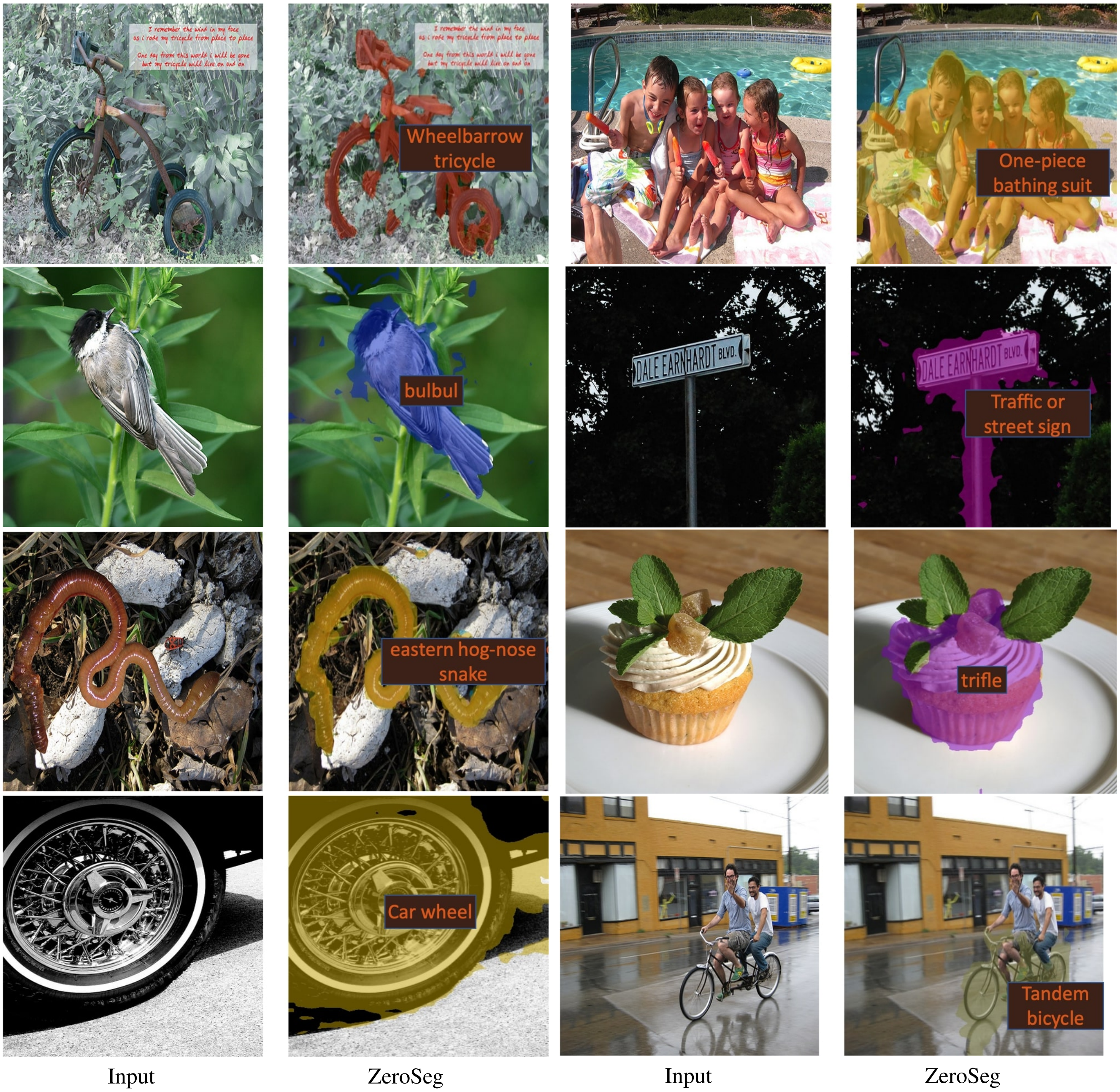

Figure 2. **More sampled example from ImageNet and Conceptual Caption val set**

References

[1] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568, 2021. 1

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1