MOS: Modeling Object-Scene Associations in Generalized Category Discovery

Supplementary Material

1. DINOv2 Result

The CUB results are shown, as follows. Our method also shows significant improvement on DINOv2.

Method	Backbone	All	Base	Novel
SimGCD [3]	DINOv2	71.5	78.1	68.3
MOS	DINOv2	81.1	82.1	80.6

2. Experiments on Ambiguity Challenge

To assess the impact of scene information in GCD, we design an observational experiment. We first categorize all unlabeled data from the CUB dataset into four distinct subsets: Base Class with Base Scene, Novel Class with Base Scene, Base Class with Novel Scene, and Novel Class with Novel Scene. The division between novel and base classes follows the settings of the SSB benchmark. For the definition of scenes, categories appearing more than fifteen times in the label set are designated as base scenes, while all remaining categories are considered novel scenes. This classification approach is employed because it is difficult to identify completely novel classes within the constraints of the SSB benchmark. We apply the trained model to each subset, calculating the accuracy for each subset individually. Notably, the categories of Novel Class with Base Scene and Base Class with Novel Scene represent the two types of ambiguity.

3. Failure Case

When handling extremely low-resolution images, such as those in the CIFAR-100 [1] dataset with a resolution of 32x32, two significant challenges arise. 1) saliency segmentation models often struggle to extract accurate object information due to the inherent limitations posed by low resolution. 2) the object features within these images are inherently blurred, making it difficult to extract meaningful information solely from the objects. Fig. 1 illustrates examples of segmentation results from CIFAR-100, clearly demonstrating the challenges associated with processing such low-resolution images.

4. Filling Methods

In our framework, the decoupled object images exhibit differences from natural images, necessitating the use of various filling methods. We explore several common techniques, including zero-padding, mean filling, and mask prompt, with the latter inspired by the OVSeg[2]. Our empirical results indicate that image mean filling delivered the most effective performance. These findings are detailed in Tab. 1.

Figure 1. **Segmentation Visualization on CIFAR-100.** It displays the segmentation results of IS-Net. Saliency segmentation models struggle to segment extremely low-resolution images effectively, and the segmented object images often lack useful information.

Table 1. **Filling Methods Comparison.**

Method	All	Base	Novel
SimGCD	61.5	65.7	59.4
Mean filling	63.1	64.7	62.2
Mask prompt	61.5	66.7	58.9
Zero-padding	62.1	67.8	59.2

References

[1] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 1

[2] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7061–7070, 2023. 1

[3] Bingchen Zhao Xin Wen and Xiaojuan Qi. Parametric classification for generalized category discovery: A baseline study. In *ICCV*, 2023. 1