Adaptive Part Learning for Fine-Grained Generalized Category Discovery: A Plug-and-Play Enhancement

Supplementary Material

In this section, we provide comprehensive information including dataset and implementation details, further experiment results, and discussions of limitations and impacts. The structure is as follows:

- Section 1 - Dataset Details
- Section 2 - Implementation Details
- Section 3 - Hyperparameters
- Section 4 - Results on DINOv2
- Section 5 - Supplement Related Works
- Section 6 - Qualitative Analysis
- Section 7 - Part Discovery Visualization
- Section 8 - Limitations and Broader Impacts

1. Dataset Details

Dataset	#Class(L/U)	#Num(L/U)
CIFAR-10 [9]	5/10	12.5K/37.5K
CIFAR-100 [9]	80/100	20.0K/30.0K
ImageNet-100 [2]	50/100	31.9K/95.3K
CUB [14]	100/200	1.5K/4.5K
Stanford-Cars [8]	98/196	2.0K/6.1K
FGVC-Aircraft [10]	50/100	1.7K/5.0K
Herbarium 19 [12]	341/683	8.9K/25.4K

Table 1. **Dataset information** for labeled (L) and Unlabelled (U) splits.

This section provides comprehensive details on all the datasets utilized in our experiments, including the number of classes and the number of samples for both labeled and unlabeled splits. Table 1 presents the specific statistics.

2. Implementation Details

To comprehensively evaluate our proposed method, we take SimGCD [16], SPTNet [15] and CMS [1]as baselines for comparison. For all three methods, we keep their original loss functions and hyperparameters unchanged but integrate our part discovery and loss. Additionally, for SPTNet, we also adopt its spatial visual prompting training scheme and freeze the first 11 layers of the backbone for fine-tuning. All experiments are conducted using a single RTX-3090 GPU, with a fixed batch size of 128 and training epochs set to 200. We use the ViT-base DINO model, where the feature dimension C is 768 and the number of heads M is 12. The number of part queries T is set to 12. Following the baseline [1, 15, 16], we assume $|C_u|$ is known in advance.

3. Hyperparameters

		All	Known	Novel
1	$T = 6$	58.5	74.3	50.9
2	$T = 9$	59.6	76.4	51.1
3	$T = 12$	60.1	77.6	51.2
4	$\epsilon = 0.8$	60.1	77.6	51.2
5	$\epsilon = 0.6$	58.9	74.2	51.4
6	$\epsilon = 0.4$	58.5	74.3	50.9

Table 2. Ablation study of hyperparameters on the Stanford Cars. T denotes the number of part queries, and ϵ represents the threshold for filtering from the attention map.

Table 2 shows the performance of our model under different hyperparameters T and ϵ. The results indicate that our method demonstrates good robustness across different ϵ values, proving that meaningful part information can indeed enhance both discrimination and generalization, and we also observe that as T increases, performance gradually improves, which indicates that having more part queries enables the model to focus on more finer-grained details in the images, which is beneficial for the final classification. In all our experiments, we set the number of part queries T to 12, ϵ to 0.8, and all temperature coefficients to 1. All other hyperparameters are inherited from the methods we integrated.

4. Results on DINOv2

We conduct experiments using our method with the DINOv2 [11] backbone and the SimGCD baseline. The results are shown in the Table 3. Combined with the main table in the text, this demonstrates that our APL consistently achieves improvements across various fine-grained datasets using different backbones.

5. Supplement Related Work

This section includes additional related works on Novel Category Discovery.

Novel Category Discovery (NCD) aims to generalize the classification of an unlabeled set by learning from both labeled and unlabeled sets, where the label space of the unlabeled set, comprising novel categories, entirely differs from that of the labeled one, known categories. The groundbreaking works [5, 6] typically involve two stages. Initially, a

Methods	CUB			Stanford Cars			FGVC-Aircraft			AVG
	All	Known	Novel	All	Known	Novel	All	Known	Novel	
GCD [13]	71.9	71.2	72.3	65.7	67.8	64.7	55.4	47.9	59.2	64.3
SimGCD [16]	71.5	78.1	68.3	71.5	81.9	66.6	63.9	69.9	60.9	69.0
Ours (SimGCD)	75.1	79.1	73.2	73.4	87.6	66.7	68.8	74.1	66.6	72.4
Δ	+3.6	+1.0	+4.9	+1.9	+5.7	+0.1	+4.9	+4.2	+5.2	+3.4

Table 3. Evaluation on the Semantic Shift Benchmark (SSB) using DINOv2 as the backbone.

feature extractor is obtained by training a classifier on the labeled set. Subsequently, the feature extractor is employed to extract features from the unlabeled data for clustering or generating pseudo-labels for training a new classifier for novel classes. Later, NCL [17] introduces contrastive loss on labeled and unlabeled data to enhance the generalization of learned representations, while OpenMix [18] mitigates the interference from incorrect pseudo-labels of unlabeled samples by mixing these pseudo-labels with the correct labels of labeled samples. UNO [3] follows the path of optimizing pseudo-labels, but it generates pseudo-labels through optimal transportation. Unlike extracting pseudo-labels solely for unlabeled data, [7] replays the pseudo latent from the labeled data to avoid overfitting to unlabeled one. Different from the above approaches, Cr-KD [4] evaluates inter-class relationships between unlabeled and labeled data, ensuring that novel class samples maintain a similar distribution as known class counterparts.

6. Qualitative Analysis

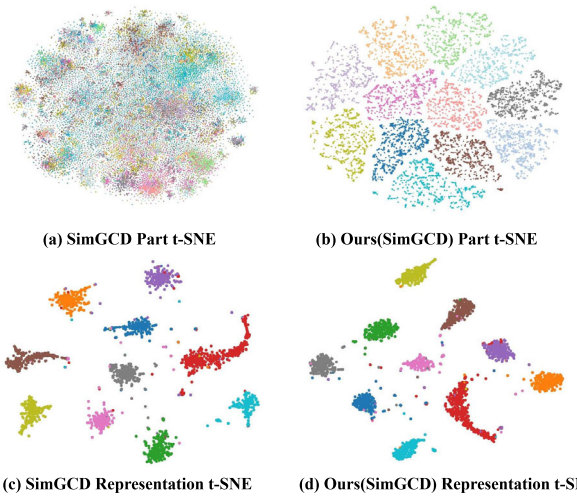

Figure 1. t-SNE visualization of part features on CUB-200 and image features on randomly sampled 10 classes on CIFAR-100. Each color represents a specific part (category) in left (right) two images.

In Figure 1, we present the t-SNE visualization of our method. Figure 1(a) shows the result obtained by filtering the local features from SimGCD using the DINO prior across the entire dataset, while Figure 1(b) demonstrates the t-SNE visualization of our resulted part features. Their comparison demonstrates that our specially designed part queries and the learning objectives for part discovery significantly enhance the differentiation of parts with different semantics, resulting in clear part boundaries in the feature space. Similarly, the comparison between Figure 1(c) and Figure 1(d) illustrates our tighter clusters, particularly evident in classes identified by brown, purple, and orange markings.

7. Part Discovery Visualization

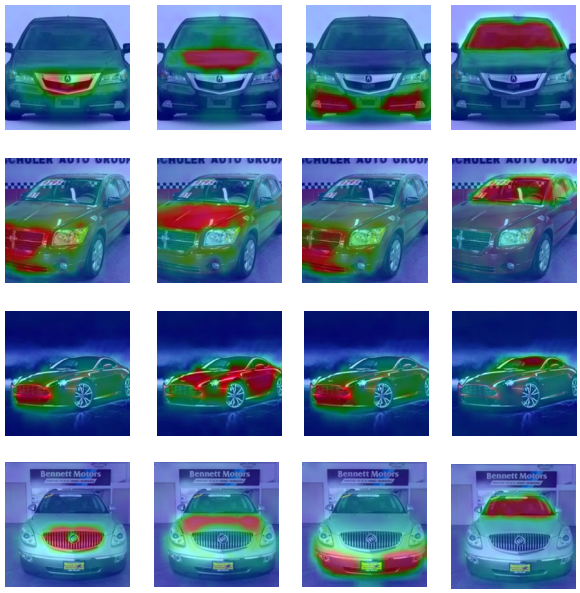

Figure 2. Some visualization of our part discovery. Each column represents the part regions attended to by the same part query.

Figure 2 presents some visualization results of the attention maps of part queries, demonstrating that our method successfully leverages the prior knowledge from DINO priors. It shows that our approach can spontaneously conduct the model to converge towards semantically meaningful directions without using part segmentation mask supervision, while also exhibiting strict correspondence properties. For additional examples, please refer to Figure 3.

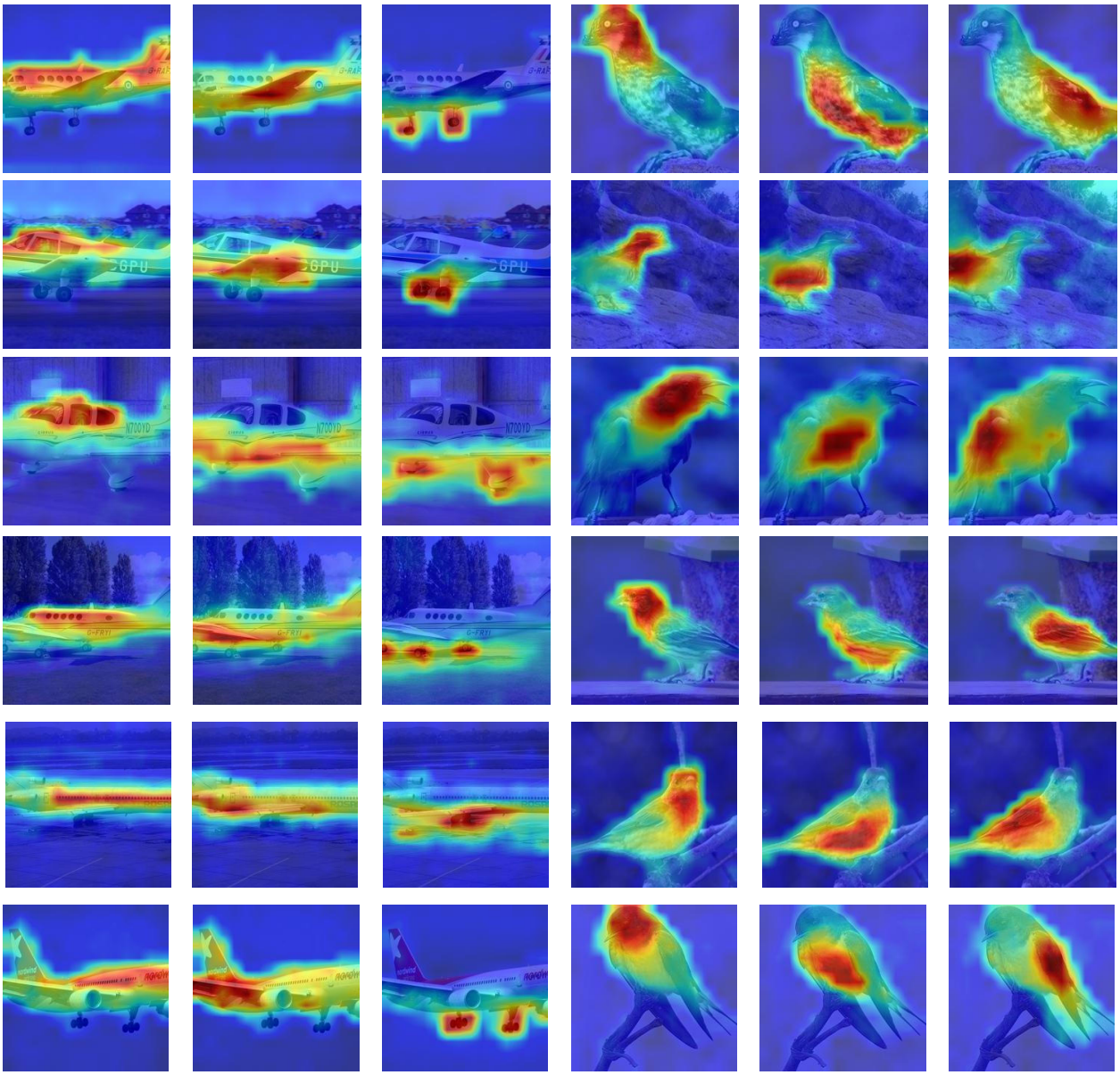

Figure 3. More visualization of our part discovery.

8. Limitations and Broader Impacts

One potential limitation of our method is that its benefits may be limited in very simplistic scenarios. This is because in such scenarios, distinguishing between different categories does not necessitate detailed part information. Additionally, in noisy environments, our method's efficacy may suffer due to complex noise, affecting part discovery. Lack of accurate part ground truth annotation can also impede performance. Furthermore, by introducing part discovery into generalized category discovery, we can achieve more nuanced and detailed image understanding, which can improve various applications such as autonomous driving, medical imaging, and surveillance, leading to safer and more effective solutions. However, if the datasets used for training contain biases, the model may inadvertently learn and reinforce these biases, leading to unfair or discriminatory outcomes.

Ethical Statements. Our research utilizes the pre-trained DINO model and it is essential to recognize that DINO may carry some biases from its training data. Removing all biases from a model is inherently difficult, and we hope users consider these possible biases when using our model.

References

[1] Sua Choi, Dahyun Kang, and Minsu Cho. Contrastive mean-shift learning for generalized category discovery. *arXiv preprint arXiv:2404.09451*, 2024. 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

[3] Enrico Fini, Enver Sangineto, Stéphane Lathuilière, Zhun Zhong, Moin Nabi, and Elisa Ricci. A unified objective for novel class discovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9284–9292, 2021. 2

[4] Peiyan Gu, Chuyu Zhang, Ruijie Xu, and Xuming He. Class-relation knowledge distillation for novel class discovery. *lamp*, 12(15.0):17–5, 2023. 2

[5] Kai Han, Andrea Vedaldi, and Andrew Zisserman. Learning to discover novel visual categories via deep transfer clustering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8401–8409, 2019. 1

[6] Kai Han, Sylvestre-Alvise Rebuffi, Sebastien Ehrhardt, Andrea Vedaldi, and Andrew Zisserman. Automatically discovering and learning new visual categories with ranking statistics. *arXiv preprint arXiv:2002.05714*, 2020. 1

[7] KJ Joseph, Sujoy Paul, Gaurav Aggarwal, Soma Biswas, Piyush Rai, Kai Han, and Vineeth N Balasubramanian. Novel class discovery without forgetting. In *European Conference on Computer Vision*, pages 570–586. Springer, 2022. 2

[8] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013. 1

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

[10] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 1

[11] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 1

[12] Kiat Chuan Tan, Yulong Liu, Barbara Ambrose, Melissa Tulig, and Serge Belongie. The herbarium challenge 2019 dataset. *arXiv preprint arXiv:1906.05372*, 2019. 1

[13] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Generalized category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7492–7501, 2022. 2

[14] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 1

[15] Hongjun Wang, Sagar Vaze, and Kai Han. Sptnet: An efficient alternative framework for generalized category discovery with spatial prompt tuning. *arXiv preprint arXiv:2403.13684*, 2024. 1

[16] Xin Wen, Bingchen Zhao, and Xiaojuan Qi. Parametric classification for generalized category discovery: A baseline study. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16590–16600, 2023. 1, 2

[17] Zhun Zhong, Enrico Fini, Subhankar Roy, Zhiming Luo, Elisa Ricci, and Nicu Sebe. Neighborhood contrastive learning for novel class discovery. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10867–10875, 2021. 2

[18] Zhun Zhong, Linchao Zhu, Zhiming Luo, Shaozi Li, Yi Yang, and Nicu Sebe. Openmix: Reviving known knowledge for discovering novel visual categories in an open world. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9462–9470, 2021. 2