Parametric Classification for Generalized Category Discovery: A Baseline Study
Supplementary Material

Xin Wen[1*] Bingchen Zhao[2*] Xiaojuan Qi[1]
[1]The University of Hong Kong [2]University of Edinburgh
{wenxin,xjqi}@eee.hku.hk bingchen.zhao@ed.ac.uk

Contents

A. Implementation Details

A.1. Experiment Setting Details

The split of labelled ('Old') and unlabelled ('New') categories follows GCD [16]. That is, 50% of all classes are sampled as 'Old' classes ($\mathcal{Y}_l$), and the rest are regarded as 'New' classes ($\mathcal{Y}_u \setminus \mathcal{Y}_l$). The exception is CIFAR100, for which 80% classes are sampled as 'Old', following the novel category discovery (NCD) literature. Regarding the sampling process, for generic object recognition datasets, the labelled classes are selected by their class index (the first $|\mathcal{Y}_l|$ ones). For the Semantic Shift Benchmark, data splits provided in [17] are adopted. For Herbarium 19 [14], the labelled classes are sampled randomly. Additionally, for ImageNet-1K [4] which is not used in [16], we follow its fashion to select the first 500 classes sorted by class id as the labelled classes. Then for all datasets, following [16], 50% of the images from the labelled classes are randomly sampled to form the labelled dataset $\mathcal{D}^l$, and all remaining images are regarded as the unlabelled dataset $\mathcal{D}^u$. All experiments are done with a batch size of 128 on a single GPU, except for ImageNet-1K, on which we train with eight GPUs, scale the learning rate with the linear scaling rule, and keep the per-GPU batch size unchanged. The inference time on ImageNet-1K is still evaluated with one GPU.

A.2. Re-implementing Previous Works

Results of GCD [16] are taken from the original paper (if available), and otherwise re-implemented with the official codebase. One exception is ImageNet-1K [4], which was not evaluated by the authors. Since naively adopting their official codebase to ImageNet-1K fails as the semi-supervised k-means procedure requires too much GPU memory and cannot be done with available hardware, we drop the k-mean++ initialisation [1] which takes the most memory, and re-implement the method with faiss [8] for speed up (otherwise the evaluation takes more than one day). The results are in the main paper, compared to our proposed strong baseline SimGCD, GCD requires significantly more time to run and more engineering efforts, and yet achieves a lower performance than SimGCD, which demonstrates the effectiveness of our proposed method. Results of UNO+ [6] and RS+ [7], which are adaptations of the original works to the GCD task, are directly taken from the GCD [16] paper. Also note that unlike UNO [6], our method does not adopt the over-clustering trick for simplicity. Results of ORCA [2] are re-implemented with the official codebase. We align the details in dataset split and backbone (ViT-B/16 [5] pre-trained with DINO [3]) with GCD [16] for a fair comparison.

A.3. Error Analysis Details

We briefly clarify the details of obtaining the four kinds of prediction errors in the main paper: we first rank the category indexes in consecutive order, such that by index, all 'Old' classes are followed by all 'New' classes. We then compute the full confusion matrix, with each element summarising how many times images of one specific class (row index) are predicted as one class (column index). All elements are divided by the number of testing samples to account for the percentage. We then reduce the diagonal terms to zero (representing correct predictions), and thus all remaining elements represent different kinds of prediction errors (*i.e.*, absolute contribution to the errors of 'All' ACC). Finally, we slice the confusion matrix into four sub-matrices

Figure 1. **Complete error analysis results of SimGCD on five representative datasets.** With appropriate entropy regularisation, the bias between 'Old'/'New' classes (see "False New" and "False Old" errors) are generally effectively alleviated, except in the long-tailed Herbarium 19 that the effect varies. Also notably, "True New" errors are consistently penalised to a considerable extent, confirming entropy regularisation's ability in helping recognise and distinguish between novel categories.

at the boundaries between the 'Old' and 'New' classes, and add all elements in each sub-matrix together, thus obtaining the final error matrix standing for the four kinds of prediction errors. Such a way of error classification helps distinguish the prediction bias between and within seen and novel categories, and thus facilitates the design of new solutions. Note that the diagonal elements, *e.g.*, 'True Old' predictions, do not stand for correct predictions, but for cases that incorrectly predicting samples of one specific 'Old' class to another wrong 'Old' class.

B. Extended Experiments And Discussions

B.1. Main Results

We present the full results of SimGCD in the main paper with error bars in Tab. 1. The results are obtained from three independent runs and thus avoid randomness.

Dataset	All	Old	New
CIFAR10 [10]	97.1±0.0	95.1±0.1	98.1±0.1
CIFAR100 [10]	80.1±0.9	81.2±0.4	77.8±2.0
ImageNet-100 [15]	83.0±1.2	93.1±0.2	77.9±1.9
ImageNet-1K [4]	57.1±0.1	77.3±0.1	46.9±0.2
CUB [18]	60.3±0.1	65.6±0.9	57.7±0.4
Stanford Cars [9]	53.8±2.2	71.9±1.7	45.0±2.4
FGVC-Aircraft [11]	54.2±1.9	59.1±1.2	51.8±2.3
Herbarium 19 [14]	44.0±0.4	58.0±0.4	36.4±0.8

Table 1. Complete results of SimGCD in three independent runs.

B.2. Unknown Category Number

In the main text, we showed that the performance of SimGCD is robust to a wide range of estimated unknown category numbers. In this section, we report the results with the number of categories estimated using an off-the-shelf method [16] (Tab. 2) or with a roughly estimated relatively big number (two times of the ground-truth K), and compare with the baseline method GCD [16].

The results on CIFAR100 [10], ImageNet-100 [4], CUB [18], and Stanford Cars [9] are available in Tabs. 3 and 4. Our method shows consistent improvements on four representative datasets when K is unknown, no matter with the category number estimated with a specialised algorithm (w/ Est.), or simply with a loose estimation that is two times the ground truth (w/ $2K$, other values are also applicable since our method is robust to a wide range of estimations). This property could ease the deployment of parametric classifiers for GCD in real-world scenarios.

	CIFAR100	ImageNet-100	CUB	SCars	Herb19
GT K	100	100	200	196	683
Est. K	100	109	231	230	520

Table 2. Number of categories K estimated using [16].

		CIFAR100			ImageNet-100		
Methods	Known K	All	Old	New	All	Old	New
GCD [16]	✓	73.0	76.2	66.5	74.1	89.8	66.3
SimGCD	✓	**80.1**	**81.2**	**77.8**	**83.0**	**93.1**	**77.9**
GCD [16]	✗(w/ Est.)	73.0	76.2	66.5	72.7	91.8	63.8
SimGCD	✗(w/ Est.)	**80.1**	**81.2**	**77.8**	**81.7**	91.2	**76.8**
SimGCD	✗(w/ $2K$)	77.7	79.5	74.0	80.9	**93.4**	74.8

Table 3. Results on generic image recognition datasets.

		CUB			Stanford Cars		
Methods	Known K	All	Old	New	All	Old	New
GCD [16]	✓	51.3	56.6	48.7	39.0	57.6	29.9
SimGCD	✓	**60.3**	**65.6**	**57.7**	**53.8**	**71.9**	**45.0**
GCD [16]	✗(w/ Est.)	47.1	55.1	44.8	35.0	56.0	24.8
SimGCD	✗(w/ Est.)	61.5	66.4	59.1	**49.1**	**65.1**	**41.3**
SimGCD	✗(w/ $2K$)	**63.6**	**68.9**	**61.1**	48.2	64.6	40.2

Table 4. Results on the Semantic Shift Benchmark [17].

B.3. Extended Analyses

In supplementary to the main paper, we present a more complete version of the analytical experiments.

In Fig. 1, we show the error analysis results of SimGCD over five representative datasets that cover coarse-grained, fine-grained, and long-tailed classification tasks. Overall, it shows that the entropy regulariser mainly helps in overcoming two types of errors: the error of misclassification between 'Old'/'New' categories, and the error of misclassification within 'New' categories. One exception is the long-

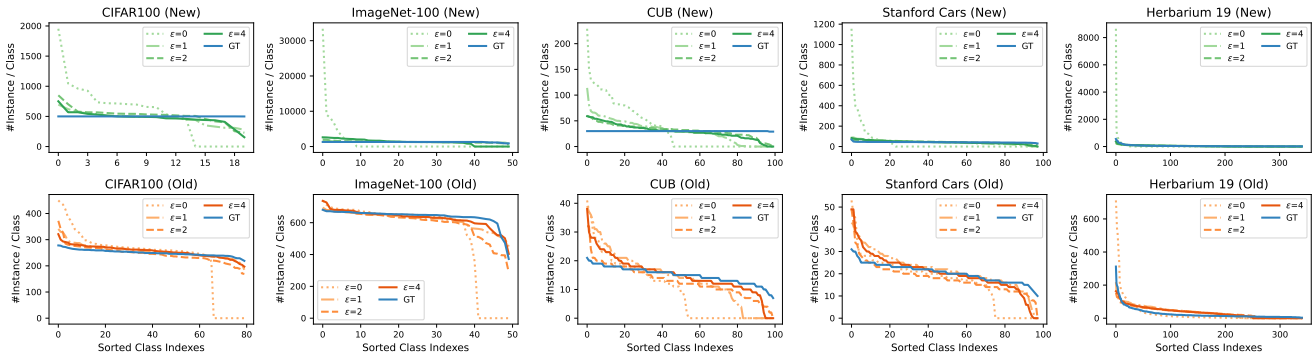

Figure 2. **Complete per-class prediction distribution results of SimGCD on five representative datasets.** Proper entropy regularisation helps overcome the prediction bias in both 'Old' classes and 'New' classes, and fits the ground-truth distribution. The conclusion is consistent across generic classification datasets, fine-grained classification datasets, and naturally long-tailed datasets.

tailed Herbarium 19 dataset, in which the models' "False Old" errors also increased, and our intuition is that the long-tailed distribution adds to the difficulty in discriminating between 'Old' and 'New' categories. Still, the gain in distinguishing between novel categories is consistent, and we provide a further analysis via per-class prediction distributions in the next paragraph.

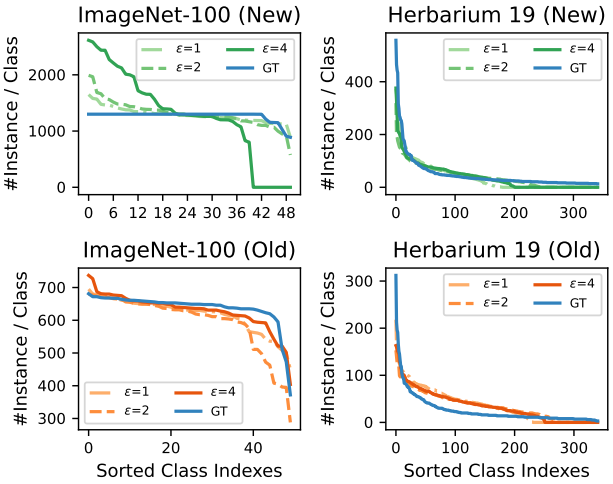

Figure 3. **A closer look at the per-class distributions.** Notably, although the entropy regularisation term is formulated to approach uniform distribution, it could make the models' predictions more biased on the class-balanced ImageNet-100 dataset when the regularisation is too strong. Interestingly, it also could help fit the distribution of the long-tailed Herbarium 19 dataset.

In Fig. 2, we show the complete per-class prediction results of SimGCD to further analyse the entropy regulariser's effect in overcoming the classification errors within 'Old' and 'New' classes, and it consistently verifies the help in alleviating the prediction bias within 'Old' and 'New' classes, and better fitting the ground-truth class distribution. In Fig. 3, we present a closer look at ImageNet-100 and Herbarium 19. The entropy regularisation term is formulated to make the model's predictions closer to the uniform distribution. But interestingly, we empirically found that it could make the models' predictions more biased on the class-balanced ImageNet-100 dataset when the regularisation is too strong. And when the dataset itself is long-tailed (Herbarium 19), it also could help fit the ground-truth distribution. We also note that the self-labelling strategy adopted by UNO [6] forces the predictions in a batch to be strictly uniform, which may account for its inferior performance.

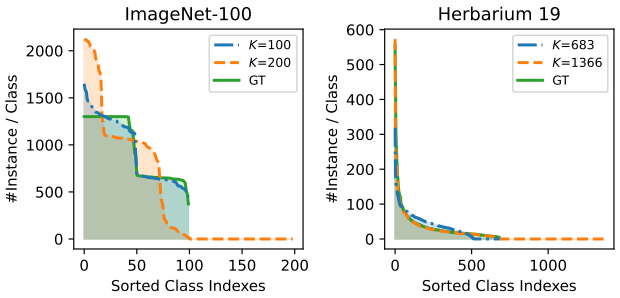

Figure 4. **Per-class prediction distributions using different category numbers on ImageNet-100 and Herbarium 19.** Our method effectively identifies the criterion for 'New' classes, thus keeping the number of active prototypes close to the ground-truth class number. Notably, a loose category number greater than the ground truth may harm fitting the class-balanced ImageNet-100 dataset, but could help fit the distribution of the long-tailed Herbarium 19 dataset.

In Fig. 4, we also show the per-class prediction distributions using different category numbers. The results on the class-balanced ImageNet-100 are consistent with the results on CIFAR100 and CUB in the main paper, using a loose category number greater than the ground truth may harm fitting the ground-truth class distribution, yet the model still manages to find the ground truth category number. Interestingly, we also find that for the long-tailed Herbarium 19 dataset, using a greater category number could in fact help fit the ground-truth distribution.

Method	Logit Adjust	CIFAR100			ImageNet-100			CUB			Stanford Cars			Herbarium 19		
		All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New
ORCA [2]	✓	69.0	77.4	52.0	73.5	92.6	63.9	35.3	45.6	30.2	23.5	50.1	10.7	20.9	30.9	15.5
DebiasPL [20]	✓	60.9	69.8	43.1	43.5	59.1	35.6	38.1	44.2	35.0	31.1	49.6	22.1	30.1	39.1	25.3
UNO+ [6]	✗	69.5	80.6	47.2	70.3	**95.0**	57.9	35.1	49.0	28.1	35.5	70.5	18.6	28.3	53.7	14.7
GCD [16]	✗	73.0	76.2	66.5	74.1	89.8	66.3	51.3	56.6	48.7	39.0	57.6	29.9	35.4	51.0	27.0
SimGCD	✗	**80.1**	**81.2**	**77.8**	**83.0**	93.1	**77.9**	**60.3**	**65.6**	**57.7**	**53.8**	**71.9**	**45.0**	**44.0**	**58.0**	**36.4**

Table 5. Comparison to imbalanced recognition-inspired methods.

B.4. Relationship to Imbalanced Recognition

Our work also shares motivation with literature in long-tailed/imbalanced recognition [12, 19, 13], in which resolving the imbalance in models' prediction is also an important issue. Technically, they commonly depend on a prior class distribution to adjust classifiers' output, which is not accessible in GCD since labels for novel classes are unknown. One could also estimate this distribution online from predictions, which is inaccurate due to its open-world nature. We note one baseline (ORCA [2]) compared in the paper also shares key intuition with these works (adaptive margin). We also reimplement one close work that operates on imbalanced semi-supervised learning, *i.e.*, DebiasPL [20], aligning representation learning with GCD, and show a comparison in Tab. 5. DebiasPL surpasses UNO+ on fine-grained classification in novel classes and verifies it could overcome the prediction imbalance to some extent. It also outperforms ORCA but still lags behind GCD and ours. We hypothesise manually altering logits may not be suitable for open-world settings. Instead, a more natural and general solution could be to regularise prediction statistics and let the model adjust via optimisation.

References

[1] David Arthur and Sergei Vassilvitskii. k-means++: the advantages of careful seeding. In *ACM-SIAM Symposium on Discrete Algorithms*, 2007. 1

[2] Kaidi Cao, Maria Brbić, and Jure Leskovec. Open-world semi-supervised learning. In *ICLR*, 2022. 1, 4

[3] 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. 1

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 1, 2

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 1

[6] Enrico Fini, Enver Sangineto, Stéphane Lathuilière, Zhun Zhong, Moin Nabi, and Elisa Ricci. A unified objective for novel class discovery. In *ICCV*, 2021. 1, 3, 4

[7] Kai Han, Sylvestre-Alvise Rebuffi, Sebastien Ehrhardt, Andrea Vedaldi, and Andrew Zisserman. Automatically discovering and learning new visual categories with ranking statistics. In *ICLR*, 2020. 1

[8] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 2019. 1

[9] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCV Workshops*, 2013. 2

[10] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. *Technical Report*, 2009. 2

[11] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 2

[12] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *ICLR*, 2021. 4

[13] Jiawei Ren, Cunjun Yu, Xiao Ma, Haiyu Zhao, Shuai Yi, and Hongsheng Li. Balanced meta-softmax for long-tailed visual recognition. In *NeurIPS*, 2020. 4

[14] Kiat Chuan Tan, Yulong Liu, Barbara Ambrose, Melissa Tulig, and Serge Belongie. The herbarium challenge 2019 dataset. *arXiv preprint arXiv:1906.05372*, 2019. 1, 2

[15] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *ECCV*, 2020. 2

[16] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Generalized category discovery. In *CVPR*, 2022. 1, 2, 4

[17] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: A good closed-set classifier is all you need? In *ICLR*, 2022. 1, 2

[18] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. Caltech-UCSD Birds 200. *Computation & Neural Systems Technical Report*, 2010. 2

[19] Xudong Wang, Long Lian, Zhongqi Miao, Ziwei Liu, and Stella X Yu. Long-tailed recognition by routing diverse distribution-aware experts. In *ICLR*, 2021. 4

[20] Xudong Wang, Zhirong Wu, Long Lian, and Stella X Yu. Debiased learning from naturally imbalanced pseudo-labels. In *CVPR*, 2022. 4