Supplementary Material for Self-Cooperation Knowledge Distillation for Novel Class Discovery

Yuzheng Wang, Zhaoyu Chen, Dingkang Yang, and Yunquan Sun, and Lizhe Qi[†]

Shanghai Engineering Research Center of AI & Robotics, Academy for Engineering & Technology, Fudan University

In this supplementary material, we provide more details of our method, organized as follows:

- In Section 1, we further discuss related but different tasks of NCD, mainly in terms of training goals and training & testing processes, including OSR and GCD tasks.
- In Section 2, we show the hyperparameter experiment, corresponding to Section 4.1 of the main body.
- In Section 3, we compare the performance of various NCD methods when facing the cross-domain dataset to measure the effectiveness of information transfer between known and novel classes.
- In Section 4, we reassess the model's performance on three fine-grained datasets without the pre-trained model.
- In Section 5, we evaluate the trade-offs between the two parts of the proposed Self-Cooperation Knowledge Distillation losses and provide readers with a reference for setting parameters according to actual needs.
- In Section 6, we discuss the differences between our and other NCD methods.
- In Section 7, we provide insights about future works.
- In Section 8, we discuss the limitations of the proposed method.

1 Discussion of Related but Distinct Tasks

Open-Set Recognition (OSR) [9] aims to detect test-time images which do not belong to one of the classes in the labeled set, but does not require any further classification amongst these detected images. Therefore, the model only needs to maintain a high degree of confidence in the known class, and does not consider the model's scalability for the novel class.

Novel Class Discovery (NCD) [4,5] not only considers the review of known classes, but also considers the discovery of unknown classes to expand the scope of application of the model. More detailed settings and definitions are shown in the main body.

Generalized Category Discovery (GCD) [10] is similar to NCD. In the training process, GCD and NCD contains the same dependencies, *i.e.*, a labeled set of known classes and an unlabeled set of unknown classes. The difference is that GCD makes no assumptions about the test set. Test samples can come from the known or novel classes.

Table 1: Analysis of hyperparameter β and α. "Train-Novel" refers to the evaluation of the unlabeled training subset, using task-aware evaluation protocol. Other results are evaluations of the testing subset, using the task-agnostic evaluation protocol.

ID	Setting		CIFAR100-50			
			Known	Novel	All	Train-Novel
(1)		0	70.3	49.7	60.0	62.6
(2)		0.01	70.7	50.6	60.7	63.3
(3)	β	0.1	**78.1**	56.4	67.3	64.7
(4)		0.5	77.9	**62.6**	**70.3**	**68.2**
(5)		1	77.3	61.2	69.3	66.3
(6)		0.01	74.5	51.8	63.2	63.6
(7)		0.05	77.5	61.2	69.4	67.2
(8)	α	0.1	**77.9**	**62.6**	**70.3**	**68.2**
(9)		0.5	75.4	59.4	67.4	65.8
(10)		1	76.4	58.5	67.5	65.4

2 Hyper-parameter Experiment

In this section, we show the test results of the hyperparameter β and α. The results are shown in Table 1. We test on the CIFAR100-50 dataset split and report the results on the training and testing subsets. For (1)-(5), we fix the α as 0.1 and show the results with various β. Compared with the original version $\beta = 0$, the proposed SCKD technology can help the model improve the performance of known and novel classes. When the value of β is too high, the model may be overly encouraged to focus on another representation space and underestimate its original pure knowledge (the labeled samples for known classes & unlabeled samples for novel classes). We recommend choosing $\beta = 0.5$ as the default value for better results. For (6)-(10), we fix the β as 0.5. Different α affects pseudo-labels' smoothness, thereby achieving different degrees of mutual information interference between known and novel classes. By default, we choose $\alpha = 0.1$.

3 Experiments on Cross-domain Dataset

In the main body of the paper, we have evaluated the performance of the algorithm on disjoint known and novel classes. In order to verify the effectiveness of the method more broadly, we conduct experiments when the known and novel classes satisfy the cross-domain yet category related. We select Office-31 [8], a well-known domain adaptation evaluation dataset with 31 categories of office objects. It contains 4652 images from three domains: Amazon (A), DSLR (D), and Webcam (W). We experiment with all six combinations (source domain $\rightarrow$ target domain): A$\rightarrow$W, D$\rightarrow$W, W$\rightarrow$D, A$\rightarrow$D, D$\rightarrow$A, W$\rightarrow$A, and report the average accuracy based for each baseline. Following KCL [4], we use ResNet-18 as the backbone pre-trained with ImageNet. Each mini-batch is constructed by 32 labeled samples from source domain and 96 unlabeled samples from target

Table 2: Cross-domain NCD on the Office-31 dataset. Avg represents the average accuracy.

Method	A→W	D→W	W→D	A→D	D→A	W→A	Avg
KCL [4]	76.7	97.3	98.2	71.2	61.0	60.5	77.5
UNO [1]	85.6	97.3	98.6	74.6	66.4	65.7	81.4
IIC [7]	89.4	98.7	99.0	79.8	70.6	71.4	84.8
SCKD	**90.4**	**99.1**	**99.8**	**84.6**	**74.4**	**75.4**	**87.3**

domain. The results are shown in Table 2, our method achieves state-of-the-art results and outperforms the previous best baseline method by an average accuracy of 2.5% to 9.8%. By representing the mutual information between spaces and the cooperation learning paradigm, our method can make more full use of limited sample information, thereby achieving competitive results.

Table 3: Experiments on fine-grained datasets without the DINO pre-trained ViT model.

Methods	Stanford Cars	CUB	FGVC-Aircraft
RankStats+ [3]	17.7	20.2	30.1
NCL [12]	27.1	24.7	36.7
UNO [1]	25.2	24.5	37.8
rKD [2]	29.2	26.5	43.8
SCKD	**31.2**	**29.7**	**48.5**

4 Experiments without Pre-trained Model

For three fine-grained datasets (*i.e.*, Stanford Cars, CUB, and FGVC-Aircraft), we choose the pre-trained ViT-B-16 model, following GCD [10] and rKD [2]. In this section, we utilize ResNet-18 as the backbone from scratch. We first pretrain the model on known classes for 200 epochs and then train on known and novel classes for 200. The results are shown in Tabel 3 For the three datasets, our method improves by 2.0%, 3.2%, and 4.7%, respectively. The results demonstrate our method's effectiveness without the pre-trained models.

5 Trade-off between Self-Cooperation Knowledge Distillation Losses

As Eq.(7) of the main body, our overall Self-Cooperation Knowledge Distillation loss consists of two parts. The information on labeled samples assists in the discovery of novel classes, and the information of unlabeled samples assists

Table 4: Experimental results with different λ on CIFAR100-50 testing subset, using task-agnostic evaluation protocol.

λ	$\mathcal{L}_{k\to n}$	$\mathcal{L}_{n\to k}$	Known	Novel	All
0.1	0.1	0.9	78.4	60.6	69.5
0.3	0.3	0.7	78.0	61.2	69.6
0.5	0.5	0.5	77.9	62.6	70.3
0.7	0.7	0.3	76.9	62.4	69.7
0.9	0.9	0.1	74.6	64.6	69.6

in the review of known classes as: $\mathcal{L}_{SCKD} = \mathcal{L}_{k\to n} + \mathcal{L}_{n\to k}$. *i.e.*, $\mathcal{L}_{SCKD} = 2 \times [\lambda \cdot \mathcal{L}_{k\to n} + (1-\lambda) \cdot \mathcal{L}_{n\to k}], \lambda = 0.5$. In this section, we use λ as an adjustable parameter to balance the two. The results are shown in Table 4. We find that adjusting λ can affect the review of known classes and the discovery of novel classes to a certain extent. Specifically, when λ is small, the model is encouraged to review known classes from unlabeled sample information, thus obtaining higher performance on known classes. When λ increases, the model pays more attention to using labeled sample information to discover novel classes. Overall, adjusting λ has little impact on the overall accuracy. For convenience, we set $\lambda = 0.5$ and encourage changing the size of λ as needed.

6 Discussion with NCDLR [11] and rKD [2]

In NCDLR [11], the author considers the long-tail distribution problem of data between novel classes and introduces an equiangular prototype representation method to provide pseudo labels for novel classes. The problem exists among novel classes without considering the distribution of the number of all training samples in the population. Unlike their motivation, our work considers another crucial real-life issue in NCD tasks. We focus on the balance of information between known and novel classes. In fact, deep learning models are often expected to continue to expand their scope of application. In this process, the problem of the imbalanced number of samples from known and novel classes is inevitable. In rKD [2], the author encourages the model to maintain the same proportional relationship between the predictions of known classes and the discovery stage of novel classes. In this process, their method weakens the information transfer process between known and novel classes. At the same time, the method does not consider the distribution differences of different categories of data. When the distribution shifts, the performance will be significantly affected (refer to Table 4 of the main body). Our method is motivated by trying to compensate for a missing perspective in current tasks, namely the imbalanced number of samples from known and novel classes. Besides, our proposed method collaboratively reviews known classes and discovers novel classes by combining information from all labeled and unlabeled samples. The motivation and method are significantly different from existing NCD baselines.

7 Future Works

In this work, we consider a real application challenge in NCD tasks, *i.e.*, the imbalanced number of samples from known and novel classes. In fact, the scalable foundation model [6] also faces similar needs and challenges. When the foundation model is required to expand new application scenarios, learning about unknown objects is often accompanied by catastrophic forgetting of known knowledge. How to fully utilize the information of all samples to enhance the performance of such large foundation models with special needs remains to be solved. We leave this section as future work.

8 Limitations

Our method is able to cope with the challenge of the imbalanced number of samples from known and novel classes and achieves considerable progress. But it is still not perfect. An obvious limitation is that our method assumes that known and novel samples can be few but not completely absent. Like other NCD methods based on information transfer, in this extreme case, our method and other methods degenerate into a simple coupling of supervised learning and unsupervised clustering. For such extreme cases, additional information banks may be needed for information storage before deploying our proposed technology.

References

1. Fini, E., Sangineto, E., Lathuiliere, S., Zhong, Z., Nabi, M., Ricci, E.: A unified objective for novel class discovery. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9284–9292 (2021) 3
2. Gu, P., Zhang, C., Xu, R., He, X.: Class-relation knowledge distillation for novel class discovery. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2023) 3, 4
3. Han, K., Rebuffi, S.A., Ehrhardt, S., Vedaldi, A., Zisserman, A.: Automatically discovering and learning new visual categories with ranking statistics. arXiv preprint arXiv:2002.05714 (2020) 3
4. Hsu, Y.C., Lv, Z., Kira, Z.: Learning to cluster in order to transfer across domains and tasks. arXiv preprint arXiv:1711.10125 (2017) 1, 2, 3
5. Hsu, Y.C., Lv, Z., Schlosser, J., Odom, P., Kira, Z.: Multi-class classification without multi-class labels. arXiv preprint arXiv:1901.00544 (2019) 1
6. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023) 5
7. Li, W., Fan, Z., Huo, J., Gao, Y.: Modeling inter-class and intra-class constraints in novel class discovery. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3449–3458 (2023) 3
8. Saenko, K., Kulis, B., Fritz, M., Darrell, T.: Adapting visual category models to new domains. In: Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV 11. pp. 213–226. Springer (2010) 2

9. Scheirer, W.J., de Rezende Rocha, A., Sapkota, A., Boult, T.E.: Toward open set recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **35**(7), 1757–1772 (2013). https://doi.org/10.1109/TPAMI.2012.256 1
10. Vaze, S., Han, K., Vedaldi, A., Zisserman, A.: Generalized category discovery. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7492–7501 (2022) 1, 3
11. Zhang, C., Xu, R., He, X.: Novel class discovery for long-tailed recognition. Transactions on Machine Learning Research (2023) 4
12. Zhong, Z., Fini, E., Roy, S., Luo, Z., Ricci, E., Sebe, N.: Neighborhood contrastive learning for novel class discovery. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10867–10875 (2021) 3