A Closer Look at Novel Class Discovery from the Labeled Set

Ziyun Li
Hasso Plattner Institute
ziyun.li@hpi.de

Jona Otholt
Hasso Plattner Institute
jona.Otholt@hpi.de

Ben Dai*
Chinese University of Hong Kong
bendai@cuhk.edu.hk

Di Hu
Renmin University of China
dihu@ruc.edu.cn

Christoph Meinel
Hasso Plattner Institute
christoph.meinel@hpi.de

Haojin Yang*
Hasso Plattner Institute
haojin.yang@hpi.de

Abstract

Novel class discovery (NCD) is to infer novel categories in an unlabeled set using prior knowledge of a labeled set comprising diverse but related classes. Existing research focuses on using the labeled set methodologically and little on analyzing it. In this study, we take a closer look at NCD from the labeled set and focus on two questions: (i) Given an unlabeled set, *what labeled set best supports novel class discovery?* (ii) A fundamental premise of NCD is that the labeled set must be related to the unlabeled set, but *how can we measure this relation?* For (i), we propose and substantiate the hypothesis that NCD could benefit from a labeled set with high semantic similarity to the unlabeled set. Using ImageNet's hierarchical class structure, we create a large-scale benchmark with variable semantic similarity across labeled/unlabeled datasets. In contrast, existing NCD benchmarks ignore the semantic relation. For (ii), we introduce a mathematical definition for quantifying the semantic similarity between labeled and unlabeled sets. We utilize this metric to validate our established benchmark and demonstrate it highly corresponds with NCD performance. Furthermore, without quantitative analysis, previous works commonly believe that label information is always beneficial. However, our experimental results counterintuitively show that using labels may lead to suboptimal outcomes in low-similarity settings. An extended paper version is available at `https://arxiv.org/abs/2209.09120`.

1 Introduction

Deep models are capable of identifying and clustering classes that are present in the training set (i.e., known/seen classes), matching or surpassing human performance. However, they lack reliable extrapolation capacity when confronted with novel classes, while humans can easily recognize the unseen categories. This encouraged researchers to establish a challenge termed novel class discovery (NCD) [10, 2, 9, 22], to identify new classes in an unlabeled dataset by utilizing information from a labeled set containing similar but disjoint classes.

Currently, most NCD research takes place at the method level, focusing on better utilizing the labeled set. Though the labeled set is essential, there is a less in-depth analysis of the labeled set itself. This lack of understanding about a crucial aspect of NCD illustrates the necessity to explore it from the labeled set's perspective. Thus, our paper concentrates on two core questions: First, given a specific unlabeled set, *what kind of labeled set can best support novel class discovery?* Second, an essential

*Corresponding author

Workshop on Distribution Shifts, 36th Conference on Neural Information Processing Systems (NeurIPS 2022).

premise of NCD is that the labeled set should be related to the unlabeled set, but *how can we measure this relation?* Based on the preceding questions, we also give insights into the importance of labeled information in NCD.

Regarding the first question, we intuitively expect that labeled sets with higher semantic similarity can provide more beneficial knowledge while the number of categories and pictures is fixed. In contrast, existing works solely use the number of labeled/unlabeled classes and images to determine NCD difficulty, e.g., [5, 2], and disregard semantic similarity. We first verify our assumption on multiple pairs of labeled and unlabeled sets with varying semantic similarity and under multiple baselines [10, 9, 5, 16]. Then, we establish a new benchmark with multiple semantic similarity levels using ImageNet's hierarchical semantic information, and more details can be found in Section 2.

Second, an essential premise of NCD is that leveraging the information of the disjoint but related labeled set improves performance on the unlabeled data. A prior work [2] points out that NCD is theoretically solvable when labeled set and unlabeled set share high-level semantic features yet without proposing any quantitative analysis. This inspires the following questions: How closely related do the sets need to be for NCD to work? How can we measure the semantic relatedness between labeled and unlabeled sets? Motivated by these questions, we propose a semantic similarity metric, called *transfer leakage*. Specifically, *transfer leakage* quantifies how much information we can leverage from the labeled dataset to help improve the performance of the unlabeled dataset, and more details are provided in Section 3.

Furthermore, we observe that labeled information may lead to sub-optimal results, contrary to the commonly held belief that labeled information is always beneficial for NCD tasks. However, it is hard to decide whether to use labeled supervised knowledge or self-supervised knowledge without labels. Thus, we provide two concrete solutions. (i) *pseudo transfer leakage*, a practical reference for what sort of data we intend to employ. (ii) A straightforward method, which smoothly combines supervised and self-supervised knowledge from the labeled set and achieves 3% and 5% improvement in both CIFAR100 and ImageNet compared to SOTA. For further information, see Section 4.

We summarize our contributions as follows: (i) We establish a comprehensive and large benchmark with varying degrees of difficulty on ImageNet and thoroughly justify the assumption that semantic similarity is a significant factor influencing NCD performance. (ii) We introduce a mathematical definition for evaluating the semantic similarity between labeled and unlabeled sets and validate it under CIFAR100 and ImageNet. (iii) We observe counterintuitive results - labeled information may lead to suboptimal performance and propose two practical applications, which achieve 3% and 5% improvement in both CIFAR100 and ImageNet compared to SOTA.

2 Assumption and Proposed Benchmarks

In this section, we address the first question: *given a specific unlabeled set, what kind of labeled set can best support novel class discovery?* We first assume that higher semantic similarity labeled sets can provide greater help compared to less similar labeled sets when the number of categories and images in the labeled sets are fixed. However, existing benchmarks were created based on the number of categories(e.g., [5]) and images(e.g.,[2]) without considering semantic similarity between the two sets. Thus, we propose a new benchmark based on the ENTITY-30 task[18] including three different semantic similarity levels (high, medium and low) by leveraging the underlying hierarchy, which contains 240 ImageNet classes in total, with 30 superclasses and 8 subclasses for each superclasses. For our benchmark, we use these classes to create NCD tasks with 90 labeled and 30 unlabeled classes each. Lastly, our hypothesis is verified on our benchmark (Table 1) and CIFAR100 (Table 4). The results demonstrate that the most similar labeled set achieves the highest performance, followed by the medium and the least similar set. Further details are provided in the Appendix A.

3 Quantifying Semantic Similarity

3.1 NCD Framework

We denote $(\mathbf{X}_l, Y_l)$ and $(\mathbf{X}_u, Y_u)$ as random samples under the *labeled/unlabeled probability measures* $\mathbb{P}_{\mathbf{X},Y}$ and $\mathbb{Q}_{\mathbf{X},Y}$, respectively. $\mathbf{X}_l \in \mathcal{X}_l \subset \mathbb{R}^d$ and $\mathbf{X}_u \in \mathcal{X}_u \subset \mathbb{R}^d$ are the labeled/unlabeled feature vectors, $Y_l \in \mathcal{C}_l$ and $Y_u \in \mathcal{C}_u$ are the true labels of labeled/unlabeled data, where $\mathcal{C}_l$ and $\mathcal{C}_u$

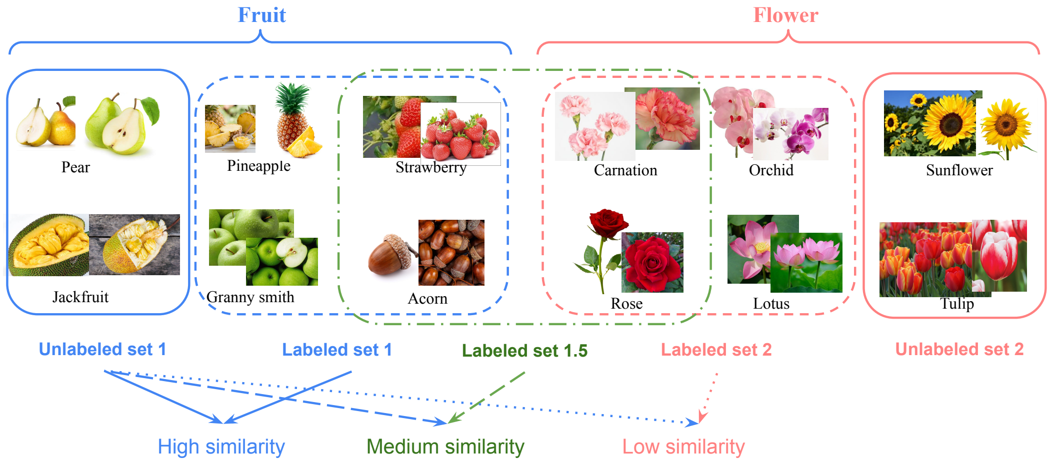

Figure 1: Illustration of how we construct the benchmark with varying levels of semantic similarity. Unlabeled set U_1 and labeled set L_1 are from the same superclass (fruit), whereas unlabeled set U_2 and labeled set L_2 belong to another superclass (flower). Labeled set $L_{1.5}$ is composed of half of L_1 and half of L_2. If both the labeled and unlabeled classes are derived from the same superclass, i.e. (U_1, L_1) and (U_2, L_2), we consider them a **high semantic similarity** split. In contrast, (U_1, L_2) and (U_2, L_1) are **low semantic similarity** splits, since the labeled and unlabeled classes are derived from distinct superclasses. In addition, we consider (U_1, $L_{1.5}$) and (U_2, $L_{1.5}$) to have **medium semantic similarity** because half of $L_{1.5}$ share the same superclass as U_1.

Table 1: Comparison of different semantic similarity settings in our proposed benchmark. L_1 is closely related to U_1 and L_2 is highly related to U_2. The third labeled set $L_{1.5}$ is constructed from half of L_1 and half of L_2, so in terms of similarity it is in between L_1 and L_2. For all splits we report the mean and the standard deviation of the clustering accuracy across multiple NCD baselines.

Methods	Unlabeled set U_1			Unlabeled set U_2		
	L_1 - high	$L_{1.5}$ - medium	L_2 - low	L_1 - low	$L_{1.5}$ - medium	L_2 - high
K-means [16]	**41.1** $\pm$ **0.4**	30.2 $\pm$ 0.4	23.3 $\pm$ 0.2	21.2 $\pm$ 0.2	29.8 $\pm$ 0.4	**45.0** $\pm$ **0.4**
DTC [10]	**43.3** $\pm$ **1.2**	35.6 $\pm$ 1.3	32.2 $\pm$ 0.8	21.3 $\pm$ 1.2	15.3 $\pm$ 1.5	**29.0** $\pm$ **0.8**
RS [8]	**55.3** $\pm$ **0.4**	50.3 $\pm$ 0.9	53.6 $\pm$ 0.6	48.1 $\pm$ 0.4	50.9 $\pm$ 0.6	**55.8** $\pm$ **0.7**
NCL [22]	**75.1** $\pm$ **0.8**	74.3 $\pm$ 0.4	71.6 $\pm$ 0.4	61.3 $\pm$ 0.1	70.5 $\pm$ 0.8	**75.1** $\pm$ **1.2**
UNO [5]	**83.9** $\pm$ **0.5**	81.0 $\pm$ 0.5	77.2 $\pm$ 0.8	77.5 $\pm$ 0.7	82.0 $\pm$ 1.7	**88.4** $\pm$ **1.2**

are the label sets under the labeled and unlabeled probability measures $\mathbb{P}_{\mathbf{X},Y}$ and $\mathbb{Q}_{\mathbf{X},Y}$, respectively. Given a labeled set $\mathcal{L}_n = (\mathbf{X}_{l,i}, Y_{l,i})_{i=1}^n$ independently drawn from the labeled probability measure $\mathbb{P}_{\mathbf{X},Y}$, and an unlabeled dataset $\mathcal{U}_m = (\mathbf{X}_{u,i})_{i=1}^m$ independently drawn from the unlabeled probability measure $\mathbb{Q}_{\mathbf{X}_u}$, our primary goal is to predict $Y_{u,i}$ given $\mathbf{X}_{u,i}$, where $Y_{u,i}$ is the label of the i-th unlabeled sample $\mathbf{X}_{u,i}$.

Definition 1 (Novel class discovery) *Let $\mathbb{P}_{\mathbf{X}_l,Y_l}$ be a labeled probability measure on $\mathcal{X}_l \times \mathcal{C}_l$, and $\mathbb{Q}_{\mathbf{X}_u,Y_u}$ be an unlabeled probability measure on $\mathcal{X}_u \times \mathcal{C}_u$, with $\mathcal{C}_u \cap \mathcal{C}_l = \emptyset$. Given a labeled dataset $\mathcal{L}_n$ sampled from $\mathbb{P}_{\mathbf{X}_l,Y_l}$ and an unlabeled dataset $\mathcal{U}_m$ sampled from $\mathbb{Q}_{\mathbf{X}_u}$, novel class discovery aims to predict the labels of the unlabeled dataset based on $\mathcal{L}_n$ and $\mathcal{U}_m$.*

3.2 Transfer Leakage

We begin with introducing Maximum Mean Discrepancy (MMD) [7], which is used to measure the discrepancy of two distributions. For example, the discrepancy of two random variables $\mathbf{Z} \sim \mathbb{P}_{\mathbf{Z}}$ and $\mathbf{Z}' \sim \mathbb{P}_{\mathbf{Z}'}$ is defined as: $\text{MMD}_{\mathcal{H}}\big(\mathbb{P}_{\mathbf{Z}}, \mathbb{P}_{\mathbf{Z}'}\big) := \sup_{\|h\|_{\mathcal{H}} \leq 1} \Big(\mathbb{E}\big(h(\mathbf{Z})\big) - \mathbb{E}\big(h(\mathbf{Z}')\big)\Big)$ where $\mathcal{H}$ is a class of functions $h : \mathcal{X}_u \to \mathbb{R}$, which is specified as a reproducing kernel Hilbert Space (RKHS) associated with a continuous kernel function $K(\cdot, \cdot)$.

In NCD, the unlabeled dataset utilizes the conditional probability $\mathbb{P}_{Y_l|\mathbf{X}_l}$ (usually presented by a pretrained neural network) from a labeled dataset. For example, if the distributions of $\mathbb{P}_{Y_l|\mathbf{X}_l=\mathbf{X}_u}$

under $Y_u = c$ and $Y_u = c'$ are significantly different, then its overall distribution discrepancy is large, yielding that more information can be leveraged in NCD. On this ground, we use MMD to quantify the discrepancy of the labeled probability measure $\mathbb{P}_{Y_l|\mathbf{X}_l}$ on $\mathbf{X}_u$ under the unlabeled probability measure $\mathbb{Q}$, namely *transfer leakage*.

Definition 2 (Transfer leakage) *The* transfer leakage *of NCD prediction under $\mathbb{Q}$ based on the labeled conditional probability $\mathbb{P}_{Y_l|\mathbf{X}_l}$ is*

$$\textit{T-Leak}(\mathbb{Q}, \mathbb{P}) = \mathbb{E}_{\mathbb{Q}}\Big(\textit{MMD}^2_{\mathcal{H}}\big(\mathbb{Q}_{\mathbf{p}(\mathbf{X}_u)|Y_u}, \mathbb{Q}_{\mathbf{p}(\mathbf{X}'_u)|Y'_u}\big)\Big), \tag{1}$$

where $(\mathbf{X}_u, Y_u), (\mathbf{X}'_u, Y'_u) \sim \mathbb{Q}$ are independent copies, the expectation $\mathbb{E}_{\mathbb{Q}}$ is taken with respect to Y_u and Y'_u under $\mathbb{Q}$, and $\mathbf{p}(\mathbf{x})$ is the conditional probability under $\mathbb{P}_{Y_l|\mathbf{X}_l}$ on an unlabeled data $\mathbf{X}_u = \mathbf{x}$, which is defined as $\mathbf{p}(\mathbf{x}) = \big(\mathbb{P}(Y_l = c \mid \mathbf{X}_l = \mathbf{x})\big)^{\intercal}_{c\in\mathcal{C}_l}$.

To summarize, *transfer leakage* measures the overall discrepancy of $\mathbf{p}(\mathbf{X}_u)$ under different new classes of the unlabeled measure $\mathbb{Q}$, which indicates the informative leakage from $\mathbb{P}$ to $\mathbb{Q}$. Next, we give a finite sample estimate of *transfer leakage*. Given an estimated probability $\widehat{\mathbb{P}}_{Y_l|\mathbf{X}_l}$ and an evaluation dataset $(\mathbf{x}_{u,i}, y_{u,i})_{i=1}^m$ under $\mathbb{Q}$, we assess $\mathbf{x}_{u,i}$ on $\widehat{\mathbb{P}}_{Y_l|\mathbf{X}_l}$ as $\widehat{\mathbf{p}}(\mathbf{x}_{u,i}) = \big(\widehat{\mathbb{P}}\big(Y_l = c|\mathbf{X}_l = \mathbf{x}_{u,i}\big)\big)^{\intercal}_{c\in\mathcal{C}_u}$, then empirical *transfer leakage* is computed as:

$$\widehat{\text{T-Leak}}(\mathbb{Q}, \mathbb{P}) = \sum_{c,c'\in\mathcal{C}_u; c\neq c'} \frac{|\mathcal{I}_{u,c}||\mathcal{I}_{u,c'}|}{m(m-1)} \widehat{\text{MMD}}^2_{\mathcal{H}}\big(\mathbb{Q}_{\widehat{\mathbf{p}}(\mathbf{X}_u)|Y_u=c}, \mathbb{Q}_{\widehat{\mathbf{p}}(\mathbf{X}'_u)|Y'_u=c}\big), \tag{2}$$

where the equality follows from the fact that $\text{MMD}^2_{\mathcal{H}}\big(\mathbb{Q}_{\mathbf{p}(\mathbf{X}_u)|Y_u=c}, \mathbb{Q}_{\mathbf{p}(\mathbf{X}'_u)|Y'_u=c}\big) = 0$. $\mathcal{I}_{u,c} = \{1 \leq i \leq m : y_{u,i} = c\}$ is the index set of unlabeled data with $y_{u,i} = c$, and $\widehat{\text{MMD}}^2_{\mathcal{H}}$ is defined in the Appendix. We use the proposed $\widehat{\text{T-Leak}}$ to quantify the difficulty of NCD in various combination of labeled and unlabeled sets. It is worth noting that the proposed *transfer leakage* and its empirical evaluation depend only on the labeled/unlabeled datasets, and it remains the same, no matter what NCD method we use. In addition, we provide *pseudo transfer leakage* (Appendix B.2), a practical evaluation of the similarity between the classified and unlabeled sets. *transfer leakage* utilizes the true label Y_u, but the *pseudo transfer leakage* utilizes the pseudo label obtained from clustering methods (e.g., k-means) on the representations. Further information is available in Appendix B.

4 Supervised Knowledge may be Harmful

The motivation behind NCD is that supervised knowledge from labeled data can enhance unlabeled data clustering. Counterintuitively, we observe that supervised information from a labeled set may result in suboptimal outcomes compared to exclusively self-supervised knowledge. Further information is provided in Appendix C.2.

4.1 Experiments and Results

To investigate, we conduct experiments in the following settings: (1) Using the unlabeled set, $\mathbf{X}_u$; (2) Using the unlabeled set and the labeled set's images without labels, $\mathbf{X}_u + \mathbf{X}_l$; (Even without labels, self-supervised learning can extract the knowledge from labeled set's image.) (3) Using the unlabeled set and the whole labeled set, $\mathbf{X}_u + (\mathbf{X}_l, Y_l)$, (i.e., standard NCD). As suggested in Table 2, NCD performance is consistently improved by incorporating more images (without labels) from a labeled set, around 10% on CIFAR100 and 6%-18% in our benchmark. By comparing (2) and (3), we can isolate the impact of the labels. Unexpectedly, on CIFAR100-50 and ImageNet with low semantic similarity, (3) performs around 2 - 8% worse than (2), yielding that "low-quality" supervised information may hurt NCD performance.

Table 2: Comparison of different data settings on CIFAR100 and our proposed benchmark. We present clustering mean and standard error on SOTA method, UNO. (1) uses only the unlabeled set, whereas (2) uses both the unlabeled set and the labeled set's images without labels. (3) represents the standard NCD setting, i.e., using the unlabeled set and the whole labeled set. Counterintuitively, in CIFAR100-50 and low similarity case of our benchmark, (2) can get greater performance than (3).

Setting	CIFAR100-50	Unlabeled set U_1		Unlabeled set U_2	
		L_1 - high	L_2 - low	L_1 - low	L_2 - high
(1) $\mathbf{X}_u$	54.9 ± 0.4	70.5 ± 1.2	70.5 ± 1.2	71.9 ± 0.3	71.9 ± 0.3
(2) $\mathbf{X}_u + \mathbf{X}_l$	$\mathbf{64.1 \pm 0.4}$	79.6 ± 1.1	$\mathbf{80.3 \pm 0.3}$	$\mathbf{85.3 \pm 0.5}$	89.2 ± 0.3
(3) $\mathbf{X}_u + (\mathbf{X}_l, Y_l)$	62.2 ± 0.2	$\mathbf{83.9 \pm 0.6}$	77.2 ± 0.8	77.5 ± 0.7	88.3 ± 1.1

4.2 Practical Applications

As shown above, supervised knowledge from the labeled set may cause damage, but it's difficult to determine whether to utilize it or self-supervised knowledge. Therefore, we offer two concrete solutions, a practical metric (i.e., *pseudo transfer leakage*) and a straightforward method.

Supervised or Self-supervised Knowledge? The proposed *pseudo transfer leakage* is a practical reference to infer what kind of data we want to employ in NCD, images-only information or image-label pairs. In Table 3 (PTL), we compute *pseudo transfer leakage* via a supervised model and a self-supervised model based on pseudo labels. As suggested in Table 3 (ACC), the *pseudo transfer leakage* is consistent with the accuracy based on various datasets. For example, in L_1-U_1, the *pseudo transfer leakage* computed on the supervised model is larger than the one computed in the self-supervised model, which is consistent with the accuracy, where the supervised method outperforms the self-supervised one.

Combining Supervised and Self-supervised Knowledge Instead of using either supervised knowledge or self-supervised knowledge from the labeled set, we propose an effective and straightforward method, which smoothly combines both of them. We combine the labeled set's ground truth labels $y_{l_{GT}}$ with self-supervised pseudo labels $y_{l_{PL}}$. $\alpha y_{l_{GT}} + (1-\alpha) y_{l_{PL}}$ is the overall classification objective, where $\alpha \in [0, 1]$ is the supervised component weight. This strategy has the same aim as UNO [5] for $\alpha = 1$, but uses self-supervised pretraining instead of supervised. As shown in Figure 3, our proposed method improves CIFAR100 and ImageNet by 3% and 5%, respectively, compared to UNO. Our method delivers significant improvements for low semantic similarity cases and competitive performances for high similarity cases.

Table 3: Results showing the link between *pseudo transfer leakage* (PTL) and accuracy on novel classes (ACC). The *pseudo transfer leakage* is computed based either on a supervised (SL) or self-supervised model (SSL), using ResNet18 in both cases. The accuracy is obtained using the standard NCD setting ($\mathbf{X}_u + (\mathbf{X}_l, Y_l)$) for supervised learning, and self-supervised NCD ($\mathbf{X}_u + \mathbf{X}_l$) for the self-supervised model.

	Model	High similarity		Medium similarity		Low similarity	
		$L_1 - U_1$	$L_2 - U_2$	$L_{1.5} - U_1$	$L_{1.5} - U_2$	$L_2 - U_1$	$L_1 - U_2$
PTL	SSL	0.96 ± 0.01	0.96 ± 0.02	$\mathbf{1.14 \pm 0.02}$	$\mathbf{1.19 \pm 0.01}$	$\mathbf{1.05 \pm 0.03}$	$\mathbf{1.25 \pm 0.03}$
	SL	$\mathbf{1.21 \pm 0.02}$	$\mathbf{1.21 \pm 0.01}$	1.03 ± 0.02	0.98 ± 0.03	0.99 ± 0.02	0.96 ± 0.01
ACC	SSL	79.6 ± 1.1	89.2 ± 0.3	79.7 ± 1.0	$\mathbf{85.2 \pm 1.0}$	$\mathbf{80.3 \pm 0.3}$	$\mathbf{85.3 \pm 0.5}$
	SL	$\mathbf{83.9 \pm 0.6}$	88.3 ± 0.5	81.0 ± 0.6	82.0 ± 1.6	77.2 ± 0.8	77.5 ± 0.7

5 Conclusion

We first offer a comprehensive ImageNet-based benchmark with varying levels of semantic similarity and show that semantic similarity affects NCD performance. Second, we present *transfer leakage*, a semantic similarity metric. Furthermore, we find that in low semantic similarity situations, labeled information may lead to inferior performance. We propose two practical applications based on these findings: (i) Using *transfer leakage* to determine what data to use. (ii) a straightforward approach that improves CIFAR100 and ImageNet by 3-5%.

References

[1] Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. Advances in Neural Information Processing Systems **33**, 9912–9924 (2020)

[2] Chi, H., Liu, F., Yang, W., Lan, L., Liu, T., Han, B., Niu, G., Zhou, M., Sugiyama, M.: Meta discovery: Learning to discover novel classes given very limited data. In: ICLR (2021)

[3] Cuturi, M.: Sinkhorn distances: lightspeed computation of optimal transport. In: Proceedings of the 26th International Conference on Neural Information Processing Systems-Volume 2, pp. 2292–2300 (2013)

[4] Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR, pp. 248–255, IEEE (2009)

[5] Fini, E., Sangineto, E., Lathuilière, S., Zhong, Z., Nabi, M., Ricci, E.: A unified objective for novel class discovery. In: ICCV, pp. 9284–9292 (2021)

[6] Garg, S., Balakrishnan, S., Lipton, Z.C.: Domain adaptation under open set label shift. arXiv preprint arXiv:2207.13048 (2022)

[7] Gretton, A., Borgwardt, K.M., Rasch, M.J., Schölkopf, B., Smola, A.: A kernel two-sample test. The Journal of Machine Learning Research **13**(1), 723–773 (2012)

[8] Han, K., Rebuffi, S.A., Ehrhardt, S., Vedaldi, A., Zisserman, A.: Automatically discovering and learning new visual categories with ranking statistics. In: International Conference on Learning Representations (2020)

[9] Han, K., Rebuffi, S.A., Ehrhardt, S., Vedaldi, A., Zisserman, A.: Autonovel: Automatically discovering and learning novel visual categories. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021)

[10] Han, K., Vedaldi, A., Zisserman, A.: Learning to discover novel visual categories via deep transfer clustering. In: CVPR (2019)

[11] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778 (2016)

[12] Hsu, Y.C., Lv, Z., Kira, Z.: Learning to cluster in order to transfer across domains and tasks. arXiv preprint arXiv:1711.10125 (2017)

[13] Hsu, Y.C., Lv, Z., Kira, Z.: Learning to cluster in order to transfer across domains and tasks. In: International Conference on Learning Representations (2018)

[14] Hsu, Y.C., Lv, Z., Schlosser, J., Odom, P., Kira, Z.: Multi-class classification without multi-class labels. In: International Conference on Learning Representations (2019)

[15] Krizhevsky, A.: Learning multiple layers of features from tiny images (2009)

[16] MacQueen, J., et al.: Some methods for classification and analysis of multivariate observations. In: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, vol. 1, pp. 281–297, Oakland, CA, USA (1967)

[17] Miller, G.A.: Wordnet: a lexical database for english. Communications of the ACM **38**(11), 39–41 (1995)

[18] Santurkar, S., Tsipras, D., Madry, A.: Breeds: Benchmarks for subpopulation shift. In: ICLR (2020)

[19] Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al.: Matching networks for one shot learning. Advances in neural information processing systems **29** (2016)

[20] Xie, J., Girshick, R., Farhadi, A.: Unsupervised deep embedding for clustering analysis. In: International conference on machine learning, pp. 478–487, PMLR (2016)

[21] Zhao, B., Han, K.: Novel visual category discovery with dual ranking statistics and mutual knowledge distillation. NIPS (2021)

[22] Zhong, Z., Fini, E., Roy, S., Luo, Z., Ricci, E., Sebe, N.: Neighborhood contrastive learning for novel class discovery. In: CVPR (2021)

[23] Zhong, Z., Zhu, L., Luo, Z., Li, S., Yang, Y., Sebe, N.: Openmix: Reviving known knowledge for discovering novel visual categories in an open world. In: CVPR (2021)

A Details of Section 2

A.1 Assumption

Existing benchmarks consider the difficulty of NCD in terms of the labeled set from two aspects: (1) the number of categories, e.g. [5] propose a more challenging benchmark called CIFAR100-50 (i.e., 50/50 classes for unlabeled/labeled set), compared to the commonly used CIFAR100-20 (i.e., 20/80 classes for unlabeled/labeled set). (2) The number of images in each category, [2] propose to use less images for each labeled's class.

However, in addition to the number of categories and images, another significant factor is the semantic similarity between the two sets. As mentioned in [2], NCD is theoretically solvable when labeled and unlabeled sets share high-level semantic features. Based on this, we conduct a further investigation with the assumption that more similar labeled sets (when the number of categories and images are fixed) can lead to better performance. Intuitively, according to Figure 1, despite the fact that the labeled (e.g., pineapple, strawberry) and unlabeled (e.g., pear, jackfruit) classes are disjoint, if they derive from the same superclass (i.e., fruit), they have a higher degree of semantic similarity. Conversely, when labeled (e.g., rose, lotus) and unlabeled (e.g., pear, jackfruit) classes are derived from distinct superclasses (i.e., labeled classes from flower while unlabeled classes from fruit), they are further apart semantically. Consequently, we construct various semantic similarity labeled/unlabeled settings based on a hierarchical class structure and evaluate our assumption on CIFAR100 and ImageNet.

A.2 Benchmark

Existing benchmarks in the field were created without regard to the semantic similarity between labeled and unlabeled set. Most works follow the standard splits introduced in [10]. In CIFAR10 [15], the labeled set is made up of the first five classes in alphabetical order, and the unlabeled set of the remaining five. A similar approach was taken with the commonly used CIFAR100-20 and CIFAR100-50 benchmarks. A benchmark based on ImageNet [4] has one labeled set, with 882 classes and three unlabeled sets. Each of these unlabeled sets contains 30 classes, which were randomly selected from the remaining non-labeled classes [19, 12, 14, 10].

To address this limitation and allow for an evaluation of our assumptions, we propose a new benchmark based on ImageNet including three different semantic similarity levels (high, medium and low). As mentioned in Section A.1, we separate labeled and unlabeled classes by leveraging ImageNet's underlying hierarchy. While ImageNet is based on the WordNet hierarchy [17], it is not well-suited for this purpose as discussed by [18]. To address these issues, they propose a modified hierarchy and define multiple hierarchical classification tasks based on it. While originally defined to measure the impact of subpopulation shift, they can also be used to define NCD tasks.

Our proposed benchmark is based on the ENTITY-30 task, which contains 240 ImageNet classes in total, with 30 superclasses and 8 subclasses for each superclasses. For example, as shown in Figure 1, we define three labeled sets L_1, L_2 and $L_{1.5}$ and two unlabeled sets U_1 and U_2. The sets L_1 and U_1 are selected from the first 15 superclasses, with 6 subclasses of each superclass assigned to L_1 and the other 2 assigned to U_1. The sets L_2 and U_2 are created from the second 15 superclasses in a similar fashion. Finally, $L_{1.5}$ is created by taking half the classes from L_1 and half of the classes from L_2. Therefore, (U_1, L_1)/(U_2, L_2) are highly related semantically, (U_1, L_2)/(U_2, L_1) belong to the low semantic cases and (U_1, $L_{1.5}$)/(U_2, $L_{1.5}$) are the medium cases. Additionally, we also create four data settings on CIFAR100, with two high semantic cases and two low semantic cases by leveraging CIFAR100 hierarchical class structure. Each case has 40 labeled classes and 10 unlabeled classes. A full list of the labeled and unlabeled sets with their respective superclasses and subclasses can be found in Appendix G.

This benchmark setup allows us to systematically investigate the influence of the labeled set on a large benchmark dataset. By keeping the unlabeled set constant and varying the used labeled set, we can isolate the influence of semantic similarity on NCD performance.

A.3 Experimental Setup and Results

To verify our assumption, we conduct experiments on four competitive baselines, including K-means [16], DTC [10], RS [9], NCL [22] and UNO [5]. We follow the baselines regarding hyperparameters and implementation details.

Results on CIFAR100 In Table 4, U_1 and U_2 represent the unlabeled sets, while L_1 and L_2 represent the labeled sets. U_1/U_2 and L_1/L_2 share the same super classes, while U_1/U_2 and L_2/L_1 belong to different super classes. We evaluate 4 different labeled/unlabeled settings in CIFAR100, with 2 high semantic cases (i.e., (U_1, L_1) and (U_2, L_2)) and low semantic cases (i.e., (U_1, L_2) and (U_2, L_1)). The gap between the high-similarity and the low-similarity settings is larger than 20% for K-means, and reaches up to 12% for more advanced methods. The strong results of UNO across all splits show that a more difficult benchmark is needed to obtain clear results for future methods.

Results on our proposed benchmark Similarly, in Table 1, the most similar labeled set generally obtains the best performance, followed by the medium and the least similar one. Under the unlabeled set U_1, L_1 achieves the highest accuracy, with around 2-17% improvement compared to L_2, and around 2-11% improvement compared to $L_{1.5}$. For the unlabeled set U_2, L_2 is the most similar set and obtains 8-14% improvement compared to L_1, and around 5-14% improvement compared to $L_{1.5}$.

Table 4: Comparison of different combinations of labeled sets and unlabeled sets consisting of subsets of CIFAR100. The unlabeled set are denoted U_1 and U_2, while the labeled sets are called L_1 and L_2. U_1 and L_1 share the same set of superclasses, similar for U_2 and L_2. Thus, the pairs (U_1, L_1) and (U_2, L_2) are close semantically, but (U_1, L_2) and (U_2, L_1) are far apart. For all splits we report the mean and standard deviation of the clustering accuracy across multiple NCD methods.

Methods	Unlabeled set U_1		Unlabeled set U_2	
	L_1 - high	L_2 - low	L_1 - low	L_2 - high
K-means [16]	**61.0** $\pm$ **1.1**	37.7 $\pm$ 0.6	33.9 $\pm$ 0.5	**55.4** $\pm$ **0.6**
DTC [10]	**64.9** $\pm$ **0.3**	62.1 $\pm$ 0.3	53.6 $\pm$ 0.3	**66.5** $\pm$ **0.4**
RS [8]	**78.3** $\pm$ **0.5**	73.7 $\pm$ 1.4	74.9 $\pm$ 0.5	**77.9** $\pm$ **2.8**
NCL [22]	**85.0** $\pm$ **0.6**	83.0 $\pm$ 0.3	72.5 $\pm$ 1.6	**85.6** $\pm$ **0.3**
UNO [5]	**92.5** $\pm$ **0.2**	91.3 $\pm$ 0.8	90.5 $\pm$ 0.7	**91.7** $\pm$ **2.2**

B Details of Section 3

B.1 The Detailed Derivation Process of Transfer Leakage

To summarize, *transfer leakage* measures the overall discrepancy of $\mathbf{p}(\mathbf{X}_u)$ under different new classes of the unlabeled measure $\mathbb{Q}$, which indicates the informative leakage from $\mathbb{P}$ to $\mathbb{Q}$.

Definition 3 (Transfer leakage) *The* transfer leakage *of NCD prediction under $\mathbb{Q}$ based on the labeled conditional probability $\mathbb{P}_{Y_l|\mathbf{X}_l}$ is*

$$\textit{T-Leak}(\mathbb{Q}, \mathbb{P}) = \mathbb{E}_{\mathbb{Q}}\Big(\textit{MMD}^2_{\mathcal{H}}\big(\mathbb{Q}_{\mathbf{p}(\mathbf{X}_u)|Y_u}, \mathbb{Q}_{\mathbf{p}(\mathbf{X}'_u)|Y'_u}\big)\Big), \tag{3}$$

where $(\mathbf{X}_u, Y_u), (\mathbf{X}'_u, Y'_u) \sim \mathbb{Q}$ are independent copies, the expectation $\mathbb{E}_{\mathbb{Q}}$ is taken with respect to Y_u and Y'_u under $\mathbb{Q}$, and $\mathbf{p}(\mathbf{x})$ is the conditional probability under $\mathbb{P}_{Y_l|\mathbf{X}_l}$ on an unlabeled data $\mathbf{X}_u = \mathbf{x}$, which is defined as $\mathbf{p}(\mathbf{x}) = \big(\mathbb{P}\big(Y_l = c \mid \mathbf{X}_l = \mathbf{x}\big)\big)^{\intercal}_{c \in \mathcal{C}_l}$.

Next, we give a finite sample estimate of *transfer leakage*. To proceed, we first rewrite *transfer leakage* as follows.

$$\text{T-Leak}(\mathbb{Q}, \mathbb{P}) = \sum_{c, c' \in \mathcal{C}_u; c \neq c'} \mathbb{Q}(Y_u = c, Y'_u = c')\text{MMD}^2_{\mathcal{H}}\big(\mathbb{Q}_{\mathbf{p}(\mathbf{X}_u)|Y_u = c}, \mathbb{Q}_{\mathbf{p}(\mathbf{X}'_u)|Y'_u = c'}\big), \tag{4}$$

where the equality follows from the fact that $\text{MMD}^2_{\mathcal{H}}\big(\mathbb{Q}_{\mathbf{p}(\mathbf{X}_u)|Y_u = c}, \mathbb{Q}_{\mathbf{p}(\mathbf{X}'_u)|Y'_u = c}\big) = 0$.

Given an estimated probability $\widehat{\mathbb{P}}_{Y_l|\mathbf{X}_l}$ and an evaluation dataset $(\mathbf{x}_{u,i}, y_{u,i})_{i=1}^m$ under $\mathbb{Q}$, we assess $\mathbf{x}_{u,i}$ on $\widehat{\mathbb{P}}_{Y_l|\mathbf{X}_l}$ as $\widehat{\mathbf{p}}(\mathbf{x}_{u,i}) = \big(\widehat{\mathbb{P}}\big(Y_l = c|\mathbf{X}_l = \mathbf{x}_{u,i}\big)\big)_{c\in\mathcal{C}_u}^\intercal$, then the empirical *transfer leakage* is computed as:

$$\widehat{\text{T-Leak}}(\mathbb{Q},\mathbb{P}) = \sum_{c,c'\in\mathcal{C}_u; c\neq c'} \frac{|\mathcal{I}_{u,c}||\mathcal{I}_{u,c'}|}{m(m-1)} \widehat{\text{MMD}}^2_{\mathcal{H}}\big(\mathbb{Q}_{\widehat{\mathbf{p}}(\mathbf{X}_u)|Y_u=c}, \mathbb{Q}_{\widehat{\mathbf{p}}(\mathbf{X}'_u)|Y'_u=c}\big), \tag{5}$$

where $\mathcal{I}_{u,c} = \{1 \leq i \leq m : y_{u,i} = c\}$ is the index set of unlabeled data with $y_{u,i} = c$, and $\widehat{\text{MMD}}^2_{\mathcal{H}}$ is defined as:

$$\begin{aligned}
\widehat{\text{MMD}}^2_{\mathcal{H}}(\mathbb{Q}_{\widehat{\mathbf{p}}(\mathbf{X}_u)|Y_u=c}, \mathbb{Q}_{\widehat{\mathbf{p}}(\mathbf{X}'_u)|Y'_u=c}) = &\frac{1}{|\mathcal{I}_{u,c}|(|\mathcal{I}_{u,c}|-1)} \sum_{i,j\in\mathcal{I}_{u,c}; i\neq j} K\big(\widehat{\mathbf{p}}(\mathbf{x}_{u,i}), \widehat{\mathbf{p}}(\mathbf{x}_{u,j})\big) \\
&+ \frac{1}{|\mathcal{I}_{u,c'}|(|\mathcal{I}_{u,c'}|-1)} \sum_{i,j\in\mathcal{I}_{u,c'}; i\neq j} K\big(\widehat{\mathbf{p}}(\mathbf{x}_{u,i}), \widehat{\mathbf{p}}(\mathbf{x}_{u,j})\big) \\
&- \frac{2}{|\mathcal{I}_{u,c}||\mathcal{I}_{u,c'}|} \sum_{i\in\mathcal{I}_{u,c}} \sum_{j\in\mathcal{I}_{u,c'}} K\big(\widehat{\mathbf{p}}(\mathbf{x}_{u,i}), \widehat{\mathbf{p}}(\mathbf{x}_{u,j})\big).
\end{aligned}$$

B.2 Pseudo Transfer Leakage

In practice, the ground-truth labels of the unlabeled set are difficult to acquire. Therefore, we utilize pseudo labels derived from clustering algorithms, e.g., k-means, rather than true labels in computing transfer leakage, and we named it as *pseudo transfer leakage*

Definition 4 (Pseudo Transfer Leakage)

$$\widehat{\textit{PT-Leak}}(\mathbb{Q},\mathbb{P}) = \sum_{c,c'\in\mathcal{C}_u; c\neq c'} \frac{|\widetilde{\mathcal{I}}_{u,c}||\widetilde{\mathcal{I}}_{u,c'}|}{m(m-1)} \widehat{\textit{MMD}}^2_{\mathcal{H}}\big(\mathbb{Q}_{\widehat{\mathbf{p}}(\mathbf{X}_u)|\widetilde{Y}_u=c}, \mathbb{Q}_{\widehat{\mathbf{p}}(\mathbf{X}'_u)|\widetilde{Y}'_u=c}\big) \tag{6}$$

where $\widetilde{\mathcal{I}}_{u,c} = \{1 \leq i \leq m : \widetilde{y}_{u,i} = c\}$ is the index set of unlabeled data with $\widetilde{y}_{u,i} = c$, $\widetilde{y}_{u,i}$ is provided based on k-means on their representations $\widehat{\mathbf{s}}(\mathbf{x}_{u,i})$, and $\widehat{\mathbf{s}}(\mathbf{x}_{u,i})$ is the representation estimated from a supervised model or a self-supervised model. $\widehat{MMD}^2_{\mathcal{H}}$ is defined similarly as in Section 3.

Pseudo transfer leakage, a practical evaluation of the similarity between the classified and unlabeled sets, could be a practical reference on whether to images-only or image-label pairs information from the labeled set.

B.3 The Lower and Upper Bounds of Transfer Leakage

Lemma 1 shows the lower and upper bounds of *transfer leakage*, and provide a theoretical justification that it is a effective quantity to measure the similarity between labeled and unlabeled datasets.

Lemma 1 *Let $\kappa := \max_{c\in\mathcal{C}_u} \mathbb{E}_{\mathbb{Q}}\big(\sqrt{K(\mathbf{p}(\mathbf{X}_u), \mathbf{p}(\mathbf{X}_u))}|Y_u = c\big) < \infty$, then $0 \leq$ T-Leak$(\mathbb{Q},\mathbb{P}) \leq 4\kappa^2$. Moreover, T-Leak$(\mathbb{Q},\mathbb{P}) = 0$ if and only if Y_u is independent with $\mathbf{p}(\mathbf{X}_u)$, that is, for any $c \in \mathcal{C}_u$:*

$$\mathbb{Q}\big(Y_u = c \mid \mathbf{p}(\mathbf{X}_u)\big) = \mathbb{Q}(Y_u = c), \tag{7}$$

yielding that $\mathbf{p}(\mathbf{X}_u)$ is useless in NCD on $\mathbb{Q}$.

Note that κ can be explicitly computed for many common used kernels, for example, $\kappa = 1$ for a Gaussian or Laplacian kernel. From Lemma 1, T-Leak$(\mathbb{Q},\mathbb{P}) = 0$ is equivalent to Y_u is independent with $\mathbf{p}(\mathbf{X}_u)$, which matches our intuition of no leakage. Alternatively, if Y_u is dependent with $\mathbf{p}(\mathbf{X}_u)$, we justifiably believe that the information of $Y_l|\mathbf{X}_l$ can be used to facilitate NCD, Lemma 1 tells that T-Leak$(\mathbb{Q},\mathbb{P}) > 0$ in this case. Therefore, Lemma 1 reasonably suggests that the proposed *transfer leakage* is an effective metric to detect if the labeling information in $\mathbb{P}$ is useful to NCD on $\mathbb{Q}$.

Proof of Lemma 1. We first show the upper bound of the transfer leakage. According to Lemma 3 in [7], we have

$$
\begin{aligned}
\text{T-Leak}(\mathbb{Q}, \mathbb{P}) &= \mathbb{E}_{\mathbb{Q}}\big(\mathrm{MMD}^2(\mathbb{Q}_{\mathbf{s}(\mathbf{X}_u)|Y_u}, \mathbb{Q}_{\mathbf{s}(\mathbf{X}'_u)|Y'_u})\big) = \mathbb{E}_{\mathbb{Q}}\Big(\big\|\mu_{\mathbb{Q}_{\mathbf{s}(\mathbf{X}_u)|Y_u}} - \mu_{\mathbb{Q}_{\mathbf{s}(\mathbf{X}'_u)|Y'_u}}\big\|_{\mathcal{H}}^2\Big) \\
&\leq \max_{c, c' \in \mathcal{C}_u} \big\|\mu_{\mathbb{Q}_{\mathbf{s}(\mathbf{X}_u)|Y_u=c}} - \mu_{\mathbb{Q}_{\mathbf{s}(\mathbf{X}'_u)|Y'_u=c'}}\big\|_{\mathcal{H}}^2 \leq 4 \max_{c \in \mathcal{C}_u} \|\mu_{\mathbb{Q}_{\mathbf{s}(\mathbf{X}_u)|Y_u=c}}\|_{\mathcal{H}}^2 \\
&= 4 \max_{c \in \mathcal{C}_u} \langle \mathbb{E}_{\mathbb{Q}}\big(K(\mathbf{s}(\mathbf{X}_u), \cdot)|Y_u = c\big), \mathbb{E}_{\mathbb{Q}}\big(K(\mathbf{s}(\mathbf{X}'_u), \cdot)|Y'_u = c\big)\rangle_{\mathcal{H}} \\
&= 4 \max_{c \in \mathcal{C}_u} \mathbb{E}_{\mathbb{Q}}\big(\langle K(\mathbf{s}(\mathbf{X}_u), \cdot), K(\mathbf{s}(\mathbf{X}'_u), \cdot)\rangle_{\mathcal{H}} | Y_u = c, Y'_u = c\big) \\
&\leq 4 \max_{c \in \mathcal{C}_u} \mathbb{E}_{\mathbb{Q}}\Big(\big\|K(\mathbf{s}(\mathbf{X}_u), \cdot)\big\|_{\mathcal{H}} \big\|K(\mathbf{s}(\mathbf{X}'_u), \cdot)\big\|_{\mathcal{H}} | Y_u = c, Y'_u = c\Big) \\
&= 4 \max_{c \in \mathcal{C}_u} \mathbb{E}_{\mathbb{Q}}\big(\sqrt{K(\mathbf{s}(\mathbf{X}_u), \mathbf{s}(\mathbf{X}_u))} | Y_u = c\big) \mathbb{E}_{\mathbb{Q}}\big(\sqrt{K(\mathbf{s}(\mathbf{X}'_u), \mathbf{s}(\mathbf{X}'_u))} | Y'_u = c\big) \leq 4\kappa^2,
\end{aligned}
$$

where $\mu_{\mathbb{Q}_{\mathbf{s}(\mathbf{X}_u)|Y_u}} := \mathbb{E}_{\mathbb{Q}}\big(K(\mathbf{s}(\mathbf{X}_u), \cdot)|Y_u\big)$ is the kernel mean embedding of the measure $\mathbb{Q}_{\mathbf{s}(\mathbf{X}_u)|Y_u}$ [7], the second inequality follows from the triangle inequality in the Hilbert space, the fourth equality follows from the fact that $\mathbb{E}_{\mathbb{Q}}$ is a linear operator, the second last inequality follows from the Cauchy-Schwarz inequality, and the last equality follows the reproducing property of $K(\cdot, \cdot)$.

Next, we show the if and only if condition for $\text{T-Leak}(\mathbb{Q}, \mathbb{P}) = 0$. Assume that $\mathbb{Q}(Y_u = c) > 0$ for all $c \in \mathcal{C}_u$. According to Theorem 5 in [7], we have

$$
\text{T-Leak}(\mathbb{Q}, \mathbb{P}) = 0 \quad \Longleftrightarrow \quad \mathbb{Q}\big(\mathbf{s}(\mathbf{x})|Y_u = c\big) = q_0(\mathbf{x}), \text{ for } c \in \mathcal{C}_u, \mathbf{x} \in \mathcal{X}_u.
$$

Note that

$$
1 = \sum_{c \in \mathcal{C}_u} \mathbb{Q}(Y_u = c | \mathbf{s}(\mathbf{x})) = \sum_{c \in \mathcal{C}_u} \frac{\mathbb{Q}(\mathbf{s}(\mathbf{x})|Y_u = c)\mathbb{Q}(Y_u = c)}{\mathbb{Q}(\mathbf{s}(\mathbf{x}))} = \sum_{c \in \mathcal{C}_u} \frac{q_0(\mathbf{x})\mathbb{Q}(Y_u = c)}{\mathbb{Q}(\mathbf{s}(\mathbf{x}))} = \frac{q_0(\mathbf{x})}{\mathbb{Q}(\mathbf{s}(\mathbf{x}))},
$$

yielding that $\mathbb{Q}\big(\mathbf{s}(\mathbf{x})|Y_u = c\big) = \mathbb{Q}(\mathbf{s}(\mathbf{x}))$, for $c \in \mathcal{C}_u, \mathbf{x} \in \mathcal{X}_u$. This is equivalent to,

$$
\mathbb{Q}\big(Y_u = c|\mathbf{s}(\mathbf{x})\big) = \frac{\mathbb{Q}\big(\mathbf{s}(\mathbf{x})|Y_u = c\big)\mathbb{Q}(Y_u = c)}{\mathbb{Q}(\mathbf{s}(\mathbf{x}))} = \mathbb{Q}(Y_u = c).
$$

This completes the proof. □

B.4 Experiments and Results

Experimental Setup and Hyperparameters To calculate the *transfer leakage/pseudo transfer leakage*, we employ ResNet18 [11] as the backbone for both datasets following [10, 9, 5]. Known-class data and unknown-class data are selected based on semantic similarity, as mentioned in Section 2. We first apply fully supervised learning to the labeled data for each data set to obtain the pretrained model. Then, we feed the unlabeled data to the pretrained model to obtain its representation. Lastly, we calculate the *transfer leakage/pseudo transfer leakage* based on the pretrained model and the unlabeled samples' representation. Speciallly, for *pseudo transfer leakage*, we apply clustering methods to generate the pseudo labels. For the first step, batch size is set to 512 for both datasets. We use an SGD optimizer with momentum 0.9, and weight decay 1e-4. The learning rate is governed by a cosine annealing learning rate schedule with a base learning rate of 0.1, a linear warmup of 10 epochs, and a minimum learning rate of 0.001. We pretrain the backbone for 200/100 epochs for CIFAR-100/ImageNet.

Results Figure 2 shows semantic similarity, *transfer leakage/pseudo transfer leakage* and NCD performance on ImageNet, in which the same color corresponds to the same unlabeled set. As expected, splits that have a higher semantic similarity yield both higher *transfer leakage* and *pseudo transfer leakage*. Alternatively, there is a consistent positive correlation between *transfer leakage/pseudo transfer leakage* and NCD accuracy, which confirms the validity of *transfer leakage/pseudo transfer leakage* as a metric in quantifying semantic similarity and the difficulty of a particular NCD problem.

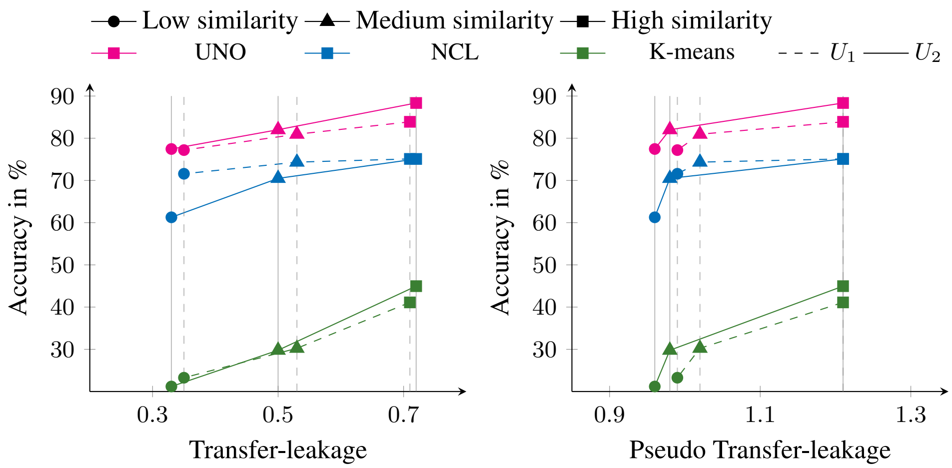

Figure 2: Experiments on *transfer leakage* and *pseudo transfer leakage*. Each line stands for one unlabeled set from our proposed ImageNet-based benchmark, and each point on the line for one labeled / unlabeled split. On the left, we measure the *transfer leakage* and the clustering accuracy obtained using UNO, NCL and k-means for each split. As expected, there is a positive correlation between semantic similarity and *transfer leakage* as well as between *transfer leakage* and accuracy. On the right, we replace *transfer leakage* with *pseudo transfer leakage* obtained using k-means clustering. The comparison shows that *pseudo transfer leakage* can in practice be used as a proxy for *transfer leakage*.

Table 5: Experiments on the stability of *pseudo transfer leakage* and *transfer leakage*. To obtain the standard deviation we recompute the *transfer leakage* and *pseudo transfer leakage* 10 times using bootstrap sampling. The results show that both *transfer leakage* has a low random variation.

Dataset	Unlabeled Set	Labeled Set	*Transfer leakage*	*Pseudo transfer leakage*
CIFAR100	U_1	L_1	0.62 ± 0.01	0.89 ± 0.06
		L_2	0.28 ± 0.01	0.74 ± 0.04
	U_2	L_1	0.33 ± 0.01	0.73 ± 0.02
		L_2	0.77 ± 0.02	1.24 ± 0.01
ImageNet	U_1	L_1	0.71 ± 0.01	1.21 ± 0.02
		$L_{1.5}$	0.54 ± 0.01	1.03 ± 0.02
		L_2	0.36 ± 0.01	0.99 ± 0.02
	U_2	L_1	0.33 ± 0.00	0.96 ± 0.01
		$L_{1.5}$	0.50 ± 0.01	0.98 ± 0.03
		L_2	0.72 ± 0.01	1.21 ± 0.01

Table 6: Experiments on *pseudo transfer leakage* under three clustering methods, i.e., k-means, GMM and agglomerative, and each setting is repeated for 10 times.

Method	Unlabeled set U_1			Unlabeled set U_2		
	L_1 - high	$L_{1.5}$ - medium	L_2 - low	L_1 - low	$L_{1.5}$ - medium	L_2 - high
k-means	1.23 ± 0.03	1.02 ± 0.03	0.99 ± 0.02	0.96 ± 0.01	0.99 ± 0.03	1.24 ± 0.02
GMM	0.79 ± 0.01	0.69 ± 0.02	0.56 ± 0.02	0.58 ± 0.02	0.68 ± 0.04	0.91 ± 0.02
Agglomerative	1.17 ± 0.00	0.96 ± 0.00	0.87 ± 0.00	0.83 ± 0.00	0.89 ± 0.00	1.15 ± 0.00

C Details of Section 4

C.1 Counterintuitive Observations

The motivation behind NCD is that prior supervised knowledge from labeled data can help improve the clustering of the unlabeled set. However, we have counterintuitive results: supervised information from a labeled set may result in suboptimal outcomes as compared to using exclusively self-supervised knowledge. To investigate, we conduct experiments in the following settings:

(1) Using the unlabeled set, $\mathbf{X}_u$
(2) Using the unlabeled set and the labeled set's images without labels, $\mathbf{X}_u + \mathbf{X}_l$
(3) Using the unlabeled set and the whole labeled set, $\mathbf{X}_u + (\mathbf{X}_l, Y_l)$, (i.e., standard NCD).

Specifically, for (1), NCD is degenerated to unsupervised learning (i.e., clustering on $\mathbf{X}_u$). For (2), even though we do not use the labels, we can still try to extract the knowledge of the labeled set via self-supervised learning. By comparing (1) and (3), we can estimate the total performance gain caused by adding the labeled set. The comparison between (1) and (2) as well as (2) and (3) allows us to further disentangle the performance gain according to the components of the labeled set.

Table 7: Comparison of different data settings on CIFAR100 and the unlabeled set U_1 of our proposed benchmark. We report the mean and the standard error of the clustering accuracy of UNO. As UNO uses multiple unlabeled heads, we report their mean accuracy, as well as that of the head with the lowest loss. Setting (1) uses only the unlabeled set, whereas (2) uses both the unlabeled set and the labeled set's images without labels. Setting (3) represents the standard NCD setting, i.e., using the unlabeled set and the whole labeled set. Counterintuitively, in CIFAR100-50 and on the low similarity case of our benchmark, we can achieve better performance without using the labeled set's labels.

UNO	Setting	CIFAR100-50	ImageNet U_1		
			L_1 - high	$L_{1.5}$ - medium	L_2 - low
Avg head	(1) $\mathbf{X}_u$	54.2 ± 0.3	69.2 ± 0.7	69.2 ± 0.7	69.2 ± 0.7
	(2) $\mathbf{X}_u + \mathbf{X}_l$	$\mathbf{63.4 \pm 0.4}$	74.9 ± 0.3	77.6 ± 0.9	$\mathbf{77.9 \pm 1.1}$
	(3) $\mathbf{X}_u + (\mathbf{X}_l, Y_l)$	61.7 ± 0.3	$\mathbf{81.7 \pm 1.0}$	$\mathbf{80.3 \pm 0.4}$	74.6 ± 0.3
Best head	(1) $\mathbf{X}_u$	54.9 ± 0.4	70.5 ± 1.2	70.5 ± 1.2	70.5 ± 1.2
	(2) $\mathbf{X}_u + \mathbf{X}_l$	$\mathbf{64.1 \pm 0.4}$	79.6 ± 1.1	79.7 ± 1.0	$\mathbf{80.3 \pm 0.3}$
	(3) $\mathbf{X}_u + (\mathbf{X}_l, Y_l)$	62.2 ± 0.2	$\mathbf{83.9 \pm 0.6}$	$\mathbf{81.0 \pm 0.6}$	77.2 ± 0.8

Table 8: Comparison of different data settings on the unlabeled set U_2 of our proposed benchmark, similar to Table 7. We report the mean and the standard error of the clustering accuracy of UNO.

UNO	Setting	ImageNet U_2		
		L_1 - low	$L_{1.5}$ - medium	L_2 - high
Avg head	(1) $\mathbf{X}_u$	68.4 ± 0.6	68.4 ± 0.6	68.4 ± 0.6
	(2) $\mathbf{X}_u + \mathbf{X}_l$	$\mathbf{81.0 \pm 0.4}$	81.6 ± 1.1	85.9 ± 0.8
	(3) $\mathbf{X}_u + (\mathbf{X}_l, Y_l)$	76.2 ± 0.6	80.0 ± 1.6	$\mathbf{87.5 \pm 1.2}$
Best head	(1) $\mathbf{X}_u$	71.9 ± 0.3	71.9 ± 0.3	71.9 ± 0.3
	(2) $\mathbf{X}_u + \mathbf{X}_l$	$\mathbf{85.3 \pm 0.5}$	$\mathbf{85.2 \pm 1.0}$	89.2 ± 0.3
	(3) $\mathbf{X}_u + (\mathbf{X}_l, Y_l)$	77.5 ± 0.7	82.0 ± 1.6	88.3 ± 1.1

Experimental Setup We conduct experiments on UNO [5], which is to the best of our knowledge the current state-of-the-art method in NCD. To perform experiments (1) and (2), we make adjustments on the framework of UNO, enabling it to run fully self-supervised. This is done by replacing the labeled set's ground truth labels $y_{l_{GT}}$ with self-supervised pseudo labels $y_{l_{PL}}$, which are obtained by applying the Sinkhorn-Knopp algorithm [3].

The standard UNO method conducts NCD in a two-step approach. In the first step, it applies a supervised pretraining on the labeled data only. The pretrained model is then used as an initialization for the second step, in which the model is trained jointly on both labeled and unlabeled data using one labeled head and multiple unlabeled heads.

To adapt UNO to the fully unsupervised setting in (1), we need to remove all parts that utilize the labeled data. Therefore, in the first step, we replace the supervised pretraining by a self-supervised one, which is trained only on the unlabeled data. For the second step, we simply remove the labeled head, thus the method is degenerated to a clustering approach based solely on the pseudo-labels generated by the Sinkhorn-Knopp algorithm. For setting (2), we apply the self-supervised pretraining based on both unlabeled and labeled images to obtain the pretrained model in the first step. In the second step, we replace the ground-truth labels for the known classes with pseudo-labels generated by the Sinkhorn-Knopp algorithm based on the logits of these classes. Taken together, the updated setup utilizes the labeled images, but not their labels.

Hyperparameters We conduct our experiments on CIFAR100 as well as our proposed ImageNet-based benchmark. All settings and hyperparameters are kept as close as possible as to the original baselines, including the choice of ResNet18 as the model architecture. We use SWAV [1] as self-supervised pretraining for all experiments. The pretraining is done using the small batch size configuration of the method, which uses a batch size of 256 and a queue size of 3840. The training is run for 800 epochs, with the queue being enabled at 60 epochs for our ImageNet-based benchmark and 100 epochs for CIFAR100. To ensure a fair comparison with the standard NCD setting, the same data augmentations were used. In the second step of UNO, we train the methods for 500 epochs on CIFAR100 and 100 epochs for each setting on our benchmark. The experiments are replicated 10 times on CIFAR100 and 5 times on the developed benchmark, and the averaged performances and their corresponding standard errors are summarized in Table 7.

Results As suggested in Table 7 and Table 8, NCD performance is consistently improved by incorporating more images (without labels) from a labeled set, the percentages of improvements in terms of accuracy are around 10% on CIFAR100 (comparing (1) and (2)). For our benchmark, the setting (2) obtains an improvement about 6 - 10% over (1) and the increase is more obvious in the lower semantic similarity cases. Similarly, by comparing (2) and (3), we can isolate the impact of the labels. For this case, the largest percentage of improvement, about 4 - 7%, is obtained in the high-similarity setting, followed by the the medium-similarity setting with 1 - 3% improvements. Interestingly, on CIFAR100-50 and ImageNet with low semantic similarity, we unexpectedly observe that (2) performs around 2 - 8% better than (3), yielding that "low-quality" supervised information may hurt NCD performance.

C.2 Practical Applications

As shown in Table 7 and Table 8, even though we find that supervised knowledge from the labeled set may cause harm rather than gain, it is still difficult to decide whether to utilize supervised knowledge with labeled data or just pure self-supervised knowledge without labels. Therefore, we offer two concrete solutions, a practical metric (i.e., *pseudo transfer leakage*) and a straightforward method.

The proposed *pseudo transfer leakage* is a practical reference to infer what sort of data we want to use in NCD, images-only information, $\mathbf{X}_u + \mathbf{X}_l$ or the image-label pairs, $\mathbf{X}_u + (\mathbf{X}_l, Y_l)$ from the labeled set. In Table 3, we compute *pseudo transfer leakage* via a supervised model and a self-supervised model based on pseudo labels. As suggested,, the *pseudo transfer leakage* is consistent with the accuracy based on various datasets. For example, in $L_1 - U_1$, the *pseudo transfer leakage* computed on the supervised model is larger than the one computed in the self-supervised model, which is consistent with accuracy, the supervised method outperforms the self-supervised one. Reversely, for L_2-U_1, L_1-U_2 and $L_{1.5} - U_1$, the *pseudo transfer leakage* computed on the self-supervised model is larger than the one computed in the supervised model, which is again consistent with their relative performance. For $L_2 - U_2$ and $L_{1.5} - U_1$, performance in the two settings is within error margins.

Also, we propose a straightforward method, which smoothly combines supervised knowledge and self-supervised knowledge. Concretely, instead of using either the labeled set's ground truth labels $y_{l_{GT}}$ or the self-supervised pseudo labels $y_{l_{PL}}$, we use a linear combination of the two. The overall classification target is $\alpha y_{l_{GT}} + (1 - \alpha) y_{l_{PL}}$, where $\alpha \in [0, 1]$ is the weight of the supervised component. This means that for $\alpha = 1$, this approach has the same target as UNO [5], and differs only in the used pretraining, which is self-supervised as opposed to supervised in UNO. As indicated in

Figure 3, our proposed method achieves 3% and 5% improvement in both CIFAR100 and ImageNet (L_2-U_1), respectively compared to UNO, full results can be found in Table 9. Our method delivers significant improvements for low semantic similarity cases and competitive performances for high semantic cases.

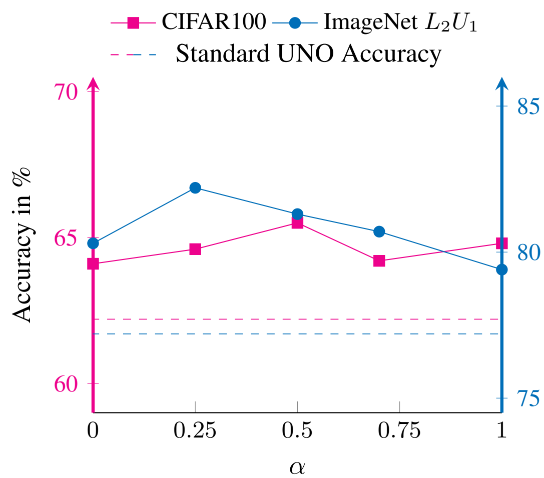

Figure 3: Experiments on combining supervised and self-supervised CIFAR100 and $L_2 - U_1$, low-similarity setting (note the different scales). α shows the weight of the supervised component. Dashed lines show the accuracy of SOTA (UNO). In low-similarity settings, a mix of supervised and self-supervised objectives outperforms either alone.

Table 9: Detailed results on combining supervised and self-supervised objectives. The α value indicates the weight of the supervised component. We see that the low similarity setting sees an improvement of up to 2.8% compared to $\alpha = 1.00$ and up to 5.0% compared to standard UNO. Standard UNO differs from the $\alpha = 1.00$ setting in that it does not use self-supervised pretraining, hence the difference in performance.

Setting	CIFAR100-50	ImageNet U_1		
		L_1 - high	$L_{1.5}$ - medium	L_2 - low
α = 0.00	64.1 ± 0.4	79.6 ± 1.1	79.7 ± 1.0	80.3 ± 0.3
α = 0.50	$\mathbf{65.5 \pm 0.5}$	82.3 ± 1.6	80.2 ± 1.6	$\mathbf{81.3 \pm 1.0}$
α = 1.00	64.8 ± 0.8	83.3 ± 0.6	81.5 ± 1.0	79.4 ± 0.4
UNO (SOTA) [5]	62.2 ± 0.2	83.9 ± 0.6	81.0 ± 0.6	77.2 ± 0.8

Table 10: Comparison of recent NCD methods with our proposed approach which combines supervised and self-supervised objectives.

Setting	CIFAR100-50	Unlabeled set U_1			Unlabeled set U_2		
		L_1 - high	$L_{1.5}$ - medium	L_2 - low	L_1 - low	$L_{1.5}$ - medium	L_2 - high
RS	39.2 ± 2.3	55.3 ± 1.0	50.3 ± 2.0	53.6 ± 1.3	48.1 ± 0.8	50.9 ± 1.3	55.8 ± 1.5
NCL	53.4 ± 0.3	75.1 ± 0.8	74.3 ± 0.4	71.6 ± 0.4	61.3 ± 0.1	70.5 ± 0.8	75.1 ± 1.2
UNO	62.2 ± 0.2	83.9 ± 0.6	81.0 ± 0.6	77.2 ± 0.8	77.5 ± 0.7	82.0 ± 1.6	88.3 ± 1.1
Ours	$\mathbf{65.5 \pm 0.5}$	83.3 ± 0.6	81.5 ± 1.0	$\mathbf{81.3 \pm 1.0}$	$\mathbf{85.8 \pm 0.8}$	$\mathbf{86.5 \pm 0.6}$	$\mathbf{91.5 \pm 1.1}$

Table 11: Comparison of different pretrained models. Self-supervised pretraining is beneficial for low semantic similarity cases.

Pretrained model	CIFAR100-50	ImageNet U_1		
		L_1 - high	$L_{1.5}$ - medium	L_2 - low
Self-supervised	**64.8** $\pm$ **0.8**	83.3 $\pm$ 0.6	81.5 $\pm$ 1.0	**79.4** $\pm$ **0.4**
Supervised	62.2 $\pm$ 0.2	83.9 $\pm$ 0.6	81.0 $\pm$ 0.6	77.2 $\pm$ 0.8

D Notations

To proceed, we summarize all notations used in the paper in Table 12.

Table 12: Notation used in the paper.

Notation	Description
$\mathbf{X}_l, \mathbf{X}_u$	labeled data / unlabeled data
y_l, y_u	label of labeled data / unlabeled data
$\mathcal{X}_l, \mathcal{X}_u$	domain of labeled data / unlabeled data
$\mathcal{C}_l, \mathcal{C}_u$	label set of labeled data / unlabeled data
$\mathbb{P}, \mathbb{Q}$	probability measure of labeled data / unlabeled data
$\mathcal{L}_n = (\mathbf{X}_{l,i}, Y_{l,i})_{i=1,\cdots,n}$	labeled dataset
$\mathcal{U}_m = (\mathbf{X}_{u,i})_{i=1,\cdots,m}$	unlabeled dataset
$\mathcal{H}$	reproducing kernel Hilbert space (RKHS)
$K(\cdot,\cdot)$	kernel function
$(\mathbf{X}', Y')$	independent copy of $(\mathbf{X}, Y)$
$\widehat{\mathbb{P}}$	estimated probability measure of labeled data
$\mathbb{E}_{\mathbb{Q}}$	expectation with respect to the probability measure $\mathbb{Q}$
$\mathbf{x}_{u,i}, y_{u,i}$	the i-th unlabeled data
$\mathcal{I}_{u,c}$	index set of unlabeled samples labeled as $y_{u,i} = c$
$\mathbf{s}(\mathbf{x}_{u,i})$	representation of the i-th unlabel data

E Related Work

Novel class discovery (NCD) is a relatively new problem proposed in recent years, aiming to discover novel classes (i.e., assign them to several clusters) by making use of similar but different known classes. Compared with unsupervised learning, NCD also requires labeled known-class data to help cluster novel-class data. NCD is first formalized in DTC [10], but the study of NCD can be dated back to earlier works, such as KCL [13] and MCL [14]. Both of these methods are designed for general task transfer learning, and maintain two models trained with labeled data and unlabeled data respectively. In contrast, DTC first learns a data embedding on the labeled data with metric learning, then employs a deep embedded clustering method based on [20] to cluster the novel-class data.

Later works further deviate from this approach. Both RS [9, 8] and [21] use self-supervised learning to boost feature extraction and use the learned features to obtain pairwise similarity estimates. Additionally, [21] improves on RS by using information from both local and global views, as well as mutual knowledge distillation to promote information exchange and agreement. NCL [22] extracts and aggregates the pairwise pseudo-labels for the unlabeled data with contrastive learning and generates hard negatives by mixing the labeled and unlabeled data in the feature space. This idea of mixing labeled and unlabeled data is also used in OpenMix [23], which mixes known-class and novel-class data to learn a joint label distribution. The current state-of-the-art, UNO [5], combines pseudo-labels with ground-truth labels in a unified objective function that enables better use of synergies between labeled and unlabeled data without requiring self-supervised pretraining. Additionally, there are a few theatrical works. Meta discovery [2] indicates that NCD is theoretically solvable if known and unknown classes share high-level semantic features and propose a solution that links NCD to meta-learning. OSLS [6] estimates the target label distribution, including the novel class and learn a target classifier.

F Discussion

The key assumption of novel class discovery is that the knowledge contained in the labeled set can help improve the clustering of the unlabeled set. Yet, what's the 'dark knowledge' transferred from the labeled set to the unlabeled set is still a mystery. Therefore, we conduct preliminary experiments to disentangle the impact of the different components of the labeled set, i.e., the images-only information and the image-label pairs. The results indicate that NCD performance is consistently improved by incorporating more images (without labels) from a labeled set while the supervised knowledge is not always beneficial. Supervised knowledge (obtained from $\mathbf{X}_l, Y_l$) can provide two types of information: classification rule and robustness. However, self-supervised information from $\mathbf{X}_l$ is primarily responsible for enhancing model robustness. In cases of high semantic similarity, the labeled and unlabeled classification rules are more similar than in cases of low semantic similarity. Thus, supervised knowledge is advantageous in high similarity cases but potentially harmful in low similarity situations while self-supervised knowledge is helpful for both cases.

G Detailed Benchmark Splits

Table 13: ImageNet class list of labeled split L_1 and unlabeled split U_1 of our proposed benchmark. As they share the same superclasses, they are highly related semantically. For each superclass, six classes are assigned to the labeled set and two to the unlabeled set. The labeled classes marked by the red box are also included in $L_{1.5}$, which shares half of its classes with L_1 and half with L_2.

Superclass	Labeled Subclasses	Unlabeled Subclasses
garment	vestment, jean, academic gown, sarong, fur coat, apron	swimming trunks, miniskirt
tableware	wine bottle, goblet, mixing bowl, coffee mug, water bottle, water jug	plate, beer glass
insect	leafhopper, long-horned beetle, lacewing, dung beetle, sulphur butterfly, fly	admiral, grasshopper
vessel	wreck, liner, container ship, catamaran, trimaran, lifeboat	yawl, aircraft carrier
building	toyshop, grocery store, bookshop, palace, butcher shop, castle	beacon, mosque
headdress	cowboy hat, bathing cap, pickelhaube, bearskin, bonnet, hair slide	crash helmet, shower cap
kitchen utensil	cocktail shaker, frying pan, measuring cup, tray, spatula, cleaver	caldron, coffeepot
footwear	knee pad, sandal, clog, cowboy boot, running shoe, Loafer	Christmas stocking, maillot
neckwear	stole, necklace, feather boa, bow tie, Windsor tie, neck brace	bolo tie, bib
bony fish	puffer, sturgeon, coho, eel, rock beauty, tench	gar, lionfish
tool	screwdriver, fountain pen, quill, shovel, screw, combination lock	torch, padlock
vegetable	spaghetti squash, cauliflower, zucchini, acorn squash, artichoke, cucumber	cardoon, butternut squash
motor vehicle	beach wagon, trailer truck, limousine, police van, convertible, school bus	garbage truck, moped
sports equipment	balance beam, rugby ball, ski, horizontal bar, racket, dumbbell	tennis ball, croquet ball
carnivore	otterhound, flat-coated retriever, Italian greyhound, Shih-Tzu, basenji, black-footed ferret	Boston bull, Bedlington terrier

Table 14: ImageNet class list of labeled split L_2 and unlabeled split U_2 of our proposed benchmark. As they share the same superclasses, they are highly related semantically. For each superclass, six classes are assigned to the labeled set and two to the unlabeled set. The labeled classes marked by the red box are also included in $L_{1.5}$, which shares half of its classes with L_1 and half with L_2.

Superclass	Labeled Subclasses	Unlabeled Subclasses
fruit	corn, buckeye, strawberry, pear, Granny Smith, pineapple	acorn, jackfruit
saurian	African chameleon, Komodo dragon, alligator lizard, agama, green lizard, Gila monster	banded gecko, American chameleon
barrier	stone wall, chainlink fence, breakwater, dam, bannister, picket fence	worm fence, turnstile
electronic equipment	cassette player, modem, printer, monitor, computer keyboard, pay-phone	dial telephone, microphone
serpentes	green snake, boa constrictor, green mamba, ringneck snake, thunder snake, king snake	rock python, garter snake
dish	hot pot, burrito, potpie, meat loaf, cheeseburger, mashed potato	hotdog, pizza
home appliance	espresso maker, toaster, washer, space heater, vacuum, microwave	dishwasher, Crock Pot
measuring instrument	wall clock, barometer, digital watch, hourglass, magnetic compass, analog clock	digital clock, parking meter
primate	indri, siamang, baboon, capuchin, chimpanzee, howler monkey	patas, Madagascar cat
crustacean	rock crab, king crab, crayfish, American lobster, Dungeness crab, spiny lobster	fiddler crab, hermit crab
musical instrument	organ, acoustic guitar, French horn, electric guitar, upright, maraca	violin, grand piano
arachnid	black and gold garden spider, wolf spider, harvestman, tick, black widow, barn spider	tarantula, scorpion
aquatic bird	dowitcher, goose, albatross, limpkin, white stork, red-backed sandpiper	drake, crane
ungulate	hippopotamus, hog, llama, hartebeest, ox, gazelle	warthog, zebra
passerine	house finch, magpie, goldfinch, indigo bunting, chickadee, brambling	bulbul, water ouzel

Table 15: Labeled Split of CIFAR100 used in Section A.2. We construct data settings based on its hierarchical class structure. U_1-L_1/U_2-L_2 are share the same superclasses.

Superclass	Labeled Subclasses (L_1)	Unlabeled Subclasses (U_1)
aquatic_mammals	dolphin, otter, seal, whale	beaver
fish	flatfish, ray, shark, trout	aquarium_fish
flower	poppy, rose, sunflower, tulip	orchids
food containers	bowl, can, cup, plate	bottles
fruit and vegetables	mushroom, orange, pear, sweet_pepper	apples
household electrical devices	keyboard, lamp, telephone, television	clock
household furniture	chair, couch, table, wardrobe	bed
insects	beetle, butterfly, caterpillar, cockroach	bee
large carnivores	leopard, lion, tiger, wolf	bear
large man-made outdoor things	castle, house, road, skyscraper	bridge
Superclass	**Labeled Subclasses (L_2)**	**Unlabeled Subclasses (U_2)**
large natural outdoor scenes	forest, mountain, plain, sea	cloud
large omnivores and herbivores	cattle, chimpanzee, elephant, kangaroo	camel
medium-sized mammals	porcupine, possum, raccoon, skunk	fox
non-insect invertebrates	lobster, snail, spider, worm	crab
people	boy, girl, man, woman	baby
reptiles	dinosaur, lizard, snake , turtle	crocodile
small mammals	mouse, rabbit, shrew, squirrel	hamster
trees	oak_tree, palm_tree, pine_tree, willow_tree	maple
vehicles 1	bus, motorcycle, pickup_truck, train	bicycle
vehicles 2	rocket, streetcar, tank, tractor	lawn-mower