Generalized Category Discovery with Hierarchical Label Smoothing

Sarah Rastegar, Yuki M. Asano, Hazel Doughty[*], Cees G. M. Snoek
University of Amsterdam

Abstract

Generalized Category Discovery seeks to cluster unknown categories while simultaneously discerning known ones. Existing approaches mostly rely on contrastive learning to produce distinctive embeddings for both labeled and unlabeled data. Yet, these methods often suffer from dispersed clusters for unknown categories due to a high rate of false negatives. To alleviate this problem, we introduce label smoothing as a hyperparameter that permits 'forgivable mistakes' for visually similar samples. We introduce a self-supervised cluster hierarchy, which allows us to control the strength of label smoothing to apply. By assigning pseudo-labels to emerging cluster candidates and using these as 'soft supervision' for contrastive learning, we effectively combine the benefits of clustering-based learning and contrastive learning. We demonstrate state-of-the-art generalized category discovery performance on various fine-grained datasets.

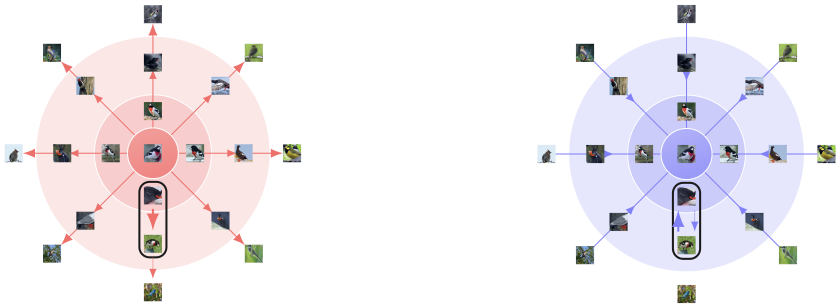
1 Introduction

As we continue to advance in computational capacity, abundant labeled datasets, and robust probabilistic models, supervised learning continues to outperform humans in classifying images within predefined categories [1–5]. Despite their superior performance in familiar, well-defined settings, deep learning models fail when faced with the unknown, demonstrating a noticeable shortfall in generalization to distinguish novel categories encountered at the test time. A primary strategy to solve this problem could involve discarding any novel category – in essence, bifurcating the data into known versus novel clusters while focusing on classifying known categories. This strategy is the well-studied problem of *open set recognition* [6–9]. An alternative methodology involves transferring classification knowledge from known to novel categories. This strategy, known as *Novel class discovery* [10–14, 14–18], aims to categorize new categories, drawing on the classification understanding gained from known categories. Nevertheless, novel class discovery harbors an inherent limitation: the presumption of mutual exclusivity of known and novel categories.

To uncover novel categories concurrently with known ones, *Generalized Category Discovery* [19–24] provides a model with unlabeled data from both novel and known categories, which can be framed as semi-supervised learning [25–29] with the uniqueness of having categories without any labeled instances. A widely accepted method to tackle semi-supervised learning challenges has been self-supervision [30–32]. Interestingly, contrastive learning [33–35], a form of self-supervision, has demonstrated its potential in unearthing new semantics that can be used for generalized category discovery [13, 36, 37]. However, a significant challenge persists due to the high occurrence of false negatives within the same category [36, 38, 39]. Supervised contrastive learning [40, 41] could potentially alleviate this issue for known categories with labeled data but not novel categories. In this paper, we propose a unique approach that harnesses the power of contrastive learning and pseudo-labeling on various hierarchy-level representations. Moreover, we employ the known labels at each stage of the hierarchy, providing more abstract supervision to the representation, which proves advantageous for recognized and novel categories. Ultimately, we utilize the labels generated at each level as pseudo labels for the supervised contrastive learning of unlabeled data.

[*]Currently at Leiden University

37th Conference on Neural Information Processing Systems (NeurIPS 2023).



(a) Unsupervised Hierarchical Contrastive Learning (b) Supervised Hierarchical Contrastive Learning

Figure 1: The motivation for hierarchical unsupervised and supervised learning **(a) Unsupervised Hierarchical Contrastive Learning.** In contrast to conventional unsupervised contrastive learning, our method implicitly divides the feature space into different zones with different strengths of repelling negative samples based on their distance to the positive sample. **(b) Supervised Hierarchical Contrastive Learning.** Our supervised contrastive learning also divides feature space into different semantic zones. If two samples are more similar semantically, even if they belong to different categories, the pull between them is stronger. In the box indicated in both Figure (a) and (b), the synergy between the two losses repels the semantically different samples and attracts the similar ones.

Our key contributions are as follows:

- Our research leverages hierarchical contrastive learning across varying degrees of supervision, providing us with the ability to adjust our supervision to different abstraction levels.
- Rather than employing clustering at every level, our approach utilizes the category cluster centers to generate more abstract labels. This strategy effectively mitigates the heavy computational expense of traditional clustering methods, particularly when dealing with many categories.
- Empirically, we demonstrate that our novel methodology facilitates effective category discovery and outperforms the existing baselines. Owing to its model-agnostic nature, our approach can be applied to other methods underpinned by contrastive learning.

2 Hierarchial Contrastive Generalized Category Discovery

Background. The *Generalized Category Discovery* problem introduced by Vaze et al. [41] tries to categorize a set of images that can be from the known categories seen during training or novel categories. Formally, we only have access to $\mathcal{Y}_{\mathcal{S}}$ or seen categories during training time, while we aim to categorize samples from novel categories or $\mathcal{Y}_{\mathcal{U}}$ during test time. *Novel Class Discovery* assumes that $\mathcal{Y}_{\mathcal{S}} \cap \mathcal{Y}_{\mathcal{U}} = \emptyset$, While *Generalized Category Discovery*, we have $\mathcal{Y}_{\mathcal{S}} \subset \mathcal{Y}_{\mathcal{U}}$. Vaze et al. [41] apply semi-supervised contrastive learning to generate a discriminative yet informative embedding for both known and unknown categories. Initially, the method employs unsupervised contrastive learning to discern an image from a multitude of others. Then, it integrates supervised contrastive learning, enabling the model to exploit the similarities among samples within the same class.

Label Smoothing. For contrastive unsupervised training, the goal is to optimize parameter θ to have $p_\theta(y = 1|\hat{x}_i, \hat{x}_j) = \delta_{ij}$ in which δ_{ij} is the delta Kronecker function which is one only when $i = j$ and zero otherwise. However, for the underlying ground truth distribution $p_{\mathcal{T}}$, this equality holds at the level of a hidden context variable, meaning $p_{\mathcal{T}}(y = 1|c_i, c_j) = \delta_{c_i c_j}$. In supplemental, it is shown that we can connect these two using Bayesian networks and label smoothing [42, 43] $p_\theta(y = 1|\hat{x}_i, \hat{x}_j) = p_{\mathcal{T}}(y = 1|c_i, c_j)(1 - \alpha) + \alpha U$ where U is the uniform distribution over all clusters, denoting the uncertainty about the ground truth y. We can approximate this α with one of the discussed strategies in supplemental. One thing to consider is that there is no specification about the level of abstraction c_i and c_j could be. They could be as fine as data samples and their augmentations or as abstract as seen vs. unseen categories. We can exploit this property to change the level of abstraction for different categories. This is the foundation for our hierarchical Supervised and unsupervised contrastive learning.

Hierarchical contrastive learning. Similar to [41], we use semi-supervised contrastive learning, but our approach deploys various levels of dependency on supervision in a hierarchical fashion, as depicted in Figure 2. The top level of this hierarchy denotes more abstract inclination, as seen vs. unseen, while the lower levels show a finer emphasis as sample vs. sample. Our hierarchical contrastive learning consists of three phases: pseudo-label extraction, unsupervised, and supervised hierarchical contrastive learning, which we will explain respectively.

Phase I - Pseudo-label extraction: Our approach introduces pseudo-labels for unlabelled samples while preserving the ground truth samples for known categories. With each iteration, these pseudo-labels are derived through a combination of clustering and linear assignment. As shown in Figure 2,

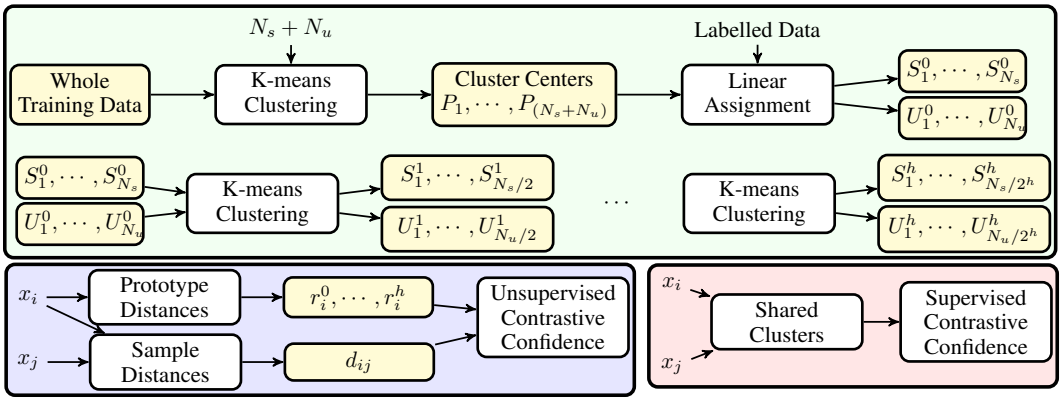


Figure 2: **Hierarchical contrastive learning framework.** Green box: pseudo-label extraction pipeline. We hierarchically use these prototypes to find confidence in supervised and unsupervised contrastive learning. Blue box: unsupervised hierarchical contrastive learning pipeline. Red box: supervised hierarchical contrastive learning pipeline.

we first apply clustering to the whole data to extract $N_s + N_u$ clusters and their corresponding centers $P_1, \cdots, P_{N_s+N_u}$ where N_s is the number of seen or known categories and N_u is the number of unknown categories. Given we have the labels for the labeled data, using linear assignment, we divide these cluster prototypes into N_s seen clusters $S_1^0, \cdots, S_N^0$, and N_u unknown clusters $U_1^0, \cdots, U_N^0$. Here, the superindex 0 shows that this is the 0th level of abstraction. Now, for any ith level of abstraction, we cluster the prototypes of $i-1$th seen abstraction level into half the amount seen categories and simultaneously $i-1$th unknown prototypes into the half amount. This ensures the prevention of over-assignment of samples to known category clusters and a few select unknown category clusters and hence reliable pseudo-labels for different levels of category abstractions.

Phase II- Unsupervised Hierarchical Contrastive Learning: In the early stages of training, the model lacks proficiency in known and unknown labels. Consequently, these pseudo-labels may be inherently noisy, leading to inaccurate model supervision. Unsupervised contrastive learning can use ground truth data to form clusters to counteract these limitations. However, since it thrives on distinguishing the augmented version of a sample from other samples, including the ones that share a semantic context with the current sample, it leads to dispersed clusters. To counteract these limitations, we employ label smoothing to manage unreliable labels. We also progressively consider more abstract categories; hence, fewer negative samples are discarded for these bigger clusters. But we also progressively decrease the importance of these "bigger cluster" negative samples. However, as we mentioned, there are multiple levels of abstraction and clusters. Hence, for the label smoothing coefficient $c_l = \alpha \sum_{i=0}^{h} \frac{1}{2^i} D^i$, we will have $\mathcal{L}_u^{total} = \mathcal{L}_u(c_l)$, where $h = \max(\lg N_s, \lg N_u)$ and $\lg$ is the logarithm base two, $\mathcal{L}_u$ is the unsupervised contrastive learning and c_l is its label smoothing matrix. Also, α is the label smoothing hyperparameter, and D^i is the distance metric at the abstraction level i. Finally, $\mathcal{L}_u^{total}$ is the unsupervised contrastive loss.

Phase III- Supervised Hierarchical Contrastive Learning: The resulting pseudo-labels from Phase I are the ground truth for the subsequent epoch. So, let us assume $\mathcal{L}_s^i$ is the supervised contrastive learning for the level of the pseudo label i. The overall supervised contrastive learning can be calculated as $\mathcal{L}_s^{total} = \sum_{i=0}^{h} \frac{1}{2^i} \mathcal{L}_s^i$. Given the availability of labeled samples from known categories, we gradually reduce the impact of label smoothing for known categories as the model becomes more adept at distinguishing them. This strategy ensures that the model generates highly accurate labels for known categories by the end of the training while still delivering valuable pseudo-labels for unknown categories. Since we can access different levels of abstraction, we use supervised contrastive learning for samples belonging to the same cluster. Finally, our total loss will be $\mathcal{L}^{total} = \mathcal{L}_s^{total} + \mathcal{L}_u^{total}$.

3 Experiments

In this section, we evaluate our method empirically. Experimental design, dataset statistics, and implementation details have been included in the supplemental.

Fine-grained image classification. Fine-grained image datasets are more aligned with a hierarchical perspective on categories. For generic datasets, visual cues aid the model in discerning the novelty of a category. On the contrary, fine-grained datasets require more nuanced attention to category-specific details. Table 1 summarizes our model's performance on the fine-grained datasets. This table shows that our model has more robust and consistent results than other methods for fine-grained datasets.

Table 1: **Comparison on fine-grained image recognition datasets.** Accuracy score for the first three methods is reported from [41] and for ORCA from [20]. Bold and underlined numbers, respectively, show the best and second-best accuracies. Our method has superior performance for the three experimental settings (*All*, *Known*, and *Novel*). This table shows that our method is especially well suited to fine-grained settings.

	CUB-200			**FGVC-Aircraft**			**Stanford-Cars**			**Oxford-IIIT Pet**			**Herbarium-19**		
Method	All	Known	Novel	All	Known	Novel	All	Known	Novel	All	Known	Novel	All	Known	Novel
k-means [46]	34.3	38.9	32.1	12.9	12.9	12.8	12.8	10.6	13.8	77.1	70.1	80.7	13.0	12.2	13.4
RankStats+ [13]	33.3	51.6	24.2	26.9	36.4	22.2	28.3	61.8	12.1	-	-	-	27.9	55.8	12.8
UNO+ [11]	35.1	49.0	28.1	40.3	56.4	32.2	35.5	70.5	18.6	-	-	-	28.3	53.7	14.7
ORCA [37]	36.3	43.8	32.6	31.6	32.0	31.4	31.9	42.2	26.9	-	-	-	24.6	26.5	23.7
GCD [41]	51.3	56.6	48.7	45.0	41.1	46.9	39.0	57.6	29.9	80.2	85.1	77.6	35.4	51.0	27.0
XCon [47]	52.1	54.3	51.0	47.7	44.4	49.4	40.5	58.8	31.7	86.7	**91.5**	84.1	-	-	-
PromptCAL [20]	62.9	64.4	62.1	52.2	52.2	**52.3**	50.2	70.1	40.6	-	-	-	-	-	-
DCCL [19]	63.5	60.8	**64.9**	-	-	-	43.1	55.7	36.2	**88.1**	88.2	**88.0**	-	-	-
SimGCD [23]	60.3	65.6	57.7	54.2	59.1	51.8	**53.8**	71.9	**45.0**	-	-	-	**44.0**	58.0	**36.4**
GPC [48]	52.0	55.5	47.5	43.3	40.7	44.8	38.2	58.9	27.4	-	-	-	-	-	-
Ours	**66.5**	**79.3**	60.1	**56.7**	**69.1**	50.5	52.2	**72.1**	42.3	87.4	91.2	85.4	41.2	**59.0**	31.6

Table 2: **Comparison with state-of-the-art for coarse-grained image classification.** Accuracy score for the three first methods is reported from [41] and for ORCA from [20] and the rest are reported from the corresponding work. Bold and underlined numbers show the best and second-best accuracies. Our method has a consistent performance for the three experimental settings (*All*, *Known*, *Novel*). Our method is especially suitable for novel categories in both datasets.

	CIFAR-10			**CIFAR-100**			**ImageNet-100**		
Method	All	Known	Novel	All	Known	Novel	All	Known	Novel
k-means [46]	83.6	85.7	82.5	52.0	52.2	50.8	72.7	75.5	71.3
RankStats+ [13]	46.8	19.2	60.5	58.2	77.6	19.3	37.1	61.6	24.8
UNO+ [11]	68.6	**98.3**	53.8	69.5	80.6	47.2	70.3	**95.0**	57.9
ORCA [37]	96.9	95.1	97.8	74.2	82.1	67.2	79.2	93.2	72.1
GCD [41]	91.5	97.9	88.2	73.0	76.2	66.5	74.1	89.8	66.3
XCon[47]	96.0	97.3	95.4	74.2	81.2	60.3	77.6	93.5	69.7
PromptCAL [20]	**97.9**	96.6	**98.5**	**81.2**	84.2	75.3	**83.1**	92.7	**78.3**
DCCL [19]	96.3	96.5	96.9	75.3	76.8	70.2	80.5	90.5	76.2
SimGCD [23]	97.1	95.1	98.1	80.1	81.2	**77.8**	83.0	93.1	77.9
GPC [48]	90.6	97.6	87.0	75.4	**84.6**	60.1	75.3	93.4	66.7
Ours	96.4	96.5	96.3	77.4	80.9	70.4	77.0	90.1	70.5

Coarse-grained image classification. We evaluate our model on three generic datasets, namely CIFAR10/100 [44] and ImageNet-100[45]. Table 2 compares our results against state-of-the-art generalized category discovery methods. Our method performs consistently well on both known and novel datasets. The generic datasets do not always have the hierarchy structure of more fine-grained datasets. Hence, the benefit of label smoothing is less substantial than fine-grained datasets.

4 Related Works

Novel Category Discovery can be traced back to [10] and [49] which solidified the novel class discovery as a new specific problem. The main goal of novel class discovery is to transfer the implicit category structure from the known categories to infer unknown categories [11–14, 14–18]. Prior to the novel class discovery, the problem of encountering new classes at the test time was investigated by open-set recognition [6–9]. However, in the open-set scenario, the model rejects the samples from novel categories, while novel class discovery aims to infer the known categories, but it still has a limiting assumption that test data only consists of novel categories. For a more realistic setting, **Generalised Category Discovery** is introduced by [41] and concurrently under the name *Open-world semi-supervised learning* by [37]. In this scenario, while the model should not lose its grasp on old categories, it must discover novel categories in test time. This adds an extra challenge because when we adapt the novel class discovery methods to this scenario, they try to be biased to either novel or old categories and miss the other group [19–24]. In this work, instead of viewing categories as separate, we take into account that there is a hidden hierarchy for the categories.

5 Conclusion

This work leverages hierarchical contrastive learning to discover unknown categories in conjunction with the known ones. To this end, we use clustering and linear assignment to extract pseudo labels for the subsequent supervised contrastive learning. The use of these pseudo-labels facilitates supervised contrastive learning for the unlabeled data, thereby enhancing the training speed and the integrity of the clusters formed for unknown categories. Since these pseudo-labels can also employed at different

levels of hierarchies, they provide informative supervision signals for different abstraction levels. Finally, by employing the label smoothing hyperparameter, we let the model adopt unsupervised contrastive learning in a more local scope and focus on finer distinctions. This in the end leads to a stronger fine-grained ability for our model.

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[2] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

[3] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *Proceedings of the International Conference on Learning Representations*, 2015.

[4] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[5] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[6] Abhijit Bendale and Terrance Boult. Towards open world recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1893–1902, 2015.

[7] Abhijit Bendale and Terrance E Boult. Towards open set deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1563–1572, 2016.

[8] Walter J Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E Boult. Toward open set recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(7):1757–1772, 2012.

[9] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: A good closed-set classifier is all you need. *Proceedings of the International Conference on Learning Representations*, 2022.

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

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

[12] Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In *International conference on machine learning*, pages 478–487. PMLR, 2016.

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

[14] Bingchen Zhao and Kai Han. Novel visual category discovery with dual ranking statistics and mutual knowledge distillation. *Advances in Neural Information Processing Systems*, 34:22982–22994, 2021.

[15] Zhun Zhong, Linchao Zhu, Zhiming Luo, Shaozi Li, Yi Yang, and Nicu Sebe. Openmix: Reviving known knowledge for discovering novel visual categories in an open world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9462–9470, 2021.

[16] KJ Joseph, Sujoy Paul, Gaurav Aggarwal, Soma Biswas, Piyush Rai, Kai Han, and Vineeth N Balasubramanian. Novel class discovery without forgetting. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIV*, pages 570–586. Springer, 2022.

[17] Subhankar Roy, Mingxuan Liu, Zhun Zhong, Nicu Sebe, and Elisa Ricci. Class-incremental novel class discovery. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIII*, pages 317–333. Springer, 2022.

[18] Mamshad Nayeem Rizve, Navid Kardan, and Mubarak Shah. Towards realistic semi-supervised learning. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXI*, pages 437–455. Springer, 2022.

[19] Nan Pu, Zhun Zhong, and Nicu Sebe. Dynamic conceptional contrastive learning for generalized category discovery. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[20] Sheng Zhang, Salman Khan, Zhiqiang Shen, Muzammal Naseer, Guangyi Chen, and Fahad Khan. Promptcal: Contrastive affinity learning via auxiliary prompts for generalized novel category discovery. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[21] Shaozhe Hao, Kai Han, and Kwan-Yee K Wong. Cipr: An efficient framework with cross-instance positive relations for generalized category discovery. *arXiv preprint arXiv:2304.06928*, 2023.

[22] Florent Chiaroni, Jose Dolz, Ziko Imtiaz Masud, Amar Mitiche, and Ismail Ben Ayed. Mutual information-based generalized category discovery. *arXiv preprint arXiv:2212.00334*, 2022.

[23] Xin Wen, Bingchen Zhao, and Xiaojuan Qi. Parametric classification for generalized category discovery: A baseline study. *arXiv preprint arXiv:2211.11727*, 2022.

[24] Wenbin An, Feng Tian, Qinghua Zheng, Wei Ding, QianYing Wang, and Ping Chen. Generalized category discovery with decoupled prototypical network. *arXiv preprint arXiv:2211.15115*, 2022.

[25] Yassine Ouali, Céline Hudelot, and Myriam Tami. An overview of deep semi-supervised learning. *arXiv preprint arXiv:2006.05278*, 2020.

[26] Xiangli Yang, Zixing Song, Irwin King, and Zenglin Xu. A survey on deep semi-supervised learning. *IEEE Transactions on Knowledge and Data Engineering*, 2022.

[27] Sylvestre-Alvise Rebuffi, Sebastien Ehrhardt, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Semi-supervised learning with scarce annotations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 762–763, 2020.

[28] Avital Oliver, Augustus Odena, Colin A Raffel, Ekin Dogus Cubuk, and Ian Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. *Advances in neural information processing systems*, 31, 2018.

[29] Olivier Chapelle, Bernhard Scholkopf, and Alexander Zien. Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. *IEEE Transactions on Neural Networks*, 20(3):542–542, 2009.

[30] Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. A survey on contrastive self-supervised learning. *Technologies*, 9(1):2, 2020.

[31] Xiaohua Zhai, Avital Oliver, Alexander Kolesnikov, and Lucas Beyer. S4l: Self-supervised semi-supervised learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1476–1485, 2019.

[32] Xiao Liu, Fanjin Zhang, Zhenyu Hou, Li Mian, Zhaoyu Wang, Jing Zhang, and Jie Tang. Self-supervised learning: Generative or contrastive. *IEEE Transactions on Knowledge and Data Engineering*, 35(1): 857–876, 2021.

[33] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[34] Junnan Li, Pan Zhou, Caiming Xiong, and Steven CH Hoi. Prototypical contrastive learning of unsupervised representations. *arXiv preprint arXiv:2005.04966*, 2020.

[35] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI*, pages 69–84. Springer, 2016.

[36] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European conference on computer vision (ECCV)*, pages 132–149, 2018.

[37] Kaidi Cao, Maria Brbic, and Jure Leskovec. Open-world semi-supervised learning. *Proceedings of the International Conference on Learning Representations*, 2022.

[38] Salar Hosseini Khorasgani, Yuxuan Chen, and Florian Shkurti. Slic: Self-supervised learning with iterative clustering for human action videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16091–16101, 2022.

[39] Tri Huynh, Simon Kornblith, Matthew R Walter, Michael Maire, and Maryam Khademi. Boosting contrastive self-supervised learning with false negative cancellation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2785–2795, 2022.

[40] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

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

[42] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[43] Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? *Advances in neural information processing systems*, 32, 2019.

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

[45] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[46] David Arthur and Sergei Vassilvitskii. K-means++ the advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035, 2007.

[47] Yixin Fei, Zhongkai Zhao, Siwei Yang, and Bingchen Zhao. Xcon: Learning with experts for fine-grained category discovery. *arXiv preprint arXiv:2208.01898*, 2022.

[48] Bingchen Zhao, Xin Wen, and Kai Han. Learning semi-supervised gaussian mixture models for generalized category discovery. *arXiv preprint arXiv:2305.06144*, 2023.

[49] Zhun Zhong, Enrico Fini, Subhankar Roy, Zhiming Luo, Elisa Ricci, and Nicu Sebe. Neighborhood contrastive learning for novel class discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10867–10875, 2021.

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

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

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

[53] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE, 2012.

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

[55] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[56] MB Wright. Speeding up the hungarian algorithm. *Computers & Operations Research*, 17(1):95–96, 1990.

A A Probabilistic Approach to Category Hierarchies

In this section, we provide insights into why hierarchical contrastive learning can provide more information about unseen categories. For contrastive supervised training, the goal is to optimize parameter θ in order to have the following equation:

$$p_\theta(y = 1|\hat{x}_i, \hat{x}_j) = \delta_{ij} \tag{1}$$

in which δ_{ij} is the delta Kronecker function which is one only when $i = j$ and zero otherwise. In reality, what we aim for the model to learn through this objective function is the equality of a hidden context variable. Hence, we can consider the following equation:

$$p_\theta(y = 1|c_i, c_j) = \delta_{c_i c_j}. \tag{2}$$

Consider the simple Bayesian network depicted in Figure 3. From this diagram, we can calculate the probability of $p(y|\hat{x}_i, \hat{x}_j)$ based on this Bayesian network.

$$p_\theta(y = 1|\hat{x}_i, \hat{x}_j) = \frac{\sum_{x_i}\sum_{x_j}\sum_{c_i}\sum_{c_j} p(y = 1|c_i, c_j)p(c_i)p(c_j)p(x_i|c_i)p(x_j|c_j)p(\hat{x}_i|x_i)p(\hat{x_j}|x_j)}{\sum_{x_i}\sum_{x_j}\sum_{c_i}\sum_{c_j}\sum_{y} p(y|c_i, c_j)p(c_i)p(c_j)p(x_i|c_i)p(x_j|c_j)p(\hat{x}_i|x_i)p(\hat{x_j}|x_j)}. \tag{3}$$

With some straightforward algebra, we can simplify this equation to

$$p_\theta(y = 1|\hat{x}_i, \hat{x}_j) = \sum_{c_i} p(c_i|\hat{x}_i)p(c_i|\hat{x}_j). \tag{4}$$

For x_i, we can assume that $\hat{x_i}$ would be a hypersphere with the radius r_{aug}. The hypersphere for cluster c_i will have a radius of R_i. We can then approximate $p(c_i|\hat{x_i})$ with the shared volume of hypersphere c_i and hypersphere $\hat{x_i}$. While this can be approximated by the shared cap volume between these two hyperspheres, we can adopt some strategies for simplifying this approximation.

Strategy one: sample distance zero one, for this strategy we simply consider that if x_i and x_j do not belong to the same cluster, then $p(c_i|\hat{x}_i)p(c_i|\hat{x_j})$ is negligible. So, for this strategy, we only consider the negative and positive samples that belong to the same cluster.

Strategy two: sample distance pairwise, for this strategy, we consider that x_i and x_j probability of being members of the same cluster will be a function of their pairwise distance.

Strategy three: cluster distance, for this strategy, instead of considering the distance to the actual data point x_j, we consider the distance of x_i to the center of the cluster containing x_j. This strategy can be seen as a combination of the previous two strategies.

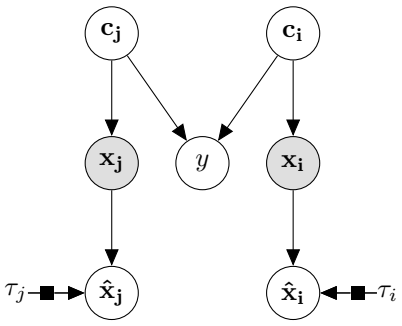


Figure 3: **Baysian Network for the Contrastive Learning Problem.** The shaded nodes are observed variables x_i and x_j which corresponds to the images i and j. The augmentations $\hat{x}_i$ and $\hat{x}_j$ created by adding a noise with parameters τ_i and τ_j to the actual images.

A.1 Experimental Setup

Eight Datasets. We evaluate our model on three generic datasets CIFAR10 [44], CIFAR100 [44] and ImageNet-100 [45] and four fine-grained datasets: CUB [50], Aircraft [52], Stanford-Cars [51] and Oxford-IIIT Pet [53]. Finally, we use the challenging Herbarium19 dataset [54], which is fine-grained and long-tailed. Following [41], we subsample the training dataset in a ratio of 50% of known categories at the train and all samples of unknown categories. For all datasets except CIFAR100, we consider 50% of categories as known categories at training time, while for CIFAR100, 80% of the categories are known during training time. A summary of dataset statistics and their train test splits is shown in Table 3.

Table 3: **Statistics of datasets and their data splits for the generalized category discovery task.** The first three datasets are coarse-grained image classification datasets, while the next four are fine-grained datasets. The Herbarium19 dataset is both fine-grained and long-tailed.

	Labelled		**Unlabelled**	
Dataset	#Images	#Categories	#Images	#Categories
CIFAR-10 [44]	12.5K	5	37.5K	10
CIFAR-100 [44]	20.0K	80	30.0K	100
ImageNet-100 [45]	31.9K	50	95.3K	100
CUB-200 [50]	1.5K	100	4.5K	200
SCars [51]	2.0K	98	6.1K	196
Aircraft [52]	3.4K	50	6.6K	100
Oxford-Pet [53]	0.9K	19	2.7K	37
Herbarium19 [54]	8.9K	341	25.4K	683

Implementation Details. Following [41], we use ViT-B/16 as our backbone, which is pre-trained by DINO [55] on unlabelled ImageNet 1K [2], we use the batch size of 128. For label smoothing, we use the α=0.5. Different from [41], we froze 10 blocks of ViT-B/16 and fine-tuned the last two blocks instead of only the last block. The code will be released.

Evaluation Metrics. Similar to [41], we use semi-supervised k-means to cluster the predicted embeddings. Then, we use the Hungarian algorithm [56] to solve the optimal assignment of emerged clusters to their ground truth labels. We report the accuracy of the model's predictions on *All*, *Known*, and *Novel* categories. Accuracy on *All* is calculated using the whole unlabelled test set, consisting of known and unknown categories. For *Known*, we only consider the samples with labels known during training. Finally, for *Novel*, we consider samples from the unlabelled categories at train time.