Foundation-Adaptive Integrated Refinement for Generalized Category Discovery

Yuwei Bian[1], **Shidong Wang**[2], **Yazhou Yao**[1], **Haofeng Zhang**[1*]

[1]School of Computer Science and Engineering, Nanjing University of Science and Technology, China
[2]School of Engineering, Newcastle University, United Kingdom
{yuwei.bian, yazhou.yao, zhanghf}@njust.edu.cn, shidong.wang@newcastle.ac.uk

Abstract

The potential of Generalized Category Discovery (GCD) lies in its ability to identify previously undiscovered patterns in both labeled and unlabeled data by leveraging insights from partially labeled training samples. However, interference can arise due to the model's dual focus on discovering both novel and known categories, often leading to conflicts that obscure true patterns in the dataset. This paper presents a divide-and-conquer framework, Foundation-Adaptive Integrated Refinement (FAIR), which fine-tunes pretrained foundational weights for various purposes, divided into Foundation (pretrained weights), Adaptive (weights fine-tuned with a variance-preserving loss), and Integrated (weights adjusted for both labeled and unlabeled data). The Adaptive utilizes a newly proposed adaptive contrastive loss that introduces variances within classes to preserve the individuality of representations. The Integrated addresses inherent estimation errors while dynamically estimating the number of categories, incorporating a cosine-based perturbation mechanism as a relaxed margin to accommodate potential ground-truth deviations, rather than relying on biased estimates. Extensive experiments on six benchmark datasets demonstrate our method's effectiveness, outperforming state-of-the-art algorithms, especially on fine-grained datasets.

Introduction

While supervised learning methods outperform humans in recognizing images using large-scale labeled data (He et al. 2016; Krizhevsky, Sutskever, and Hinton 2012; Dosovitskiy et al. 2021), its deployability to a wider range of domains is hindered by the labor-intensive annotation process. This challenge is especially pronounced in fine-grained recognition tasks, where more effort is required to distinguish subtle differences. The recently proposed Generalized Category Discovery (GCD) (Vaze et al. 2022), a challenging setting of deep clustering (Wang et al. 2022b; Han, Vedaldi, and Zisserman 2019), makes these assumptions more realistic by leveraging labeled data to inform the clustering of unlabeled data, requiring the model to maintain a dual focus on recognizing both known and novel categories.

However, enforcing a model to equitably focus on discovering patterns underlying both known and novel categories is challenging. Concretely, current methods (Choi, Kang, and Cho 2024; Pu, Zhong, and Sebe 2023; Wang, Vaze, and Han 2024; Wen, Zhao, and Qi 2023) typically fine-tune the pretrained weights (Caron et al. 2021; Song et al. 2023; Wang et al. 2022a; Huang et al. 2022), completely disregarding the interference introduced by unsupervised learning of unlabeled data. The adaptability of the pretrained model, in this scenario, cannot be guaranteed to be balanced and even may lead to conflicts as parameter updates account for both labeled and unlabeled data without distinction, obscuring true patterns and detrimentally affecting the performance.

Introducing strong supervision on labeled data effectively prevents the model from learning overly independent features during unsupervised training, thereby enhancing cluster awareness. However, this strategy also introduces a new challenge: when learning category concepts solely from labeled data, training a prototype classifier with cross-entropy loss tends to make the model overly confident in known classes, which exacerbates prediction bias and hinders the discovery of novel categories.

The key to mitigating the interference caused by imbalanced attention lies in enhancing the distinctiveness of labeled sample representations, while ensuring that shared latent features can effectively guide the clustering of unlabeled data. To this end, it presents a divide-and-conquer scheme, Foundation-Adaptive Integrated Refinement (FAIR, as illustrated in Figure 1), which addresses the dual focus problem distinctly to improve the balance between discovering known and novel categories. Specifically, the model parameters are initialized with pretrained weights to ensure generic knowledge is included as the `Foundation` for subsequent training. It then selectively increases the intra-class variance, enhancing the individuality of prototypical representations while remaining flexible during the `Adaptive` stage. Lastly, a relaxed margin through a cosine-based perturbation mechanism, based on the inherent error between the estimate and the ground truth, ensures dynamic estimation of category numbers and achieves `Integrated` weight adjustments for both labeled and unlabeled data, surpassing the capabilities of current methods (Pu, Zhong, and Sebe 2023; Vaze et al. 2022; Wang, Vaze, and Han 2024; Wen, Zhao, and Qi 2023) where the number of clusters is prior. Briefly, the contributions can be summarized as:

- A divide-and-conquer strategy named Foundation-

*corresponding Author.

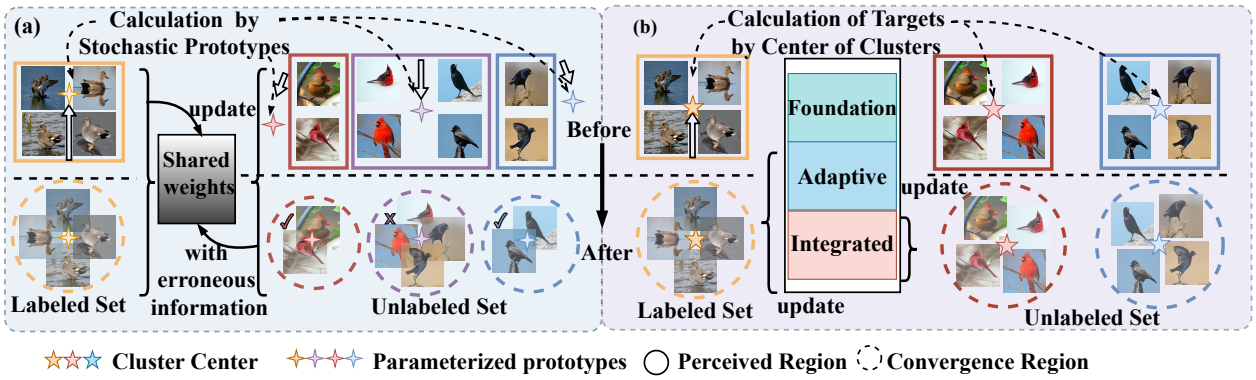

Figure 1: Difference between General Strategies and FAIR. The left illustrates the general strategies. Intuitively, the stochastically initialized prototypes are prone to local optima. The right diagram presents the FAIR. In `adaptive` parameter group, FAIR avoids enforcing the convergence of prototypes to a single representation. And in the `integrated` parameter group, the information in the labeled data is migrated through the cluster center.

Adaptive Integrated Refinement (FAIR) is introduced to alleviate the effects of interference caused by the dual focus on discovering both known and novel categories in challenging GCD tasks.

- A new contrastive learning loss is specifically designed to selectively increase intra-class variance, enhancing the model's ability to distinguish subtle differences within labeled classes and better cluster unlabeled data.
- A relaxed margin through a cosine-based perturbation mechanism is introduced to realize the dynamic estimation of category numbers, with a relaxed assumption that the true number of categories falls within the margin of the estimate.

Related Work

Category Discovery

The goal of the novel category discovery (NCD) is to utilize labeled data to guide the clustering of unlabeled data. Unlike traditional weakly supervised tasks, NCD not only addresses the classification of known categories but also requires the identification and clustering of unknown categories. Currently, pseudo-label-based methods for NCD can be broadly divided into two categories. The first (Chi et al. 2022; Zhong et al. 2021) involves designing algorithms to generate pairwise pseudo-labels to guide learning. The second type employs other pseudo-labeling algorithms, such as SK algorithm (Cuturi 2013; Asano, Rupprecht, and Vedaldi 2020), to directly obtain multi-class pseudo-labels, which are used to guide model training (Fini et al. 2021).

The NCD task is extended into GCD task (Vaze et al. 2022) by proposing a more realistic setting that discards the assumption that unlabeled data contains only unknown categories. PromptCL (Zhang et al. 2023) and SPTNet (Wang, Vaze, and Han 2024) utilized prompt learning to assist in feature representation for unlabeled data. Mutual information (Chiaroni et al. 2023) has also been employed to enhance the consistency of representations by maximizing the mutual information between features and pseudo-labels. DCCL (Pu, Zhong, and Sebe 2023) and CMS (Choi, Kang, and Cho 2024) apply Infomap (Rosvall and Bergstrom 2008) and mean-shift (Fukunaga and Hostetler 1975) to learning to represent unlabeled data, respectively. SimGCD (Wen, Zhao, and Qi 2023) employs parametric classification that incorporates cluster assignment into the training process via self-distillation, and SPTNet also follows this approach using parametric classification (Wang, Vaze, and Han 2024). However, parametric classification requires the number of categories to be predefined as prior knowledge and the assumption of a known number of categories is unrealistic for practical applications.

Contrastive Learning

The primary objective of contrastive learning is to train a model by defining positive pairs and negative pairs, such that similar samples close to each other in the embedding space, while dissimilar samples are pushed apart. SimCLR (Chen et al. 2020) generates positive and negative pairs through data augmentation and introduces a projection head to further separate learned features. Instead, MoCo (He et al. 2020) uses momentum updates to alleviate the computational bottleneck of storing a large number of sample representations through dictionary updates, and still has good performance when there is no need to set up large batches. All the above methods follow the idea of representation learning based on positive and negative pairs. SimSiam (Chen and He 2021) uses a simple Siamese Network for unsupervised representation learning by stop-gradient operation. SwAV (Caron et al. 2020) combines contrastive learning with online clustering to force different views of the same image to be consistent between cluster assignments.

Method

Problem Definition

We consider the training set, represented as $\mathcal{D}$, which consists of labeled set $\mathcal{D}_l$ and unlabeled set $\mathcal{D}_u$. The labeled set $\mathcal{D}_l$ is defined as $\mathcal{D}_l = \{(x_i, y_i)|i = 1, \cdots, |\mathcal{D}_l|\}$, where y_i belongs to a known class space $\mathcal{Y}_l$. The unlabeled set $\mathcal{D}_u$ is represented as $\mathcal{D}_u = \{(x_j)|j = 1, \cdots, |\mathcal{D}_u|\}$. The category

corresponding to $x_j \in \mathcal{D}_u$ belongs to the complete class space $\mathcal{Y}$. The known class space $\mathcal{Y}_l$ is a subset of the complete class space $\mathcal{Y}$. The unknown class space $\mathcal{Y}_u$ and the known class space $\mathcal{Y}_l$ are semantically related. Furthermore, the L_2 normalized visual feature $v = \mathcal{H}(x)$. Ultimately, the task is to leverage the labeled set $\mathcal{D}_l$ to guide the representation of the unlabeled set $\mathcal{D}_u$, ensuring that samples within $\mathcal{D}_u$ are correctly mapped to the corresponding categories.

Overview

The core of our proposed framework is to address the imbalanced focus of the model in discovering known and novel categories, which is caused by interference, such as noise and ambiguity present in both labeled and unlabeled data. We introduce Foundation-Adaptive Integrated Refinement (FAIR), which is not simply a learning scheme but also embodies our commitment to a divide-and-conquer approach with pretrained weights, in order to find the best balance between differentiating labeled data and using it to guide category discovery from unlabeled data. As illustrated in Figure 2, it introduces a new supervised contrastive learning loss when fine-tuning the foundational weights to ensure fair use of labeled data, and a dynamic category number estimation method that is embedded into the integrated weight updates when both labeled and unlabeled data are involved.

Foundation-Adaptive Fine-tuning

To mitigate overfitting to known categories during supervised fine-tuning, we propose the **Adaptive Contrastive Learning (ACL)** loss. Unlike conventional contrastive learning, ACL explicitly preserves intra-class variability by contrasting each sample against other same-class instances from previous training epochs. This helps maintain semantic flexibility, which is critical for generalizing to novel categories in GCD.

Specifically, for each labeled image x_i, ACL constructs a positive pair using the current embedding $\mathbf{v}_i$ and a historical embedding $\mathbf{v}'_i$ stored before the last model update. All other historical embeddings from the same class serve as negative pairs. Given a batch B and a feature memory bank $\mathcal{D}_{\mathcal{V}}^{l}$, the ACL loss is formulated as:

$$\mathcal{L}_{ACL} = \frac{1}{|B|} \sum_{i=1}^{|B|} -\log \frac{\exp(\mathbf{v}_i^\top \mathbf{v}'_i/\tau)}{\sum_{m \in \mathcal{P}_i} \exp(\mathbf{v}_i^\top \mathbf{v}'_m/\tau)}, \quad (1)$$

where $\mathcal{P}_i$ includes all stored features from the same class as x_i. We also apply a standard cross-entropy loss with a linear classifier to ensure label supervision:

$$\mathcal{L}_{CLS} = -\frac{1}{|B|} \sum_{i=1}^{|B|} \sum_{k=1}^{\mathcal{C}_l} y_k^{(i)} \log p_k^{(i)}. \quad (2)$$

The total loss for the adaptive stage is:

$$\mathcal{L}_{adaptive} = \mathcal{L}_{ACL} + \alpha \cdot \mathcal{L}_{CLS},$$

where α balances structure preservation and classification supervision.

Theoretical Justification. A core objective of ACL is to preserve representation diversity while ensuring sufficient inter-class separability—thereby avoiding premature feature collapse. To formalize this intuition, we adopt a hyperspherical embedding assumption and derive a margin–concentration trade-off under the von Mises–Fisher (vMF) (Banerjee et al. 2005) class-wise model. Specifically, we show that the cosine similarity between class prototypes is upper-bounded as a function of intra-class angular variance and a desired minimum margin. This result links ACL's effect on variance preservation to its ability to prevent overlapping class distributions.

Theorem 1 (Margin–Concentration Trade-off)**.** *Assume that feature embeddings lie on the unit hypersphere and that features of class c are drawn from a von Mises–Fisher distribution $\mathrm{vMF}(\mu_c, \kappa_c)$ centered at prototype μ_c. For any two distinct classes c, c', let $\theta_{cc'} = \arccos(\mu_c^\top \mu_{c'})$ denote the inter-class angular distance, and let $\theta_c, \theta_{c'}$ be the angular radii such that with probability at least $1-\epsilon$, a sample from class c (or c') lies within an angle θ_c (or $\theta_{c'}$) from its prototype. Then, to ensure class-level separability with high probability, it suffices that:*

$$\theta_{cc'} \geq \theta_c + \theta_{c'} + m, \quad (3)$$

where m is a desired minimal angular margin.

Moreover, with $\theta_c \leq \arccos\left(1 - \frac{\log(1/\epsilon)}{\kappa_c}\right)$, this implies the following bound on prototype similarity:

$$\mu_c^\top \mu_{c'} \leq \cos\Big(\arccos\left(1 - \frac{\log(1/\epsilon)}{\kappa_c}\right) + \arccos\left(1 - \frac{\log(1/\epsilon)}{\kappa_{c'}}\right) + m\Big). \quad (4)$$

Implication. This result reveals how ACL implicitly regulates a trade-off between intra-class density (governed by κ_c) and inter-class margins (reflected in $\theta_{cc'}$). By preserving temporal embedding diversity and reducing prototype collapse, ACL maintains high intra-class variance, which in turn necessitates wider angular separation between class centroids. A detailed proof is provided in Appendix.

Integrated Refinement

The integrated refinement presents a reliable way to leverage the accumulated knowledge to guide the discovery of novel categories in unlabeled data. Dynamically estimating the number of classes is a critical challenge, as opposed to fixing a prior constant as adopted by most existing methods. However, the inherent error between the estimated and ground-truth is unavoidable, and relying solely on the estimated value as the number of clusters may lead to suboptimal convergence. Given that the exact number of categories cannot be predicted, it is reasonable to assume that it falls within a certain range. To address this, we introduce a cosine-based perturbation mechanism as a relaxation margin. This mechanism transforms the task of predicting the number of classes as a scalar into one of 'capturing' the ground truth within the perturbation range, thereby effectively mitigating the impact of estimation errors.

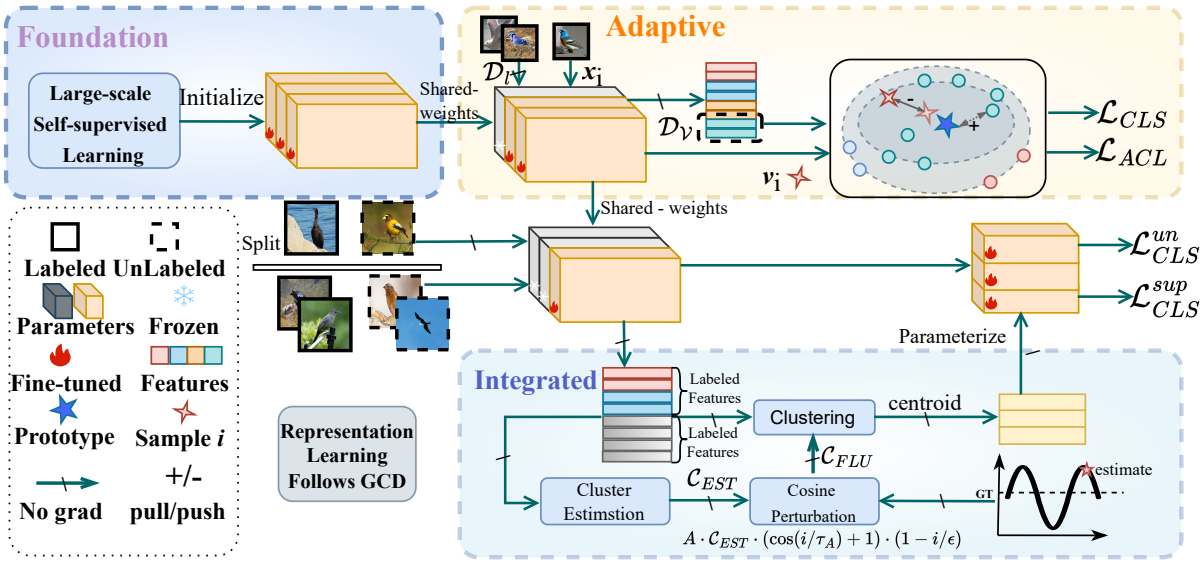

Figure 2: Overview of the proposed FAIR approach. The process begins with the pretrained weights (`Foundation`). Next, the `Adaptive` tuning step utilizes labeled data to refine the model's representation, enhancing its individuality by increasing intra-class distances. This is followed by `Integrated Refinement`, which aims to estimate the number of categories and learn unlabeled data representations. A cosine perturbation is specifically introduced to fine-tune the category count. Hierarchical clustering is also applied to determine cluster centers, which are parameterized and then used as initial weights for the classifier.

Cosine Perturbation: Estimating the number of categories inherently involves uncertainty, which can introduce bias, especially in parametric classification methods. Here, we provide a gradient-based explanation for why FAIR introduces the proposed cosine perturbation in the context of category number estimation. The widely adopted self-distillation loss can be expressed as follows:

$$\mathcal{L}_{CE} = \sum \mathbb{Q}(k|x) \log \mathbb{P}(k|x) = \sum \mathcal{T}(\mathbb{P}(k|x)) \log \mathbb{P}(k|x), \tag{5}$$

where $k = 1, 2 \ldots, \mathcal{C}$ and $\mathbb{P}(Y = k|x)$ represents the probability distribution, $\mathcal{T}$ is designed to compute pseudo-labels $\mathbb{Q}(Y = k|x)$.

Now, suppose the model is composed of a single fully-connected layer. This simplification helps reduce the complexity of the derivatives that the model could have. Then, we have $\mathbb{P}(k|\mathbf{x}) = \texttt{softmax}(\mathbf{x}\mathbf{W}^\top)_k$, and $\sum_{k=1}^{\mathcal{C}} \mathbb{Q}(k|\mathbf{x}) = 1$, where the involved weights W can be updated using the chain rule as follows:

$$W_k = W_k - \eta \cdot \nabla_{W_k} \mathcal{L}_{CE} = W_k - \eta \cdot (p_k - q_k) x^\top, \tag{6}$$

where p_k and q_k are the k-th values of original probability distribution $\mathbb{P}$ and the target probability distribution $\mathbb{Q}$, respectively.

The above equations suggest that the gradient direction is significantly influenced by the magnitude gap between $\mathbf{p}_k$ and $\mathbf{q}_k$. In practice, this often causes multiple category prototypes, particularly those for novel classes in unlabeled data, to collapse into a single representation. This leads the model to underestimate the number of categories in the absence of prior knowledge about category count, which in turn degrades the quality of prototype generation and significantly hinders subsequent self-distillation learning. To address this issue, we introduce a cosine perturbation strategy that dynamically adjusts the estimated number of categories after each iteration, reducing the risk of the model being misled by inaccurate estimates.

The estimation of the number of categories, denoted as $\mathcal{C}_{EST}$, begins by traversing values within a certain range and selecting the one with the highest clustering accuracy (Choi, Kang, and Cho 2024). To enhance this estimation, we introduce a cosine perturbation that iterates with the training epochs. This involves adding a fluctuation, $\mathcal{C}_{FLU}$, to the estimated number of categories, $\mathcal{C}_{EST}$, which varies with the training progress. The fluctuation, $\mathcal{C}_{FLU}$, is calculated using a cosine function, with its amplitude linearly decreasing as the training progresses:

$$\mathcal{C}_{FLU} = A \cdot \mathcal{C}_{EST} \cdot (\cos(i/\tau_A) + 1) \cdot (1 - i/\epsilon), \tag{7}$$

where i denotes the epoch currently in progress, ϵ is the total number of training rounds, A is a hyperparameter controlling the magnitude, and τ_A determines the rate of cycle change.

Prototype Generation: Learning representative prototypes plays a crucial role in discovering novel categories. A common method for achieving this is by incorporating nearest neighbor information into contrastive learning. However, this approach carries the risk of misclassifying neighbors that belong to different categories, potentially leading to incorrect associations. To address this, we propose a prototype generation method that involves clustering features and using the cluster centers as initial parameters for the classifier.

We perform hierarchical clustering using the Ward linkage criterion on the feature set $\mathcal{D}_\mathcal{V}$ at the end of each training epoch. This process iteratively merges the two clos-

est clusters until the number of clusters matches the predefined number $\mathcal{C} = \mathcal{C}_{EST} + \mathcal{C}_{FLU}$. The cluster centers are then computed based on the clustering results $\{\mathcal{D}_i\}_{i=1}^{\mathcal{C}} = \texttt{clustering}(\mathcal{D}_{\mathcal{V}})$ and used as the initial classifier parameters $\boldsymbol{W}$ for the subsequent training epoch.

To ensure correct alignment between clusters and labels, we use the Hungarian optimal assignment algorithm (Kuhn 1955). This algorithm enables the optimal matching of clusters to known categories, resulting in the final classifier parameters $\boldsymbol{W}$, which can be expressed as,

$$\boldsymbol{W} = \left[\frac{1}{|\mathcal{D}_{\mathrm{p}(1)}|} \sum_{\boldsymbol{v}_j \in \mathcal{D}_{\mathrm{p}(1)}} \boldsymbol{v}_j, \cdots, \frac{1}{|\mathcal{D}_{\mathrm{p}(\mathcal{C})}|} \sum_{\boldsymbol{v}_j \in \mathcal{D}_{\mathrm{p}(\mathcal{C})}} \boldsymbol{v}_j \right], \quad (8)$$

where $\mathrm{p}(i)$ denotes the optimal permutation between predicted clusters and labeled categories. It is important to note that only labels corresponding to labeled data are available during the clustering and label assignment process. The remaining clusters, after the assignment of those corresponding to labeled data, are randomly ordered. In essence, our approach converts nearest neighbor information into learnable classifier parameters $\boldsymbol{W}$, ensuring that prototype generation guides the discovery of novel classes.

Composite Loss

We adopt a variant of weighted self-distillation inspired by (Huang, Zhang, and Zhang 2024), which emphasizes high-confidence samples. Unlike (Huang, Zhang, and Zhang 2024), where weights are derived from EMA-based pseudo-labels, our pseudo-labels are sharpened predictions as in SimGCD. Given that most predictions are already one-hot-like—due to our prototype initialization from clustering centroids—the original weighting scheme becomes ineffective. To address this, we instead use the maximum predicted probability $w_i = \max_k p_k^{(i)}$ as a confidence-aware weight, ensuring adaptiveness under our setting. The unsupervised classification loss is:

$$\mathcal{L}_{CLS}^{un} = -\frac{1}{\sum w_i} \sum_{i=1}^{N} w_i \sum_{k=1}^{\mathcal{C}_l} q_k^{(i)} \log p_k^{(i)}, \quad (9)$$

where $\mathcal{C}_l$ is the number of clusters, and $p_k^{(i)}$ denotes the predicted probability of sample i belonging to cluster k.

To complement this, we include cross-entropy loss $\mathcal{L}CLS^{sup}$ for labeled data, mean-entropy maximization $\mathcal{L}MMR$ (Wen, Zhao, and Qi 2023; Assran et al. 2022) to promote uncertainty on unlabeled samples, and representation loss $\mathcal{L}_{REP}$ (Vaze et al. 2022) from an auxiliary MLP head. The total loss is:

$$\mathcal{L} = \lambda \mathcal{L}_{CLS}^{sup} + (1-\lambda)\mathcal{L}_{CLS}^{un} + \mathcal{L}_{REP} + \rho \mathcal{L}_{MMR}, \quad (10)$$

where λ and ρ are balancing coefficients.

Experiments

Experimental Setup

Datasets and Evaluation Protocol. The performance of proposed method is evaluated across multiple benchmark datasets (CIFAR-10/100 (Krizhevsky, Hinton et al. 2009), ImageNet-100 (Russakovsky et al. 2015), CUB-200 (Reed et al. 2016), Stanford-Cars (Krause et al. 2013) and Herbarium19 (Tan et al. 2019)) to demonstrate the effectiveness. The partitioning of labeled and unlabeled sets follows the the criteria described in (Vaze et al. 2022). Unless otherwise specified, all experimental results are obtained under the assumption that the number of categories is known, and clustering accuracy (ACC) is used as the evaluation metric.

Implementation Details. We utilize a ViT-B/16 model (Dosovitskiy et al. 2021) pre-trained with DINO (Caron et al. 2021) as the backbone, using the [CLS] token as the global visual feature for prototype generation and classification. The hyper-parameters λ is set to 0.35, like in (Vaze et al. 2022; Wen, Zhao, and Qi 2023). The proposed ESC only fine-tunes the last block of the backbone. The balance hyper-parameter α is set to 0.5 to adjust the intra-class variance of the labeled data, and the fluctuation range A is set to 0.2 to control the number of prototypes generated by the model. τ_A is set to 1 to control the speed of fluctuations. To accelerate prototype generation, a random sampling method is employed to select 20,000 samples for prototype generation in large-scale coarse-grained datasets. All our experiments are conducted using a NVIDIA GeForce RTX 3090 GPU.

Comparison with State-of-the-art

To demonstrate the performance of the proposed algorithm under both known and unknown category numbers, we compare our method with the state-of-the-art approaches (Han et al. 2021; Zhao, Wen, and Han 2023; Fini et al. 2021; Cao, Brbic, and Leskovec 2021; Vaze et al. 2022; Pu, Zhong, and Sebe 2023; Wen, Zhao, and Qi 2023; Zhang et al. 2023; Choi, Kang, and Cho 2024; Wang, Vaze, and Han 2024; Chiaroni et al. 2023) assuming known number of categories in Table 1(a). Additionally, as shown in Table 1(b), we evaluate it against several algorithms (Vaze et al. 2022; Zhao, Wen, and Han 2023; Chiaroni et al. 2023; Choi, Kang, and Cho 2024) that account for the more realistic scenario where the number of categories is unknown,to evaluate its performance in practical applications.

Comparison on Coarse-grained Datasets. We validate the effectiveness of FAIR on three coarse-grained datasets, demonstrating its more balanced performance in predicting both old and new classes, as shown in Figure 3. FAIR significantly reduces the predictive bias between novel and old classes, enhancing its potential for practical applications. On CIFAR-100, FAIR achieves an accuracy of 86.6% on New classes, outperforming the second-best method by 8.8%. Although it do not achieve the state-of-the-art result on CIFAR-10, it ranks second, just 0.3% behind promptCAL, which only excells on CIFAR-10, with discrepancy in ACC between Old and New classes of just 1.4%, as shown in Table 1(a).

In scenarios where the category number is unknown and the estimated number of categories is shown in Table 2, FAIR outperforms all existing methods across all generic datasets as depicted in Table 1(b). Furthermore, its performance remains close to that of algorithms with access to the ground-truth number of categories. On CIFAR-100, even without known number of categories, FAIR surpasses all methods that used

Methods	CIFAR-10			CIFAR-100			ImageNet-100			CUB-200			Stanford-Cars			Herbarium19		
	All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New
(a) The ground-truth number of categories is known																		
RankStats+	46.8	19.2	60.5	58.2	77.6	19.3	37.1	61.6	24.8	33.3	51.6	24.2	28.3	61.8	12.1	27.9	55.8	12.8
UNO+	68.6	**98.3**	53.8	69.5	80.6	47.2	70.3	95.0	57.9	35.1	49.0	28.1	35.5	70.5	18.6	28.3	53.7	14.7
ORCA	81.8	86.2	79.6	69.0	77.4	52.0	73.5	92.6	63.9	35.3	45.6	30.2	23.5	50.1	10.7	22.0	31.8	17.1
GCD	91.5	97.9	88.2	73.0	76.2	66.5	74.1	89.8	66.3	51.3	56.6	48.7	39.0	57.6	29.9	35.4	51.0	27.0
GPC	92.2	98.2	89.1	77.9	85.0	63.0	76.9	94.3	71.0	55.4	58.2	53.1	42.8	59.2	32.8	-	-	-
PIM	94.7	97.4	93.3	78.3	84.2	66.5	83.1	95.3	77.0	62.7	75.7	56.2	43.1	66.9	31.6	42.3	56.1	34.8
DCCL	96.3	96.5	96.9	75.3	76.8	70.2	80.5	90.5	76.2	63.5	60.8	64.9	43.1	55.7	36.2	-	-	-
SimGCD	97.1	95.1	98.1	80.1	81.2	77.8	83.0	93.1	77.9	60.3	65.6	57.7	53.8	71.9	45.0	44.0	58.0	36.4
PromptCAL	**97.9**	96.6	98.5	81.2	84.2	75.3	83.1	92.7	78.3	62.9	64.4	62.1	50.2	70.1	40.6	37.0	52.0	28.9
μGCD	-	-	-	-	-	-	-	-	-	65.7	68.0	64.6	56.5	68.1	50.9	45.8	**61.9**	39.2
selEx	95.9	**98.1**	94.8	82.3	85.3	76.3	83.1	93.6	77.8	73.6	75.3	72.8	58.5	75.6	50.3	39.6	54.9	31.3
LegoGCD	97.1	94.3	98.5	81.8	81.4	82.5	**86.3**	94.5	**82.1**	63.8	71.9	59.8	57.3	75.7	48.4	45.1	57.4	38.4
CMS	**-**	-	-	82.3	**85.7**	75.5	84.7	**95.6**	79.2	68.2	**76.5**	64.0	56.9	76.1	47.6	36.4	54.9	26.4
SPTNet	97.3	95.0	**98.6**	81.4	84.3	75.6	85.4	93.2	81.4	65.8	68.8	65.1	59.0	**79.2**	49.3	43.4	58.7	35.2
ProtoGCD	97.3	95.3	98.2	81.9	82.9	80.0	84.0	92.2	79.9	63.2	68.5	60.5	53.8	73.7	44.2	44.5	59.4	36.5
FAIR	97.6	96.7	98.1	**84.2**	83.0	**86.6**	86.0	94.4	81.7	**73.7**	75.3	**73.0**	**63.6**	70.4	**60.4**	**49.1**	59.1	**43.8**
(b) The ground-truth number of categories is unknown																		
GCD	-	-	-	70.8	77.6	57.0	77.9	91.1	51.1	56.4	48.4	48.7	39.1	58.6	29.7	37.2	51.7	29.4
GPC	90.6	**98.2**	87.1	75.7	**84.7**	60.9	75.7	93.4	66.8	52.1	55.4	45.7	38.9	58.9	28.6	-	-	-
PIM	94.7	97.4	93.3	75.6	81.6	63.6	83.0	95.3	76.9	62.0	75.7	55.1	42.4	65.3	31.3	42.0	55.5	34.7
CMS	**-**	-	-	79.6	83.2	72.3	81.3	**95.6**	74.2	64.4	68.2	62.4	51.7	68.9	43.4	37.4	56.5	27.1
ProtoGCD	-	-	-	81.9	82.9	80.0	84.8	90.9	81.8	61.4	66.2	58.8	52.7	71.1	43.8	-	-	-
FAIR	**97.4**	94.9	**98.7**	**83.2**	82.6	**84.4**	**84.8**	92.9	**80.7**	**72.9**	**71.1**	**73.8**	**60.9**	**72.4**	**55.4**	**47.6**	**56.9**	**42.7**

Table 1: Results on six benchmark datasets. Bolding indicates the best result and underlining indicates the second ranked result.

Methods	CIFAR-10	CIFAR-100	ImageNet-100	CUB-200	Stanford-Cars	Herbarium19
Ground truth	10	100	100	200	196	683
GCD(Vaze et al. 2022)	9	100	109	231	230	520
DCCL(Pu, Zhong, and Sebe 2023)	14	146	129	172	192	-
PIM(Chiaroni et al. 2023)	10	95	102	227	169	563
GPC(Zhao, Wen, and Han 2023)	9	100	103	212	201	-
CMS(Choi, Kang, and Cho 2024)	-	97	116	170	156	666
ProtoGCD(Ma et al. 2025)	10	100	106	211	205	603
FAIR	10	100	97	191	184	567

Table 2: Estimated number of classes on six benchmark datasets.

the number of categories as a prior.

Comparison on Fine-grained Datasets. The experiments are further extended to cover three fine-grained datasets, as shown in Table 1. FAIR achieves the best performance on both All and New classes. Specifically, it reaches an accuracy of 73.7% on CUB-200, which is 5.5% higher than the CMS. Similarly, FAIR performs exceptionally well on Stanford-Cars and Herbarium19, with improvements of 4.6% and 5.1%, respectively, over the SOTA methods on All classes. Particularly for New classes, FAIR outperforms the existing methods by 11.4% and 7.4% on the two datasets. These results demonstrate that FIAR significantly enhances performance on New classes,due to the effective generalization of knowledge learned from labeled data to unlabeled data through the designed multi-stage strategy.

Even when the categories number is unknown, FAIR still performs excellent, surpassing SOTA algorithms across the board, as depicted in Table 1(b). Notably, even without the known number of categories, FAIR achieves improvements in ACC compares with these methods that consider the number of categories as prior, with gains of 4.7%, 1.9%, and 2.6% on CUB-200, Stanford-Cars, and Herbarium19, respectively. The algorithm effectively handles the scenario of an unknown number of categories, avoiding the performance degradation typically caused by the predicted category number, making it more suitable for real-world applications.

index	component					CIFAR-100			CUB-200		
	AFT	ACL	GP	COS	w.	All	Old	New	All	Old	New
base						80.1	81.2	77.8	60.3	65.6	57.7
(a)	✓					67.3	73.5	54.9	49.4	39.9	54.6
(b)		✓				64.8	76.5	41.5	47.2	38.4	51.6
(c)	✓	✓				69.5	76.5	55.5	52.5	61.2	48.2
(d)			✓			81.4	81.6	80.9	65.1	69.9	62.7
(e)			✓	✓		81.6	82.7	79.9	66.0	70.4	63.9
(f)	✓	✓	✓		✓	83.2	82.6	84.4	71.7	73.6	70.7
(g)	✓		✓	✓	✓	82.5	83.0	81.4	71.6	70.1	71.6
(h)	✓	✓	✓	✓		83.1	83.2	82.7	72.8	74.9	71.7
(i)	✓	✓	✓	✓	✓	84.2	83.0	86.6	73.7	75.3	73.0

Table 3: Effectiveness of each component of framework.

Ablation Study

In this section, we systematically evaluate the effectiveness of the FAIR's individual modules and assess their impact on clustering performance. Seven sets of experiments are designed, each aims at validating the performance of different modules on both CUB-200 and the CIFAR-100. These two

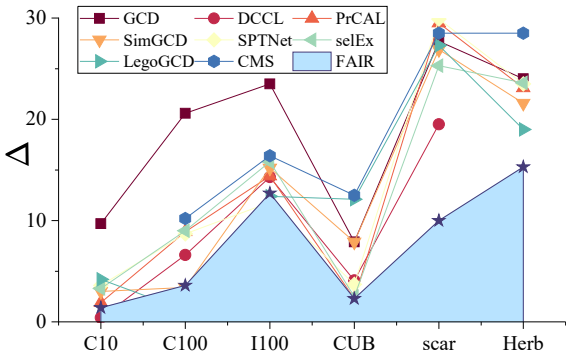

Figure 3: Comparison of absolute errors of new and old category accuracies.

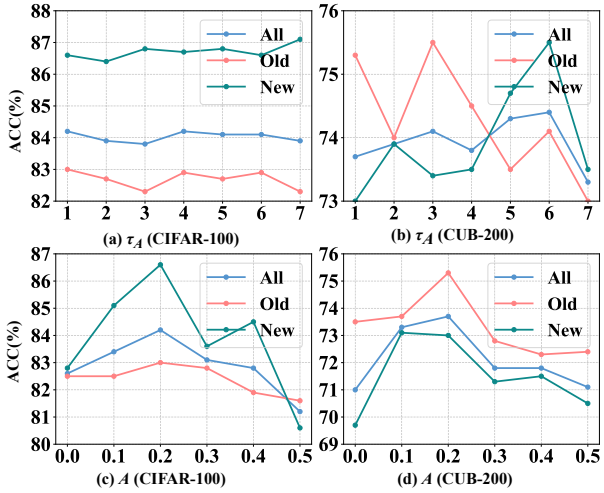

Figure 4: Impact of hyper-parameters. The ACC on CIFAR-100 and CUB-200 are reported.

datasets are representative of typical fine-grained and coarse-grained classification tasks, respectively.

Effectiveness of Adaptive Fine-tuning and Adaptive Contrastive Learning. The purpose of AFT is to prevent unlabeled data from interfering with class information derived from labeled data. As illustrated in Table 3, experiment (a)(b)(d)(e) show that AFT must be combined with the proposed prototype generation to function optimally. Our analysis suggests that the performance drop without prototype generation stems from their use of randomly initialized prototypes, where small-step optimization fails to achieve convergence. In contrast, our approach first obtains a better initialization through clustering and then applies gradual optimization, leading to more stable convergence and improved performance.

Effectiveness of Cosine Perturbation. By comparing the results of experiment (d) and (e), as well as (f) and (g), we conclude that applying a positive cosine perturbation to the estimated number of categories can mitigate the collapse of class representations into a few dominant categories. This phenomenon is similar to the effects observed in over-clustering methods commonly used in unsupervised representation learning tasks. In the supplementary materials, we provide further analysis and detailed explanations, supported by visualizations, of the rationale behind introducing the positive cosine perturbation.

In this section, we conduct parameter testing by dividing validation subsets from a coarse-grained dataset, CIFAR-100, and a fine-grained dataset, CUB-200. We discuss the impact of two hyper-parameters in our work: The speed rate of cosine cyclicality τ_A and the fluctuation range of the cosine perturbation A.

Effectiveness of Prototype Generation. By comparing the results of experiment (d) with the baseline, we found that prototype generation significantly improves the model's performance on new categories. Based on the analysis of experiment f, we observed that after AFT captures higher-level general semantic features, the prototype generation method is able to produce more representative prototypes for each category, leading to a notable enhancement in performance.

Effectiveness of Weighted Self-distillation. As shown in Table 3, weighted self-distillation learning outperforms traditional self-distillation learning across all metrics. This is because, compared to traditional self-distillation, weighted self-distillation significantly increases the trust in high-confidence representations (high-weighted), rather than uniformly optimizing all representations based solely on the loss function, which effectively mitigates the interference from hard samples in unlabeled data.

Hyper-Parameter Analysis

Impact of the Speed Rate of Cosine Cyclicality: To evaluate the effect of the speed rate of cosine cyclicality, we set the range of τ_A from 1 to 7, as illustrated in Figure 4(a) and Figure 4(b). Experimental results show that CIFAR-100 is not sensitive to τ_A, whereas on CUB-200 it can be improved by almost 1% by controlling τ_A.

Impact of the Fluctuation Range of Cosine Perturbation: Figure 4(c) and Figure 4(d) demonstrate the influence of the fluctuation range, where both excessively large and small ranges fail to achieve satisfactory performance, particularly in the case of excessive ranges. Consequently, we set the maximum fluctuation range to 20% of the number of new categories in all experiments.

Conclusion

To address the issue of unlabeled data interfering with the information of labeled data, we propose the Foundation-Adaptive Integrated Refinement (FAIR), which aims to consolidate the shared knowledge between labeled and unlabeled data, effectively mitigating the interference caused by unlabeled data. Furthermore, to overcome the limitation of existing methods that assume the number of categories as prior knowledge, we propose to leverage clustering algorithm to obtain cluster centers. These centers are then parameterized and updated to retain the neighborhood information inherent in the clustering process while also benefiting from the adaptability of parametric methods. Extensive experiments demonstrate that FAIR outperforms existing methods.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under the Grants No. 62371235 and No. U25A20444, in part by the Key Research and Development Plan of Jiangsu Province under Grant No. BE2023008-2.

References

Asano, Y. M.; Rupprecht, C.; and Vedaldi, A. 2020. Self-labelling via simultaneous clustering and representation learning. In *International Conference on Learning Representations (ICLR)*.

Assran, M.; Caron, M.; Misra, I.; Bojanowski, P.; Bordes, F.; Vincent, P.; Joulin, A.; Rabbat, M.; and Ballas, N. 2022. Masked Siamese Networks for Label-Efficient Learning. In Avidan, S.; Brostow, G.; Cissé, M.; Farinella, G. M.; and Hassner, T., eds., *Computer Vision – ECCV 2022*, 456–473.

Banerjee, A.; Dhillon, I. S.; Ghosh, J.; Sra, S.; and Ridgeway, G. 2005. Clustering on the Unit Hypersphere using von Mises-Fisher Distributions. *Journal of Machine Learning Research*, 6(9).

Cao, K.; Brbic, M.; and Leskovec, J. 2021. Open-world semi-supervised learning. In *International Conference on Learning Representations (ICLR)*.

Caron, M.; Misra, I.; Mairal, J.; Goyal, P.; Bojanowski, P.; and Joulin, A. 2020. Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. In *Advances in Neural Information Processing Systems*, volume 33, 9912–9924.

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging Properties in Self-Supervised Vision Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 9650–9660.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A Simple Framework for Contrastive Learning of Visual Representations. In *Proceedings of the 37th International Conference on Machine Learning*, 1597–1607.

Chen, X.; and He, K. 2021. Exploring Simple Siamese Representation Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15750–15758.

Chi, H.; Liu, F.; Han, B.; Yang, W.; Lan, L.; Liu, T.; Niu, G.; Zhou, M.; and Sugiyama, M. 2022. Meta discovery: Learning to discover novel classes given very limited data. In *International Conference on Learning Representations (ICLR)*.

Chiaroni, F.; Dolz, J.; Masud, Z. I.; Mitiche, A.; and Ben Ayed, I. 2023. Parametric information maximization for generalized category discovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 1729–1739.

Choi, S.; Kang, D.; and Cho, M. 2024. Contrastive Mean-Shift Learning for Generalized Category Discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 23094–23104.

Cuturi, M. 2013. Sinkhorn Distances: Lightspeed Computation of Optimal Transport. In *Advances in Neural Information Processing Systems*, volume 26.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ICLR*.

Fini, E.; Sangineto, E.; Lathuilière, S.; Zhong, Z.; Nabi, M.; and Ricci, E. 2021. A Unified Objective for Novel Class Discovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 9284–9292.

Fukunaga, K.; and Hostetler, L. 1975. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on Information Theory*, 21(1): 32–40.

Han, K.; Rebuffi, S.-A.; Ehrhardt, S.; Vedaldi, A.; and Zisserman, A. 2021. AutoNovel: Automatically Discovering and Learning Novel Visual Categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*.

Han, K.; Vedaldi, A.; and Zisserman, A. 2019. Learning to Discover Novel Visual Categories via Deep Transfer Clustering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum Contrast for Unsupervised Visual Representation Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Huang, L.; You, S.; Zheng, M.; Wang, F.; Qian, C.; and Yamasaki, T. 2022. Learning Where to Learn in Cross-View Self-Supervised Learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14431–14440.

Huang, L.; Zhang, C.; and Zhang, H. 2024. Self-Adaptive Training: Bridging Supervised and Self-Supervised Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(3): 1362–1377.

Krause, J.; Stark, M.; Deng, J.; and Fei-Fei, L. 2013. 3d object representations for fine-grained categorization. In *Proceedings of the International Conference on Computer Vision workshops*, 554–561.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images. Technical report, University of Toronto.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, volume 25.

Kuhn, H. W. 1955. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2): 83–97.

Ma, S.; Zhu, F.; Zhang, X.-Y.; and Liu, C.-L. 2025. ProtoGCD: Unified and Unbiased Prototype Learning for Generalized Category Discovery. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(7): 6022–6038.

Pu, N.; Zhong, Z.; and Sebe, N. 2023. Dynamic Conceptional Contrastive Learning for Generalized Category Discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7579–7588.

Reed, S.; Akata, Z.; Lee, H.; and Schiele, B. 2016. Learning deep representations of fine-grained visual descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 49–58.

Rosvall, M.; and Bergstrom, C. T. 2008. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105: 1118–1123.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115: 211–252.

Song, K.; Xie, J.; Zhang, S.; and Luo, Z. 2023. Multi-Mode Online Knowledge Distillation for Self-Supervised Visual Representation Learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11848–11857.

Tan, K. C.; Liu, Y.; Ambrose, B.; Tulig, M.; and Belongie, S. 2019. The herbarium challenge 2019 dataset. *arXiv preprint arXiv:1906.05372*.

Vaze, S.; Han, K.; Vedaldi, A.; and Zisserman, A. 2022. Generalized Category Discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7492–7501.

Wang, G.; Tang, Y.; Lin, L.; and Torr, P. H. 2022a. Semantic-Aware Auto-Encoders for Self-supervised Representation Learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9654–9665.

Wang, H.; Vaze, S.; and Han, K. 2024. SPTNet: An Efficient Alternative Framework for Generalized Category Discovery with Spatial Prompt Tuning. In *International Conference on Learning Representations (ICLR)*.

Wang, J.; Ma, Z.; Nie, F.; and Li, X. 2022b. Progressive Self-Supervised Clustering With Novel Category Discovery. *IEEE Transactions on Cybernetics*, 52(10): 10393–10406.

Wen, X.; Zhao, B.; and Qi, X. 2023. Parametric Classification for Generalized Category Discovery: A Baseline Study. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 16590–16600.

Zhang, S.; Khan, S.; Shen, Z.; Naseer, M.; Chen, G.; and Khan, F. S. 2023. PromptCAL: Contrastive Affinity Learning via Auxiliary Prompts for Generalized Novel Category Discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3479–3488.

Zhao, B.; Wen, X.; and Han, K. 2023. Learning Semi-supervised Gaussian Mixture Models for Generalized Category Discovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 16623–16633.

Zhong, Z.; Fini, E.; Roy, S.; Luo, Z.; Ricci, E.; and Sebe, N. 2021. Neighborhood Contrastive Learning for Novel Class Discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10867–10875.