Learning a Fix and Explore Framework for Continuous Generalized Category Discovery

Chunming Li[1], **Shidong Wang**[2], **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
chunmingli@njust.edu.cn, shidong.wang@newcastle.ac.uk, zhanghf@njust.edu.cn

Abstract

To address the limitations of transductive learning in evolving real-world scenarios where unknown categories may continuously emerge, Continual Generalized Category Discovery (C-GCD) presents a novel paradigm that extends conventional category discovery frameworks. Unlike traditional static learning environments, C-GCD requires models to incrementally discover novel categories across multiple operational phases while maintaining discrimination capabilities for previously learned classes, posing significant challenges in balancing stability and plasticity. Prior approaches typically employ parameter-level knowledge distillation from historical models to alleviate catastrophic forgetting, which effectively preserves prior knowledge and optimizes computational efficiency. However, our analysis reveals that the persistent availability of samples from previous stages enables more sophisticated knowledge preservation strategies. Specifically, we present a Fix and Explore strategy that employs distinct learning methodologies for different types of potential data, aiming to preserve the features of old categories as much as possible and gradually exploring the potential distribution of new class latent spaces, we can enhance the model's ability to discover novel categories. This paper investigates this effect and introduces a novel heuristic paradigm to solve the C-GCD problem, called Fix and Explore (FaE), which aims to provide sufficient imaginative space for new classes while preserving the classification ability for old tasks. We conducted experiments across multiple datasets and performed detailed comparisons. The results demonstrate that our method achieves state-of-the-art performance at each stage across all datasets.

Introduction

In recent years, the rapid development of deep supervised learning has been attributed mainly to the accumulation of large-scale datasets (Krizhevsky, Sutskever, and Hinton 2012). With sufficient annotated data, deep learning models are capable of surpassing human-level performance in many important computer vision tasks (He et al. 2016, 2022; Dosovitskiy et al. 2021). Annotating large-scale data is labor-intensive and impractical due to the vast number of categories in open-world scenarios. To mitigate this, New Category Discovery (NCD) (Han, Vedaldi, and Zisserman 2019; Gu et al. 2023; Li et al. 2023; Fini et al. 2021) explores leveraging labeled data to cluster or classify unlabeled samples from novel categories. Generalized Category Discovery (GCD) (Vaze et al. 2022; Wen, Zhao, and Qi 2023; Pu, Zhong, and Sebe 2023; Rastegar et al. 2024) further relaxes assumptions by allowing unlabeled samples to originate from both seen and unseen classes. However, both NCD and GCD adopt a transductive setting (Vaze et al. 2022), restricting them to single-stage inference and limiting their ability to detect emerging categories dynamically (Ma et al. 2024; Cendra, Zhao, and Han 2024). Recent extensions—Continuous NCD (C-NCD) and Continuous GCD (C-GCD)—address this by enabling multi-stage discovery. C-NCD assumes each stage contains entirely new categories, while C-GCD permits samples from previously seen classes.

Like typical continual learning, both C-GCD and C-NCD face the severe issue of catastrophic forgetting (Roy et al. 2022; Zhang et al. 2022). Common approaches to mitigate forgetting include data replay, dynamic networks, and knowledge distillation (Zhou et al. 2024). Data replay and dynamic networks tend to outperform knowledge distillation in terms of performance but they face challenges related to privacy concerns or storage limitations.

As shown in Fig. 1, the model is required to recognize both old classes from previous stages and new classes it has never seen before, while facing a large number of unlabeled samples at each stage (Ma et al. 2024). To resist forgetting, Happy (Ma et al. 2024) proposed cosine distance knowledge distillation for all samples in each session. PA-CGCD (Kim et al. 2023) proposed using proxies to classify samples into new and old classes, performing feature distillation only on old classes. GM (Zhang et al. 2022) dynamically classifies all samples as coming from novel categories and applies cosine distance loss knowledge distillation to all samples. Compared to knowledge distillation on all samples, this approach may be more conducive to discovering new categories, which inspired us to explore different distillation strategies for C-GCD.

This paper addresses the C-GCD task of "More learning stages with more new classes" proposed in Happy (Ma et al. 2024), analyzing the impact of different distillation strategies and introducing the Fix and Explore (FaE) framework.

*corresponding Author.

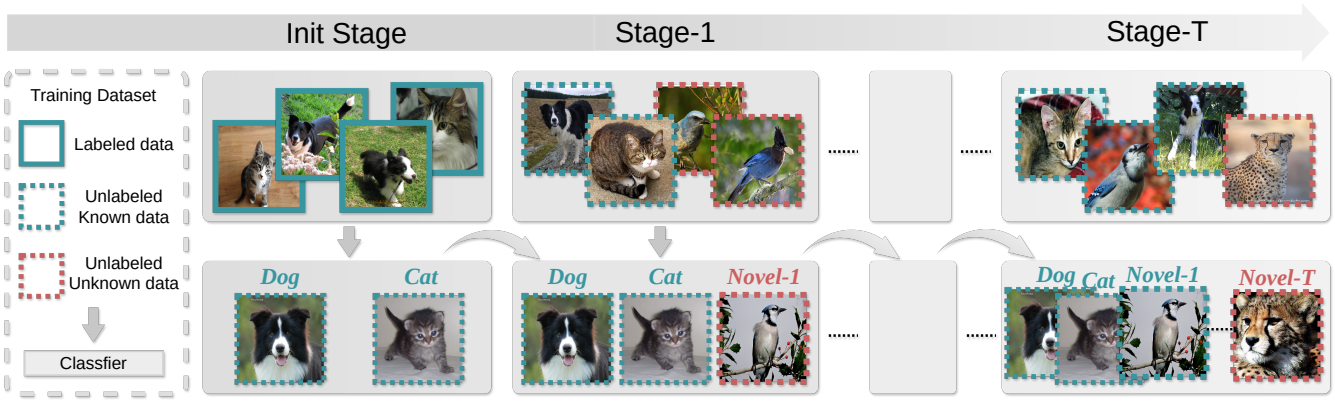


Figure 1: Illustration of C-GCD setting. In the initialization phase, the model is pre-trained on labeled data. During the continual learning phase, the model receives unlabeled images belonging to both known and novel classes. Finally, the model is evaluated on a test set comprising all encountered classes.

While existing methods mitigate catastrophic forgetting (Ma et al. 2024), they still suffer from negative optimization in later stages due to growing coupling between new and old class feature spaces. We attribute this to the lack of mechanisms for preserving old-class information and exploring new-class potential. L2 norm-based distillation ensures strict feature alignment, whereas graph-based distillation allows slight deviations while preserving structure (Zhou et al. 2024). Building on these insights, FaE heuristically distinguishes potential old and new samples via clustering, applying distinct distillation strategies to extract information of different hardness levels from the previous model. It further employs regularization to reduce prediction bias and an improved Hardness-aware Reweighed Feature Reply to alleviate forgetting. To enhance novel category discovery, FaE observes centroid drift in new classes during training and introduces a graph regularization based on visible categories, iteratively separating new from old distributions.

Our contributions can be summarized as follows:

- We propose a Fix and Explore framework FaE for C-GCD, which utilizes a **Fix** term to stabilize and an **Explore** term to better adapt the latent space for new categories.
- We introduce **Known-Aware Graph Regularization** and employ distinct distillation losses for data from different categories to reduce control over the feature space distribution of new categories and mitigate catastrophic forgetting.
- Extensive experiments on multiple datasets demonstrate that our method significantly outperforms others in both category discovery and forgetting mitigation.

Related Works

Category Discovery

Category Discovery addresses the challenge of classifying unlabeled samples from novel categories with limited labeled data. Novel Category Discovery (NCD) (Han, Vedaldi, and Zisserman 2019) assumes that unlabeled data only contain unseen classes, distinguishing it from semi-supervised learning. Early methods like DTC use clustering and prototypes, while AutoNovel (Han et al. 2020) adopts self-supervised pretraining and pseudo-labeling. Recent works extend NCD to tasks like semantic segmentation, incremental learning, and Generalized Category Discovery (GCD), which includes both known and novel classes. Representative approaches include PromptCAL (Zhang et al. 2023), SimGCD (Wen, Zhao, and Qi 2023), and SPTNet (Wang, Vaze, and Han 2024).

Continual Novel Category Discovery (C-NCD) (Joseph et al. 2022) further enables models to discover new classes from streaming data. It involves learning from labeled data and incrementally processing unlabeled samples. Techniques such as feature replay (FRoST (Roy et al. 2022)), self-supervised learning (MSc-iNCD (Liu et al. 2023b)), and Growing-Merging frameworks (GM (Zhang et al. 2022)) have been proposed. Other works like PA-CGCD (Kim et al. 2023) and MetaGCD (Wu et al. 2023) focus on mitigating forgetting via pseudo-labeling and meta-learning. PromptCCD (Cendra, Zhao, and Han 2024) employs dynamic prompting and Gaussian Mixture Prompting to enhance representation learning.

Graph Knowledge Distillation

Graph Knowledge Distillation (a.k.a. Relational Distillation) was first applied in GNNs (Liu et al. 2023a), utilizing a "Teacher-Student" framework to transfer the parameters of a high-capacity model to a lower-capacity model, thereby improving its performance. Similarly, in deep neural networks, many approaches suggest that student models can directly extract rich inter-sample relational knowledge learned by the teacher model through constructed graphs (Yim et al. 2017; Passalis, Tzelepi, and Tefas 2020). This ability to align inter-model distributions naturally lends itself to incremental learning. In incremental learning, graph knowledge distillation often involves constructing triplet relations (Zhou et al.

2024) (anchor, positive, and negative samples) and preserving the triplet relationship distances between the old and new models (Gao et al. 2022). Compared to traditional instance-based distillation in incremental learning, graph knowledge distillation reveals more structural information about the distribution (Zhou et al. 2024).

Methodology

Problem Definition

The C-GCD consists of two phases. During the initialization phase, the model is provided with a labeled training dataset $\mathcal{D}^0_{train} = \{(\mathbf{x}^0_i, \mathbf{y}^0_i)\}^{N_0}_{i=1}$, consisting of input-output pairs $(\mathbf{x}^0_i, \mathbf{y}^0_i)$ from a set of known categories $\mathcal{C}_0 = \{1, 2, ..., K^0\}$. The objective is for the model to learn generalizable representations from these categories. In the subsequent continual learning phase ($Stage$-$1 \rightarrow Stage$-T), at $Stage$-$t (t \in [1, T])$, the model is given a sequence of unlabeled datasets $\mathcal{D}^t_{train} = \{x^t_i\}^{N_t}_{i=1}$, containing samples from $\mathcal{C}_t = \{1, 2, ..., K^t\}$, where $\mathcal{C}_t = \mathcal{C}_{t-1} \cup \{K^{t-1} + 1, K^{t-1} + 2, ..., K^t\}$. For convenience of description, we define the number of classes in each phase of the continual learning process as K^{new}. Finally, after T discovery stages, the model will be tested on the dataset $\mathcal{D}_U$ containing all categories $\mathcal{C}_T$.

Distillation in Traditional Continual Learning

Assume that at $Stage$-t, we have the set of all samples $\{\mathbf{x}, \mathbf{y}\}$ from the current stage and the current stage model $\phi^t = \{\mathrm{f}^t, \mathrm{h}^t, \mathrm{g}^t\}$, which can be decomposed into a feature extractor $\mathrm{f}^t(\cdot)$, a classifier $\mathrm{h}^t(\cdot)$, and a projection head $\mathrm{g}^t(\cdot)$. In the process of incremental learning, it is common to leverage the model from the previous stage to guide the training of the current model and a process often referred to as knowledge distillation. The loss used in this context is typically divided into the loss for mitigating forgetting and the loss for learning new classes. Formally, this can be summarized as:

$$\mathcal{L} = \mathcal{L}_1(\phi^t(\mathbf{x}), \mathbf{y}) + \mathcal{L}_2(\phi^t(\mathbf{x}), \phi^{t-1}(\mathbf{x})), \tag{1}$$

where $\mathcal{L}_1$ is the cross-entropy loss for learning new samples, and $\mathcal{L}_2$ is distillation loss and used to prevent forgetting. $\mathcal{L}_2$ can take various forms, such as logit distillation or graph-based knowledge distillation. $\phi^{t-1}(\cdot)$ denotes the model from the previous stage, which is often frozen during the current training. The distillation loss used in methods like LwF is defined as $\mathcal{L}^{LwF}_2 = \sum -\mathcal{S}(\phi^{t-1}(\mathbf{x})) \log \mathcal{S}(\phi^t(\mathbf{x}))$, where $\mathcal{S}$ denotes Softmax activation. The main goal is to establish a mapping between the predicted probabilities of the old and new models, ensuring that their outputs converge to similar values.

Similarly, graphical knowledge distillation can be summarised as:

$$\mathcal{L}_2 = \mathrm{D}(\omega_s(\phi^t(\mathbf{x}_i), \phi^t(\mathbf{x}_j)), \omega_t(\phi^{t-1}(\mathbf{x}_i), \phi^{t-1}(\mathbf{x}_j))), \tag{2}$$

where $\omega_s(\cdot, \cdot)$ and $\omega_t(\cdot, \cdot)$ denote the similarity/distance function for student and teacher networks, and $\mathrm{D}(\cdot, \cdot)$ represents the objective function used to minimize the distance between the distributions. Compared to instance-level distillation, graph knowledge distillation offers greater flexibility. As illustrated in Fig. 2, the proposed FaE framework leverages diverse distillation strategies to preserve existing knowledge while actively exploring accessible information.

Initialization Phase

To ensure fairness in comparing results, the model is trained on $\mathcal{D}^0_{train}$ during the initialization phase and adapts the loss used in SimGCD (Wen, Zhao, and Qi 2023) and Happy (Ma et al. 2024). Precisely, to extract more generalizable features, cross-entropy loss $\mathcal{L}_{cls}$, supervised contrastive learning (Khosla et al. 2020) $\mathcal{L}^l_{con}$ and self-supervised contrastive learning (Chen et al. 2020) $\mathcal{L}^u_{con}$ can be formulated as:

$$\mathcal{L}_{cls} = \frac{1}{|B|} \sum_{i \in B} -\mathbf{y}_i \log \mathbf{p}_i, \tag{3}$$

$$\mathcal{L}^l_{\text{con}} = -\frac{1}{|B|} \sum_{i \in B} \frac{1}{|P(i)|} \sum_{q \in P(i)} \log \frac{\exp\left(\mathbf{z}_i^\top \mathbf{z}_q / \tau_c\right)}{\sum_{n \neq i} \exp\left(\mathbf{z}_i^\top \mathbf{z}_n / \tau_c\right)}, \tag{4}$$

$$\mathcal{L}^u_{\text{con}} = -\frac{1}{|B|} \sum_{i \in B} \log \frac{\exp\left(\mathbf{z}_i^\top \mathbf{z}'_i / \tau_c\right)}{\sum_{n \neq i} \exp\left(\mathbf{z}_i^\top \mathbf{z}_n / \tau_c\right)}, \tag{5}$$

where $\mathbf{p}_i = \mathrm{h}(\mathrm{f}(\mathbf{x}_i))$ is the logit prediction and $\mathbf{z} = \mathrm{g}(\mathrm{f}(\mathbf{x}_i))$ is the projection. The final loss for initialization phase can be summarized as:

$$\mathcal{L}_{init} = \mathcal{L}_{cls} + \mathcal{L}^l_{con} + \mathcal{L}^u_{con}. \tag{6}$$

Continual Learning Phase

Classifier Initialization. During the training phase, FaE needs to distinguish possible samples from new classes. Using a fixed threshold to identify new classes will make the model cumbersome and increase the difficulty of model tuning. To avoid setting a threshold and the potential instability caused by a randomly initialized classifier during training, we use clustering to initialize the classifier. Specifically, we cluster the data of each phase into K^t classes and select the top K^{new} centers that are farthest from the known class prototypes as the new class classifiers.

The initialized classifier then serves as the new class detector and participates in the subsequent training process. When the classifier predicts a sample as a new class, we temporarily treat it as a potential new class. For simplicity, in $Stage$-t, the samples detected as potentially new classes are marked as $B^t_{new} = \{\mathbf{x}^b_{new}\}^{|B^t_{new}|}_{b=1}$, while the remaining samples are marked as $B^t_{old} = \{\mathbf{x}^k_{old}\}^{|B^t_{old}|}_{k=1}$ and $B^t = B^t_{new} \cup B^t_{old}$.

Fix Term: Different Distillations Strategies. As mentioned earlier, different distillation strategies should be adopted for various types of data. In contrast to existing methods that apply a one-size-fits-all approach, where information is distilled from individual samples, such methods are no longer suitable due to the limited information they provide. However, for the old category samples, the old model still performed well as a teacher model. Therefore, for samples that may belong to old classes, using a loss

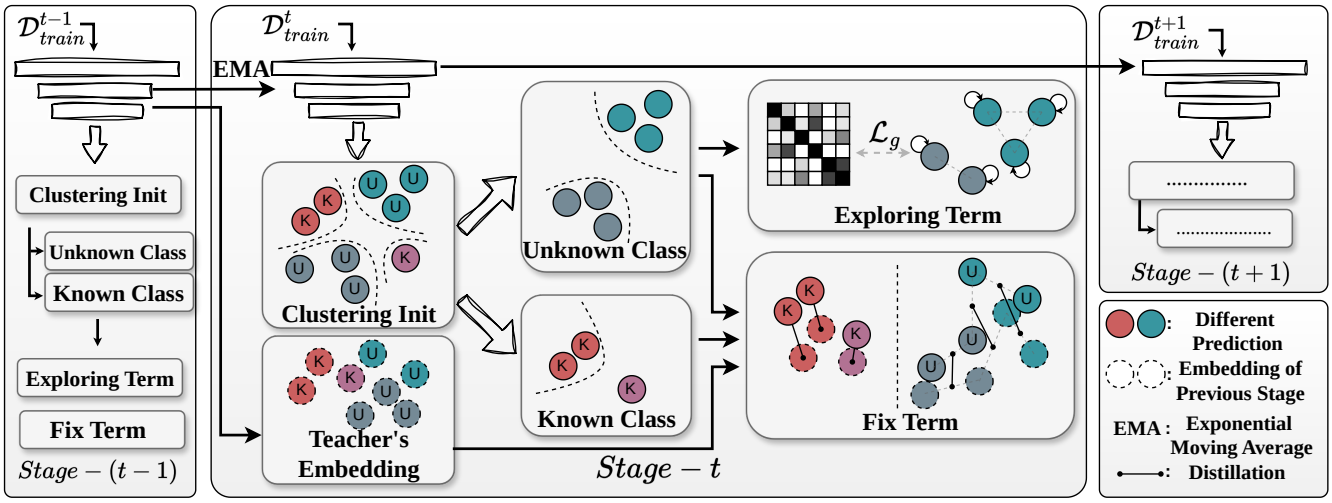


Figure 2: Main architecture of FaE. Firstly, we use K-means to initialize the classifier, providing a rough estimation of potential unknown class samples. In the Fix Term, different distillation strategies are employed to mitigate forgetting, while in the Explore Term, all unknown samples are gradually uncovered. The model adopts an EMA (Exponential Moving Average) update mechanism to prevent it from prematurely converging to a local optimum.

function based on the cosine distance is more effective than traditional methods:

$$\mathcal{L}_{old} = \frac{1}{|B^t_{old}|} \sum_{k=1}^{|B^t_{old}|} 1 - \cos(\mathrm{f}^t(\mathbf{x}^k_{old}), \mathrm{f}^{t-1}(\mathbf{x}^k_{old})), \quad (7)$$

where $\mathbf{x}_{old} \in B^t_{old}$ are the samples predicted to old classes, and $\cos(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a}\cdot\mathbf{b}}{\|\mathbf{a}\|_2\|\mathbf{b}\|_2}$ denotes cosine similarity.

Using the same distillation strategy may hinder new classes from converging to a unified representation space. Incorporating the previously introduced graph-based knowledge distillation helps retain semantic information from the teacher network, enabling gradual discovery and learning of new classes. We construct the graph using cosine distance between samples and employ MSE as the distillation objective, as shown in Eq. 2:

$$\mathcal{L}_{new} = \frac{1}{|B^t_{new}|} \sum_{i=1}^{|B^t_{new}|} \| \cos(\mathrm{f}^t(\mathbf{x}^i_{new}), \mathrm{f}^t(\mathbf{x}^j_{new})) - \cos(\mathrm{f}^{t-1}(\mathbf{x}^i_{new}), \mathrm{f}^{t-1}(\mathbf{x}^j_{new}))\|^2_F. \quad (8)$$

The overall distillation loss can be summarized as:

$$\mathcal{L}_{distill} = \mathcal{L}_{old} + \mathcal{L}_{new}. \quad (9)$$

Explore Term: Known-Aware Graph Regularization. As stated in the introduction, continual learning should not only mitigate forgetting but also facilitate new class discovery. To this end, we enhance the FaE framework with graph-based learning. We identify two key limitations in existing methods: (1) Self-supervised contrastive learning is class-agnostic, often forcing different representations for same-class images—undesirable for classification; (2) Strong correlation between feature distribution and predictions can cause class overlap, leading to incorrect hard pseudo-labels and impaired class discovery.

Inspired by Comatch (Li, Xiong, and Hoi 2021), we propose using logit predictions on old classes to generate pseudo-label graphs and guide the learning of representation. Due to the shift in the new category centroids obtained from the clustering initialization and the continuous movement of the centroids during training, we employ a more stable classifier from the old classes to guide the model's representation learning.

Given the logit prediction set $P_{new} = \{\mathbf{p}^t_b\}^{|B_{new}|}_{b=1}$, where $\mathbf{p}^t_b = \mathcal{S}(\mathrm{h}^t_{old}(f^t(x_b)))$ represents the prediction classifier of the old class and $\mathbf{x}_b \in \{\mathbf{x}^b_{new}\}^{|B_{new}|}_{b=1}$. We construct pseudo-label graphs based on the similarity of the predictions:

$$W^t_{bj} = \begin{cases} 1, & \text{if } b = j, \\ \mathbf{p}^t_b \cdot \mathbf{p}^t_j, & \text{if } b \neq j \\ 0, & \text{otherwise.} \end{cases} \quad (10)$$

The pseudo-labels obtained from the logit predictions serve as the target for the projection layer during training. To construct the projection graph, we apply image augmentation $\mathrm{Aug}(\cdot)$ to each sample of a possible new category and construct the embedding graph:

$$W^z_{bj} = \begin{cases} \exp(\mathbf{z}_b \cdot \mathbf{z}'_b/\tau), & \text{if } b = j, \\ \exp(\mathbf{z}_b \cdot \mathbf{z}_j/\tau), & \text{otherwise,} \end{cases} \quad (11)$$

where $\mathbf{z}_b = \mathrm{g}^t(\mathrm{f}^t(x^b))$, $\mathbf{z}'_b = \mathrm{g}^t(\mathrm{f}^t(\mathrm{Aug}(x^b)))$ and $\mathbf{x}^b \in B^t_{new}$.

Finally, we normalize the two graphs W^z_{bj} and W^t_{bj} as $\hat{W_{bj}} = W_{bj}/\sum_j W_{bj}$, ensuring that the similarity of each row sums to 1. We then minimize the cross-entropy between

the two normalized graphs and define the contrastive loss as:

$$\mathcal{L}_g = \frac{1}{B^t_{new}} \sum_{b=1}^{B^t_{new}} H\left(\hat{W}^t_{bj}, \hat{W}^z_{bj}\right), \tag{12}$$

where $H(\cdot,\cdot)$ is the contrastive loss function, and we employ cross-entropy here.

In the initialization phase, the dense distribution of labeled samples leads to fewer detected new-category samples. The graph regularization loss $\mathcal{L}_g$ encourages samples and newly discovered ones with similar pseudo-labels to share embeddings. Unlike computing loss over all samples, this relaxes constraints on labeled data and enhances latent space exploration.

Other Loss

In addition to the aforementioned loss, we also employ methods from prior work. Standard entropy regularization typically performs poorly when faced with imbalanced sample distributions, especially when the number of samples from old classes exceeds that of new classes, leading to a bias towards the old classes. To address this issue, we introduce **Group-wise Soft Entropy Regularization**:

$$\begin{aligned}\mathcal{L}_{entropy} = \bar{\mathbf{p}}_{old} \log \bar{\mathbf{p}}_{old} + \bar{\mathbf{p}}_{new} \log \bar{\mathbf{p}}_{new} \\ + \sum_{c \in C^t_{\text{old}}} \bar{\mathbf{p}}^{(c)} \log \bar{\mathbf{p}}^{(c)} + \sum_{c \in C^t_{\text{new}}} \bar{\mathbf{p}}^{(c)} \log \bar{\mathbf{p}}^{(c)},\end{aligned} \tag{13}$$

where $\bar{\mathbf{p}} = \frac{1}{|B^t|} \sum_{x \in B^t} \mathrm{h}^t \circ \mathrm{f}^t(\mathbf{x}) \in \mathbb{R}^{K^t}$, $\bar{\mathbf{p}}_{\text{old}} = \sum_{c \in C^t_{\text{old}}} \bar{\mathbf{p}}^{(c)} \in \mathbb{R}$ and $\bar{\mathbf{p}}_{\text{new}} = \sum_{c \in C^t_{\text{new}}} \bar{\mathbf{p}}^{(c)} \in \mathbb{R}$. The first two terms control the marginal distribution balance between the old and new classes because of $\bar{\mathbf{p}}_{\text{new}} + \bar{\mathbf{p}}_{\text{old}} = 1$, while the last two terms control the marginal distribution balance within the old and new classes, respectively.

To maintain the ability to predict old classes and mitigate forgetting while placing more focus on harder samples, we adopt the **Hardness-aware Prototype Sampling base feature reply**. Given means and variances $\mu_c = \frac{1}{N_c} \sum_{y_i=c} f^t_\theta(x_i)$, $\Sigma_c = \frac{1}{N_c} \sum_{y_i=c} (f^t_\theta(x_i) - \mu_c)(f^t_\theta(x_i) - \mu_c)^\top$. In each stage, data is generated using a normal distribution $\mathcal{N}(\mu_c, \Sigma)$ to combat forgetting, where $\Sigma = \sqrt{\sum_{c \in \mathcal{C}_0} Tr\Sigma_c/(dK^0)}$.

The hardness is defined as $p^{(i)}_{\text{hardness}} = \sigma(h_i/\tau_h)$, where $h_i = \frac{1}{K^t_{\text{old}}-1} \sum_{j=1, j\neq i}^{K^t_{\text{old}}} \cos(\mu_i, \mu_j)$, and the final loss can be expressed as:

$$\mathcal{L}_{\text{hap}} = \mathbb{E}_{c \sim p_{\text{hardness}}} \mathbb{E}_{\mathbf{z}_c \sim \mathcal{N}(\mu_c, rI)} - y_c \log \sigma(g^t_\phi(\mathbf{z}_c)/\tau_p). \tag{14}$$

Subsequently, after introducing the self-distillation loss (Caron et al. 2021a) $\mathcal{L}_{\text{self}} = \frac{1}{2|B|} \sum_{i \in B} \ell(\mathbf{q}'_i, \mathbf{p}_i) + \ell(\mathbf{q}_i, \mathbf{p}'_i)$ with the same hyper-parameter in Happy (Ma et al. 2024). the sum of the other losses is given by:

$$\mathcal{L}_h = \mathcal{L}_{hap} + \mathcal{L}_{\text{self}} + \mathcal{L}_{entropy}. \tag{15}$$

Finally, to avoid overfitting when the model learns new categories and to preserve the knowledge it has already acquired, we applied the Exponential Moving Average (EMA) method during the learning process. Hence, the overall loss can be written as follows:

$$\mathcal{L} = \mathcal{L}_{distill} + \mathcal{L}_g + \mathcal{L}_h. \tag{16}$$

Experiment

Datasets

We evaluate our methods on two well-known large-scale datasets: CIFAR-100 (Krizhevsky, Hinton et al. 2009) ,Tiny-ImageNet (Le and Yang 2015) and ImageNet-100 (Krizhevsky, Sutskever, and Hinton 2012), as well as a fine-grained datasets CUB-200 (Reed et al. 2016). Each dataset is split into two sets: labeled data is used for the initial phase ($Stage$-0), and unlabeled data is used for continual learning ($Stage$-$1 \rightarrow Stage$-T). More statistics on the partitioning of adopted datasets are presented in the Appendix.

Evaluation Protocol

After the completion of training $Stage$-t on $\mathcal{D}^t_{train}$, the model is evaluated on the test set $\mathcal{D}^t_{test}$, which contains samples from $\mathcal{C}_t$. With the prediction $\hat{\dagger}$ and the ground truth y. The accuracy is calculated as: $ACC = \max_{p \in \mathcal{P}(\hat{\mathcal{Y}})} \left(\frac{1}{M} \sum_{i=1}^{M} \mathbb{I}(y_i = p(\hat{y}_i))\right)$, where $M = |\mathcal{D}^t_{test}|$, and $\mathcal{P}(\hat{\mathcal{Y}})$ defines how the predicted labels for test samples are matched to the true labels. At the same time, we adapt the maximum forgetting metric ($\mathcal{M}_f$) and the final discovery metric ($\mathcal{M}_d$) as extra metrics.

Implement Details

The backbone employed in our experiments was ViT-B/16 (Dosovitskiy et al. 2021) pre-trained on the Imagenet-1K dataset using the DINO self-supervised learning approach (Caron et al. 2021b) and only fine-tuned the last block of the ViT model. At $Stage$-0, models are trained for 100 epochs. At the continual learning stage, models are trained for 30 epochs. All experiments were conducted using an NVIDIA GeForce RTX A6000 GPU.

Main Results

We compared adaptive methods from other tasks, such as K-means (MacQueen 1967), the recent C-GCD method: VanillaGCD (Vaze et al. 2022) and SimGCD (Wen, Zhao, and Qi 2023) and SimGCD+LwF (Li and Hoiem 2017), C-GCD methods: GM (Zhang et al. 2022) and FRoST (Roy et al. 2022), and MetaGCD (Wu et al. 2023), and the current state-of-the-art method Happy (Ma et al. 2024). As shown in Table 1, FaE outperforms the best baseline across multiple datasets in the 5-stage continual learning task. With more learning stages, its performance steadily improves and the gains become more pronounced—for example, improvements on CIFAR-100 and Tiny-ImageNet rise from 3.2% and 0.58% in $Stage$-1 to 5.21% and 9.09% in $Stage$-5.

We further compared GM metrics in Table 2, including maximum forgetting ($\mathcal{M}_f$) and final discovery ($\mathcal{M}_d$), reflecting the model's ability to retain old-class performance

Datasets	Methods	S-0	Stage-1			Stage-2			Stage-3			Stage-4			Stage-5		
		All	All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New
C100	KMeans	66.16	40.27	41.76	32.80	37.14	38.33	30.00	36.20	37.63	26.20	36.66	38.30	23.50	35.69	36.79	25.80
	VanillaGCD	90.82	72.32	78.50	41.40	67.04	72.50	34.30	57.99	62.26	28.10	56.60	59.55	33.00	51.36	53.70	30.30
	SimGCD	90.36	73.37	86.44	8.00	62.56	72.43	3.30	54.17	61.61	2.10	47.62	53.37	1.60	43.53	47.86	4.60
	SimGCD+	90.36	75.93	87.04	20.40	67.07	75.33	17.50	58.45	64.33	17.30	54.31	58.71	19.10	50.49	53.90	19.80
	FRoST	90.36	76.87	79.58	63.30	65.31	68.88	43.90	58.01	61.09	36.50	49.27	50.90	36.20	48.03	48.17	46.80
	GM	90.36	76.58	79.80	60.50	71.10	74.52	50.60	63.51	68.16	31.00	59.74	62.51	37.60	54.11	54.74	48.40
	MetaGCD	90.82	76.12	83.60	38.70	69.40	72.82	48.90	61.95	65.76	35.30	58.22	61.21	34.30	55.78	58.47	31.60
	Happy	90.36	80.40	85.26	56.10	74.13	78.27	49.30	68.23	70.86	49.80	62.26	63.75	**50.30**	59.99	60.96	51.30
	FaE (Ours)	90.36	**83.60**	**86.38**	**69.70**	**78.84**	**83.10**	**53.30**	**73.05**	**76.17**	**51.20**	**67.76**	**70.00**	49.80	**65.20**	**66.44**	**54.00**
IN100	KMeans	85.56	54.90	57.04	44.20	54.73	56.37	44.90	54.67	56.66	40.80	54.63	56.25	41.70	53.92	56.18	33.60
	VanillaGCD	95.96	70.13	72.92	56.20	69.37	73.47	44.80	68.50	70.63	53.60	65.56	67.85	47.20	64.54	67.44	38.40
	SimGCD	96.20	79.67	91.68	19.60	70.23	78.83	18.60	61.90	67.43	23.20	56.67	60.92	22.60	52.90	56.40	21.40
	SimGCD+	96.20	83.07	95.16	22.60	74.57	83.47	21.20	67.60	73.57	25.80	62.09	66.83	24.20	57.62	61.47	23.00
	FRoST	96.20	87.50	92.96	60.20	79.63	83.37	57.20	76.78	77.00	75.20	66.18	68.65	46.40	63.82	66.40	40.60
	GM	96.20	89.53	95.04	62.00	82.34	86.93	54.80	77.97	79.17	69.60	72.80	74.65	58.00	71.08	71.76	65.00
	MetaGCD	95.96	75.27	78.20	60.60	73.79	75.93	54.90	69.35	72.20	49.40	67.22	70.10	44.20	66.68	69.31	43.00
	Happy	96.20	91.20	95.36	70.40	87.83	90.83	69.80	85.22	86.40	77.00	81.93	83.00	**73.40**	78.58	79.11	73.80
	FaE (Ours)	96.20	**92.33**	**96.20**	**73.00**	**89.46**	**92.13**	**73.40**	**87.50**	**88.86**	**78.00**	**83.00**	**86.08**	58.00	**80.88**	**81.64**	**74.00**
Tiny	KMeans	61.70	35.42	35.46	35.20	34.99	35.75	30.40	34.80	36.07	25.90	34.77	35.90	24.90	34.62	35.63	25.50
	VanillaGCD	84.20	55.93	58.92	41.00	54.96	58.58	33.20	52.82	55.74	32.40	48.81	51.46	27.60	45.94	48.06	26.90
	SimGCD	85.86	66.95	79.94	2.00	57.81	66.98	2.80	52.70	59.83	2.77	45.01	50.29	2.80	41.59	45.79	3.80
	SimGCD+	85.86	70.38	81.80	13.30	62.47	70.75	12.80	54.55	60.46	13.20	47.98	52.49	11.90	42.98	46.46	12.70
	FRoST	85.86	75.15	78.56	58.10	65.64	67.83	52.50	51.32	54.31	30.40	48.22	52.14	16.90	40.15	42.73	16.90
	GM	85.86	76.42	82.40	46.50	68.87	73.82	39.20	58.68	63.43	25.40	52.86	57.21	18.10	46.90	50.62	13.40
	MetaGCD	84.20	60.88	64.90	40.80	57.20	61.03	34.20	54.36	57.19	34.60	50.83	53.59	28.80	48.14	50.16	30.00
	Happy	85.86	78.85	82.40	**61.10**	71.34	76.18	42.30	64.68	68.70	36.50	58.49	60.64	41.30	54.56	56.66	35.70
	FaE (Ours)	85.86	**79.43**	**83.46**	59.30	**75.26**	**76.72**	**66.50**	**69.60**	**72.13**	**51.90**	**63.99**	**66.50**	**43.90**	**63.65**	**66.12**	**41.40**
CUB	KMeans	43.93	32.54	30.76	41.18	31.19	30.53	35.20	29.28	27.46	42.09	29.19	28.13	37.61	28.17	27.01	38.53
	VanillaGCD	89.20	64.47	67.06	51.93	58.15	60.65	42.91	54.10	56.40	37.91	49.98	51.33	39.32	46.84	46.58	49.14
	SimGCD	90.26	73.84	84.54	22.02	63.36	72.35	8.58	55.63	61.95	11.13	49.31	54.55	7.86	44.72	48.69	9.25
	SimGCD+	90.26	75.62	85.55	25.97	65.32	73.93	13.68	57.40	63.28	16.26	51.11	55.72	14.27	45.79	49.29	14.28
	FRoST	90.26	77.03	83.95	43.53	50.77	53.46	34.33	46.42	49.31	26.09	39.40	41.47	23.08	34.55	35.12	29.45
	GM	90.26	76.17	80.23	56.51	67.91	73.38	34.58	61.12	66.53	23.00	55.90	57.49	43.38	51.96	54.40	30.10
	MetaGCD	89.20	67.08	70.21	51.92	60.77	62.39	50.86	57.53	59.33	37.78	51.90	52.22	49.40	49.60	49.96	46.38
	Happy	90.26	**81.40**	85.06	**63.70**	74.27	76.03	**63.57**	67.09	71.06	39.13	62.25	63.83	**49.74**	59.39	60.49	**49.52**
	FaE (Ours)	90.26	80.80	**88.73**	38.15	**74.30**	**77.38**	55.52	**69.95**	**73.75**	**43.13**	**66.41**	**68.93**	46.50	**63.72**	**66.47**	39.21

Table 1: Performance (%) for 5 Stages continual learning on CIFAR-100 (C100), ImageNet-100 (IN100), Tiny-ImageNet (Tiny) and CUB-200 (CUB).

Method	C100		Tiny	
	$\mathcal{M}_f \downarrow$	$\mathcal{M}_d \uparrow$	$\mathcal{M}_f \downarrow$	$\mathcal{M}_d \uparrow$
VanillaGCD	17.10	33.42	22.20	32.22
FRoST	22.82	45.34	21.62	34.96
MetaGCD	16.56	37.76	19.30	33.68
Happy	11.22	51.36	9.75	43.38
FaE	**6.82**	**55.60**	**5.43**	**52.60**

Table 2: Forgetting & discovery.

and discover new classes. FaE achieves a 5–9% overall improvement, maintaining balanced enhancement. This indicates that FaE stabilizes old class distributions while exploring features beneficial for new class discovery. In a more challenging 10-stage continual setting (Table 3), FaE consistently outperforms others, with performance gaps expanding to 6.26% and 11.14%, demonstrating superior stability in complex scenarios.

Ablation Study

Components Analysis

FaE comprises two core components: distinct distillation strategies and knowledge-aware graph regularization. As shown in Table 4, "Old distill" denotes the distillation strategy for potential old-class samples. Using cosine loss-based distillation for old classes and graph-based distillation for new classes yields notable gains. Moreover, incorporating graph regularization further balances old–new class performance, improving results by 4.34% on CIFAR-100 and 4.79% on Tiny-ImageNet.

Effects of Training Stragey

Notably, Happy adopts the best-performing model from all epochs at each stage, which is impractical in real-world settings. Without a validation set, using the last epoch's output is often more feasible. On datasets like Tiny-ImageNet and CUB-200, extended training across stages leads to negative optimization, mainly due to latent space confusion and poor retention of old category distributions—challenges FaE is designed to address. As shown in Fig. 3, comparing the best-epoch model with the last-epoch model as input to the next stage, our method achieves more stable performance across long-term training.

Unknown Class Number

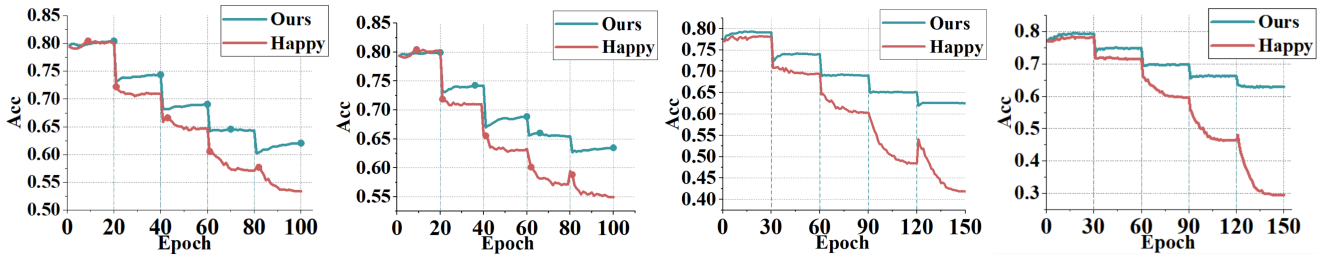
We assume the number of categories per stage is known, though this is often unavailable in practice. Since stage in-

Datasets	Methods	0	1	2	3	4	5	6	7	8	9	10
C100	VanillaGCD	90.82	78.42	75.68	70.35	66.64	64.29	61.05	58.33	57.14	56.23	55.15
	MetaGCD	90.82	81.07	76.55	74.26	67.64	64.45	61.58	59.13	60.13	56.91	56.51
	Happy	90.36	85.62	81.88	79.82	74.01	71.81	68.46	64.05	62.14	61.38	57.81
	FaE (Ours)	90.36	**86.96**	**82.75**	**80.03**	**76.53**	**74.91**	**71.41**	**68.00**	**67.03**	**64.74**	**64.07**
Tiny	VanillaGCD	84.20	65.15	64.63	60.94	59.46	56.52	55.47	51.65	50.66	49.83	48.56
	MetaGCD	84.20	68.87	65.48	62.92	60.81	58.21	56.16	54.68	52.58	50.57	48.92
	Happy	85.86	80.75	76.92	73.34	69.77	66.33	62.75	57.56	54.73	53.02	50.69
	FaE (Ours)	85.86	**82.95**	**79.82**	**76.62**	**74.10**	**71.55**	**68.90**	**68.07**	**64.21**	**62.53**	**61.83**

Table 3: Performance (%) for 10 Stages continual learning.

	Old Distill		New Distill		Graph-Reg	C100			Tiny		
	Cosine	Graph	Cosine	Graph		ALL	OLD	NEW	ALL	OLD	NEW
(1)	✔	✗	✔	✗	✗	69.00	71.82	51.36	65.58	68.92	43.38
(2)	✔	✗	✔	✗	✔	73.16	74.97	**62.22**	67.59	71.16	43.70
(3)	✗	✔	✗	✔	✗	69.95	73.95	41.94	69.03	72.93	43.30
(4)	✔	✗	✗	✔	✗	72.44	75.47	52.42	70.35	73.33	51.62
(5)	✔	✗	✗	✔	✔	**73.34**	**76.05**	55.39	**70.39**	**72.99**	**52.60**

Table 4: Results (%) for the effectiveness of different distillation strategies and graph regularization.



(a) CUB-200 (b) CUB-200 (c) Tiny-ImageNet (d) Tiny-ImageNet

Figure 3: Performance (%) with different training strageies. (a, c): Selecting the optimal model from all epochs in the stage as the output, while the dot represents the optimal point. (b, d): Selecting the final model as the output.

	C100		
Method	ALL	OLD	NEW
GCD	58.72	62.66	32.92
MetaGCD	63.28	67.65	34.94
Happy	68.80	72.40	45.74
Ours	**71.45**	**74.79**	**53.12**

Table 5: Performance (%) with estimated the estimated number of categories on CIFAR-100.

puts are unlabeled, accuracy-based estimation is infeasible. Following Happy, we estimate category numbers using the silhouette coefficient, selecting the k-means result with the highest score. As shown in Table 5, our method consistently outperforms others.

Conclusion

In this paper, we propose a Fix and Explore learning framework, FaE, to address the problem of Continual Generalized Category Discovery. FaE effectively balances the retention of knowledge about old categories and the discovery of new categories through a divide-and-conquer strategy. Specifically, the framework employs a clustering algorithm to differentiate between potential old and new category samples, adopts distinct knowledge distillation strategies for different types of samples, and introduces a known-category-aware graph regularization method to isolate new categories progressively. Results on multiple datasets demonstrate that FaE consistently achieves outstanding performance across all stages, excelling in category discovery while significantly mitigating catastrophic forgetting.

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

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021a. Emerging properties in self-supervised vision transformers. In *ICCV*, 9650–9660.

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021b. Emerging properties in self-supervised vision transformers. In *ICCV*, 9650–9660.

Cendra, F. J.; Zhao, B.; and Han, K. 2024. PromptCCD: Learning Gaussian Mixture Prompt Pool for Continual Category Discovery. In *ECCV*.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *ICML*, 1597–1607. PMLR.

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

Fini, E.; Sangineto, E.; Lathuiliere, S.; Zhong, Z.; Nabi, M.; and Ricci, E. 2021. A unified objective for novel class discovery. In *ICCV*, 9284–9292.

Gao, Q.; Zhao, C.; Ghanem, B.; and Zhang, J. 2022. R-dfcil: Relation-guided representation learning for data-free class incremental learning. In *ECCV*, 423–439. Springer.

Gu, P.; Zhang, C.; Xu, R.; and He, X. 2023. Class-relation Knowledge Distillation for Novel Class Discovery. In *ICCV*, 16474–16483.

Han, K.; Rebuffi, S.-A.; Ehrhardt, S.; Vedaldi, A.; and Zisserman, A. 2020. Automatically Discovering and Learning New Visual Categories with Ranking Statistics. In *ICLR*.

Han, K.; Vedaldi, A.; and Zisserman, A. 2019. Learning to discover novel visual categories via deep transfer clustering. In *ICCV*, 8401–8409.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked Autoencoders Are Scalable Vision Learners. In *CVPR*, 15979–15988.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *CVPR*, 770–778.

Joseph, K. J.; Paul, S.; Aggarwal, G.; Biswas, S.; Rai, P.; Han, K.; and Balasubramanian, V. N. 2022. Novel Class Discovery Without Forgetting. In *ECCV*, 570–586. Cham: Springer Nature Switzerland.

Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; and Krishnan, D. 2020. Supervised contrastive learning. In *NeurIPS*, volume 33, 18661–18673.

Kim, H.; Suh, S.; Kim, D.; Jeong, D.; Cho, H.; and Kim, J. 2023. Proxy Anchor-based Unsupervised Learning for Continuous Generalized Category Discovery. In *ICCV*, 16688–16697.

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

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *NeurIPS*.

Le, Y.; and Yang, X. 2015. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7): 3.

Li, J.; Xiong, C.; and Hoi, S. C. 2021. Semi-supervised Learning with Contrastive Graph Regularization. In *ICCV*, 9475–9484.

Li, W.; Fan, Z.; Huo, J.; and Gao, Y. 2023. Modeling Inter-Class and Intra-Class Constraints in Novel Class Discovery. In *CVPR*, 3449–3458.

Li, Z.; and Hoiem, D. 2017. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12): 2935–2947.

Liu, J.; Zheng, T.; Zhang, G.; and Hao, Q. 2023a. Graph-based Knowledge Distillation: A survey and experimental evaluation. *arXiv preprint arXiv:2302.14643*.

Liu, M.; Roy, S.; Zhong, Z.; Sebe, N.; and Ricci, E. 2023b. Large-scale pre-trained models are surprisingly strong in incremental novel class discovery. *arXiv preprint arXiv:2303.15975*.

Ma, S.; Zhu, F.; Zhong, Z.; Liu, W.; Zhang, X.-Y.; and Liu, C.-L. 2024. Happy: A Debiased Learning Framework for Continual Generalized Category Discovery. In *NeurIPS*.

MacQueen, J. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability*.

Passalis, N.; Tzelepi, M.; and Tefas, A. 2020. Probabilistic knowledge transfer for lightweight deep representation learning. *IEEE Transactions on Neural Networks and Learning Systems*, 32(5): 2030–2039.

Pu, N.; Zhong, Z.; and Sebe, N. 2023. Dynamic Conceptional Contrastive Learning for Generalized Category Discovery. In *CVPR*, 7579–7588.

Rastegar, S.; Salehi, M.; Asano, Y. M.; Doughty, H.; and Snoek, C. G. M. 2024. SelEx: Self-Expertise in Fine-Grained Generalized Category Discovery. In *ECCV*.

Reed, S.; Akata, Z.; Lee, H.; and Schiele, B. 2016. Learning deep representations of fine-grained visual descriptions. In *CVPR*, 49–58.

Roy, S.; Liu, M.; Zhong, Z.; Sebe, N.; and Ricci, E. 2022. Class-incremental novel class discovery. In *ECCV*, 317–333.

Vaze, S.; Han, K.; Vedaldi, A.; and Zisserman, A. 2022. Generalized Category Discovery. In *CVPR*, 7492–7501.

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

Wen, X.; Zhao, B.; and Qi, X. 2023. Parametric Classification for Generalized Category Discovery: A Baseline Study. In *ICCV*, 16590–16600.

Wu, Y.; Chi, Z.; Wang, Y.; ; and Feng, S. 2023. MetaGCD: Learning to Continually Learn in Generalized Category Discovery. In *ICCV*.

Yim, J.; Joo, D.; Bae, J.; and Kim, J. 2017. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *CVPR*, 4133–4141.

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 *CVPR*, 3479–3488.

Zhang, X.; Jiang, J.; Feng, Y.; Wu, Z.-F.; Zhao, X.; Wan, H.; Tang, M.; Jin, R.; and Gao, Y. 2022. Grow and Merge: A Unified Framework for Continuous Categories Discovery. In *NeurIPS*.

Zhou, D.-W.; Wang, Q.-W.; Qi, Z.-H.; Ye, H.-J.; Zhan, D.-C.; and Liu, Z. 2024. Class-Incremental Learning: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12): 9851–9873.