Active Generalized Category Discovery

Shijie Ma[1,2], Fei Zhu[3], Zhun Zhong[4,5], Xu-Yao Zhang[1,2*], Cheng-Lin Liu[1,2]
[1]MAIS, Institute of Automation, Chinese Academy of Sciences, China
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences, China
[3]Centre for Artificial Intelligence and Robotics, HKISI-CAS, China
[4]School of Computer Science and Information Engineering, Hefei University of Technology, China
[5]School of Computer Science, University of Nottingham, NG8 1BB Nottingham, UK
{mashijie2021, zhufei2018}@ia.ac.cn, zhunzhong007@gmail.com, {xyz, liucl}@nlpr.ia.ac.cn

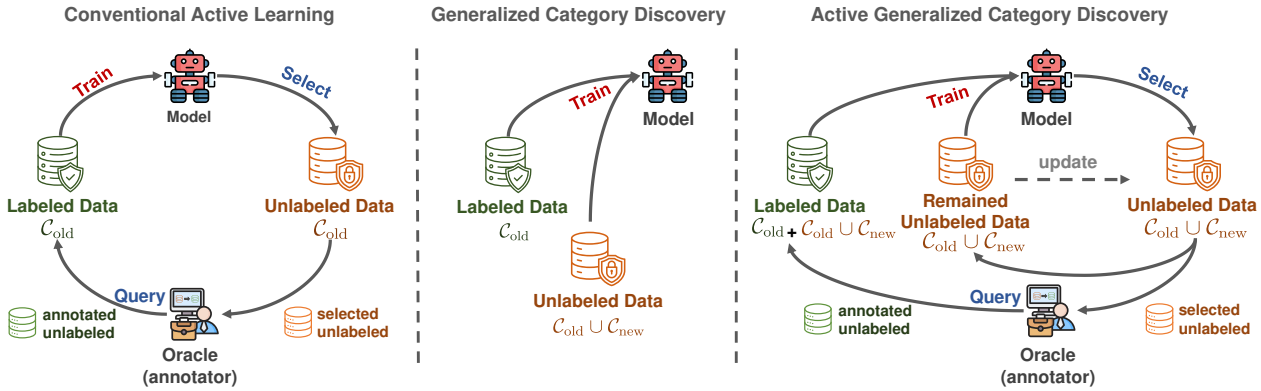

Figure 1. The diagram of three settings. Left: Conventional AL is a closed-world setting, where labeled and unlabeled classes are identical. Middle: GCD requires no active labeling and suffers from severe issues. Right: AGCD is an open-world extrapolated version of AL, where unlabeled data contains novel categories, and models are trained on both labeled and unlabeled data to cluster both old and new classes.

Abstract

Generalized Category Discovery (GCD) is a pragmatic and challenging open-world task, which endeavors to cluster unlabeled samples from both novel and old classes, leveraging some labeled data of old classes. Given that knowledge learned from old classes is not fully transferable to new classes, and that novel categories are fully unlabeled, GCD inherently faces intractable problems, including imbalanced classification performance and inconsistent confidence between old and new classes, especially in the low-labeling regime. Hence, some annotations of new classes are deemed necessary. However, labeling new classes is extremely costly. To address this issue, we take the spirit of active learning and propose a new setting called Active Generalized Category Discovery (AGCD). The goal is to improve the performance of GCD by actively selecting a limited amount of valuable samples for labeling from the oracle. To solve this problem, we devise an adaptive sampling strategy, which jointly considers novelty, informativeness and diversity to adaptively select novel samples with proper uncertainty. However, owing to the varied orderings of label indices caused by the clustering of novel classes, the queried labels are not directly applicable to subsequent training. To overcome this issue, we further propose a stable label mapping algorithm that transforms ground truth labels to the label space of the classifier, thereby ensuring consistent training across different active selection stages. Our method achieves state-of-the-art performance on both generic and fine-grained datasets. Our code is available at https://github.com/mashijie1028/ActiveGCD

1. Introduction

Humans could transfer previously acquired knowledge while learning new concepts [29]. For example, once children have been taught to recognize "cats" and "dogs" based on external contours, they can group "birds" and "bears" according to the same rule. However, due to species disparities, this classification criteria is limited. Children may confuse "zebras" with "horses", and "huskies" with "wolves", and they still need guidance to focus on fine-grained fea-

*Corresponding author.

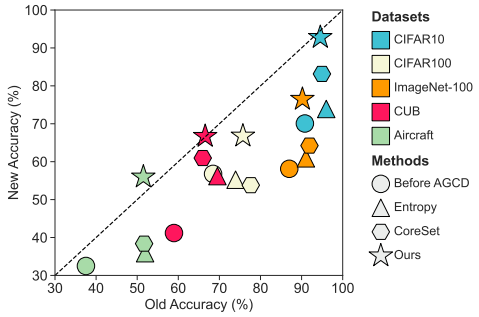

Figure 2. Accuracy of old and new classes in AGCD with different methods (shapes) on various datasets (colors). The closer to the diagonal, the more balanced accuracy between old and new classes. In each dataset (color), our method (star) achieves not only the best overall performance but also more balanced accuracy.

tures like stripes and eyes. Thus, proper guidance is indispensable when acquiring new knowledge. However, seeking help every time is impractical. Instead, they are supposed to actively query some confusing samples [7].

Deep learning is also inspired by the cognitive processes above, and could be endowed with the abilities of knowledge transfer [18, 32, 58] and active learning [34, 38, 50], especially in the open-environments [14, 28, 43, 48, 57] containing unlabeled novel categories. In this paper, we study the task of Generalized Category Discovery (GCD) [33, 42, 47, 53], which aims to transfer knowledge from some labeled samples of old classes to cluster novel categories in the unlabeled data. In addition, models should also be able to classify old classes present in the unlabeled data. Pioneer works [12, 42] leverage supervised [23] and unsupervised contrastive learning [8] with non-parametric K-Means [30] for clustering. Later works [33, 52] further exploit underlying cross-instance relationships. Wen *et al*. [47] rethink the failure of parametric classifiers and propose a simple method to achieve impressive results.

Although great progress has been made, GCD still faces intractable problems, including imbalanced accuracy (see Fig. 3) and inconsistent confidence (see Fig. 4) between old and new classes, especially in low-labeling regimes. In essence, these issues arise from the nature of the GCD task itself. As old knowledge is not fully transferable to the new one, and novel classes are fully unlabeled, models would encounter inherent challenges, and could not rectify errors by themselves without the supervision of confusing categories. Therefore, we argue that some annotations of new categories [55] are necessary. However, due to the computational cost of annotation, it is not practical to label all the novel classes. This raises a question: *Can deep learning models actively select a small number of unlabeled samples for labeling to remarkably enhance category discovery?*

In this work, we try to answer this question and propose a new setting, namely Active Generalized Category Discovery (AGCD) as in Fig. 1. During training, models actively select a limited number of samples in unlabeled data, which contains both old and new classes, and query their labels from the oracle, these newly-labeled data are then incorporated into labeled data for the next training round. Through human-in-the-loop interaction, models actively select informative novel samples, acquire knowledge that could not be obtained via pure unsupervised learning, and rectify previous errors and biases. AGCD is a realistic setting, which addresses the problems of GCD and largely enhances the performance, requiring very limited annotations. As in Fig. 2, we improve the new accuracy of GCD by 25.52%/23.49% on CUB/Air with only ~ 2.5 samples labeled per class.

In the task of AGCD, one could inevitably encounter two challenges, which we aim to address in this paper: (1) Conventional AL methods do not take novel categories into consideration, which makes them not applicable to AGCD and leads to sub-optimal results. (2) Considering the clustering nature of GCD, the queried ground truth labels could not be directly used by parametric classifiers due to the different ordering of indices. To solve the first problem, we take novelty, informativeness and diversity into consideration and propose an adaptive sampling strategy called `Adaptive-Novel`, which adaptively chooses samples within appropriate uncertainty intervals according to the clustering performance. To alleviate the second problem, we propose to perform label mapping on the queried samples which "translates" ground-truth labels to the labels the model could understand, however, considering the scarcity of labeled data, we devise a stable label mapping method with the model exponential moving average [17, 40, 56].

Our contributions are summarized as follows: (1) We propose a new task called Active Generalized Category Discovery (AGCD) considering the inherent issues in GCD, and establish its pipeline and metrics. (2) We propose an adaptive query strategy called `Adaptive-Novel` to select valuable novel samples for labeling and address the problems of GCD with affordable budgets. (3) We devise a stable label mapping method to obtain credible mapping and alleviate the issue of different label ordering in clustering. (4) Extensive experiments show our method achieves state-of-the-art performance among various strategies on generic and fine-grain datasets, as in Fig. 2.

2. Related Works

Novel Category Discovery (NCD) [41] was first formalized as deep transfer clustering [18] to discover unlabeled new classes using the knowledge of labeled classes. Han *et al*. [18, 19] utilize self-supervision for representation learning and ranking statistics for knowledge transfer. Zhong *et al*. [54] propose to mixup [51] old and new classes to prevent overfitting. UNO [13] is a unified objective to handle old and new classes jointly via swapped prediction [5]. NCD assumes all unlabeled data are from new classes.

Generalized Category Discovery (GCD) [4, 42] removes the limited assumption and aims to simultaneously cluster old and new classes in the unlabeled data, given some labeled samples of old classes. Pioneer works [12, 42] conduct supervised [23] and unsupervised contrastive learning [8], and employ semi-supervised K-means [30, 42] clustering. Later works [33, 52] exploit underlying relationships for better feature representation. Zhao *et al.* [53] propose an EM-like framework alternating between contrastive learning [27] and class number estimation. These methods predominantly rely on non-parametric classifiers. By contrast, recent works [9, 47] propose to avoid prediction biases to achieve remarkable results with parametric classifiers. Although GCD has made great advancements [15, 44], it inherently suffers from issues like imbalanced accuracy and confidence between old and new classes, which is intractable due to the incompletely transferable knowledge and unlabeled nature of new classes. In this paper, we propose AGCD to address them with affordable labeling budgets.

Active Learning (AL) [38] aims to maximize models' performance with a limited labeling budget. We focus on pool-based AL [49]. Sampling strategies contain two types. Uncertainty-based methods select samples with high predictive uncertainty, *e.g.*, entropy [46], least confidence [46] and margin [35]. Diversity-based methods select samples that could represent the entire dataset. Typical works include KMeans [30], CoreSet [37] and BADGE [3]. Hybrid methods [1, 21, 22] combine the two types for further improvements. In principle, AL is in a close-world setting, where labeled and unlabeled data share classes. While in AGCD, unlabeled data contain more categories than labeled data, and models are expected to classify all the classes, not limited to the old classes present in labeled data.

3. Preliminaries and Analysis

Here, we briefly introduce the setting and methods of Generalized Category Discovery (GCD) (Sec. 3.1) and give empirical results to reveal inherent issues (Sec. 3.2), which motivates us to propose our setting AGCD in Sec. 4.

3.1. Setup and Training Methods of GCD

Problem definition of GCD. Given a labeled dataset $\mathcal{D}_l = \{(\mathbf{x}_i^l, y_i^l)\} \subset \mathcal{X} \times \mathcal{Y}_l$ and an unlabeled dataset $\mathcal{D}_u = \{(\mathbf{x}_i^u, y_i^u)\} \subset \mathcal{X} \times \mathcal{Y}_u$. $\mathcal{D}_l$ only contains old classes, while $\mathcal{D}_u$ contains both old and new classes, *i.e.*, $\mathcal{Y}_l = \mathcal{C}_{old}, \mathcal{Y}_u = \mathcal{C}_{old} \cup \mathcal{C}_{new}$. Models are required to cluster both old and new classes in $\mathcal{D}_u$. The number of novel classes K_{new} is known a-prior or estimated [33, 42, 53]. $f(\cdot)$ and $g(\cdot)$ are feature extractor and projection head for contrastive learning respectively. $\mathbf{h}_i = f(\mathbf{x}_i)$ and $\mathbf{z}_i = g(\mathbf{h}_i)$ are ℓ-2 normalized feature and projected embeddings respectively.

Related Training Methods. Vaze *et al.* [42] propose to employ supervised [23] and self-supervised [8] contrastive learning on labeled B^l and whole mini-batch B:

$$\mathcal{L}_{con}^l = \frac{1}{|B^l|} \sum_{i \in B^l} \frac{1}{|\mathcal{N}(i)|} \sum_{q \in \mathcal{N}(i)} -\log \frac{\exp(\mathbf{z}_i^\top \mathbf{z}_q'/\tau_c)}{\sum_{n \neq i} \exp(\mathbf{z}_i^\top \mathbf{z}_n'/\tau_c)}, \tag{1}$$

$$\mathcal{L}_{con}^u = \frac{1}{|B|} \sum_{i \in B} -\log \frac{\exp(\mathbf{z}_i^\top \mathbf{z}_i'/\tau_c)}{\sum_{n \neq i} \exp(\mathbf{z}_i^\top \mathbf{z}_n'/\tau_c)}. \tag{2}$$

The overall contrastive loss $\mathcal{L}_{con} = (1-\lambda)\mathcal{L}_{con}^u + \lambda \mathcal{L}_{con}^l$.

SimGCD [47] employs a parametric prototypical classifier $\mathcal{C} = \{\mathbf{c}_1, \cdots, \mathbf{c}_K\}$, where $K = K_{old} + K_{new}$. The posterior probability could be expressed as:

$$\mathbf{p}_i^{(k)} = \frac{\exp(\mathbf{h}_i^\top \mathbf{c}_k)/\tau_p}{\sum_{k'} \exp(\mathbf{h}_i^\top \mathbf{c}_k')/\tau_p}. \tag{3}$$

SimGCD implements self-distillation on two views along with an entropy $H(\cdot)$ regularization across all samples:

$$\mathcal{L}_{cls}^u = \frac{1}{|B|} \ell(\mathbf{q}_i', \mathbf{p}_i) - \lambda_e H(\overline{\mathbf{p}}), \tag{4}$$

where $\mathbf{q}_i'$ is a sharpened probability of another view, and $\overline{\mathbf{p}} = \frac{1}{2|B|} \sum_{i \in B} (\mathbf{p}_i + \mathbf{p}_i')$, $\ell(\cdot)$ denotes cross-entropy loss. The supervised loss is also employed on $\mathcal{D}_l$ with labels y_i:

$$\mathcal{L}_{cls}^l = \frac{1}{|B^l|} \sum_{i \in B^l} \ell(y_i, \mathbf{p}_i). \tag{5}$$

3.2. Inherent Problems in GCD

SimGCD [47] is the state-of-the-art (SOTA) and effective GCD method with a parametric classifier, we thus use it for the analysis of confidence [16] and accuracy. To find the problems of GCD in real scenarios, we train models with SimGCD in a practical low-label condition as in Table 1.

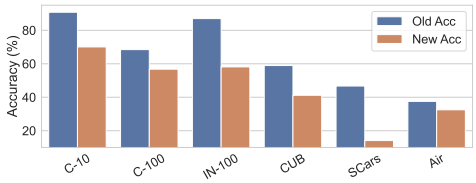

Figure 3. Accuracy of old and new classes on six datasets.

GCD suffers from severe performance mismatch between old and new classes. As Fig. 3 shows, the accuracy of old classes largely surpasses the new ones, which is imbalanced. For example, the performance gap is 28.88% and 32.59% on IN-100 and SCars. The underlying reason is the inherent label condition imbalanced, *i.e.*, old classes are partially labeled while new classes are fully unlabeled, which is the essential issue of the GCD task itself.

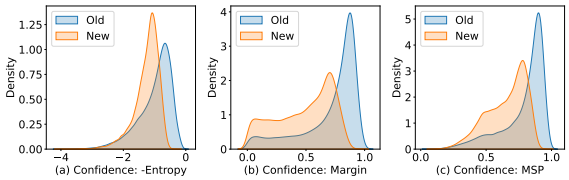

Figure 4. Confidence distribution on ImageNet-100. Three measures: -entropy (a), margin (b) and MSP (c).

Models tend to have inconsistent confidence between old and new classes. We plot the distribution of the model's predictive confidence, with three metrics, *i.e.*, (minus) entropy [39], margin [35] and maximum softmax probability (MSP) [20]. From Fig. 4, we observe that the confidence distribution between old and new classes is inconsistent. The confidence of old classes is relatively high, which aligns with intuition because some old samples are labeled, while novel classes are learned with soft targets in Eq. (4), resulting in ambiguous predictions.

4. Active Generalized Category Discovery

As in Sec. 3.2, models perform poorly on the fully unlabeled new classes across various datasets and are unable to correct errors. We argue that some annotations on new classes are deemed indispensable. To avoid excessive labeling costs and make the setting practical, we aim to enhance GCD with limited annotation budgets and propose Active Generalized Category Discovery (AGCD). Firstly, we introduce the problem definition, analyze its challenges and distinguish the differences from AL (Sec. 4.1). Then, we provide metrics considering the accuracy and novelty of samples (Sec. 4.2). Next, we elaborate on the proposed sampling strategy `Adaptive-Novel` (Sec. 4.3) and stable label mapping algorithm (Sec. 4.4).

Overview. We provide the framework of AGCD in Fig. 5. (a) illustrates the pipeline and dataflow of AGCD, and models are trained with off-the-shelf SimGCD [47]. (b) shows the proposed `Adaptive-Novel` query strategy. (c) demonstrates the *labeling mapping* method to transform ground truth labels to models' label space.

4.1. The Task of AGCD and Basic Analysis

Problem definition of AGCD. Initially, the model is trained on both $\mathcal{D}_l^{init} = \mathcal{D}_l^0 = \{(\mathbf{x}_i^l, y_i^l)\} \subset \mathcal{X} \times \mathcal{Y}_l$ and $\mathcal{D}_u^{init} = \mathcal{D}_u^0 = \{(\mathbf{x}_i^u, y_i^u)\} \subset \mathcal{X} \times \mathcal{Y}_u$ with off-the-shelf GCD training method SimGCD [47]. The initial data splits are similar to GCD [42] as in Table 1. After this base training stage, AGCD could have multiple rounds (denoted as n in total) like AL. At round t, the model first selects a batch of b samples (budget size) from unlabeled data $\mathcal{D}_u^{t-1}$ and queries its labels to obtain $\mathcal{D}_q^t = \{(\mathbf{x}_i^q, y_i^q)\}$. Then labeled and unlabeled data are updated as $\mathcal{D}_l^t = \mathcal{D}_l^{t-1} \cup \mathcal{D}_q^t \subset \mathcal{X} \times \mathcal{Y}_l^t$, $\mathcal{D}_u^t = \mathcal{D}_u^{t-1} \setminus \mathcal{D}_q^t$. Models are then trained on $\mathcal{D}_l^t \cup \mathcal{D}_u^t$ with off-the-shelf SimGCD [47]. Note that initial labeled data only contains old classes, *i.e.* $\mathcal{Y}_l^0 = \mathcal{C}_{old}$, but after querying in AGCD, the labeled data $\mathcal{D}_l^t$ could contain some classes of all novel ones $\mathcal{C}_{new}$. The total budget size is $b \times n$. For the queried data $\mathcal{D}_q^t$, models are trained with supervised loss $\mathcal{L}_{con}^l$ in Eq. (1) and $\mathcal{L}_{cls}^l$ in Eq. (5).

Table 1. The default setting of 3 generic datasets and 3 fine-grained datasets in AGCD benchmark. $|\mathcal{Y}_l^{\text{init}}| = K_{old}$, $|\mathcal{Y}_u^{\text{init}}| = K_{old} + K_{new}$ denote the initial number of classes in $|\mathcal{D}_l^{\text{init}}|$ and $|\mathcal{D}_u^{\text{init}}|$. The number of queries across all rounds is displayed in both the total count and average count per class.

Dataset	Labeled $\mathcal{D}_l^{\text{init}}$		Unlabeled $\mathcal{D}_u^{\text{init}}$		#Rounds	#Query (total)	#Query (per class)
	$\lvert\mathcal{D}_l^{\text{init}}\rvert$	$\lvert\mathcal{Y}_l^{\text{init}}\rvert$	$\lvert\mathcal{D}_u^{\text{init}}\rvert$	$\lvert\mathcal{Y}_u^{\text{init}}\rvert$			
CIFAR10 (C-10) [25]	2,000	2	48,000	10	1	100	10
CIFAR100 (C-100) [25]	5,000	50	45,000	100	5	500	5
ImageNet-100 (IN-100) [10]	12,744	50	114,371	100	5	500	5
CUB (CUB) [45]	599	100	5,395	200	5	500	2.5
Stanford Cars (SCars) [24]	800	98	7,344	196	5	500	2.5
FGVC-Aircraft (Air) [31]	666	50	6,001	100	5	500	5

Two challenges in AGCD. (1) Directly employing conventional AL methods (*e.g.*, Entropy) results in sub-optimal performance. This is because they do not consider new classes, and the confidence and feature distribution of old and new classes are inconsistent, as discussed in Sec. 3.2. (2) Considering the clustering of GCD, the queried labels could not be directly sent to models due to the different ordering of indices of new classes. For example, considering new classes "birds" and "lions", the model assigns "8", "7" to them while the ground truth is "9", "6", then we should obtain a mapping as: "9"→"8", "6"→"7", and map the ground truth to the model's label space for supervision.

Distinguishing between AL and AGCD. (1) AGCD could be viewed as an open-world extrapolated version of AL requiring models to classify both old and new classes, and the unlabeled data could contain new classes. (2) In conventional AL, models are not trained on $\mathcal{D}_u$, which is only used for sample selection and only the selected samples engage in training. In contrast, in AGCD, models not only select samples in $\mathcal{D}_u$ but are also trained on it.

4.2. Evaluation and Metrics

Accuracy Evaluation. GCD adopts a *transductive* evaluation on unlabeled training data $\mathcal{D}_u$. By contrast, we adopt an *inductive* evaluation for AGCD, *i.e.*, we test models on the unseen and disjoint test dataset. The reason is that models query some labels in $\mathcal{D}_u$, making it unfair for evaluation. The accuracy is calculated using ground truth labels y_i and models' predictions $\hat{y}_i$ as follows:

$$ACC = \max_{p \in \mathcal{P}(\mathcal{Y}_u)} \frac{1}{M} \sum_{i=1}^{M} \mathbb{1}(y_i = p(\hat{y}_i)), \tag{6}$$

$M = |\mathcal{D}_u|$ and $\mathcal{P}(\mathcal{Y}_u)$ is the set of all permutations across all classes $\mathcal{C}_{old} \cup \mathcal{C}_{new}$. The maximum value is computed by the Hungarian optimal assignment algorithm [26] across all $(K_{old} + K_{new})$ classes, which is the same as GCD [42].

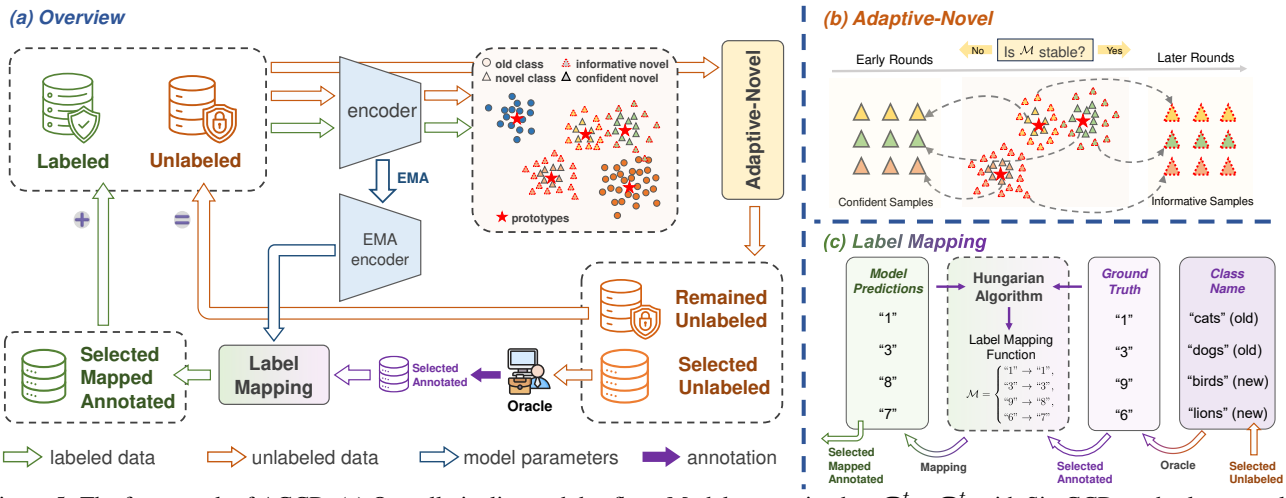

Figure 5. The framework of AGCD. (a) Overall pipeline and dataflow. Models are trained on $\mathcal{D}_l^t \cup \mathcal{D}_u^t$ with SimGCD, and select samples in $\mathcal{D}_u^t$. (b) The proposed `Adaptive-Novel` sampling strategy. Here $\mathcal{M}$ denotes the *label mapping function*. Stable $\mathcal{M}$ means that at the initial and end epochs of the current round, $\mathcal{M}$ does not change largely. *Confident novel samples* are sampled at early rounds, when $\mathcal{M}$ is stable, we select the *informative* ones. (c) Illustration of *label mapping* computed by model predictions and ground truth on $\mathcal{D}_l^{t-1} \cup \mathcal{D}_q^t$.

Novelty Evaluation. We also evaluate the selected samples' category attribution with the following metrics: (1) Novelty Coverage `Nov-C` that measures the coverage of new classes. (2) Novelty Ratio `Nov-R` that measures the ratio of the selected samples belonging to new classes. (3) Novelty Uniformity `Nov-U` that measures the uniformity of the coverage across new classes. (4) Novelty Information `Nov-I`, which considers both ratio and uniformity, a high value indicates one could neither randomly select samples across old and new classes nor from very few new classes. Specifically, these metrics are formulated as:

$$\texttt{Nov-C} = |\mathcal{C}_{new,select}|/K_{new}, \tag{7}$$

$$\texttt{Nov-R} = \sum_{i=1}^{N_{select}} \frac{\mathbb{1}(y_i \in \mathcal{C}_{new})}{N_{select}}, \tag{8}$$

$$\texttt{Nov-U} = -\sum_{c=1}^{K_{new}} \frac{N_{new,i}}{N_{select}} \log \frac{N_{new,i}}{N_{select}} / \log K_{new}, \tag{9}$$

$$\texttt{Nov-I} = \texttt{Nov-R} \times \texttt{Nov-U}, \tag{10}$$

where N_{select} and $N_{new,i}$ are the numbers of samples in total and those belonging to the i-th new class respectively. $\mathcal{C}_{new,select}$ denotes the selected new classes.

4.3. Adaptive Novel Sampling

In AGCD, we simultaneously consider three aspects: *novelty*, *informativeness* and *diversity* of samples and propose an adaptive sampling strategy called `Adaptive-Novel`, as in Fig. 5 (b). (1) For the aspect of novelty, as the initial labeling condition is severely imbalanced, we should give priority to selecting samples from new classes. Models' predictions $\hat{y}_i$ are proxies of samples' novelty. (2) For the aspect of diversity, we uniformly select samples from novel classes, *i.e.*, at each round, we select $\lfloor b/K_{new} \rfloor$ samples in each new class based on the model's prediction. (3) For the aspect of informativeness, we choose Margin [35] as the uncertainty metric. Selecting the most uncertain or informative samples [36, 38] has been a consensus in the literature of AL. However, new classes are initially fully unlabeled, and clusters of new classes might be unstable and biased [2], please refer to visualizations of the appendix, thus including difficult samples at early rounds to biased clusters hinders training. Here, we name two types of samples of novel classes, the most ambiguous and informative samples are named *informative novel samples* while the most certain ones are called *confident novel samples*. We devise an adaptive mechanism, where models select *confident novel samples* with minimum uncertainties to rectify and stabilize novel clusters at initial rounds. While at later rounds, *informative novel samples* with maximum uncertainties are selected to refine decision boundaries and further improve the performance.

In our method, we are expected to capture which type of samples are more important to the current model. To achieve this goal, we offer a heuristic criterion, where we regard the stability of label mapping $\mathcal{M}$ (as in Sec. 4.4) between model predictions and ground truth as the stability of the clusters, especially for new classes. If at round k, the change of $\mathcal{M}$ between the start and end epochs is negligible, the clustering is deemed stable and we can transfer to sample *informative novel samples* from round $k+1$.

4.4. Stable Label Mapping Algorithm

As the queried labels could not be directly used, one should perform *label mapping* to "translate" ground truth labels to the model's label space. as in Fig. 5 (c). We propose to calculate the mapping function from ground truth to the model's perspective via Hungarian algorithm [26], similar

Table 2. Comparative results of various methods with 5 rounds of active category discovery on generic datasets. Our method outperforms several uncertainty-based (Unc.) and representative/diversity-based (Rep./Div.) methods. Mean results over three runs are reported.

Type	AL Strategies	CIFAR10			CIFAR100			ImageNet-100		
		All	Old	New	All	Old	New	All	Old	New
Baseline	w/o AGCD	74.22	90.80	70.07	62.62	68.46	56.78	72.56	87.00	58.12
	`Random`	82.74	93.05	80.16	67.28	74.52	60.04	79.16	89.40	68.92
Unc.	`Entropy` [46]	76.25	95.55	71.43	64.59	73.94	55.24	75.96	91.04	60.88
	`LeastConf` [46]	78.32	**96.00**	73.90	65.63	**76.74**	54.52	76.82	91.92	61.72
	`Margin` [35]	92.34	94.35	91.84	69.08	75.58	62.58	80.46	92.40	68.52
Rep./Div.	`KMeans` [30]	91.18	93.10	90.70	66.70	72.66	60.74	78.18	90.08	66.28
	`CoreSet` [37]	85.51	94.95	83.15	65.72	77.64	53.80	78.08	91.92	64.24
	`BADGE` [3]	92.31	94.75	91.70	67.22	73.70	60.74	81.48	**92.68**	70.28
Ours	`Adaptive-Novel`	**93.15**	94.55	**92.80**	**71.25**	75.72	**66.78**	**83.34**	90.20	**76.48**

to Eq. (6). However, we can only perform label mapping using accessible labeled data, *i.e.*, $\mathcal{D}_l^t = \mathcal{D}_l^{t-1} \cup \mathcal{D}_q^t$ at round t, which is very limited, especially for new classes, and could bring about unstable results. To alleviate this, we maintain an exponential moving average (EMA) [17, 40, 56] of the model, and compute the *label mapping function* utilizing the EMA model's predictions on $\mathcal{D}_l^t = \mathcal{D}_l^{t-1} \cup \mathcal{D}_q^t$:

$$\mathcal{M}^t = \underset{m \in \mathcal{P}(\mathcal{C}_{all})}{\arg\max} \frac{1}{|\mathcal{D}_l^t|} \sum_{i \in \mathcal{D}_l^t} \mathbb{1}(m(y_i) = \hat{y}_i^{ema}), \qquad (11)$$

where y_i and $\hat{y}_i^{ema}$ are ground truth and predicted labels by EMA models. $\mathcal{P}(\mathcal{C}_{all})$ is the permutation across all classes $\mathcal{C}_{old} \cup \mathcal{C}_{new}$. $\mathcal{M}^t$ is a one-to-one mapping between two sets of classes. Then the mapped label of each query is $y_i^{map} = \mathcal{M}(y_i)$. Let $\mathcal{D}_{q,map}^t$ denote the query dataset after *label mapping*, the mapped labeled data is $\mathcal{D}_l^t = \mathcal{D}_l^{t-1} \cup \mathcal{D}_{q,map}^t$.

5. Experiments

5.1. Experimental Setup

Datasets. We construct AGCD on six datasets as shown in Table 1. For each dataset, K_{old} classes are selected as "old" classes, while the remaining K_{new} classes are "new" classes. We then sub-sample 20% of the training samples in K_{old} as the initial labeled dataset $\mathcal{D}_l^{\text{init}}$, while all remaining samples constitute the initial unlabeled part $\mathcal{D}_u^{\text{init}}$ for querying in subsequent rounds. The construction of $\mathcal{D}_l^{\text{init}}$ and $\mathcal{D}_u^{\text{init}}$ is similar to the literature of GCD [42, 47], but with fewer labeling ratio which is closer to a real-world scenario.

Evaluation. We compare the accuracy of AGCD with various query strategies. For fair comparisons, we use off-the-shelf SimGCD [47] for training as SimGCD is effective and the SOTA method in GCD. We employ model EMA for all query methods. Models are evaluated on disjoint test data using Eq. (6) in an *inductive* setting.

Query strategies for comparison. We compare our method `Adaptive-Novel` with various AL strategies [34, 38, 50], *e.g.*, Random Sampling (`Random`), uncertainty-based and representative/diversity-based sampling methods. For the uncertainty-based methods, we compare Maximum Entropy (`Entropy`) [46], and Least Confidence (`LeastConf`) [46], Least Margin (`Margin`) [35]. The representative-based methods include KMeans Clustering (`KMeans`) [30], Core-Set (`CoreSet`) [37], and Batch Active learning by Diverse Gradient Embeddings (`BADGE`) [3]. More details are in the Appendix.

AGCD pipeline and implementation details. Following GCD [42, 47], we employ ViT-B/16 [11] pre-trained by DINO [6] as the backbone, and fine-tune only the last transformer block for all experiments. The output of `[CLS]` token is chosen as the feature representation. The batch size for the original dataset $\mathcal{D}_l$ and $\mathcal{D}_u$ is 128. For queried samples, we use a smaller batch size $\mathcal{B}_q = 8$. We implement the base training stage like GCD for 200 epochs and choose models as initialization for AGCD. At each round, we train models on $\mathcal{D}_l^t$ and $\mathcal{D}_u^t$ by various query strategies for 15 epochs. All selection methods are trained using SimGCD with an initial learning rate of 0.1 and a cosine annealed schedule both in the base training stage and subsequent AGCD stage. All experiments are conducted on NVIDIA RTX A6000 GPUs. More details are in the appendix.

5.2. Comparative Results

`Adaptive-Novel` achieves stronger overall performance. As in Table 2 and Table 3, our method outperforms others consistently on various generic and fine-grained datasets. For example, on CIFAR100, our method outperforms `Random` by 3.97%/6.74%, and `CoreSet` by 5.53%/12.98% in terms of accuracy of all/new classes. Overall, our method exhibits an obvious advantage, especially in the accuracy of new classes.

Table 3. Comparative results of various methods with 5 rounds of active category discovery on fine-grained datasets. Our method outperforms several uncertainty-based (Unc.) and representative/diversity-based (Rep./Div.) methods. Mean results over three runs are reported.

Type	Query Strategies	CUB			Stanford Cars			FGVC-Aircraft		
		All	Old	New	All	Old	New	All	Old	New
Baseline	w/o AGCD	50.17	58.95	41.18	30.12	46.71	14.12	35.01	37.53	32.49
	`Random`	62.74	64.88	60.62	44.12	53.44	35.13	50.41	51.38	49.43
Unc.	`Entropy` [46]	62.82	**69.52**	56.19	42.40	53.75	31.44	43.89	51.92	35.86
	`LeastConf` [46]	61.48	66.12	56.87	45.82	55.32	36.65	44.91	50.42	39.40
	`Margin` [35]	65.08	68.41	61.79	46.03	57.67	34.79	51.37	**52.46**	50.27
Rep./Div.	`KMeans` [30]	61.30	68.27	54.40	40.79	52.99	29.03	51.58	51.08	52.07
	`CoreSet` [37]	63.44	65.95	60.96	42.52	52.00	33.37	45.03	51.68	38.38
	`BADGE` [3]	65.84	69.00	62.71	45.82	54.41	37.53	52.03	51.68	52.37
Ours	`Adaptive-Novel`	**66.62**	66.54	**66.70**	**48.36**	**57.73**	**39.34**	**53.74**	51.50	**55.98**

Table 4. Novelty metrics of all the selected data over 5 rounds on CIFAR100 and Stanford Cars.

AL Strategies	CIFAR100				Stanford Cars			
	`Nov-C`	`Nov-R`	`Nov-U`	`Nov-I`	`Nov-C`	`Nov-R`	`Nov-U`	`Nov-I`
`Random`	**1.00**	0.52	0.97	0.50	0.93	0.57	0.96	0.55
`Entropy`	0.90	0.44	0.91	0.40	0.85	0.64	0.92	0.59
`Margin`	0.96	0.63	0.95	0.60	0.90	0.66	0.93	0.61
`CoreSet`	0.96	0.61	0.94	0.57	0.89	**0.69**	0.94	0.65
`BADGE`	**1.00**	0.63	**0.98**	0.62	0.95	0.64	**0.97**	0.62
Ours	**1.00**	**0.71**	**0.98**	**0.70**	**0.98**	**0.69**	**0.97**	**0.67**

`Adaptive-Novel` achieves more balanced results between old and new classes. For all six datasets, the difference in accuracy between old and new classes of our method is minimal, indicating that our method effectively addresses the imbalanced issue in Sec. 3.2. One of the key insights is to prioritize samples from new classes for annotation, which helps to alleviate the inherent imbalanced labeling condition of GCD. For example in CUB, the divergence of old and new accuracy is 0.16%, while for other methods ranging from $\sim$ 4% to $\sim$ 14%. And in ImageNet-100, our method reduces the gap from $\sim$ 28% to 13.72%.

`Adaptive-Novel` significantly improves GCD with a limited budget size. As in Table 3, when selecting only $\sim$ 2.5 samples per class for annotation, the accuracy of new classes improves by 25.52%/25.22% on CUB/SCars, showcasing the efficiency and practicality of AGCD.

Do more novel samples necessarily lead to better performance? By comparing the results in Table 2, Table 3 and Table 4, our method generally samples more novel samples with more comprehensive class coverage, which contributes to the remarkable results. However, on Scars, `BADGE` selects more novel samples than `Margin`, but its performance is worse. As a result, solely sampling more novel class samples does not necessarily work fine, it is also important to consider the value of different samples.

Table 5. Ablations on three key factors, *i.e.*, novelty, informativeness and diversity for sample selection in AGCD.

ID	Novelty	Informativeness	Diversity	CIFAR100			CUB		
				All	Old	New	All	Old	New
(a)	✗	✗	✗	67.28	74.52	60.04	62.74	64.88	60.62
(b)	✓	✗	✗	69.33	72.76	65.90	63.58	63.31	63.85
(c)	✓	✓	✗	69.65	75.56	63.74	64.28	65.33	63.23
(d)	✓	✓	✓	**71.25**	**75.72**	**66.78**	**66.62**	**66.54**	**66.70**

5.3. Ablation Studies

In this section, we implement extensive experiments to validate the effectiveness of each component, including three aspects for sample selection in Table 5, the adaptive mechanism in Fig. 6 and model EMA in Fig. 7.

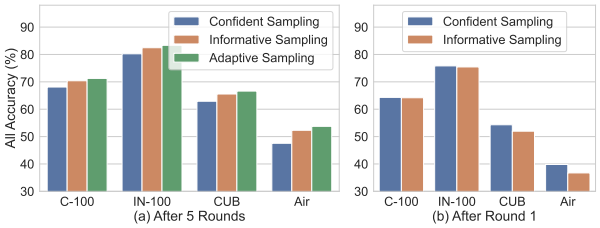

Figure 6. Ablation on adaptive sampling.

The effect of three factors: Novelty & Informativeness & Diversity. In Table 5, (a) denotes `Random`, (b) is random selection within new classes according to predictions, (c) denotes sampling informative instances from the entirety of samples predicted as new classes, and (d) is our method. The difference between (c) and (d) is that (d) samples informative instances in a class-wise manner, with $\lfloor b/K_{new} \rfloor$ of each new class to ensure diversity. (b) outperforms (a) by 2.05% and 0.84% on two datasets. (c) is slightly better than (b) for considering informativeness. Compared with (c), our method obtains consistent improvements in both old and new classes (+1.21% and +3.47% in old and new on CUB), indicating the importance of diversity.

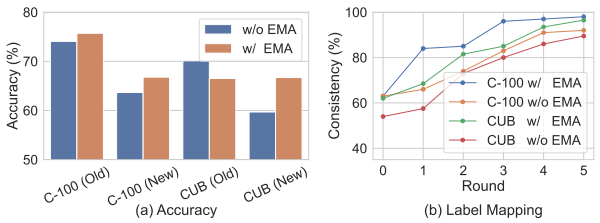

Figure 7. Ablation on model EMA.

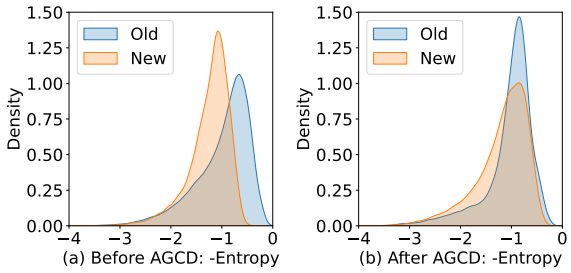

Figure 8. Confidence distribution before and after AGCD.

The effect of adaptive sampling. Our adaptive mechanism selects *confident novel samples* at early rounds while *informative samples* at later rounds. We compare our results with two singular baseline strategies in Fig. 6. Ours consistently outperforms the other two strategies. For baseline methods, informative sampling outperforms confident sampling after five rounds. However, as shown in Fig. 6 (b), confident sampling works better at early rounds, which aligns with our rationale in Sec. 4.3. That is, models require more confident samples for early learning of stable clusters.

The effect of model EMA. As in Fig. 7 (b), we compute the consistency of *label mapping function* computed on limited $\mathcal{D}_l^t$ and the whole test data. Results validate that EMA provides a more stable and consistent $\mathcal{M}$, which is suitable for *inductive* evaluation, and obtains better results on various datasets in Fig. 7 (a).

5.4. Further Analysis

Confidence consistency. Fig. 8 reveals that AGCD improves confidence consistency between old and new classes. Before AGCD, there was a noticeable gap between the confidence distributions of new and old classes, with a peak difference of ~ 0.5. After AGCD, the peak gap is almost reduced to zero. As a result, our method effectively addresses two issues of GCD including imbalanced accuracy and confidence with an affordable annotation budget.

Table 6. Results of All Acc on CUB of various initial label ratios.

label ratio	0	0.01	0.05	0.1	0.2	0.3
w/o AGCD	14.15	18.12	27.68	37.49	50.17	58.49
`Random`	31.46	32.00	45.88	53.16	62.74	66.45
`Entropy`	32.02	36.95	46.32	53.78	62.82	60.55
Ours	**33.36**	**38.73**	**46.88**	**55.30**	**66.62**	**69.47**

Table 7. Results on CUB with various budge sizes per round b. Models are trained with 3 AGCD rounds.

b	30	50	100	300	500
`Random`	52.92	56.44	58.54	66.78	72.68
`Entropy`	52.90	54.47	58.87	68.69	70.87
Ours	**53.40**	**56.50**	**59.86**	**69.59**	**73.56**

Table 8. Results of AGCD with an unknown class number (estimated class number) on ImageNet-100 and CUB.

Strategies	CUB			ImageNet-100		
	All	Old	New	All	Old	New
`Random`	60.68	63.31	58.08	77.86	**90.34**	65.38
`Entropy`	60.48	64.84	56.15	76.12	71.52	60.72
Ours	**64.14**	**66.09**	**62.20**	**82.46**	89.84	**70.64**

Various label ratios and budget sizes. We conduct experiments with different settings, including various initial labeling ratios in Table 6 and various budget sizes b in Table 7. The proposed strategy `Adaptive-Novel` consistently outperforms others across various settings, which showcases the sample selection aspects are general and robust to different settings in AGCD.

Unknown class number. We also consider the scenarios with unknown class number K_{new} in Table. 8. We perform an off-the-shelf number estimation algorithm [42] to get an estimation of K_{new} in advance, and use it to construct classifiers. Results in Table 8 show that `Adaptive-Novel` is robust to the unknown class numbers, indicating the superiority of our method. Details are shown in the appendix.

6. Conclusions

In this paper, we propose a new setting of Active Generalized Category Discovery (AGCD) to address the inherent and intractable issues of GCD. Moreover, we pose two unique challenges in AGCD, *i.e.*, new classes in unlabeled data and the clustering nature of GCD. To solve these challenges, we propose an adaptive query strategy `Adaptive-Novel` considering novelty, informativeness and diversity, which adaptively selects samples with proper uncertainty. Besides, we further propose a stable *label mapping* algorithm to address the issue of different ordering of label indices between ground truth labels and models' label space. Experiments show that our method achieves state-of-the-art performance across different scenarios.

Acknowledgements This work has been supported by the National Science and Technology Major Project (2022ZD0116500), National Natural Science Foundation of China (U20A20223, 62222609, 62076236), CAS Project for Young Scientists in Basic Research (YSBR-083), and Key Research Program of Frontier Sciences of CAS (ZDBS-LY-7004).

References

[1] Sharat Agarwal, Himanshu Arora, Saket Anand, and Chetan Arora. Contextual diversity for active learning. In *European Conference on Computer Vision*, pages 137–153, 2020. 3

[2] Eric Arazo, Diego Ortego, Paul Albert, Noel E O'Connor, and Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2020. 5

[3] Jordan T. Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. In *International Conference on Learning Representations*, 2020. 3, 6, 7

[4] Kaidi Cao, Maria Brbic, and Jure Leskovec. Open-world semi-supervised learning. In *International Conference on Learning Representations*, 2022. 3

[5] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020. 2

[6] 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 International Conference on Computer Vision (ICCV)*, 2021. 6

[7] Rui Castro, Charles Kalish, Robert Nowak, Ruichen Qian, Tim Rogers, and Jerry Zhu. Human active learning. *Advances in neural information processing systems*, 21, 2008. 2

[8] 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. 2, 3

[9] Florent Chiaroni, Jose Dolz, Ziko Imtiaz Masud, Amar Mitiche, and Ismail Ben Ayed. Parametric information maximization for generalized category discovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1729–1739, 2023. 3

[10] 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. 4

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 6

[12] Yixin Fei, Zhongkai Zhao, Siwei Yang, and Bingchen Zhao. Xcon: Learning with experts for fine-grained category discovery. In *British Machine Vision Conference (BMVC)*, 2022. 2, 3

[13] Enrico Fini, Enver Sangineto, Stéphane Lathuiliere, 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. 2

[14] Chuanxing Geng, Sheng-Jun Huang, and Songcan Chen. Recent advances in open set recognition: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43 (10):3614–3631, 2021. 2

[15] Peiyan Gu, Chuyu Zhang, Ruijie Xu, and Xuming He. Class-relation knowledge distillation for novel class discovery. *lamp*, 12(15.0):17–5, 2023. 3

[16] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017. 3

[17] Yuxin Guo, Shijie Ma, Hu Su, Zhiqing Wang, Yuhao Zhao, Wei Zou, Siyang Sun, and Yun Zheng. Dual mean-teacher: An unbiased semi-supervised framework for audio-visual source localization. In *Advances in Neural Information Processing Systems*, pages 48639–48661, 2023. 2, 6

[18] 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. 2

[19] Kai Han, Sylvestre-Alvise Rebuffi, Sebastien Ehrhardt, Andrea Vedaldi, and Andrew Zisserman. Automatically discovering and learning new visual categories with ranking statistics. In *International Conference on Learning Representations*, 2020. 2

[20] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017. 4

[21] Wei-Ning Hsu and Hsuan-Tien Lin. Active learning by learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2015. 3

[22] Sheng-Jun Huang, Rong Jin, and Zhi-Hua Zhou. Active learning by querying informative and representative examples. *Advances in neural information processing systems*, 23, 2010. 3

[23] 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. 2, 3

[24] 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. 4

[25] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 4

[26] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955. 4, 5

[27] Junnan Li, Pan Zhou, Caiming Xiong, and Steven Hoi. Prototypical contrastive learning of unsupervised representations. In *International Conference on Learning Representations*, 2021. 3

[28] Shijie Ma, Fei Zhu, Zhen Cheng, and Xu-Yao Zhang. Towards trustworthy dataset distillation. *arXiv preprint arXiv:2307.09165*, 2023. 2

[29] Cathlin Macaulay. Transfer of learning. In *Transfer of learning in professional and vocational education*, pages 17–22. Routledge, 2002. 1

[30] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, pages 281–297. Oakland, CA, USA, 1967. 2, 3, 6, 7

[31] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 4

[32] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009. 2

[33] Nan Pu, Zhun Zhong, and Nicu Sebe. Dynamic conceptional contrastive learning for generalized category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7579–7588, 2023. 2, 3

[34] Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B Gupta, Xiaojiang Chen, and Xin Wang. A survey of deep active learning. *ACM computing surveys (CSUR)*, 54(9):1–40, 2021. 2, 6

[35] Dan Roth and Kevin Small. Margin-based active learning for structured output spaces. In *Machine Learning: ECML 2006: 17th European Conference on Machine Learning Berlin, Germany, September 18-22, 2006 Proceedings 17*, pages 413–424. Springer, 2006. 3, 4, 5, 6, 7

[36] Tobias Scheffer, Christian Decomain, and Stefan Wrobel. Active hidden markov models for information extraction. In *International symposium on intelligent data analysis*, pages 309–318. Springer, 2001. 5

[37] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*, 2018. 3, 6, 7

[38] Burr Settles. Active learning literature survey. 2009. 2, 3, 5, 6

[39] Claude Elwood Shannon. A mathematical theory of communication. *ACM SIGMOBILE mobile computing and communications review*, 5(1):3–55, 2001. 4

[40] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017. 2, 6

[41] Colin Troisemaine, Vincent Lemaire, Stéphane Gosselin, Alexandre Reiffers-Masson, Joachim Flocon-Cholet, and Sandrine Vaton. Novel class discovery: an introduction and key concepts. *arXiv preprint arXiv:2302.12028*, 2023. 2

[42] 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. 2, 3, 4, 6, 8

[43] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: A good closed-set classifier is all you need. In *International Conference on Learning Representations*, 2022. 2

[44] Sagar Vaze, Andrea Vedaldi, and Andrew Zisserman. No representation rules them all in category discovery. In *Advances in Neural Information Processing Systems*, pages 19962–19989, 2023. 3

[45] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 4

[46] Dan Wang and Yi Shang. A new active labeling method for deep learning. In *2014 International joint conference on neural networks (IJCNN)*, pages 112–119. IEEE, 2014. 3, 6, 7

[47] Xin Wen, Bingchen Zhao, and Xiaojuan Qi. Parametric classification for generalized category discovery: A baseline study. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16590–16600, 2023. 2, 3, 4, 6

[48] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *arXiv preprint arXiv:2110.11334*, 2021. 2

[49] Xueying Zhan, Huan Liu, Qing Li, and Antoni B Chan. A comparative survey: Benchmarking for pool-based active learning. In *IJCAI*, pages 4679–4686, 2021. 3

[50] Xueying Zhan, Qingzhong Wang, Kuan-hao Huang, Haoyi Xiong, Dejing Dou, and Antoni B Chan. A comparative survey of deep active learning. *arXiv preprint arXiv:2203.13450*, 2022. 2, 6

[51] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018. 2

[52] Sheng Zhang, Salman Khan, Zhiqiang Shen, Muzammal Naseer, Guangyi Chen, and Fahad Shahbaz Khan. 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*, pages 3479–3488, 2023. 2, 3

[53] Bingchen Zhao, Xin Wen, and Kai Han. Learning semi-supervised gaussian mixture models for generalized category discovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16623–16633, 2023. 2, 3

[54] 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. 2

[55] Fei Zhu, Xu-Yao Zhang, Chuang Wang, Fei Yin, and Cheng-Lin Liu. Prototype augmentation and self-supervision for incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5871–5880, 2021. 2

[56] Fei Zhu, Zhen Cheng, Xu-Yao Zhang, and Cheng-Lin Liu. Rethinking confidence calibration for failure prediction. In *European Conference on Computer Vision*, pages 518–536. Springer, 2022. 2, 6

[57] Fei Zhu, Shijie Ma, Zhen Cheng, Xu-Yao Zhang, Zhaoxiang Zhang, and Cheng-Lin Liu. Open-world machine learning: A review and new outlooks. *arXiv preprint arXiv:2403.01759*, 2024. 2

[58] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2020. 2