Active Generalized Category Discovery

Shijie Ma\textsuperscript{1,2}, Fei Zhu\textsuperscript{3}, Zhun Zhong\textsuperscript{4,5}, Xu-Yao Zhang\textsuperscript{1,2,}\textsuperscript{*}, Cheng-Lin Liu\textsuperscript{1,2}

\textsuperscript{1}MAIS, Institute of Automation, Chinese Academy of Sciences, China
\textsuperscript{2}School of Artificial Intelligence, University of Chinese Academy of Sciences, China
\textsuperscript{3}Centre for Artificial Intelligence and Robotics, HKISI-CAS, China
\textsuperscript{4}School of Computer Science and Information Engineering, Hefei University of Technology, China
\textsuperscript{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

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-
categories in the unlabeled data. In addition, models should actively query some confusing samples. Therefore, we argue that some annotations of new classes are fully unlabeled, models would 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. 

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) is a unified objective to handle old and new classes jointly via swapped prediction. NCD assumes all unlabeled data are from new classes.

![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.](image)

features 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. Deep learning is also inspired by the cognitive processes above, and could be endowed with the abilities of knowledge transfer and active learning, especially in the open-environments. K-Means and unsupervised contrastive learning containing unlabeled novel categories. In this paper, we study the task of Generalized Category Discovery (GCD) 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 leverage supervised learning and unsupervised contrastive learning with non-parametric K-Means for clustering. Later works further exploit underly the cross-instance relationships. We 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 and inconsistent confidence. 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 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.
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 KMMeans [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 \( D_l = \{(x_i^l, y_i^l)\} \subset \mathcal{X} \times \mathcal{Y}_l \) and an unlabeled dataset \( D_u = \{(x_i^u, y_i^u)\} \subset \mathcal{X} \times \mathcal{Y}_u \), \( D_l \) only contains old classes, while \( D_u \) 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 \( D_u \). The number of novel classes \( |\mathcal{C}_{new}| \) is known a-prior or estimated [33, 42, 53]. \( f(\cdot) \) and \( g(\cdot) \) are feature extractor and projection head for contrastive learning respectively. \( h_i = f(x_i) \) and \( z_i = g(h_i) \) are \( \ell_2 \) 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 \( B \):

\[
L_{con}^l = \frac{1}{|B_l|} \sum_{i \in B_l} \frac{1}{|\mathcal{N}(i)|} \sum_{q \in \mathcal{N}(i)} - \log \frac{\exp(z_i^l z_q^l / \tau_c)}{\sum_{n \neq i} \exp(z_i^l z_n^l / \tau_c)},
\]

\[
L_{con}^u = \frac{1}{|B|} \sum_{i \in B} - \log \frac{\exp(z_i^u z_i^l / \tau_c)}{\sum_{n \neq i} \exp(z_i^u z_n^l / \tau_c)}.
\]

The overall contrastive loss \( L_{con} = (1 - \lambda)L_{con}^u + \lambda L_{con}^l \).

SimGCD [47] employs a parametric prototypical classifier \( \mathcal{C} = \{c_1, \cdots, c_K\} \), where \( K = K_{old} + K_{new} \). The posterior probability could be expressed as:

\[
p_i^{(k)} = \frac{\exp(h_i^\top c_k)}{\sum_{u} \exp(h_i^\top c_u)} / \tau_p.
\]

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

\[
L_{cls}^u = \frac{1}{|B|} \ell(q_i^l, p_i^l) - \lambda \cdot H(\overline{p}),
\]

where \( q_i^l \) is a sharpened probability of another view, and \( \overline{p} = \frac{1}{|B|} \sum_{i \in B} (p_i^l + p_i^u) \). \( \ell(\cdot) \) denotes cross-entropy loss. The supervised loss is also employed on \( D_l \) with labels \( y_i^l \):

\[
L_{cls}^l = \frac{1}{|B|} \sum_{i \in B_l} \ell(y_i^l, p_i^l).
\]

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.](image)

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.
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). First, 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 $D^{init}_l = \{(x^l_i, y_i^l)\} \subseteq \mathcal{X} \times \mathcal{Y}_l$ and $D^{init}_u = \{(x^u_i, y_i^u)\} \subseteq \mathcal{X} \times \mathcal{Y}_u$ 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 $D_{u,t-1}$ and queries its labels to obtain $D_q^{t} = \{(x^u_i, y^u_i)\}$. Then labeled and unlabeled data are updated as $D^{l}_t = D^{l}_{t-1} \cup D_q^{t} \subseteq \mathcal{X} \times \mathcal{Y}_l$, $D^{u}_t = D^{u}_{t-1} \setminus D_q^{t}$. Models are then trained on $D^{l}_t \cup D^{u}_t$ with off-the-shelf SimGCD [47]. Note that initial labeled data only contains old classes, i.e., $\mathcal{Y}_t^u = \mathcal{C}_{old}$, but after querying in AGCD, the labeled data $D^{l}_q$ could contain some classes of all novel ones $\mathcal{C}_{new}$. The total budget size is $b \times n$. For the queried data $D^{l}_q$, models are trained with supervised loss $\mathcal{L}_{con}$ in Eq. (1) and $\mathcal{L}_{cls}$ in Eq. (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 $D_u$, 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 $D_u$ but are also trained on it.

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

| Dataset | Labeled $|\mathcal{D}^l_t|$ | Unlabeled $|\mathcal{D}^u_t|$ | #Rounds | #Query (total) | #Query (per class) |
|---------|----------------|----------------|--------|----------------|-------------------|
| CIFAR10 (C-10) [23] | 2,000 | 48,000 | 10 | 100 | 10 |
| CIFAR100 (C-100) [23] | 5,000 | 45,000 | 100 | 5 | 500 | 5 |
| ImageNet-100 (IN-100) [10] | 12,744 | 114,371 | 100 | 5 | 500 | 5 |
| SUN (CUB) [45] | 599 | 5,395 | 200 | 5 | 500 | 2.5 |
| Stanford Cars (SCars) [24] | 800 | 7,344 | 196 | 5 | 500 | 2.5 |
| FGVC-Aircraft (Av) [13] | 666 | 6,001 | 100 | 5 | 500 | 5 |

4.2. Evaluation and Metrics

Accuracy Evaluation. GCD adopts a transductive evaluation on unlabeled training data $D_u$. 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 $D_u$, making it unfair for evaluation. The accuracy is calculated using ground truth labels $y_i$ and models’ predictions $\hat{y}_i$ as follows:

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

$M = |D_u|$ 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].
**Novelty Evaluation.** We also evaluate the selected samples’ category attribution with the following metrics: (1) Novelty Coverage $\text{Nov-C}$ that measures the coverage of new classes. (2) Novelty Ratio $\text{Nov-R}$ that measures the ratio of the selected samples belonging to new classes. (3) Novelty Uniformity $\text{Nov-U}$ that measures the uniformity of the coverage across new classes. (4) Novelty Information $\text{Nov-I}$, 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:

\[
\text{Nov-C} = \left| C_{\text{new},\text{select}} \right| / K_{\text{new}},
\]

\[
\text{Nov-R} = \frac{1}{N_{\text{select}}} \sum_{i=1}^{N_{\text{select}}} \mathbb{1}(y_i \in C_{\text{new}}),
\]

\[
\text{Nov-U} = -\sum_{c=1}^{K_{\text{new}}} \frac{N_{\text{new},i}}{N_{\text{select}}} \log \frac{N_{\text{new},i}}{N_{\text{select}}} / \log K_{\text{new}},
\]

\[
\text{Nov-I} = \text{Nov-R} \times \text{Nov-U},
\]

where $N_{\text{select}}$ and $N_{\text{new},i}$ are the numbers of samples in total and those belonging to the $i$-th new class respectively. $C_{\text{new},\text{select}}$ 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 $[b/K_{\text{new}}]$ 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 $M$ (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 $M$ 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...
to Eq. (6). However, we can only perform label mapping using accessible labeled data, i.e., $D_t^l = D_t^{old} \cup D_t^{old}$ 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 $D_t^l = D_t^{old} \cup D_t^{old}$:

$$M^t = \arg \max_{m \in \mathcal{P}(C_{all})} \frac{1}{|D_t^l|} \sum_{i \in D_t^l} 1\{m(y_i) = \hat{y}_i^{ema}\},$$ \hspace{1cm} (11)

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

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 $D_{init}^l$, while all remaining samples constitute the initial unlabeled part $D_{init}^u$ for querying in subsequent rounds. The construction of $D_{init}^l$ and $D_{init}^u$ 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 $D_t^l$ and $D_t^u$ is 128. For queried samples, we use a smaller batch size $B_q = 8$. 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 $D_t^l$ and $D_t^u$ 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>
<thead>
<tr>
<th>Type</th>
<th>AL Strategies</th>
<th>CIFAR10 All</th>
<th>CIFAR10 Old</th>
<th>CIFAR10 New</th>
<th>ImageNet-100 All</th>
<th>ImageNet-100 Old</th>
<th>ImageNet-100 New</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline w/o AGCD</td>
<td>74.22 90.80 70.07</td>
<td>62.62 68.46 56.78</td>
<td>72.56 87.00 58.12</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Random</td>
<td>82.74 93.05 80.16</td>
<td>67.28 74.52 60.04</td>
<td>79.16 89.40 68.92</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unc.</td>
<td>Entropy [46] 76.25 95.55 71.43</td>
<td>64.59 73.94 55.24</td>
<td>75.96 91.04 60.88</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>LeastConf [46] 78.32 96.00 73.90</td>
<td>65.63 76.74 54.52</td>
<td>76.82 91.92 61.72</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Margin [35] 92.34 94.35 91.84</td>
<td>69.08 75.58 62.58</td>
<td>80.46 92.40 68.52</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rep./Div.</td>
<td>KMeans [30] 91.18 93.10 90.70</td>
<td>66.70 72.66 60.74</td>
<td>78.18 90.08 66.28</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>CoreSet [37] 85.51 94.95 83.15</td>
<td>65.72 77.64 53.80</td>
<td>78.08 91.92 64.24</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>BADGE [3] 92.31 94.75 91.70</td>
<td>67.22 73.70 60.74</td>
<td>81.48 92.68 70.28</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ours</td>
<td>Adaptive-Novel 93.15 94.55 92.80</td>
<td>71.25 75.72 66.78</td>
<td>83.34 90.20 76.48</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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 ~28% to 13.72%.

Adaptive-Novel significantly improves GCD with a limited budget size. As in Table 3, when selecting only ~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.

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.

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.

<table>
<thead>
<tr>
<th>Type</th>
<th>Query Strategies</th>
<th>CUB</th>
<th>Stanford Cars</th>
<th>FGVC-Aircraft</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>All</td>
<td>Old</td>
<td>New</td>
</tr>
<tr>
<td>Baseline</td>
<td>w/o AGCD</td>
<td>50.17</td>
<td>58.95</td>
<td>41.18</td>
</tr>
<tr>
<td></td>
<td>Random</td>
<td>62.74</td>
<td>64.88</td>
<td>60.62</td>
</tr>
<tr>
<td>Unc.</td>
<td>Entropy [46]</td>
<td>62.82</td>
<td>69.52</td>
<td>56.19</td>
</tr>
<tr>
<td></td>
<td>LeastConf [46]</td>
<td>61.48</td>
<td>66.12</td>
<td>56.87</td>
</tr>
<tr>
<td></td>
<td>Margin [35]</td>
<td>65.08</td>
<td>68.41</td>
<td>61.79</td>
</tr>
<tr>
<td>Rep./Div.</td>
<td>KMeans [30]</td>
<td>61.30</td>
<td>68.27</td>
<td>54.40</td>
</tr>
<tr>
<td></td>
<td>CoreSet [37]</td>
<td>63.44</td>
<td>65.95</td>
<td>60.96</td>
</tr>
<tr>
<td></td>
<td>BADGE [3]</td>
<td>65.84</td>
<td>69.00</td>
<td>62.71</td>
</tr>
<tr>
<td>Ours</td>
<td>Adaptive-Novel</td>
<td>66.62</td>
<td>65.54</td>
<td>66.70</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>ID Strategies</th>
<th>CIFAR100</th>
<th>Stanford Cars</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Nov-C</td>
<td>Nov-R</td>
</tr>
<tr>
<td>Random</td>
<td>1.00</td>
<td>0.52</td>
</tr>
<tr>
<td>Entropy</td>
<td>0.90</td>
<td>0.44</td>
</tr>
<tr>
<td>Margin</td>
<td>0.96</td>
<td>0.63</td>
</tr>
<tr>
<td>CoreSet</td>
<td>0.96</td>
<td>0.61</td>
</tr>
<tr>
<td>BADGE</td>
<td>1.00</td>
<td>0.63</td>
</tr>
<tr>
<td>Ours</td>
<td>1.00</td>
<td>0.71</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>ID</th>
<th>Novelty</th>
<th>Informativeness</th>
<th>Diversity</th>
</tr>
</thead>
<tbody>
<tr>
<td>CIFAR100</td>
<td>All</td>
<td>Old</td>
<td>New</td>
</tr>
<tr>
<td>(a)</td>
<td>✓</td>
<td>✓</td>
<td>✗</td>
</tr>
<tr>
<td>(b)</td>
<td>✓</td>
<td>✗</td>
<td>✓</td>
</tr>
<tr>
<td>(c)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>(d)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

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 \[b/K_{new}\] 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.
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 D^l and the whole test data. Results validate that EMA provides a more stable and consistent 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.

<table>
<thead>
<tr>
<th>label ratio</th>
<th>0</th>
<th>0.01</th>
<th>0.05</th>
<th>0.1</th>
<th>0.2</th>
<th>0.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>w/o AGCD</td>
<td>14.15</td>
<td>18.12</td>
<td>27.68</td>
<td>37.49</td>
<td>50.17</td>
<td>58.49</td>
</tr>
<tr>
<td>Random</td>
<td>31.46</td>
<td>32.00</td>
<td>45.88</td>
<td>53.16</td>
<td>62.74</td>
<td>66.45</td>
</tr>
<tr>
<td>Entropy</td>
<td>32.02</td>
<td>36.95</td>
<td>46.32</td>
<td>53.78</td>
<td>62.82</td>
<td>60.55</td>
</tr>
<tr>
<td>Ours</td>
<td>33.36</td>
<td>38.73</td>
<td><strong>46.88</strong></td>
<td><strong>55.30</strong></td>
<td><strong>66.62</strong></td>
<td><strong>69.47</strong></td>
</tr>
</tbody>
</table>

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


[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.


