Active Generalized Category Discovery

Supplementary Material

Overview

In this appendix, we provide additional descriptions of the following contents:

- Fundamental principles of AGCD are discussed in Appendix A, including the relationship to some similar settings and the main concern of this paper.
- More details of the proposed sampling strategy are elaborated in Appendix B.
- We provide more experimental details in Appendix C.
- Additional quantitative and qualitative results are provided in Appendix D and Appendix E, respectively.

A. Fundamental Principles of AGCD

In this section, we compare several related settings, and distinguish between sample selection strategies and training methods, to clarify the positioning of our contribution.

Comparisons to related settings. We compare conventional Active Learning (AL), Generalized Category Discovery (GCD) and the proposed Active Generalized Category Discovery (AGCD):

- AL aims to greatly improve models' performance with affordable labeling budgets. The spirit of AL is to select valuable and informative samples for labeling and incorporate newly annotated samples into the training set for subsequent training, which helps disambiguate confusing samples and correct previous errors. AL principally follows a closed-world setting, where unlabeled data contains the same set of classes as the labeled data.
- GCD is an open-world task, which aims to classify old classes and discover novel classes in the unlabeled data, given some labeled samples from old classes. However, GCD itself is not a fully learnable task in that the knowledge from old classes is not fully transferable to new classes and new classes are purely unlabeled, resulting in intractable problems including imbalanced accuracy and confidence between old and new classes, as in Sec. 3.2.
- AGCD is proposed to address the intractable problems and largely enhance the performance of GCD with minimal annotation costs. AGCD adopts a similar spirit to AL. Additionally, AGCD is a more general setting and generalizes AL to the open-world setting. We could refer to GCD as an extrapolated version of AL, where new classes could exist in unlabeled data, and models are required to classify both old and new classes. With additional labeled data, especially those from new classes, models could rectify previously inevitable errors and biases, and clarify the decision boundaries. Unlike GCD, models in AGCD are evaluated in *inductive settings* with disjoint test datasets.

Sample selection strategies and training methods. AGCD mainly involves two components, including sample selection strategies and training methods. They are complementary to each other. Models first select informative samples to obtain annotations from the oracle, and are subsequently trained on them using specific training methods.

Main concern and focus of this paper. In this paper, our main concern is how to select valuable samples from the data perspective and why conventional AL strategies are not applicable to the task of AGCD, and we propose a sample selection strategy called `Adaptive-Novel`. To validate its effectiveness, we compare `Adaptive-Novel` with various strategies in the literature of AL, as elaborated in Appendix B. For training methods, we directly employ the off-the-shelf method SimGCD [19] for all strategies for fair comparison, considering that SimGCD [19] is the SOTA method in GCD and it employs parametric classifiers which are efficient for training and convenient for evaluation including accuracy and confidence. By contrast, non-parametric methods [11, 16] require independent K-Means [7] clustering, which is inefficient and non-trivial to analyze confidence and evaluate in *inductive* settings.

Although the training methodology, *i.e.*, the loss function, is not the focus of this paper, we still need to address specific issues over the course of AGCD training, *i.e.*, different ordering of label indices in clustering problems. Specifically, in GCD, we implement clustering on new classes, the core is to cluster samples of the same novel class together and separate samples from different classes. The specific assignment of the labels to each cluster is not crucial. As a result, the label indices of new classes differ across various experiments and are unpredictable. That is the reason why it is common to employ the Hungarian algorithm [6] to evaluate the performance of GCD, as in Eq. (6). For example, the model might assign "8" and "9" to "birds" in two runs. This brings about a challenging issue in AGCD, *i.e.*, the queried ground truth labels could not be directly used by the model due to the different ordering between the model's predictions and ground truth labels. As a result, it is necessary to obtain a *mapping function* to convert the ground truth labels to the label space of the classifier in advance, as discussed in Sec. 4.4. Considering that the labeled data $\mathcal{D}_l^t$ is limited, especially for new classes, this would introduce instability to the computation of *label mapping* $\mathcal{M}^t$. To overcome this issue, we propose a stable labeling mapping method. Specifically, we perform Hungarian optimal assignment algorithm [6] between the EMA model's predictions $\hat{y}_i^{ema}$ and the ground truth labels

y_i over the accessible labeled data of the current round, *i.e.*, $\mathcal{D}_l^t = \mathcal{D}_l^{t-1} \cup \mathcal{D}_q^t$, as in Eq. (11). The EMA model could offer more stable predictions during training. Additionally, *label mapping* $\mathcal{M}^t$ is computed in every iteration, so do the mapped ground truth labels $\mathcal{M}(y_i)$ and $\mathcal{D}_{l,map}^{t-1} \cup \mathcal{D}_{q,map}^t$, because the model's parameters are continuously updated, especially at the beginning of each round, we should also keep the *label mapping* up-to-date to fit the changing model.

B. More Details about `Adaptive-Novel`

Adaptive sampling mechanism. In `Adaptive-Novel` strategy, we propose an adaptive mechanism, *i.e.*, at early rounds, we sample *confident novel samples* to stabilize new clusters while *informative novel samples* at later rounds to refine decision boundaries. We propose to transfer from the former to the latter when the clusters of new classes are stable. Technically, we compare the *label mapping* at the initial $\mathcal{M}_{init}^t$ and final epochs $\mathcal{M}_{final}^t$ of current round t, when the difference is lower than a pre-defined threshold δ, the clusters are deemed stable, and we transfer to seek for *informative novel samples* at the next round $t+1$, the difference is computed as follows:

$$\text{diff} = \frac{\sum_{i=1}^{K} \mathbb{1}(\mathcal{M}_{init}^t[i] \neq \mathcal{M}_{final}^t[i])}{K} \tag{S1}$$

where $K = K_{old} + K_{new}$ denotes the total number of classes, and $\mathcal{M}^t[i]$ denotes the mapping $\mathcal{M}^t$ from the i-th label index of the ground truth labels to the $\mathcal{M}^t[i]$-th classifier's predictive label index. In all our experiments, we set $\delta = 0.1$ for all six datasets.

Uncertainty metrics. We mainly consider three uncertainty/confidence metrics, including maximum softmax probability (MSP [5]) $p(\hat{y}_1|\mathbf{x})$, margin $p(\hat{y}_1|\mathbf{x}) - p(\hat{y}_2|\mathbf{x})$ and entropy $-\sum_{i=1}^{K} p(y = i|\mathbf{x}) \log p(y = i|\mathbf{x})$, where $\hat{y}_1 = \arg\max_y p(y|\mathbf{x})$ and $\hat{y}_2 = \arg\max_{y \neq \hat{y}_1} p(y|\mathbf{x})$ represent two most likely labels of sample $\mathbf{x}$. Informative samples refer to those with maximum uncertainty, *i.e.*, minimum MSP value, minimum margin value and maximum entropy value. In the setting AGCD, we found that margin is more robust and could select more novel samples, as a result, we choose margin as the uncertainty metric in `Adaptive-Novel`. We further provide results of different metrics in Appendix D.

`Adaptive-Novel` sampling algorithm. Here we give the exact algorithm of `Adaptive-Novel` as in Algorithm 1. We highlight three aspects for sample selection, including novelty, informativeness and diversity, with red colors. Note that the *label mapping* is performed in each iteration when the model is updated.

C. More Experimental Details

Comparative strategies. We compare the proposed `Adaptive-Novel` with various query strategies in the literature of conventional AL, including uncertainty-based and diversity/representative-based methods:

- `Random`: a baseline that randomly selects samples from the unlabeled dataset.
- `Entropy` [18]: an uncertainty-based method that selects samples with the highest entropy over all classes $-\sum_{i=1}^{K} p(y = i|\mathbf{x}) \log p(y = i|\mathbf{x})$.
- `LeastConf` [18]: an uncertainty-based method that selects samples with the lowest MSP [5] $p(\hat{y}_1|\mathbf{x})$.
- `Margin` [12]: an uncertainty-based method that selects samples with the lowest margin $p(\hat{y}_1|\mathbf{x}) - p(\hat{y}_2|\mathbf{x})$.
- `KMeans` [7]: a diversity-based method that selects samples closest to the centroids of K-Means, which is implemented in the embedding space in a cluster-wise manner.
- `CoreSet` [13] picks up unlabeled samples with the greatest distances to their nearest labeled neighbor, and obtains representative samples of unlabeled data.
- `BADGE` [1] is short for Batch Active learning by Diverse Gradient Embeddings and could be viewed as a hybrid method to query centroids from K-Means clustering over the gradient embeddings.

Other implementation details. We adopt all training parameters from SimGCD [19]. The weight of supervised loss λ is 0.35, and the weight λ_e depends on specific datasets, we set $\lambda_e = 1$ for CIFAR10, Aircraft and Stanford Cars, and 2 for ImageNet-100 and CUB, while 4 for CIFAR100. The temperature τ_c, τ_p are 0.07 and 0.1 respectively. The sharpened temperature in self-distillation in Eq. (4) is the same as SimGCD [19], *i.e.*, ramp-up schedule from 0.07 to 0.04. As for the hyper-parameters related to `Adaptive-Novel`, the EMA decay rate $\beta = 0.9$, and we set the threshold δ for justification stability of *label mapping function* to be 0.1 for all datasets. We first train models with SimGCD for 200 epochs as the base training stage to initialize models for subsequent AGCD. At each round of AGCD, we train models for 15 epochs. The default setting is to query 100 samples per round, and five rounds in total, CIFAR10 is an exception with only one round. During AGCD, we separately train all the queried data till the current round $\mathcal{D}_{q,map}^{all,t} = \mathcal{D}_{q,map}^1 \cup \mathcal{D}_{q,map}^2 \cup \cdots \cup \mathcal{D}_{q,map}^t$ and the original data $\mathcal{D}_l^0, \mathcal{D}_u^0$. We choose a smaller batch size 8 for $\mathcal{D}_{q,map}^{all,t}$ to acquire more update iterations and keep batch size 128 for the original dataset. The parameters above are applied to all query strategies for fair comparisons. All labeled data including $\mathcal{D}_l^0$ and $\mathcal{D}_{q,map}^{all,t}$ are trained with supervised objectives $\mathcal{L}_{con}^l$ Eq. (1) and $\mathcal{L}_{cls}^l$ in Eq. (5). We also employ unsupervised loss $\mathcal{L}_{con}^u$ in Eq. (2) and $\mathcal{L}_{cls}^u$ in Eq. (4) on both labeled and unlabeled data, which is consistent with SimGCD.

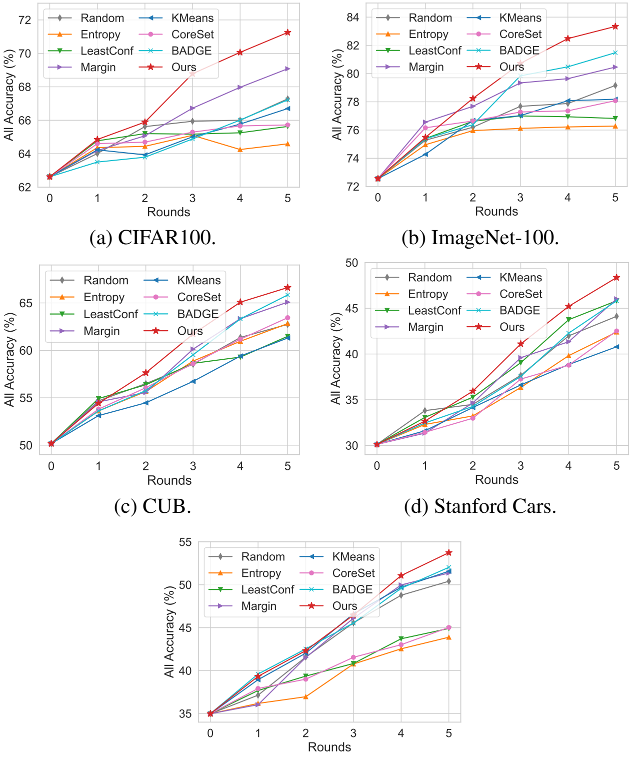
Algorithm 1 Adaptive-Novel Sampling Strategy for AGCD

```
Input: Initial labeled dataset $\mathcal{D}_l^0$ and unlabeled dataset $\mathcal{D}_u^0$.
Input: Total rounds $N$ and labeling budget per round $b$.
Input: Total class number $K = K_{old} + K_{new}$ (The ground truth or estimated).
Input: Stable mapping threshold $\delta$, initial transfer scalar $\mathcal{T} =$ False.
Input: Total epochs $E$ of each round.
Input: EMA decay parameter $\beta$.
Initialize $\mathcal{D}_l^0$ and $\mathcal{D}_u^0$ as in Table 1.
for current round $t = 1 \rightarrow N$ do
    for $c = 1 \rightarrow K_{new}$ do                                  # Class-wise sampling for Diversity
        if $\mathcal{T}$ then               # Adaptive Informativeness: informative sampling at later rounds
            ▷ Select $\lfloor b/K_{new} \rfloor$ samples with minimum margin from the $c$-th predictive novel class ($\hat{y} = K_{old} + c$)    # Novelty
              and query their labels to obtain $\mathcal{D}_{q,c}^t$
        else                                 # Adaptive Informativeness: confident sampling at early rounds
            ▷ Select $\lfloor b/K_{new} \rfloor$ samples with maximum margin from the $c$-th predictive novel class ($\hat{y} = K_{old} + c$)    # Novelty
              and query their labels to obtain $\mathcal{D}_{q,c}^t$
        end if
    end for
    ▷ All the queried data of the current round:
        $\mathcal{D}_q^t = \mathcal{D}_{q,1}^t \cup \mathcal{D}_{q,2}^t \cdots \cup \mathcal{D}_{q,K_{new}}^t$
    ▷ Update labeled and unlabeled datasets:
        $\mathcal{D}_l^t = \mathcal{D}_l^{t-1} \cup \mathcal{D}_q^t$
        $\mathcal{D}_u^t = \mathcal{D}_u^{t-1} \setminus \mathcal{D}_q^t$
    for current epoch $e = 1 \rightarrow E$ do                                   # Training at the current round
        ▷ Obtain label mapping function $\mathcal{M}^t$ between ground truth labels $y_i$ and the EMA model predictions $\hat{y}_i^{ema}$ in Eq. (11) on $\mathcal{D}_l^t$
        ▷ Perform label mapping on $\mathcal{D}_l^t$:
            $\mathcal{D}_{l,map}^t = \mathcal{D}_{l,map}^{t-1} \cup \mathcal{D}_{q,map}^t = \mathcal{M}^t(\mathcal{D}_l^{t-1}) \cup \mathcal{M}^t(\mathcal{D}_q^t)$
        ▷ Train the model on $\mathcal{D}_{l,map}^t$ with supervised loss $\mathcal{L}_{con}^l$ in Eq. (1) and $\mathcal{L}_{cls}^l$ in Eq. (5) and unsupervised loss,
          and on $\mathcal{D}_u^t$ with purely unsupervised loss
        ▷ Update the EMA model with decay rate $\beta$
    end for
    ▷ Compute the difference between $\mathcal{M}_{init}^t$ and $\mathcal{M}_{final}^t$ of this round as in Eq. (S1)
    if diff $< \delta$ then   # Mapping is stable and we transfer to informative sampling from round $t+1$
        $\mathcal{T} =$ True
    end if
end for
Output: The trained model and datasets $\mathcal{D}_l^N$, $\mathcal{D}_u^N$ after $N$ AGCD rounds.
```

D. Additional Quantitative Results

In this section, we provide additional quantitative results beyond the main text.

Details of AGCD performance across five rounds. In the main text, we mainly report the performance after all rounds of AGCD in Table 2 and Table 3. Here, we provide more detailed results performance over the course of different AGCD rounds, as shown in Fig. S1, where mean results over three runs are plotted.



(a) CIFAR100.

(b) ImageNet-100.

(c) CUB.

(d) Stanford Cars.

(e) Aircraft.

Figure S1. All accuracy of different strategies over 5 rounds.

Results of transfer rounds. As discussed in Sec. 4.3 and Appendix B, when *label mapping function* is stable at round t, *i.e.*, diff $< \delta$, we transfer from *confident novel sampling* to *informative novel sampling* from round $t+1$. Results of the transfer round $t+1$ are shown in Table S1, including five datasets CIFAR100 (C-100), ImageNet-100 (IN-100), CUB, Stanford Cars (SCars) and FGVC-Aircraft (Air). As for CIFAR10, the default setting has only one round, and we adopt *informative novel sampling* on it.

Table S1. Transfer round $t+1$ from *confident novel sampling* to *informative novel sampling* on various datasets.

Datasets	C-100	IN-100	CUB	SCars	Air
Round $t+1$	2	2	4	2	2

`Adaptive-Novel` with different uncertainty metrics. As introduced in Sec. 4.3, we choose margin as the metric for uncertainty/confidence for `Adaptive-Novel`, owing to the fact that margin is more robust in the task of AGCD and could select more sample from new classes when compared with MSP and entropy. From Fig. 4 we can observe that there are more samples from novel categories in the low confidence regime of Margin compared with Entropy and MSP. The results in Table 4 also show that we could select more new samples when using `Margin` as a metric for uncertainty-based methods. Here, we also conduct experiments using our `Adaptive-Novel` strategy but with different uncertainty metrics, as shown in Table S2. Results indicate the superiority of Margin benefiting from selecting more samples from novel categories.

Table S2. Results of `Adaptive-Novel` when applying different uncertainty metrics, including Entropy, MSP and Margin.

Metrics	CIFAR100			CUB		
	All	Old	New	All	Old	New
Ours w/ Entropy	67.30	**76.94**	57.66	63.75	**69.22**	58.34
Ours w/ MSP	67.63	75.18	60.08	63.70	67.37	60.07
Ours w/ Margin	**71.25**	75.72	**66.78**	**66.62**	66.54	**66.70**

Unknown class number scenarios. SimGCD [19] is a parametric classifier, it requires the number of classes K or K_{new} is known *a-prior* before training, here it could be the ground truth or estimated with off-the-shelf methods. In this paper, we adopt the off-the-shelf method `Max-ACC` [16] to estimate the number of the new classes. `Max-ACC` performs K-Means clustering on the entire dataset with various number of new classes, and choose the value as an estimation corresponding to the maximum clustering accuracy of labeled data. The estimation results on several datasets are shown in Table S3. Then we directly use the estimated number $\widehat{K}$ to set the prototypical classifier $\mathcal{C} = \{\mathbf{c}_1, \mathbf{c}_2, \cdots, \mathbf{c}_{\widehat{K}}\}$. In our main text, we have provided results on ImageNet-100 and CUB as in Table 8. Here we show results on more datasets in Table S4.

Table S3. Estimation of total class number $\widehat{K} = K_{old} + \widehat{K}_{new}$ using `Max-ACC` [16].

Datasets	CIFAR100	ImageNet-100	CUB	Stanford Cars
Ground Truth	100	100	200	196
Estimated	100	109	231	230

Results about novelty metrics. In the main text, results about novelty metrics on CIFAR100 and Stanford Cars are shown in Table 4. Here we provide the results on all six datasets as in Table S5 and Table S6. Our method consistently selects more samples evenly distributed across novel categories, leading to better AGCD performance.

Table S4. Results of AGCD with unknown class number (estimated class number) on various datasets.

Strategies	CIFAR100			ImageNet-100			CUB			Stanford Cars		
	All	Old	New	All	Old	New	All	Old	New	All	Old	New
Random	67.28	74.52	60.04	77.86	**90.34**	65.38	60.68	63.31	58.08	42.22	55.24	29.66
Entropy	64.59	73.94	55.24	76.12	71.52	60.72	60.48	64.84	56.15	41.77	56.36	27.71
Ours	**71.25**	**75.72**	**66.78**	**82.46**	89.84	**70.64**	**64.14**	**66.09**	**62.20**	**43.92**	**58.05**	**30.30**

Table S5. Novelty metrics of selected data on generic datasets.

AL Strategies	CIFAR10				CIFAR100				ImageNet-100			
	Nov-C	Nov-R	Nov-U	Nov-I	Nov-C	Nov-R	Nov-U	Nov-I	Nov-C	Nov-R	Nov-U	Nov-I
Random	**1.00**	0.84	0.96	0.81	**1.00**	0.52	0.97	0.50	**1.00**	0.56	0.97	0.54
Entropy	**1.00**	0.62	0.90	0.56	0.90	0.44	0.91	0.40	0.88	0.40	0.90	0.36
Margin	**1.00**	0.89	0.90	0.80	0.96	0.63	0.95	0.60	0.98	0.78	0.95	0.74
CoreSet	**1.00**	0.83	0.96	0.80	0.96	0.61	0.94	0.57	0.98	0.56	0.96	0.54
BADGE	0.88	0.77	0.87	0.67	**1.00**	0.63	**0.98**	0.62	**1.00**	0.71	0.97	0.69
Ours	**1.00**	**0.90**	**0.97**	**0.87**	**1.00**	**0.71**	**0.98**	**0.70**	**1.00**	**0.82**	**0.98**	**0.80**

Table S6. Novelty metrics of selected data on fine-grained datasets.

AL Strategies	CUB				Stanford Cars				FGVC-Aircraft			
	Nov-C	Nov-R	Nov-U	Nov-I	Nov-C	Nov-R	Nov-U	Nov-I	Nov-C	Nov-R	Nov-U	Nov-I
Random	0.95	0.63	0.96	0.60	0.93	0.57	0.96	0.55	0.96	0.53	0.98	0.52
Entropy	0.78	0.58	0.91	0.53	0.85	0.64	0.92	0.59	0.92	0.57	0.92	0.52
Margin	0.87	0.59	0.95	0.56	0.90	0.66	0.93	0.61	**1.00**	0.66	0.98	0.65
CoreSet	0.93	0.71	0.95	0.67	0.89	**0.69**	0.94	0.65	0.94	0.58	0.95	0.55
BADGE	0.98	0.57	0.97	0.55	0.95	0.64	**0.97**	0.62	1.00	0.59	0.98	0.58
Ours	**1.00**	**0.75**	**0.98**	**0.74**	**0.98**	**0.69**	**0.97**	**0.67**	**1.00**	**0.76**	**0.99**	**0.75**

Performance on the long-tailed dataset. In the real world, the class distributions often follow a long-tailed distribution, where the head classes have significantly more samples than the tail classes. Several methods [2, 20] have explored this issue in the task of GCD. In the long-tailed settings, the uniform constraint $H(\overline{\mathbf{p}})$ of the SimGCD [19] training procedure could be less applicable, as a result, we assign a small value to the weight λ_e to balance between addressing imbalanced class distribution and avoiding trivial solutions. In this paper, we also test our methods in this realistic scenarios. Specifically, we conduct experiments on the long-tailed Herbarium19 [14] dataset. Table S7 shows that our strategy outperforms other competitors, demonstrating the strong applicability of our method.

Table S7. Performance on the long-tailed Herbarium19 dataset.

Strategies	All	Old	New
w/o AGCD	46.55	62.67	29.49
Entropy	52.44	**65.36**	38.79
BADGE	53.31	61.53	44.62
Ours	**54.50**	63.33	**45.16**

Performance under other GCD training procedures. In the main manuscript, we compare different sample selection strategies with the GCD training procedure SimGCD [19]. Here we present results under two recent GCD training methods, *i.e.*, μGCD [17] and PIM [3]. As shown in Table S8, Adaptive-Novel works across various GCD training methods, indicating the superiority of our strategy consistency.

Table S8. Performance with another two GCD training procedures, μGCD and PIM.

Strategies	μGCD			PIM		
	All	Old	New	All	Old	New
w/o AGCD	48.69	57.00	40.45	48.84	55.34	43.64
Entropy	59.04	65.36	52.78	60.96	67.79	54.19
BADGE	61.08	62.66	59.52	62.32	**68.94**	55.77
Ours	**63.00**	**65.88**	**60.14**	**63.74**	65.50	**61.99**

Comparison with more recent and open-set AL strategies. We adapt recent methods to AGCD, including LfOSA [8], MQ-Net [9], ConAL [4], and one additional

Table S9. Comparison with open-set and standard AL methods.

Methods	Venue	CUB			SCars		
		All	Old	New	All	Old	New
LfOSA	CVPR'22	59.39	62.34	56.46	42.66	53.60	32.10
ConAL	T-PAMI'22	62.00	64.25	59.76	46.54	**58.89**	34.62
MQ-Net	NIPS'22	64.69	**66.75**	62.65	45.21	55.12	36.65
ALFA-Mix	CVPR'22	63.32	66.64	60.03	46.81	54.67	39.22
Ours	This Work	**66.62**	66.54	**66.70**	**48.36**	57.73	**39.34**

standard AL method ALFA-Mix [10], results are shown in Table S9. Because open-set AL merely cares about 'Old ACC', and treats new classes as noise/outliers, it aims to detect/filter them and mainly query old classes. Instead, our approach further clusters new classes. As a result, open-set AL generally selects fewer samples from new classes, resulting in even worse performance than standard AL baselines in AGCD.

E. Visualization of Feature Spaces

We visualize features of CIFAR10 using t-SNE [15] in Fig. S2. Original GCD (Left) suffers from severe confusion classes (*e.g.*, "deer" and "horses"), while Random and Entropy struggle to select informative samples, resulting in overlapped cluster boundaries. Instead, our approach achieves clear inter-class separation. We further visualize the feature space on CIFAR100 and CUB using t-SNE [15]. For CIFAR100, we randomly selected 10 classes (5 old classes and 5 new classes), while for CUB, 20 classes in total (10 old classes and 10 new classes) for visualization, results are shown in Fig. S3 and Fig. S4. As in Fig. S3, there are many confusing classes on which the model behaves ambiguously before AGCD, *e.g.*, "beaver", "squirrel" and "bear". When the model is trained on the newly labeled data queried by our method, it could achieve relatively more separated clusters among the classes. Additionally, it is observable in Fig. S3 and Fig. S4 that our method could generally bring about more compact class-wise clusters on both datasets.

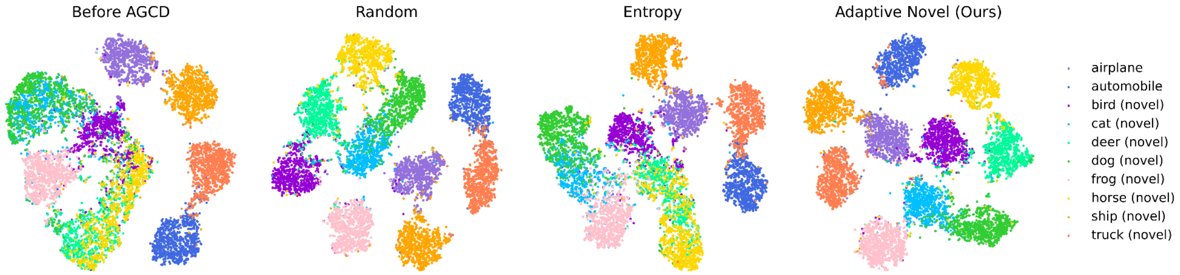


Figure S2. t-SNE [15] feature visualization of different strategies on CIFAR10.

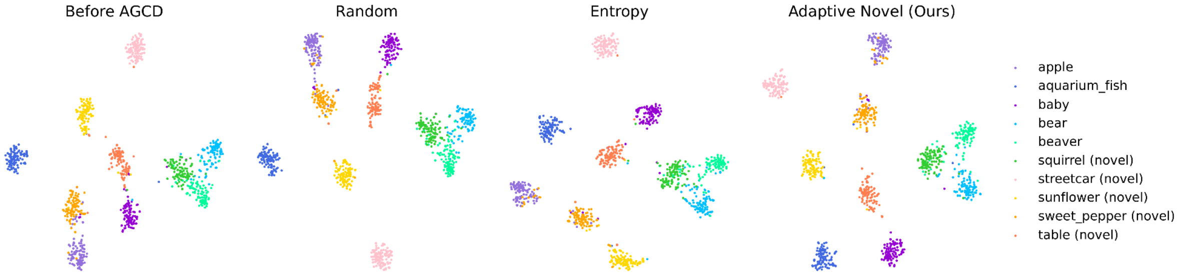


Figure S3. t-SNE [15] feature visualization of different strategies on CIFAR100.

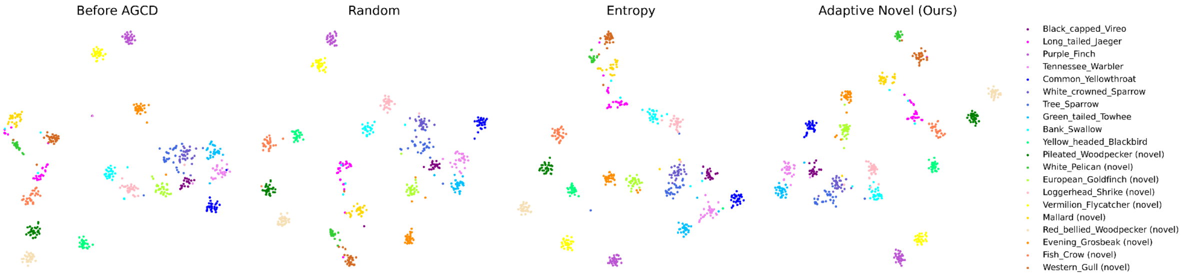


Figure S4. t-SNE [15] feature visualization of different strategies on CUB.

References

[1] 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. 2

[2] Jianhong Bai, Zuozhu Liu, Hualiang Wang, Ruizhe Chen, Lianrui Mu, Xiaomeng Li, Joey Tianyi Zhou, YANG FENG, Jian Wu, and Haoji Hu. Towards distribution-agnostic generalized category discovery. In *Advances in Neural Information Processing Systems*, pages 58625–58647, 2023. 5

[3] 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. 5

[4] Pan Du, Hui Chen, Suyun Zhao, Shuwen Chai, Hong Chen, and Cuiping Li. Contrastive active learning under class distribution mismatch. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4260–4273, 2022. 5

[5] 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. 2

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

[7] 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. 1, 2

[8] Kun-Peng Ning, Xun Zhao, Yu Li, and Sheng-Jun Huang. Active learning for open-set annotation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 41–49, 2022. 5

[9] Dongmin Park, Yooju Shin, Jihwan Bang, Youngjun Lee, Hwanjun Song, and Jae-Gil Lee. Meta-query-net: Resolving purity-informativeness dilemma in open-set active learning. *Advances in Neural Information Processing Systems*, 35:31416–31429, 2022. 5

[10] Amin Parvaneh, Ehsan Abbasnejad, Damien Teney, Gholamreza Reza Haffari, Anton Van Den Hengel, and Javen Qinfeng Shi. Active learning by feature mixing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12237–12246, 2022. 6

[11] 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. 1

[12] 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. 2

[13] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*, 2018. 2

[14] Kiat Chuan Tan, Yulong Liu, Barbara Ambrose, Melissa Tulig, and Serge Belongie. The herbarium challenge 2019 dataset. *arXiv preprint arXiv:1906.05372*, 2019. 5

[15] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (11), 2008. 6, 7

[16] 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. 1, 4

[17] 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. 5

[18] 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. 2

[19] 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. 1, 2, 4, 5

[20] Chuyu Zhang, Ruijie Xu, and Xuming He. Novel class discovery for long-tailed recognition. *Transactions on Machine Learning Research*, 2023. 5