Contrastive Open-Set Active Learning-Based Sample Selection for Image Classification

Zizheng Yan, *Member, IEEE*, Delian Ruan, Yushuang Wu, *Graduate Student Member, IEEE*, Junshi Huang, Zhenhua Chai, Xiaoguang Han, *Member, IEEE*, Shuguang Cui, *Fellow, IEEE*, and Guanbin Li, *Member, IEEE*

***Abstract*— In this paper, we address a complex but practical scenario in Active Learning (AL) known as open-set AL, where the unlabeled data consists of both in-distribution (ID) and out-of-distribution (OOD) samples. Standard AL methods will fail in this scenario as OOD samples are highly likely to be regarded as uncertain samples, leading to their selection and wasting of the budget. Existing methods focus on selecting the highly likely ID samples, which tend to be easy and less informative. To this end, we introduce two criteria, namely contrastive confidence and historical divergence, which measure the possibility of being ID and the hardness of a sample, respectively. By balancing the two proposed criteria, highly informative ID samples can be selected as much as possible. Furthermore, unlike previous methods that require additional neural networks to detect the OOD samples, we propose a contrastive clustering framework that endows the classifier with the ability to identify the OOD samples and further enhances the network's representation learning. The experimental results demonstrate that the proposed method achieves state-of-the-art performance on several benchmark datasets.**

***Index Terms*— Image recognition, active learning, contrastive learning.**

I. Introduction

DEEP Neural Networks (DNNs) have emerged as a promising solution for a wide range of applications, including image recognition [1], recommendation systems [2], and biomedical imaging [3]. However, the data-hungry nature of DNNs and the high cost of labeling large volumes of data have presented significant challenges. Meanwhile, Active Learning (AL) [4], [5] has gained popularity as an approach to resolve the problem of expensive annotation costs by utilizing machine learning models to identify informative subsets of data for annotation and subsequently train models on the selected subset. As a result, numerous works have been proposed to tackle the AL problem in the field of deep learning [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18].

Although effective, standard Active Learning (AL) methods operate under the implicit assumption that unlabeled data solely consists of in-distribution (ID) samples. This assumption is unrealistic since many real-world scenarios are open-set and often contain numerous out-of-distribution (OOD) samples. Specifically, the label sets of ID and OOD samples differ, *e.g.*, they belong to different classes, and the OOD classes are irrelevant to the task. In practice, standard closed-set AL algorithms fail in the open-set AL context since they primarily

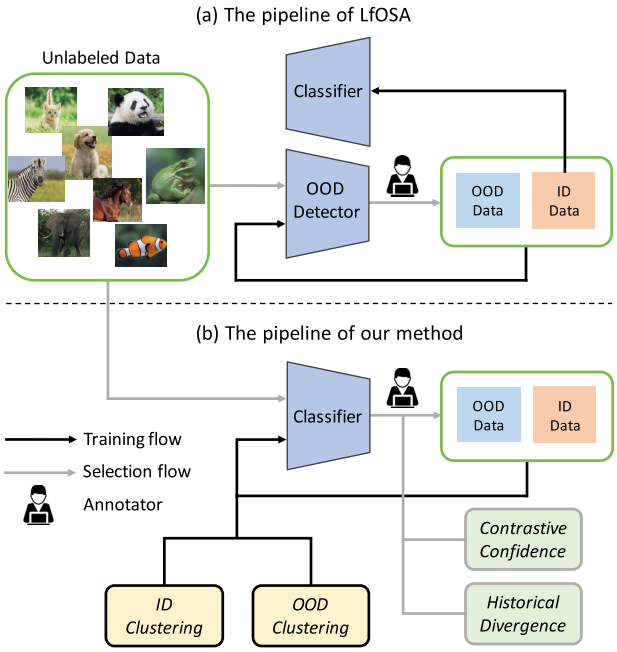


Fig. 1. The comparison between our proposed method and the previous method LfOSA [19]. LfOSA trains an additional OOD detector to select the samples most likely to be ID for annotation and trains the task classifier solely on the labeled ID samples. In contrast, our method leverages both ID and OOD data for training the task classifier and has the capability to select highly informative ID samples without the need for additional OOD detectors.

Received 8 October 2023; revised 6 April 2024 and 20 July 2024; accepted 28 July 2024. Date of publication 5 September 2024; date of current version 4 October 2024. This work was supported in part by NSFC under Grant 62293482; in part by the Basic Research Project of Hetao Shenzhen-HK Science and Technology Cooperation Zone under Grant HZQB-KCZYZ-2021067; in part by the National Natural Science Foundation of China under Grant 62322608; in part by the Fundamental Research Funds for the Central Universities under Grant 22lgqb25; in part by Shenzhen Science and Technology Program under Grant JCYJ20220530141211024; in part by the Open Project Program of the Key Laboratory of Artificial Intelligence for Perception and Understanding, Liaoning Province (AIPU), under Grant 20230003; in part by the National Key Research and Development Program of China under Grant 2018YFB1800800, and in part by Shenzhen Outstanding Talents Training Fund under Grant 202002. The associate editor coordinating the review of this article and approving it for publication was Dr. Laura Boucheron. *(Corresponding author: Guanbin Li.)*

Zizheng Yan, Yushuang Wu, Xiaoguang Han, and Shuguang Cui are with the Shenzhen Future Network of Intelligence Institute, the School of Science and Engineering, and the Guangdong Provincial Key Laboratory of Future Networks of Intelligence, Chinese University of Hong Kong at Shenzhen, Shenzhen 518172, China.

Delian Ruan, Junshi Huang, and Zhenhua Chai are with Meituan, Beijing 100102, China.

Guanbin Li is with the School of Computer Science and Engineering, Research Institute of Sun Yat-sen University in Shenzhen, Sun Yat-sen University, Guangzhou 510008, China (e-mail: liguanbin@mail.sysu.edu.cn).

Digital Object Identifier 10.1109/TIP.2024.3451928

focus on selecting highly uncertain samples for labeling. This results in the selection of numerous OOD samples as the task neural networks [20] are usually not confident when dealing with such samples. Therefore, the most straightforward solution for open-set AL is to select the samples that are most likely to be ID for labeling, thereby creating a clean query set.

The pioneering work LfOSA [19] proposes to use an additional neural network to detect the OOD samples and avoid selecting them. Specifically, given ID samples of K classes and a few OOD samples, a $(K+1)$-way classifier is trained along with the task classifier. After each active learning round, the samples with high maximum ID activation values are selected for labeling. The maximum ID activation value is defined as the maximum of the K-dimensional vector for ID classes of the final feature.

Although LfOSA [19] can effectively filter out OOD samples, we argue that it has the following shortcomings. Firstly, it is not optimal to only focus on the "cleanness" of the query set. LfOSA only selects the ID samples that are most differentiable from OOD samples. Thus, the selected samples tend to be "easy" and less informative. Secondly, diversity is not considered, as all the samples are sorted together according to the maximum activation value. This might bring the class imbalance issue as some classes may be highly differentiable while some may be ambiguous to the OOD classes. Finally, an additional classifier needs to be trained since the classification performance of the OOD detector is low, because all the OOD samples are forcibly classified as a single class, while classifying the samples belonging to multiple categories into a single class will hurt the representation learning of the network.

In this paper, we address the aforementioned shortcomings by a selection method that balances the cleanliness and hardness of the query set using two proposed criteria: *contrastive confidence* and *historical divergence*. Specifically, the selection method has three components: i) measuring the possibility that a sample belongs to ID by contrastive confidence, ii) measuring the hardness of a sample by historical divergence, and iii) a class-wise selection module that ensures diversity of the query set. The contrastive confidence and historical divergence are balanced by a round-adaptive parameter. Furthermore, we propose a contrastive clustering method that fully leverages both the ID and OOD samples, endowing the classifier with the ability to identify OOD samples without training an OOD detector and enhancing the representation learning of the network. The contrastive clustering consists of two components: (i) An ID clustering method that contrastively clusters the ID samples in a supervised manner and pushes away ID samples from OOD samples in the feature space. (ii) An OOD clustering method that clusters OOD samples into compact clusters to benefit the network's representation learning, which is achieved by contrasting the prototypical predictions of two augmented views in a mini-batch. A comparison between our method and LfOSA [19] is shown in Figure 1.

In summary, our work has the following contributions:

- We propose a sample selection method that selects as many highly informative ID samples as possible by balancing the query set's cleanness and hardness.
- We propose an ID clustering method that enhances the representation learning of the network and pushes ID samples away from OOD samples in feature space so that ID samples can be more discriminative from OOD samples.
- In addition to the ID clustering, we propose an OOD clustering method that fully leverages the selected OOD samples and further improves the network's performance.
- We perform extensive experiments on several benchmark datasets to verify the effectiveness of our proposed method, and the results show that our method outperforms previous state-of-the-art methods by clear margins.

II. Related Work

A. Active Learning

The existing active learning methods usually focus on designing various sample selection strategies that can broadly be grouped into uncertainty-based [6], [11], [21], [22], [23], [24], [25], [26], [27], [28], [29] and diversity-based methods [7], [8], [10], [12], [14], [18], [30], [31], [32], [33], [34]. Uncertainty-based methods mainly aim to select samples that are ambiguous about the predictions. Standard approaches use the posterior probability to measure the uncertainty, *e.g.*, least softmax confidence [21], margin [22], and entropy [6], [23]. In addition, many other approaches have been proposed to estimate the uncertainty, especially for DNNs, *e.g.*, Yoo and Kweon [11] estimates the training loss of unlabeled samples and select the samples with large losses, Ducoffe and Precioso [24] proposes to using the adversarial robustness as the uncertainty criterion, and Liu et al. [25] proposes a measure called influence function to assist sample selection. Diversity-based methods mainly aim to select samples that are representative in the feature space so that the distribution of this selected subset can become close to the original unlabeled samples. CoreSet [10] proposes to formulate the problem to a k-center problem and greedily select the samples with the greatest distance to their nearest neighbors. Furthermore, Agarwal et al. [13] replace the Euclidean distance with context-aware KL-divergence. VAAL [12] uses a variational autoencoder [35] to approximate the distribution of labeled data and a discriminator to discriminate the labeled and unlabeled data. Moreover, many approaches [7], [8], [18], [30], [30], [36], [37] have been proposed to use both uncertainty and diversity as the selection criterion to select more diverse and informative samples. BADGE [7] proposes to cluster the gradient embeddings regarding to the pseudo labels to ensure both diversity and uncertainty. ALFA-Mix [8] identifies unlabelled instances with sufficiently distinct features by seeking inconsistencies in predictions resulting from their representation mixup [38].

The above methods are no longer feasible in the case of the open-set setting as OOD samples are highly likely to be regarded as uncertain samples, and they are easy to show great diversity in the feature space. As a result, some OOD samples will be selected as hard samples, and the potentially really-hard samples will be left out, thereby wasting the budget.

In this paper, we propose a sample selection method to select as many hard ID samples as possible.

B. Open-Set Recognition

Open-set recognition (OSR) has emerged as a critical problem in the field of machine learning that addresses the challenge of learning with both known and unknown categories. The foundational framework for OSR, proposed by Scheirer et al. [39], posits that the training set comprises known classes, whereas the testing set includes both known and unknown classes. The objective is to classify the known classes while rejecting the unknown ones during testing. A lot of approaches have been introduced to address the OSR problem [40], [41], [42], [43], [44], [45], [46], [47], [48]. Building upon this framework, various related settings have been proposed, including open-set semi-supervised learning [49], [50], [51], [52], [53], few-shot OSR [54], [55], [56], [57], [58], [59], and novel category discovery [60], [61], [62], [63], [64]. Open-set semi-supervised learning presumes the presence of unknown classes within the unlabeled data, aiming to classify known classes. Few-shot OSR extends few-shot learning by requiring the model to not only classify novel classes but also reject unknown class samples during testing. Novel category discovery delves deeper into handling unknown classes by assuming that the training set contains both labeled known classes and unlabeled unknown classes, with the goal of clustering the unknown classes into meaningful semantic cluters.

However, the above paradigms focus on learning in the presence of unknown classes, whereas open-set active learning centers on the strategies of sample selection. Given an unlabeled dataset containing both known and unknown classes, the goal is to select the most informative samples for labeling in order to enhance the classification of known classes. This diverges from the typical settings outlined above and presents unique challenges within the domain of active learning.

C. Open-Set Active Learning

There are few works that address the problem of open-set active learning [19], [65], [66], [67]. SIMILAR [66] employs submodular functions to select the samples farthest from the identified OOD samples while nearest to all unlabeled samples. However, such an operation tends to select easy ID samples. CCAL [65] pre-trains two OOD detector networks, each for measuring the informativeness and the possibility of being ID of a given sample. However, the detector networks are pre-trained in an unsupervised manner, then are frozen during active learning rounds, which do not interact with the task classifier during training, thereby the estimation of the informativeness and IDness is biased. Unlike CCAL [65], LfOSA [19] does not discard the selected OOD samples. In contrast, LfOSA [19] uses them along with the ID samples to train a $(K+1)$-way OOD detector network on-the-fly during active learning rounds. However, only the possibility of belonging to ID samples is considered for the selection, which tends to select low informative samples. MQNet [67] leverages a self-supervised model and a meta-net to balance the purity and informativeness for selection. However, the training of the self-supervised model and the meta-net brings additional computational costs.

Compared with the above methods, our method jointly models the hardness and the possibility of being ID of the unlabeled samples using the task classifier, which can select highly informative ID samples. Furthermore, our method does not require additional OOD detector networks as the proposed contrastive clustering endows the classifier with the ability to identify OOD samples.

D. Generalized Category Discovery

Generalized Category Discovery (GCD) [62], [68], [69], [70] addresses the challenge of clustering a pool of unlabeled data that includes both ID and OOD samples. The goal is to effectively cluster all unlabeled samples, encompassing both ID and OOD categories. The pioneering work by Vaze et al. [62] employs semi-supervised K-Means to tackle this task. Additionally, several subsequent studies have explored the use of parametric classifiers [69], [70]. Recently, an extended setting of GCD, termed Active Generalized Category Discovery (AGCD), has been proposed to address the GCD task within the constraints of affordable labeling budgets. AGCD introduces an adaptive sampling strategy that jointly considers novelty, informativeness, and diversity to adaptively select novel samples with appropriate uncertainty.

In contrast to AGCD, our approach focuses solely on classifying ID samples during inference, which eliminates the necessity of considering the novelty of samples during sampling. Furthermore, results from AGCD indicate that while accounting for novelty can improve the clustering performance of OOD samples, it can also negatively impact the performance on ID samples.

III. Method

In this section, we first present the preliminaries including the definition of open-set active learning and associated notations and then detail the proposed framework.

A. Preliminaries

In the open-set active learning problem, a small-size initial labeled set $\mathcal{D}_L = \{(\mathbf{x}_i^l, \mathbf{y}_i^l)\}_{i=1}^{n_l}, \mathbf{y}_i^l \in \mathcal{C}_{ID}$ and a large-size unlabeled set $\mathcal{D}_U = \{(\mathbf{x}_i^u)\}_{i=1}^{n_u}$ is given, where $\mathcal{C}_{ID}$ is the label set of $\mathcal{D}_L$, consisting of the in-distribution (ID) target classes that we aim to classify. In contrast, the label set of $\mathcal{D}_U$ consists of not only the ID classes but also the out-of-distribution (OOD) classes, *e.g.*, $\mathcal{C}_U = \mathcal{C}_{ID} \cup \mathcal{C}_{OOD}$. Active Learning (AL) aims to employ the model to select a set of informative unlabeled samples X^{query} for querying labels from the annotators, and this process iterates over R rounds. In each round r, the labeled set $\mathcal{D}_L^r$ is used to train the model. In the open-set scenario, the annotator will label the ID samples with their ground truth and the OOD samples as a single unknown class.

The existing work LfOSA [19] addresses the problem by selecting the unlabeled samples that are most likely to belong to the ID classes. An OOD detector is trained to classify the labeled ID and OOD samples to $K+1$ classes and

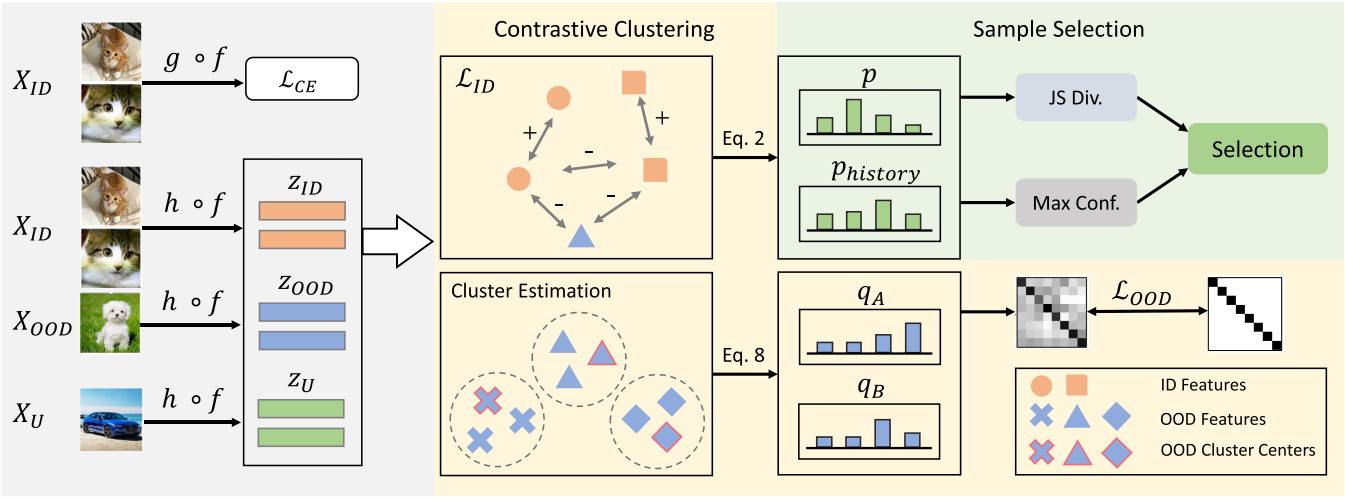


Fig. 2. Illustration of our proposed framework. f is the backbone, g is the classification head, and h is the projection head for contrastive learning. Given the ID features z_{ID} and the unlabeled features z_U, the prototypical prediction $\mathbf{p}$ and the historical prediction of z_U are first calculated. Then, two criteria, contrastive confidence (Eq. 2) and historical divergence (Eq. 4), are calculated and then balanced to obtain the final score (Eq. 5) for sample selection. The contrastive clustering consists of two parts: i) $\mathcal{L}_{ID}$ (Eq. 6) that pulls the positive pairs (*e.g.*, same class samples) and pushes the negative pairs (*e.g.*, different class samples and OOD samples) away from each other, and ii) $\mathcal{L}_{OOD}$ (Eq. 8) that first estimates the OOD clusters, then contrasts the batch prototypical predictions of two augmented views, $\mathbf{q}_A$ and $\mathbf{q}_B$, enforcing the correlation matrix close to the identity matrix. Note that the unlabeled data are only used for selection, not training.

select the samples with high maximum ID activation values. Although LfOSA can select most ID samples, there are several drawbacks. Firstly, the selected samples are too easy and less informative, thereby contributing little to the model training. Secondly, the method does not consider diversity, which may lead to class imbalance. Finally, the classifier does not fully leverage the OOD samples, as they are only used to train the OOD detector.

Accordingly, we propose a contrastive framework for the open-set active learning problem to address the above issues. Our framework has two components: i) a sample selection method that leverages contrastive confidence and the historical divergence to select hard and diverse ID samples; ii) a contrastive clustering method that fully leverages both ID and OOD samples, pushing ID features away from OOD ones, and helps the network to learn compact and highly discriminative features for both ID and OOD samples that benefit the sample selection in return. The overview of our method is shown in Figure 2.

B. Sample Selection

Given the small-sized initial labeled set $\mathcal{D}_L$, the trained neural network may easily mis-classify OOD samples as ID samples due to poorly learned representations. To address this issue, we propose first training a self-supervised network, *i.e.*, SimCLR [71], using both the labeled data $\mathcal{D}_L$ and the unlabeled data $\mathcal{D}_U$. With the contrastive features of the pre-trained network, samples similar to those in $\mathcal{D}_L$ are more likely to be ID samples. However, this criterion tends to select easy ID samples, as hard samples are usually not well-clustered in the embedding space. To balance the trade-off between "hardness" and the quantity of ID samples, we propose leveraging the contrastive confidence and the historical divergence of the unlabeled sample. We also propose selecting samples in a class-wise manner to ensure diversity. Our selection method is detailed as follows:

1) Contrastive Confidence: This step aims to differentiate ID and OOD samples in $\mathcal{D}_U$ with *contrastive confidence*. Formally, let $f : \mathcal{X} \rightarrow \mathbb{R}^d$ denote the backbone network. After each active learning round r, given the labeled set $\mathcal{D}_L^r$, we first compute the prototypes $\{v_c\}_{c=1}^{\mathcal{C}_{ID}}$ of ID classes:

$$v_c = \frac{1}{N_c}\sum_{i}^{N_c} f(\mathbf{x}_i)/||\frac{1}{N_c}\sum_{i}^{N_c} f(\mathbf{x}_i)||_2, \tag{1}$$

where N_c denotes the number of samples of class c in $\mathcal{D}_L^r$. We normalize the prototypes as the original contrastive features $f(\mathbf{x}_i)$ are $l2$ normalized [71]. With the computed prototypes, we define the prototypical prediction $\mathbf{p}_i \in \mathbb{R}^{|\mathcal{C}_{ID}|}$ and the *contrastive confidence* p_i as follows:

$$\mathbf{p}_i^j = \frac{\exp(v_j \cdot f(\mathbf{x}_i)/\tau)}{\sum_{c=1}^{|\mathcal{C}_{\mathcal{ID}}|}\exp(v_c \cdot f(\mathbf{x}_i)/\tau)},$$
$$p_i = \max \mathbf{p}_i, \tag{2}$$

where $\mathbf{x}_i \in \mathcal{D}_U^r$, τ is a temperature parameter. It can be observed that the contrastive confidence p_i is higher when the sample is closer to the ID prototypes, implying that the sample is more likely to be an ID sample.

2) Historical Divergence: After obtaining the prototypical predictions of the current active learning round r, we compute the divergence between them and the historical prototypical predictions of round $r-1$. As storing the historical predictions of all previous rounds and computing the divergence is inefficient, we propose to update the historical prediction with the exponential moving average. Formally, the historical prediction of the ith sample at round r is defined as follows:

$$\bar{\mathbf{p}}_{i,r} = \frac{1}{2}\left(\bar{\mathbf{p}}_{i,r-1} + \mathbf{p}_{i,r}\right). \tag{3}$$

Then we define the *historical divergence* as the Jensen-Shannon (JS) divergence between the prototypical prediction and the historical prediction:

$$d_{i,r} = JSD(\bar{\mathbf{p}}_{i,r-1}, \mathbf{p}_{i,r}). \tag{4}$$

It can be observed that the historical divergence of a sample measures how the network prediction changes over different active learning rounds. In other words, the higher $d_{i,r}$, the "harder" the sample as the prediction is more unstable.

3) Class-Wise Selection: Given the contrastive confidence $p_{i,r}$ and historical divergence $d_{i,r}$ of the ith sample at round r, we derive its final score as follows:

$$s_{i,r} = \frac{r}{\alpha R} d_{i,r} + (1 - \frac{r}{\alpha R}) p_{i,r}, \tag{5}$$

where $r/(R \cdot \alpha) \in (0, 1)$ is a weighting parameter to balance the ID possibility and "hardness", R is the total number of rounds and $\alpha \geq 1$ is a hyper-parameter. It can be observed that $r/(R \cdot \alpha) \in (0, 1)$ is small in early rounds and large in later rounds. The motivation is that the network lacks ID samples in early rounds, so selecting more ID-like samples will benefit the training. In contrast, the network is more well-trained in later rounds and can more accurately select "hard" ID samples.

Subsequently, the samples are sorted according to their final score s in descending order, and the first b samples are chosen, where b represents the budget size. In order to ensure diversity, the selection is performed in a class-wise manner: for each class, the samples that are predicted to that class are sorted, and $\lceil b/|\mathcal{C}_{ID}| \rceil$ samples are chosen. If there are insufficient samples in some classes, the remaining samples are randomly selected to fulfill the budget.

C. Contrastive Clustering

Given the labeled set, the ground truth of ID samples is fully available, and all the OOD samples have the same "unknown" label. To fully exploit the label information of ID samples, we propose two different clustering losses for ID and OOD samples, respectively.

1) Contrastive Clustering for ID Samples: Similar to sample selection, we exploit the features extracted by the projection head h. For the ID samples, clustering is performed with the guidance of the ground truth. Conversely, instead of simply discarding OOD samples, we propose training them using a self-supervised contrastive loss and pushing them away from ID samples. The ID and OOD samples are treated as negative pairs of each other.

Given a training batch with size N, we perform two random augmentations to create two views of the batch and stack them together. Let $j(i)$ denote the index of the augmented counterpart of the ith sample in the batch. Then, the loss for the ith sample is defined as follows:

$$\mathcal{L}_{ID} = -\frac{1}{|T(i)|} \sum_{t \in T(i)} \log \frac{\exp(\mathbf{z}_i \cdot \mathbf{z}_t / \tau)}{\sum_{a \in A(i)} \exp(\mathbf{z}_i \cdot \mathbf{z}_a / \tau)}, \tag{6}$$

where $\mathbf{z}_* = f(\mathbf{x}_*)$, $A(i) = \{1, \ldots, 2N\} \setminus i$ is the set of indices other than i, $T(i) = T(i)_{ID} \cup T(i)_{OOD}$, $T(i)_{ID} = \{t \in A(i) : \mathbf{y}_t = \mathbf{y}_i\}$ and $T(i)_{OOD} = \{j(i)\}$ are the sets of indices of positives in the multi-view batch for ID and OOD samples, respectively. The above loss aims to pull together the ID samples belonging to the same class, push apart the ID samples from different classes, and push ID samples away from OOD samples.

2) Contrastive Clustering for OOD Sample: Although $\mathcal{L}_{ID}$ helps the network to learn good representations, the OOD samples are not fully leveraged as they are only trained to pull together with the augmented counterpart. To this end, we propose to cluster the OOD samples into compact and discriminative clusters, which will benefit the representation learning of the network. The clustering process has two steps: i) estimate the number of clusters and the initial cluster centers, and ii) cluster the samples dynamically. The first step is performed after each active learning round, and then the cluster labels are exploited in the next round. Our clustering method is detailed as follows:

Step 1. Cluster estimation. A straightforward approach to estimate the number of clusters is to cluster all the selected data $\mathcal{D}_L^r$ into K clusters and evaluate the cluster accuracy solely on the ID samples, then find the optimal K that have the highest ID cluster accuracy. However, the brute-force estimation of K is challenging as the OOD data may contain many classes. Therefore, we propose dividing the OOD samples into subsets based on their predicted labels and then performing the aforementioned estimation for each subset. We use k-means to perform clustering and find the optimal k using Brent's algorithm.[1] Specifically, we treat k as the variable and the ID clustering accuracy as the function value. Finally, we merge the clustering results of each subset and assign each OOD sample a cluster label.

Step 2. Contrastive clustering. Since the network is re-initialized at the beginning of each round, we re-compute the OOD cluster centroids $\{o_c\}_{c=1}^M$ by the cluster labels. Note that the cluster centroids are also $l2$ normalized. Then, for each OOD sample $\mathbf{x}_i$, we calculate its prototypical prediction $\mathbf{q}_i \in \mathbb{R}^M$ against the cluster centroids,

$$\mathbf{q}_i^j = \frac{\exp(o_j \cdot f(\mathbf{x}_i)/\tau)}{\sum_{c=1}^M \exp(o_c \cdot f(\mathbf{x}_i)/\tau)}. \tag{7}$$

It is worth noting that when a sample is very close to a centroid, its corresponding vector $\mathbf{q}_i$ will be sharp, *e.g.*, it will be a one-hot vector. If the OOD samples are well-clustered, each one will have a sharp $\mathbf{q}_*$, and the predictions will be diverse. This observation motivates us to develop the clustering loss. Additionally, we leverage consistency regularization between the multi-view augmented samples, represented as $\mathbf{q}_A$ and $\mathbf{q}_B$. We obtain a batch of prototypical predictions, denoted as $\mathbf{Q}_A$ and $\mathbf{Q}_B$, for views A and B, respectively. Here, $\mathbf{q} \in \mathbb{R}^M$ and $\mathbf{Q} \in \mathbb{R}^{N \times M}$, where N is the batch size, and M is the number of OOD cluster centroids. Finally, we compute the correlation matrix $\mathbf{R}$: $\mathbf{R} = \mathbf{Q}_A^T \mathbf{Q}_B$, where $\mathbf{R} \in \mathbb{R}^{M \times M}$. Then the clustering loss is defined as follows,

$$\mathcal{L}_{OOD} = -\frac{1}{K} \mathrm{Tr}\big(\phi(\mathbf{R} + \mathbf{R}^\top)\big), \tag{8}$$

where $\mathrm{Tr}(\cdot)$ is the trace of a matrix, $\phi(\cdot)$ is a row normalization operation where each element is divided by the row sum. As $\mathbf{R}$

[1] https://en.wikipedia.org/wiki/Brent%27s_method

is asymmetric and the row summations differ, we convert it to be symmetric before row normalization. Minimizing the above loss function not only maximizes the diagonal values but also reduces the off-diagonal values of $\mathbf{R}$. It can be shown that the optimal solutions fulfill the following conditions: i) $\mathbf{Q}_A = \mathbf{Q}_B$, ii) $\mathbf{Q}_A$ and $\mathbf{Q}_B$ have sharp (one-hot-like) and diverse rows (each row is a $\mathbf{q}_*$). This indicates that OOD samples are clustered into compact clusters. Finally, the OOD cluster centroids are updated with gradients. More detailed analysis of this loss can be found in Secs. IV-D.9–IV-D.11.

D. Overall Pipeline

In addition to the contrastive clustering losses, a cross-entropy loss $\mathcal{L}_{CE}$ is applied to the classification head g with the labeled ID samples. Therefore, the overall loss can be formulated as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{CE} + \mathcal{L}_{ID} + \mathcal{L}_{OOD}. \tag{9}$$

The overall active learning pipeline iterates the following steps in each round: (i) training with the above loss functions, and (ii) sample selection and cluster estimation.

IV. Experiments

A. Datasets

Following [19], we evaluate our method on three benchmark datasets, including CIFAR-10 [72], CIFAR-100 [72], and Tiny-Imagenet [73]. CIFAR-10 consists of 50,000 training images and 10,000 test images from 10 classes, where the image size is 32×32. CIFAR-100 has the same size as CIFAR-10 while containing 100 classes. TinyImagenet is a subset of the Imagenet [74] dataset consisting of 200 natural image classes, each containing 500 training images and 50 test images, where the image size is 64×64. For ease of implementation, we resize them to 32×32. In addition, we conduct the experiments on the settings with the ID classes ratio $\xi = 20\%, 30\%$ and 40%, where ξ is defined as $\xi = |\mathcal{C}_{\mathcal{ID}}|/|\mathcal{C}_{\mathcal{ID}} \cup \mathcal{C}_{\mathcal{OOD}}|$. Specifically, for each dataset, the samples from the first $|\mathcal{C}_{\mathcal{ID}}|$ classes are defined as the ID samples while the rest are defined as the OOD samples. Similar to [19], we perform evaluations on the test set samples from the ID classes and report the accuracy.

B. Implementation Details

Similar to [19], we use Resnet-18 as the backbone network. Different from [19] that sets the backbone feature dimension as 2, we set it to the original dimension, *i.e.*, 512. For the self-supervised pre-training, we employ SimCLR [71] as the training framework. Specifically, a 2-layer MLP is utilized as the projection head h, which first projects the backbone feature to 512 dimensions and then 128 dimensions. We train the network for 10 rounds with an initial budget size of 4% of the whole dataset and a query budget size of 500. For each round, we fine-tune from the self-supervised pre-trained network using SGD with a learning rate 1e-3 and batch size 128. We report the mean and standard error of the results over 4 runs for each experiment. The temperature τ of the prototypical prediction and the contrastive clustering is set to 0.1 for all experiments. In addition, the α for the trade-off between contrastive confidence and historical divergence is set to 8.

C. Comparison Experiments

We compare our method to the following methods: (1) Random sampling. (2) Least confidence - lowest max softmax output. (3) Uncertainty - highest entropy of softmax outputs. (4) MSP - the max softmax posterior probability. (5) BADGE [7] - employing both uncertainty and diversity of samples. (6) CCAL [65] - using two pre-trained OOD detectors to select more ID samples. (7) LfOSA [19] - using a $K+1$ way classifier to differentiate ID samples from OOD ones. (8) MQNet [67] - using a mete-network to assist the selection. All of the above methods are trained on the same protocol, *e.g.*, starting from a SimCLR pre-trained model at each round.

Results: The comparison results are presented in Figure 3. The results indicate that our proposed method outperforms all previous methods across all datasets with varying ID ratios, providing evidence of its efficacy. Remarkably, the model attains satisfactory performance at early training stages, even at the first active round, implying that our proposed contrastive clustering approach can potentially reduce the number of active rounds required. Additionally, it is notable that our method exhibits a smaller variance than most other methods, indicating greater stability.

D. Ablation Studies

We perform several ablation studies to validate our proposed method.

1) Sample Selection: To verify the efficacy of each component of the proposed sample selection method, we perform an ablation study on CIFAR100 with ID ratio $\xi = 40\%$:

a) Contrastive confidence: Contrastive confidence (Eq. 2) aims to measure the possibility that a sample belongs to ID. In contrast, the absence of contrastive confidence leads to suboptimal performance, as evidenced in Figure 4a. We conjecture that this is due to the selection criterion being based solely on the historical divergence, resulting in the selection of more uncertain samples, which are more likely to be OOD, as illustrated in Figure 4b. Therefore, contrastive confidence plays a vital role in identifying ID samples. Additionally, it is worth highlighting that, even in the absence of contrastive confidence, our method can select more ID samples than random selection, indicating that the historical divergence can implicitly identify the ID samples.

b) Historical divergence: The goal of historical divergence (Eq. 4) is to measure the uncertainty of a sample. The higher the historical divergence, the higher the uncertainty and informativeness. The result of not using historical divergence is shown in Figure 4a. It can be observed that the performance drops by 3% - 5% when historical divergence is not used. Interestingly, as Figure 4b shows, many more ID samples are selected in this case, almost twice as many as when historical divergence is used. However, selecting more ID samples does not improve performance. We conjecture that there are two

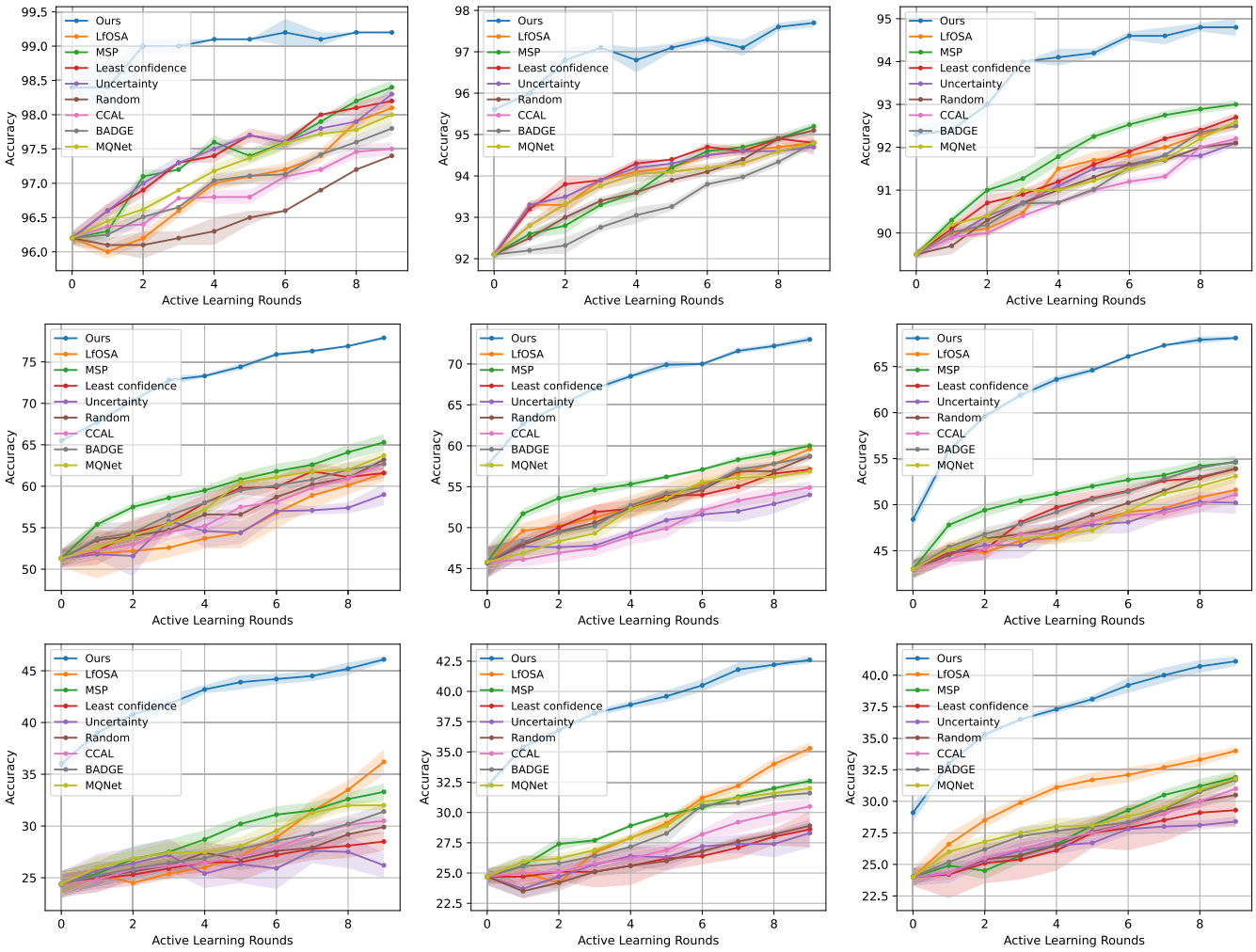
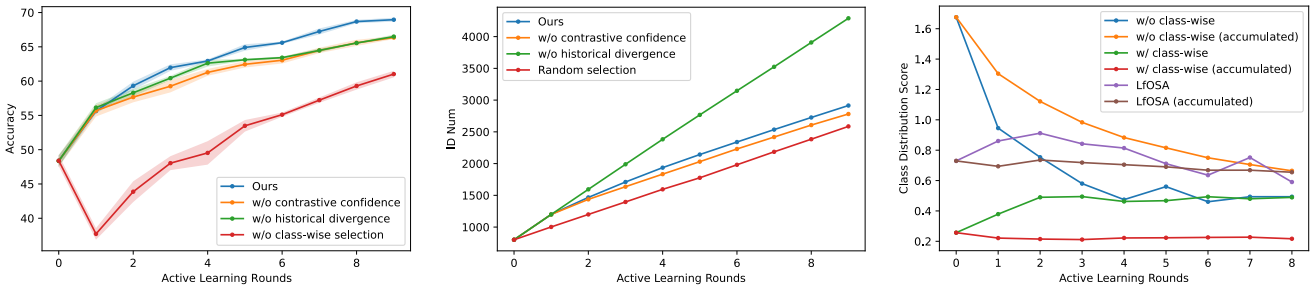


Fig. 3. Comparison results on CIFAR-10 (first row), CIFAR-100 (second row) and TinyImagenet (third row). The ID ratio ξ is 20% (first column), 30% (second column) and 40% (third column). Note that all of the above methods start from a SimCLR pre-trained model at each round.



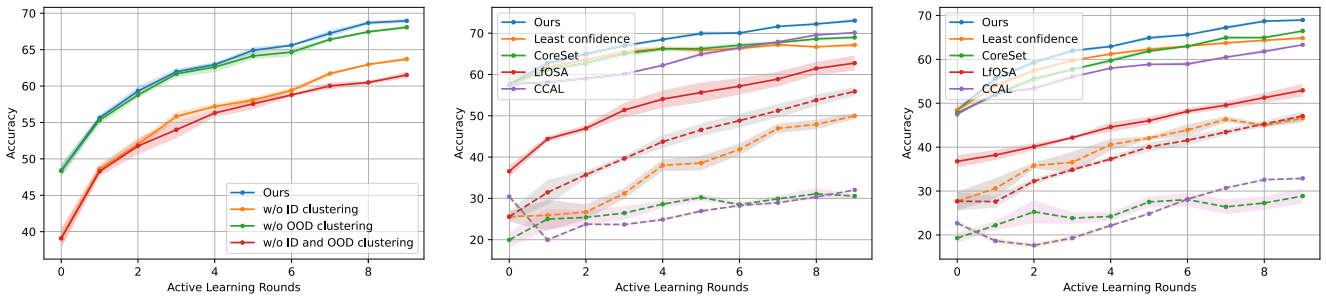
(a) (b) (c)

Fig. 4. (a) Ablation study on sample selection. (b) The accumulated number of ID samples during active learning. (c) The class distribution score (CDS, Eq. 10) of selected ID samples and the accumulated ID samples. CDS $\in [0, 2-2/n]$ and lower is better.

reasons: i) those samples are more likely to be easy samples that lack informativeness, and ii) the contrastive clustering can handle the OOD samples properly, which benefits network training.

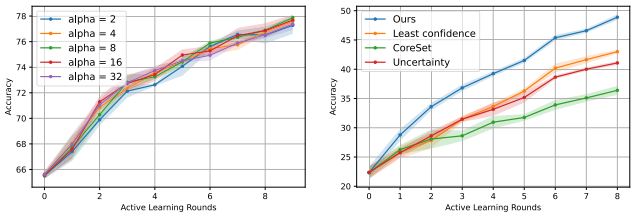
c) Class-wise selection: This module aims to select class-balanced samples. Without class-wise selection, all samples are sorted based on their final scores and selected in descending order. As Figure 4a shows, the model performs poorly in the early rounds when class-wise selection is not utilized. While the performance improves in the subsequent rounds, it remains far from the performance achieved with class-wise selection. To quantify the degree of class balance, we introduce a metric called class distribution score (CDS), which measures the $l1$ distance between the class distribution of the selected ID samples and a uniform distribution. Formally,

$$\text{CDS} = \mathbf{1}^\top |\mathbf{y} - \bar{\mathbf{y}}|, \tag{10}$$



(a) (b) (c)

Fig. 5. (a) Ablation study on the proposed contrastive clustering framework. (b) Plug-and-play evaluation on CIFAR-100 with ID ratio $\xi = 30\%$. The dashed lines represent the performance of baseline methods without using contrastive clustering. (c) Plug-and-play evaluation on CIFAR-100 with ID ratio $\xi = 40\%$.



(a) (b)

Fig. 6. (a) Analysis on trade-off parameters. (b) Performance on CIFAR-100 closed-set setting.

where $\mathbf{y} \in \mathbb{R}^n_{\geq 0}$ is the class distribution with a summation of 1, $\bar{\mathbf{y}}$ is the uniform distribution, and $\mathbf{1}$ is a vector of ones. Note that CDS $\in [0, 2 - 2/n]$, and lower is better. As Figure 4 shows, the class-wise selection can effectively select more class-balanced samples, demonstrating its efficacy despite the simple implementation. Furthermore, we compare the CDS of LfOSA [19] and find that the samples selected by LfOSA [19] exhibit severe class imbalance problems.

2) Contrastive Clustering: We conducted an ablation study on CIFAR-100 with an ID ratio of $\xi = 40\%$ to evaluate the effectiveness of each component of the proposed contrastive clustering. Specifically, we performed the following experiments: i) without OOD clustering $\mathcal{L}_{OOD}$, ii) without ID clustering $\mathcal{L}_{ID}$, and iii) without contrastive clustering $\mathcal{L}_{ID}$ and $\mathcal{L}_{OOD}$. As Figure 5a shows, both $\mathcal{L}_{ID}$ and $\mathcal{L}_{OOD}$ contribute significantly to the performance improvement when used independently. Furthermore, combining $\mathcal{L}_{ID}$ and $\mathcal{L}_{OOD}$ results in further performance gains. These findings demonstrate the efficacy of the proposed contrastive clustering approach. Additionally, our selection method outperforms other baselines even without contrastive clustering, as demonstrated in Figure 3.

3) Plug-and-Play Evaluation: We evaluate other AL selection methods on our contrastive clustering framework. The experiments are conducted on CIFAR-100 with ID ratio $\xi = 30\%$ and $\xi = 40\%$. As illustrated in Figures 5b and 5c, the proposed contrastive clustering consistently enhances the performance of the baseline methods considerably, indicating that it consistently aids the network in learning better representations. Furthermore, our selection method outperforms the other methods, providing further evidence for its efficacy. Notably, the performance improvement achieved by LfOSA [19] is not as significant as other methods. This may be due to the highly imbalanced selection of samples by LfOSA, as illustrated in Figure 4c.

4) Trade-Off Parameters: In this study, we investigate the sensitivity of our proposed method to the hyper-parameter α (Eq. 5), which controls the trade-off between contrastive confidence and historical divergence. A larger value of α assigns greater importance to contrastive confidence, implying that historical divergence is not utilized when $\alpha \to \infty$. Our experiments are performed on CIFAR-100 with ID ratio $\xi = 20\%$. As illustrated in Figure 6a, the proposed method achieves optimal performance when $\alpha = 8$. Additionally, the results reveal that our method is not significantly sensitive to the choice of α.

5) Closed-Set Active Learning: While our focus is on addressing the open-set active learning problem, it is interesting to investigate the performance of our methods in the closed-set setting. We perform such experiments on CIFAR-100, where all classes are considered in-distribution (ID). In this scenario, only the ID clustering loss $\mathcal{L}_{ID}$ is utilized since there is no OOD data. The compared methods are also trained with $\mathcal{L}_{ID}$. As depicted in Figure 6, our proposed method outperforms baseline methods by a significant margin, demonstrating the versatility of our sample selection method and the ID clustering loss.

6) Time Cost of Sample Selection: We conducted a comparison of our sample selection method's time cost for a single query with that of other methods. Table I presents the results, indicating that our method is one of the most efficient among the compared methods. While LfOSA [19] has similar efficiency in sample selection, its OOD detector training incurs significantly more time costs. Consequently, our method outperforms the baselines in both performance and efficiency.

7) Varying Initial and Budget Size: In this study, we investigate the robustness of our method against the initial and query budget size. We conducted experiments on CIFAR-100 with an ID ratio of $\xi = 30\%$, and the accuracy of the last round is reported. Figures 7a and 7b illustrate the results, indicating that our method outperforms the baseline methods for various initial and query budget sizes, demonstrating its robustness. Notably, the performance gap is more significant when the budget size is smaller, underscoring that our method can yield more cost savings in realistic scenarios involving limited annotation resources.

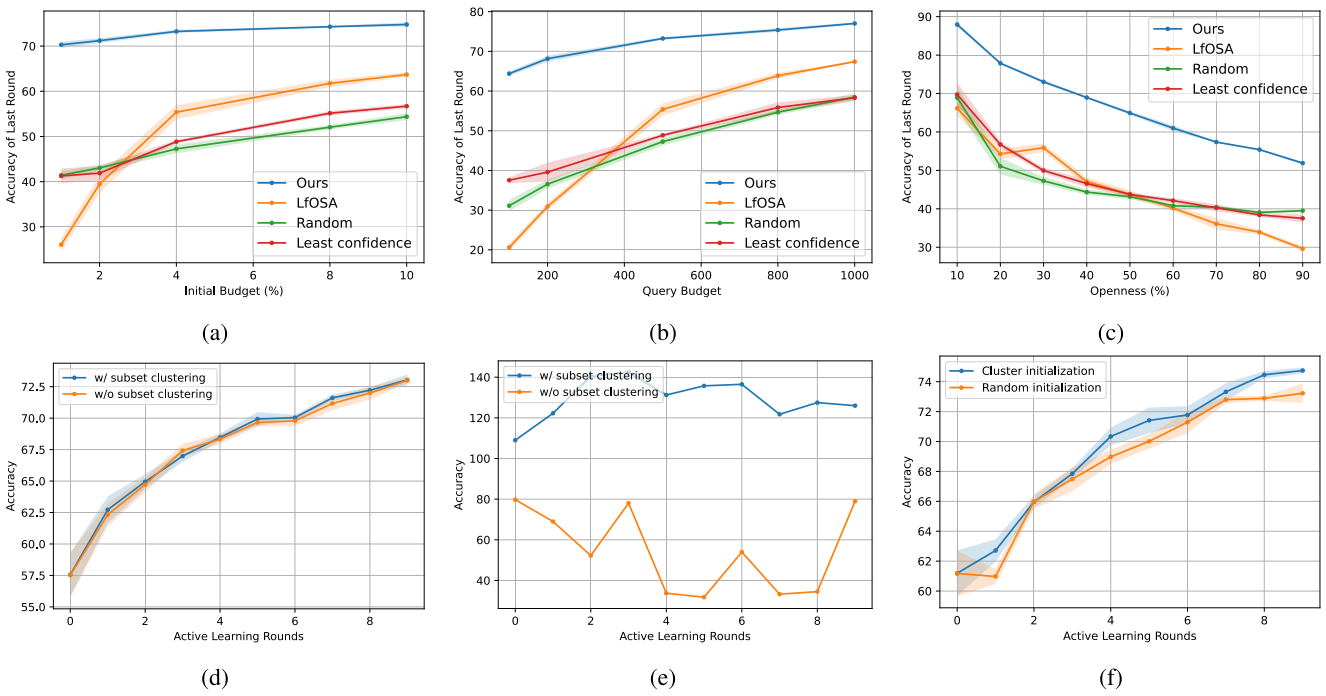


Fig. 7. (a) Varying initial budget. (b) Varying query budget. (c) Varying openness. (d) Comparison between the subset clustering and the whole set clustering in the cluster estimation step. (e) The number of estimated clusters in the cluster estimation step. (f) Comparison between the centroids initialization strategies in OOD clustering.

TABLE I

COMPARISON OF TIME COST FOR SAMPLE SELECTION. NOTE THAT WE DO NOT INCLUDE THE TIME COST FOR OBTAINING THE PREDICTIONS AND FEATURES OF UNLABELED SAMPLES

Ours	Random	Uncertainty	LfOSA	MSP	CCAL	BADGE
$0.015s$	$0s$	$0.012s$	$0.016s$	$0.012s$	$635s$	$1010s$

8) Varying Openness: In the main paper, we have demonstrated that our method can achieve satisfactory performance in the setting of closed-set active learning. To further explore its effectiveness across different levels of openness, we conducted additional experiments on CIFAR-100 with ID ratios ξ ranging from 10% to 90%, as depicted in Figure 7c. The results indicate that our method consistently outperforms the other methods by a significant margin, demonstrating the versatility of our approach.

9) Subset Clustering: As mentioned in the section on OOD clustering of the main paper, we first divide the selected OOD samples into subsets according to their predictions, then perform cluster estimation in each subset. The rationale stems from the fact that performing clustering estimation in the whole set is time-consuming. Accordingly, we compare the subset and whole set clustering on CIFAR-100 with ID ratio $\xi = 30\%$. As Figure 7d shows, the subset clustering slightly outperforms the whole set clustering. Moreover, the whole set clustering costs almost twice the time as the subset clustering, demonstrating the efficiency of the subset clustering.

10) Number of Estimated Clusters: In addition to the efficiency and performance of the subset clustering, we also show the number of estimated clusters. There are 70 OOD classes, and the estimated number of clusters is presented in Figure 7e. Notably, the subset clustering estimates more clusters than the whole set clustering. The whole set clustering, on the other hand, underestimates the number of clusters, while the subset clustering overestimates them. As noted in prior research [63], [75], over-clustering is beneficial to clustering, while under-clustering is not. These findings indicate that subset clustering produces better representations than whole set clustering.

11) Centroids Initialization: The OOD contrastive clustering loss $\mathcal{L}_{OOD}$ relies on prototypical predictions versus OOD cluster centroids, which are obtained by computing the cluster labels estimated at the end of the previous active learning round. During training, the centroids are updated using gradient descent. To assess the effectiveness of utilizing the cluster labels as an initialization for the centroids, we conducted experiments on randomly initialized centroids. The results, as illustrated in Figure 7f, indicate that cluster label initialization outperforms random initialization, highlighting the efficacy of the cluster estimation step. It is worth noting that even with random initialization, the performance achieved is decent, further emphasizing the versatility of the OOD contrastive clustering method.

12) Exclude OOD Samples: Our proposed method employs selected OOD samples to enhance representation learning. An interesting question is to investigate the effect of our approach in the absence of OOD samples. To address this, we conducted an experiment on CIFAR-100 with ID ratios of $\xi = 30\%$ and $\xi = 40\%$. The results presented in Table II demonstrate that our method significantly improves performance by leveraging OOD samples, thereby confirming its efficacy in enhancing representation learning.

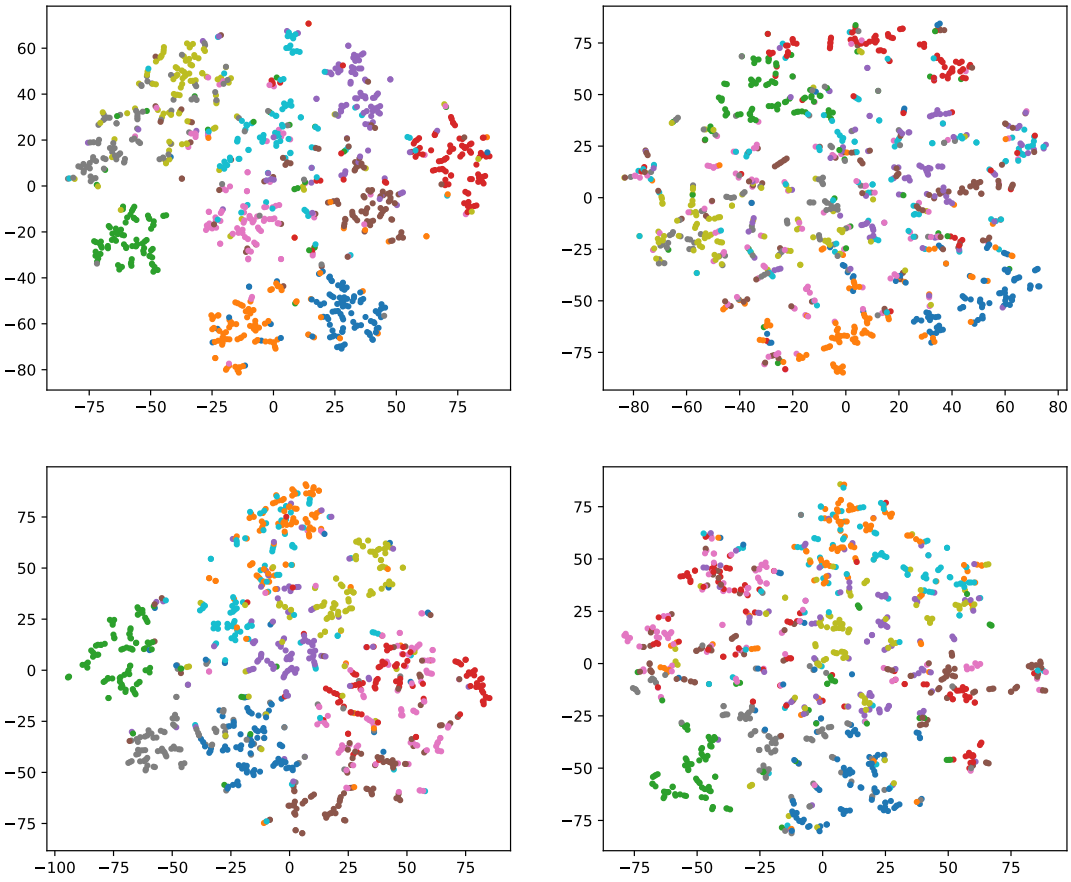


Fig. 8. First row: our ID and LfOSA ID samples. Second row: ours OOD and LfOSA OOD samples.

TABLE II

RESULT OF EXCLUDING OOD SAMPLES FROM TRAINING. THE ACCURACY OF THE FINAL ROUND IS REPORTED

Method	$\xi = 30\%$	$\xi = 40\%$
w/o OOD	69.5	65.3
w/ OOD	72.1	69.0

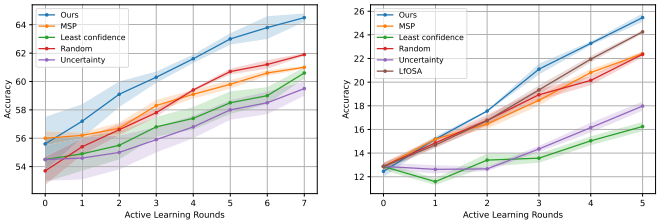
TABLE III

RESULT OF USING FIXED λ

λ	0.2	0.4	0.8	Adaptive
Final Accuracy	67.9	67.2	66.5	69.0

13) Adaptive Scheduling: In order to evaluate the effectiveness of the adaptive scheduling of λ, an experiment was conducted with λ held constant and set to various values on CIFAR100 with $\xi = 40\%$. The results presented in Table III showcase the superiority of our proposed adaptive scheduling method.

14) Feature Visualization: We visualize the feature learned by our method and LfOSA (LfOSA is trained with our clustering method) via t-SNE [76]. Specifically, we sample 10 classes for both ID and OOD classes. As Figure 8 shows, our method can learn more discriminative and well-clustered features for ID classes. Furthermore, it is worth noting that our method can also learn better representations for OOD classes.



(a) (b)

Fig. 9. (a) Results on Imagenet-1k. (b) Cross dataset experiment.

15) Results on Large-scale Dataset: In addition to CIFAR and TinyImageNet, we also evaluated our method on a large-scale dataset, namely, ImageNet-1K. The ImageNet-1K dataset is a large-scale, highly diverse collection of over 1 million annotated images across 1000 classes. We use the first 200 classes as the in-distribution (ID) classes and the remaining classes as the out-of-distribution (OOD). Since performing clustering on a million-scale sample set is very challenging, methods that require clustering are infeasible for this dataset, for example, LfOSA [19], BADGE [7], and our contrastive clustering component. Therefore, we only evaluate our sample selection component on ImageNet-1K. We use ResNet-50 as the backbone network. As shown in Figure 9a, our method still outperforms other methods by significant margins, demonstrating the versatility of our method on large-scale datasets.

TABLE IV
RESULT OF ROUND REINITIALIZATION. THE ACCURACY OF THE FINAL ROUND IS REPORTED

Method	$\xi = 30\%$	$\xi = 40\%$
w/o Reinitialization	71.8	69.1
w/ Reinitialization	72.1	69.0

16) Cross Dataset Experiment: In the previous experiments, although the ID and OOD samples were from different classes, they belonged to the same dataset. To further verify the effectiveness of our method, we performed an experiment that involved ID and OOD samples from different datasets. Since the CIFAR-10/100 and TinyImageNet datasets have a substantial overlap in classes, we employed two additional datasets: Stanford Cars [77] and Oxford Flowers [78]. Specifically, the Stanford Cars dataset was regarded as the ID, and the Oxford Flowers dataset was regarded as the OOD. The results are shown in Figure 9. It can be observed that our method still shows superiority over the other methods compared.

17) Round Reinitialization: In this paper, we reinitialize the network with the pre-trained weights at the beginning of each round, following the convention of active learning. However, it is interesting to consider how performance would change if we were to continue training the network without reinitialization. We conducted an experiment on CIFAR100 with ID ratios of $\xi = 30\%$ and $\xi = 40\%$. Table IV shows that there is no significant difference between the two choices. We conjecture that the reason for this is that the pre-trained backbone has already learned highly discriminative representations, and continual training will no longer improve its capacity.

V. CONCLUSION

In this paper, we introduce a novel approach to tackle the open-set active learning problem. Our proposed method consists of a simple selection method and a contrastive clustering method. In contrast to previous approaches that focus on selecting only highly likely ID samples, we consider three aspects: (i) the possibility of being ID, (ii) the hardness of the query set, and (iii) the diversity of the query set. To this end, we present two criteria and a selection operation, namely contrastive confidence, historical divergence, and class-wise selection, that can effectively address the above aspects. Moreover, unlike prior approaches that require an OOD detector to filter out OOD samples, we propose a contrastive clustering method that leverages both selected ID and OOD samples, empowering the network to detect OOD samples while simultaneously enhancing representation learning. Experimental results demonstrate that our method achieves state-of-the-art performance.

REFERENCES

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Proc. ECCV*, 2016, pp. 630–645.

[2] P. Covington, J. Adams, and E. Sargin, "Deep neural networks for YouTube recommendations," in *Proc. RecSys*, Sep. 2016, pp. 191–198.

[3] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. MICCAI*, 2015, pp. 234–241.

[4] B. Settles, "Active learning literature survey," Univ. Wisconsin–Madison, Madison, WI, USA, Rep. 1648, 2009.

[5] P. Ren, "A survey of deep active learning," *ACM Comput. Surv.*, vol. 54, no. 9, pp. 1–40, 2021.

[6] Y. Gal, R. Islam, and Z. Ghahramani, "Deep Bayesian active learning with image data," in *Proc. ICML*, 2017, pp. 1183–1192.

[7] J. T. Ash, C. Zhang, A. Krishnamurthy, J. Langford, and A. Agarwal, "Deep batch active learning by diverse, uncertain gradient lower bounds," 2019, *arXiv:1906.03671*.

[8] A. Parvaneh, E. Abbasnejad, D. Teney, R. Haffari, A. Van Den Hengel, and J. Q. Shi, "Active learning by feature mixing," in *Proc. CVPR*, Jun. 2022, pp. 12227–12236.

[9] K. Wang, D. Zhang, Y. Li, R. Zhang, and L. Lin, "Cost-effective active learning for deep image classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 12, pp. 2591–2600, Dec. 2017.

[10] O. Sener and S. Savarese, "Active learning for convolutional neural networks: A core-set approach," in *Proc. Int. Conf. Learn. Represent.*, 2018. [Online]. Available: https://openreview.net/forum?id=H1aIuk-RW

[11] D. Yoo and I. S. Kweon, "Learning loss for active learning," in *Proc. CVPR*, Jun. 2019, pp. 93–102.

[12] S. Sinha, S. Ebrahimi, and T. Darrell, "Variational adversarial active learning," in *Proc. ICCV*, Oct. 2019, pp. 5971–5980.

[13] S. Agarwal, H. Arora, S. Anand, and C. Arora, "Contextual diversity for active learning," in *Proc. ECCV*, 2020, pp. 137–153.

[14] G. Citovsky et al., "Batch active learning at scale," in *Proc. NeurIPS*, vol. 34, 2021, pp. 11933–11944.

[15] M. Gao, Z. Zhang, G. Yu, S. Ö. Arık, L. S. Davis, and T. Pfister, "Consistency-based semi-supervised active learning: Towards minimizing labeling cost," in *Proc. ECCV*, 2020, pp. 510–526.

[16] R. Caramalau, B. Bhattarai, and T.-K. Kim, "Sequential graph convolutional network for active learning," in *Proc. CVPR*, Jun. 2021, pp. 9578–9587.

[17] T. Yuan et al., "Multiple instance active learning for object detection," in *Proc. CVPR*, Jun. 2021, pp. 5326–5335.

[18] G. Hacohen, A. Dekel, and D. Weinshall, "Active learning on a budget: Opposite strategies suit high and low budgets," 2022, *arXiv:2202.02794*.

[19] K.-P. Ning, X. Zhao, Y. Li, and S.-J. Huang, "Active learning for open-set annotation," in *Proc. CVPR*, Jun. 2022, pp. 41–49.

[20] S. Vaze, K. Han, A. Vedaldi, and A. Zisserman, "Open-set recognition: A good closed-set classifier is all you need?" in *Proc. Int. Conf. Learn. Represent.*, 2022. [Online]. Available: https://openreview.net/forum?id=5hLP5JY9S2d

[21] D. D. Lewis and J. Catlett, "Heterogeneous uncertainty sampling for supervised learning," in *Proc. Mach. Learn.*, 1994, pp. 148–156.

[22] D. Roth and K. Small, "Margin-based active learning for structured output spaces," in *Proc. ECCV*, 2006, pp. 413–424.

[23] A. J. Joshi, F. Porikli, and N. Papanikolopoulos, "Multi-class active learning for image classification," in *Proc. CVPR*, Jun. 2009, pp. 2372–2379.

[24] M. Ducoffe and F. Precioso, "Adversarial active learning for deep networks: A margin based approach," 2018, *arXiv:1802.09841*.

[25] Z. Liu, H. Ding, H. Zhong, W. Li, J. Dai, and C. He, "Influence selection for active learning," in *Proc. ICCV*, Oct. 2021, pp. 9254–9263.

[26] N. Houlsby, F. Huszár, Z. Ghahramani, and M. Lengyel, "Bayesian active learning for classification and preference learning," 2011, *arXiv:1112.5745*.

[27] M. Kampffmeyer, A.-B. Salberg, and R. Jenssen, "Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks," in *Proc. CVPR*, Jun. 2016, pp. 680–688.

[28] T. Wang et al., "Boosting active learning via improving test performance," in *Proc. AAAI*, 2022, pp. 8566–8574.

[29] J.-J. Zhu and J. Bento, "Generative adversarial active learning," 2017, *arXiv:1702.07956*.

[30] F. Zhdanov, "Diverse mini-batch active learning," 2019, *arXiv:1901.05954*.

[31] Y. Geifman and R. El-Yaniv, "Deep active learning over the long tail," 2017, *arXiv:1711.00941*.

[32] E. Bıyık, K. Wang, N. Anari, and D. Sadigh, "Batch active learning using determinantal point processes," 2019, *arXiv:1906.07975*.

[33] D. Gissin and S. Shalev-Shwartz, "Discriminative active learning," 2019, *arXiv:1907.06347*.

[34] C. Shui, F. Zhou, C. Gagné, and B. Wang, "Deep active learning: Unified and principled method for query and training," in *Proc. AISTATS*, 2020, pp. 1308–1318.

[35] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," 2013, *arXiv:1312.6114*.
[36] C. Yin et al., "Deep similarity-based batch mode active learning with exploration-exploitation," in *Proc. ICDM*, Nov. 2017, pp. 575–584.
[37] X. Zhan, Q. Li, and A. B. Chan, "Multiple-criteria based active learning with fixed-size determinantal point processes," 2021, *arXiv:2107.01622*.
[38] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," 2017, *arXiv:1710.09412*.
[39] W. J. Scheirer, L. P. Jain, and T. E. Boult, "Probability models for open set recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 11, pp. 2317–2324, Nov. 2014.
[40] A. Bendale and T. E. Boult, "Towards open set deep networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1563–1572.
[41] L. Neal, M. Olson, X. Fern, W.-K. Wong, and F. Li, "Open set learning with counterfactual images," in *Proc. ECCV)*, 2018, pp. 613–628.
[42] Z. Ge, S. Demyanov, Z. Chen, and R. Garnavi, "Generative OpenMax for multi-class open set classification," 2017, *arXiv:1707.07418*.
[43] S. Kong and D. Ramanan, "OpenGAN: Open-set recognition via open data generation," in *Proc. ICCV*, Oct. 2021, pp. 793–802.
[44] R. Yoshihashi, W. Shao, R. Kawakami, S. You, M. Iida, and T. Naemura, "Classification-reconstruction learning for open-set recognition," in *Proc. CVPR*, Jun. 2019, pp. 4011–4020.
[45] P. Oza and V. M. Patel, "C2AE: Class conditioned auto-encoder for open-set recognition," in *Proc. CVPR*, Jun. 2019, pp. 2302–2311.
[46] X. Sun, Z. Yang, C. Zhang, K.-V. Ling, and G. Peng, "Conditional Gaussian distribution learning for open set recognition," in *Proc. CVPR*, Jun. 2020, pp. 13477–13486.
[47] G. Chen et al., "Learning open set network with discriminative reciprocal points," in *Proc. ECCV*, 2020, pp. 507–522.
[48] G. Chen, P. Peng, X. Wang, and Y. Tian, "Adversarial reciprocal points learning for open set recognition," 2021, *arXiv:2103.00953*.
[49] K. Saito, D. Kim, and K. Saenko, "OpenMatch: Open-set semi-supervised learning with open-set consistency regularization," in *Proc. NeurIPS*, vol. 34, 2021, pp. 25956–25967.
[50] Y. Fan, A. Kukleva, D. Dai, and B. Schiele, "SSB: Simple but strong baseline for boosting performance of open-set semi-supervised learning," in *Proc. ICCV*, Oct. 2023, pp. 16022–16032.
[51] J. Huang et al., "Trash to treasure: Harvesting OOD data with cross-modal matching for open-set semi-supervised learning," in *Proc. ICCV*, Oct. 2021, pp. 8290–8299.
[52] E. Wallin, L. Svensson, F. Kahl, and L. Hammarstrand, "Improving open-set semi-supervised learning with self-supervision," in *Proc. WACV*, vol. 32, Jan. 2024, pp. 2345–2354.
[53] M. N. Rizve, N. Kardan, S. Khan, F. S. Khan, and M. Shah, "OpenLDN: Learning to discover novel classes for open-world semi-supervised learning," in *Proc. ECCV*, 2022, pp. 382–401.
[54] T. Ahmad et al., "Variable few shot class incremental and open world learning," in *Proc. CVPR*, Jun. 2022, pp. 3687–3698.
[55] S. Huang, J. Ma, G. Han, and S.-F. Chang, "Task-adaptive negative envision for few-shot open-set recognition," in *Proc. CVPR*, Jun. 2022, pp. 7161–7170.
[56] B. Li, S. Luo, J. Wang, L. Tian, and D. Chen, "Few-shot object recognition based on three-way decision and active learning," *Vis. Comput.*, vol. 37, Jan. 2022.
[57] C. Peng, K. Zhao, T. Wang, M. Li, and B. C. Lovell, "Few-shot class-incremental learning from an open-set perspective," in *Proc. ECCV*, 2022, pp. 382–397.
[58] B. Liu, H. Kang, H. Li, G. Hua, and N. Vasconcelos, "Few-shot open-set recognition using meta-learning," in *Proc. CVPR*, Jun. 2020, pp. 8795–8804.
[59] M. Jeong, S. Choi, and C. Kim, "Few-shot open-set recognition by transformation consistency," in *Proc. CVPR*, Jun. 2021, pp. 12561–12570.
[60] K. Han, A. Vedaldi, and A. Zisserman, "Learning to discover novel visual categories via deep transfer clustering," in *Proc. ICCV*, Oct. 2019, pp. 8400–8408.
[61] B. Zhao and K. Han, "Novel visual category discovery with dual ranking statistics and mutual knowledge distillation," in *Proc. NeurIPS*, vol. 34, 2021, pp. 22982–22994.
[62] S. Vaze, K. Hant, A. Vedaldi, and A. Zisserman, "Generalized category discovery," in *Proc. CVPR*, Jun. 2022, pp. 7482–7491.
[63] E. Fini, E. Sangineto, S. Lathuilière, Z. Zhong, M. Nabi, and E. Ricci, "A unified objective for novel class discovery," in *Proc. ICCV*, Oct. 2021, pp. 9264–9272.
[64] H. Chi et al., "Meta discovery: Learning to discover novel classes given very limited data," 2021, *arXiv:2102.04002*.
[65] P. Du, S. Zhao, H. Chen, S. Chai, H. Chen, and C. Li, "Contrastive coding for active learning under class distribution mismatch," in *Proc. ICCV*, Oct. 2021, pp. 8907–8916.
[66] S. Kothawade, N. Beck, K. Killamsetty, and R. Iyer, "Similar: Submodular information measures based active learning in realistic scenarios," in *Proc. NeurIPS*, vol. 34, 2021, pp. 18685–18697.
[67] D. Park, Y. Shin, J. Bang, Y. Lee, H. Song, and J.-G. Lee, "Meta-query-net: Resolving purity-informativeness dilemma in open-set active learning," in *Proc. NeurIPS*, vol. 35, 2022, pp. 31416–31429.
[68] N. Pu, Z. Zhong, and N. Sebe, "Dynamic conceptional contrastive learning for generalized category discovery," in *Proc. CVPR*, Jun. 2023, pp. 7579–7588.
[69] X. Wen, B. Zhao, and X. Qi, "Parametric classification for generalized category discovery: A baseline study," in *Proc. ICCV*, Oct. 2023, pp. 16590–16600.
[70] F. Chiaroni, J. Dolz, Z. I. Masud, A. Mitiche, and I. B. Ayed, "Parametric information maximization for generalized category discovery," in *Proc. ICCV*, Oct. 2023, pp. 1729–1739.
[71] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. ICML*, 2020, pp. 1597–1607.
[72] A. Krizhevsky et al., "Learning multiple layers of features from tiny images," Univ. Toronto, Toronto, ON, Canada, 2009.
[73] Y. Le and X. Yang, "Tiny ImageNet visual recognition challenge," Stanford Univ., Stanford, CA, USA, 2015.
[74] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. CVPR*, Jun. 2009, pp. 248–255.
[75] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," in *Proc. NeuIPS*, 2020, pp. 9912–9924.
[76] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 2579–2605, 2008.
[77] T. Gebru, J. Krause, Y. Wang, D. Chen, J. Deng, and L. Fei-Fei, "Fine-grained car detection for visual census estimation," in *Proc. AAAI*, 2017, pp. 4502–4508.
[78] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *Proc. ICVGIP*, Dec. 2008, pp. 722–729.

Zizheng Yan (Member, IEEE) received the B.Eng. degree from The Chinese University of Hong Kong at Shenzhen in 2019, where he is currently pursuing the Ph.D. degree. His research interests include computer vision and deep learning and their applications.

Delian Ruan received the B.Eng. and M.Eng. degrees from Xiamen University in 2018 and 2021, respectively. She is currently a Researcher with Meituan. Her research interests include computer vision and deep learning.

Yushuang Wu (Graduate Student Member, IEEE) received the B.Sc. degree from the Bell Honor School, Nanjing University of Posts and Telecommunications, in 2019. He is currently pursuing the Ph.D. degree with The Chinese University of Hong Kong at Shenzhen. His research interests include 3D reconstruction and point cloud representation.

Junshi Huang received the Ph.D. degree from the National University of Singapore. He is currently a Research Scientist with Meituan and the Leader of the Image Understanding and Generation Group. His major research interests include computer vision, machine learning, and multimedia analysis.

Zhenhua Chai received the B.E. degree (Hons.) in automation from the Central University of Nationality in 2008 and the Ph.D. degree in computer application technology from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, in 2013. He is currently a Research Expert with Meituan. His research interests include AutoDL, model compression, self-supervised learning, and applications on computer vision.

Xiaoguang Han (Member, IEEE) is currently an Assistant Professor with The Chinese University of Hong Kong at Shenzhen. He already published over 100 papers in top-tier conferences, such as CVPR, ICCV, ECCV, SIGGRAPH, and SIGGRAPH Asia. His research interests include computer vision and computer graphics. He is serving as an Area Chair for NeurIPS, CVPR, and ECCV, a Program Committee Member for SIGGRAPH Asia, and an Associate Editor for IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS and *Computer & Graphics*.

Shuguang Cui (Fellow, IEEE) received the Ph.D. degree in electrical engineering from Stanford University, Stanford, CA, USA, in 2005. He has been an Assistant Professor, an Associate Professor, a Full Professor, and a Chair Professor of electrical and computer engineering with The University of Arizona, Texas A&M University, UC Davis, and The Chinese University of Hong Kong at Shenzhen, respectively. His current research interests include the merging between AI and communication networks. He has also been serving as the Area Editor for *IEEE Signal Processing Magazine*, an Associate Editor for IEEE TRANSACTIONS ON BIG DATA, IEEE TRANSACTIONS ON SIGNAL PROCESSING, and IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, and the Editor-in-Chief for IEEE TRANSACTIONS ON MOBILE COMPUTING. He was selected as the Thomson Reuters Highly Cited Researcher and listed in the World's Most Influential Scientific Minds by ScienceWatch in 2014.

Guanbin Li (Member, IEEE) received the Ph.D. degree from The University of Hong Kong in 2016. He is currently a Full Professor with the School of Computer Science and Engineering, Sun Yat-sen University. He has authored or co-authored more than 150 papers in top-tier academic journals and conferences. His current research interests include computer vision, image processing, and deep learning. He was a recipient of the ICCV 2019 Best Paper Nomination Award. He serves as an Area Chair for CVPR 2024. He has been serving as a reviewer for numerous academic journals and conferences, such as IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, *International Journal of Computer Vision*, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON CYBERNETICS, CVPR, ICCV, ECCV, and NeurIPS.