Entropic Open-Set Active Learning

Bardia Safaei[1], Vibashan VS[1], Celso M. de Melo[2], Vishal M. Patel[1]

[1]Johns Hopkins University, Baltimore, MD, USA
[2]DEVCOM Army Research Laboratory, Adelphi, MD, USA
{bsafaei1, vvishnu2}@jhu.edu, celso.m.demelo.civ@army.mil, vpatel36@jhu.edu

Abstract

Active Learning (AL) aims to enhance the performance of deep models by selecting the most informative samples for annotation from a pool of unlabeled data. Despite impressive performance in closed-set settings, most AL methods fail in real-world scenarios where the unlabeled data contains unknown categories. Recently, a few studies have attempted to tackle the AL problem for the open-set setting. However, these methods focus more on selecting known samples and do not efficiently utilize unknown samples obtained during AL rounds. In this work, we propose an Entropic Open-set AL (EOAL) framework which leverages both known and unknown distributions effectively to select informative samples during AL rounds. Specifically, our approach employs two different entropy scores. One measures the uncertainty of a sample with respect to the known-class distributions. The other measures the uncertainty of the sample with respect to the unknown-class distributions. By utilizing these two entropy scores we effectively separate the known and unknown samples from the unlabeled data resulting in better sampling. Through extensive experiments, we show that the proposed method outperforms existing state-of-the-art methods on CIFAR-10, CIFAR-100, and TinyImageNet datasets. Code is available at https://github.com/bardisafa/EOAL.

Introduction

In recent years, deep learning methods have shown remarkable performance in a large number of complex computer vision tasks such as classification (He et al. 2016; Radford et al. 2021), segmentation(Chen et al. 2017; Kirillov et al. 2023) and object detection (Ren et al. 2015; Redmon et al. 2016). However, the success of these deep learning models in solving these complex tasks heavily relies on the availability of extensive labeled data (VS et al. 2023; Vs et al. 2022). Obtaining labeled data is generally labor-intensive, and expensive (Wei, Iyer, and Bilmes 2015; VS, Oza, and Patel 2023). Active Learning (AL) tackles this huge data labeling issue by strategically selecting a subset of informative samples for annotation, rather than labeling the entire data. Primarily, there are two types of AL techniques: a) uncertainty-based methods, and b) diversity-based methods. Uncertainty-based techniques (Seung, Opper, and Sompolinsky 1992) leverage model uncertainty on unlabeled samples to select the most informative ones, while diversity-based methods (Nguyen and Smeulders 2004) focus on enhancing model learning by carefully choosing samples that show maximal diversity.

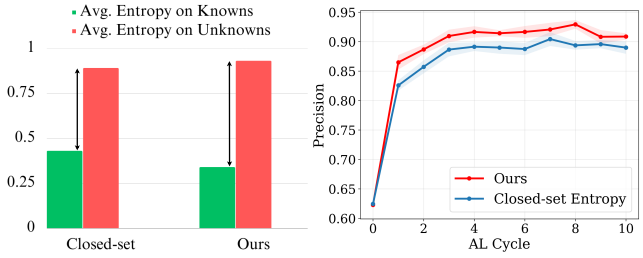

Figure 1: *Left*: At the 5-th AL round, we plot the difference between the average entropy scores of known and unknown samples using a closed-set classifier and our method. Our method utilizes two entropy scores which enhance the separation between known and unknown samples resulting in a better sampling of the knowns. *Right*: Active sampling precision graph for a closed-set classifier and our method. Because of better separation between known and unknown samples, our method tends to have better precision in sampling known samples at each AL cycle. (CIFAR-10, 40% mismatch ratio)

In general, AL methods produce promising results in closed-set settings where the unlabeled data contains only known classes. However, in real-world scenarios, this assumption does not hold as the unlabeled data contains both known and unknown samples. As a result, the performance of these closed-set AL methods significantly declines (Ning et al. 2022). One main reason for this phenomenon is that existing uncertainty- and diversity-based methods choose the unknown samples as the most informative samples for human annotation, thereby wasting the annotation budget. Human annotators would disregard these unknown samples because they are unnecessary for the target task. Therefore, it is important to address the problem of active learning under an open-set scenario where unknown class samples might appear in the unlabeled data during sampling.

A straightforward open-set AL approach is to train a closed-set classifier and utilize entropy scores to separate

known and unknown samples. Later, we annotate the low entropy samples because the closed-set classifier produces low entropy for known samples (see Fig. 1). However, acquiring unknown samples together with know-class samples is unavoidable, especially when their presence is extensive. To address this challenge, LfOSA (Ning et al. 2022) proposes to utilize unknown samples and train a classifier to reject the unknown samples and focus more on selecting known samples. MQNet (Park et al. 2022) proposes a more agnostic approach where the model leverages meta-learning to separate known and unknown samples. Despite their promising performance, these methods focus more on known-class sampling and do not efficiently utilize unknown samples obtained during AL rounds.

To this end, we propose a novel AL framework designed to enhance the selection of informative samples from both known and unknown categories during the training process. Our approach aims to enhance the separation between known and unknown samples effectively. We achieve this by incorporating two distinct entropy scores into our framework. The first entropy score is computed using outputs from known classifiers. This score quantifies the uncertainty of samples with respect to the distribution of known classes. The second entropy score operates on a distance-based principle. Specifically, the unknown samples obtained from AL rounds are used to model the unknown distribution. Following this, the second score quantifies the uncertainty of samples with respect to the data distribution of unknown classes. Finally, utilizing the combined entropy score for all unlabeled samples we perform active sampling. Furthermore, to ensure diversity, we adaptively cluster and perform sampling on each cluster according to the AL budget. Extensive experiments show that our method outperforms existing state-of-the-art methods on CIFAR-10, CIFAR-100 and TinyImageNet datasets.

The contributions of this paper are as follows.

- We introduce an AL framework that leverages both known and unknown distributions to select informative samples during AL rounds.
- Specifically, we propose two entropy scores which separate the known and unknown samples for precise AL sampling.
- Our experimental results show that the proposed method outperforms existing state-of-the-art methods on CIFAR-10, CIFAR-100, and TinyImageNet datasets.

Related Work

Active Learning (AL). Active learning aims at maximizing the performance improvement of a model by choosing the most beneficial samples from a set of unlabeled data, labeling them, and incorporating them into the supervised training process. Uncertainty-based AL approaches attempt to select the samples that the model is most uncertain about via various uncertainty measures, such as entropy (Luo, Schwing, and Urtasun 2013), mutual information (Kirsch, Van Amersfoort, and Gal 2019) and confidence margin (Balcan, Broder, and Zhang 2007). On the other hand, diversity-based approaches (Sener and Savarese 2017; Xu et al. 2003; Nguyen and Smeulders 2004) cluster the unlabeled samples and select representative samples from each cluster to better model the underlying distribution of unlabeled data. Query-by-Committee methods (Seung, Opper, and Sompolinsky 1992; Hino and Eguchi 2022) employ a measure of disagreement between an ensemble of models as the sample selection criterion. Recently, some methods have achieved enhanced AL performance by combining multiple sample selection criteria (Ash et al. 2019; Wei, Iyer, and Bilmes 2015; Parvaneh et al. 2022). For example, (Ash et al. 2019) a combination of diversity and uncertainty is employed to achieve improved performance, using the model's gradient magnitude as a measure of uncertainty. However, while these standard AL methods excel in typical AL scenarios, they cannot perform as effectively in the open-set setting with a class distribution mismatch between labeled and unlabeled data.

Open-set Recognition (OSR). The problem of open-set recognition was first formulated in (Scheirer et al. 2012) and has gained significant traction in recent years. In (Bendale and Boult 2016), the authors introduce a method called OpenMax that trains a model with an extra class denoting the probability that a sample belongs to an open-set class. They utilize Extreme Value Theory (EVT) to calibrate the network's output for better OSR performance. (Ge et al. 2017; Neal et al. 2018; Moon et al. 2022) use Generative Adversarial Networks (GANs) (Goodfellow et al. 2014) to synthesize images that resemble open-set classes. These images are then used to model the open-set space. Another line of work studies reconstruction-based approaches (Oza and Patel 2019; Sun et al. 2020; Yoshihashi et al. 2019) for open-set recognition. For example, C2AE (Oza and Patel 2019) trains class-conditioned auto-encoders and models the reconstruction errors using EVT. During inference, the minimum value of class-wise reconstruction errors is compared with a predefined threshold to detect open-set samples. (Lu et al. 2022; Chen et al. 2020; Yang et al. 2020; Shu et al. 2020) design different prototype-based learning mechanisms for OSR. (Chen et al. 2020) proposes a strategy called *reciprocal point learning* that learns reciprocal points to represent *otherness* of each known class. It attempts to push known samples far from reciprocal points and bound the open space of known classes. (Safaei et al. 2023; Cheng, Zhang, and Liu 2023) propose the use of one-versus-all classifiers to enhance OSR performance. In general, the OSR task differs from the task of open-set AL in two critical aspects. First, in OSR, the entire set of known classes is labeled and accessible to the model during training, whereas in open-set AL, only a few labeled samples are available initially. Second, OSR algorithms lack access to the real unknown data during training, while open-set AL approaches must be specifically designed to fully utilize the knowledge obtained from unknown samples gathered in later rounds of AL. These differences highlight the importance of devising approaches tailored to the challenges posed by open-set AL.

Open-set Active Learning. Recently, some approaches have studied the AL problem in the presence of open-set classes (Kothawade et al. 2021; Du et al. 2021; Park et al. 2022; Ning et al. 2022). MQNet (Park et al. 2022) addresses the purity-informativeness trade-off in open-set AL by train-

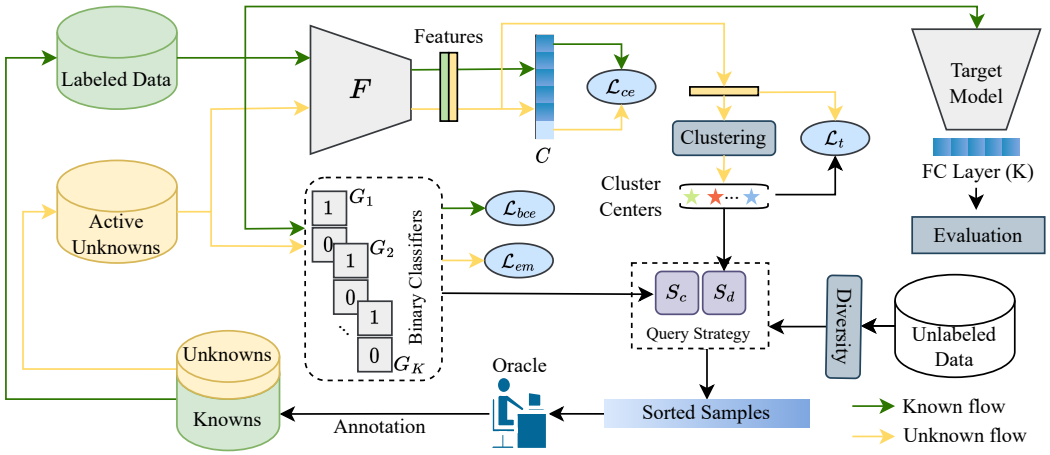

Figure 2: Overview of the proposed method for open-set active learning. At each AL cycle, we begin by training $F, C, \{G_i\}_{i=1}^{K}$ via minimizing Eq. 9 on labeled data and active unknown data. Two distinct entropy scores are then computed to establish the query strategy S as given in Eq. 8. Next, we cluster the unlabeled samples into K clusters and select the $\frac{b}{K}$ samples with the lowest S values for annotation, where b is the per-cycle annotation budget. Labeled and Active Unknown datasets are updated based on the annotated samples, and the target model is trained using the updated labeled dataset.

ing an MLP that receives one open-set score and one AL score as input and outputs a balanced meta-score for sample selection. LfOSA (Ning et al. 2022) attempts to construct a pure query set of known samples by modeling the maximum activation values of labeled data through class-wise GMMs and rejecting samples with lower probabilities as unknowns. However, as previously mentioned, many of these approaches do not fully utilize the availability of actual unknown data queried in AL rounds.

Methodology

In this section, we first present the problem of open-set AL, and then elaborate on the proposed approach in detail.

Problem Formulation. In open-set AL, we consider active learning for a K-way classification problem in an open-set setting, where K denotes the number of classes of interest (known classes). In this setting, we are initially given a small labeled dataset $\mathcal{D}_L = \left\{\left(x_i^L, y_i^L\right)\right\}_{i=1}^{N_L}$ of *known* samples that belong to the label space $\mathcal{K} = \{j\}_{j=1}^{K}$, and a large pool of unlabeled data $\mathcal{D}_U = \left\{x_i^U\right\}_{i=1}^{N_U}$ ($N_L \ll N_U$) which contains a mixture of *known* and *unknown* samples, where *unknown* samples belong to the label space $\mathcal{U}$, and $\mathcal{K} \cap \mathcal{U} = \emptyset$.

At each AL cycle, the query strategy selects a batch of b samples X^{active} from the unlabeled data, and their labels are queried from an oracle. X^{active} consists of *known* queried samples X^k and *unknown* queried samples X^u. The samples in X^u are labeled as the *open-set* class (class 0) by the oracle and referred to as *active unknowns* in this paper. We add X^u samples to the set of active unknowns $\mathcal{D}_{AU}$ and update the labeled dataset as $\mathcal{D}_L = \mathcal{D}_L \cup X^k$. The updated $\mathcal{D}_L$ is utilized to enhance the performance of a target model $T(\cdot)$ for an intended classification task.

Overview. Our AL framework utilizes two entropy scores that effectively differentiate between known and unknown samples, making them suitable for selecting the most valid samples for annotation. First, the *Closed-set Entropy* is calculated based on the outputs of K class-aware binary classifiers (BC) trained on $\mathcal{D}_L$. This entropy quantifies the uncertainty of a sample with respect to the distributions of the known classes, which tends to be low for known samples and high for unknown samples. Second, the *Distance-based Entropy* is utilized to prioritize the selection of samples that stand apart from distributions of unknown classes. To compute this entropy for a given sample, we start by clustering the Convolutional Neural Network (CNN) based features of $\mathcal{D}_{AU}$ samples and determining the cluster centers. The distances between the sample and these cluster centers are then used to measure the entropy of the samples. Fig. 2 provides an overview of our approach.

Training for Closed-set Entropy Scoring

To quantify closed-set entropy score (S_c), we employ (1) a CNN-based feature extractor $F(\cdot)$, (2) K class-aware binary classifiers $G_i(\cdot)$, $i \in \{1, 2, ..., K\}$, and (3) a fully-connected layer $C(\cdot)$ that produces a probability vector $\in \mathbb{R}^{K+1}$ for $(K+1)$-way classification on $\mathcal{D}_L \cup \mathcal{D}_{AU}$. The parameters of F and C are updated using a standard cross-entropy loss ($\mathcal{L}_{ce}$) on $\mathcal{D}_L \cup \mathcal{D}_{AU}$.

Training Binary Classifiers. We train each G_i with the samples in the i-th known class as positives, and the remaining known samples as negatives. For a given image $\mathbf{x}$, let $\mathbf{f} = F(\mathbf{x})$ denote the extracted features of $\mathbf{x}$, and $p^i = \sigma(G_i(\mathbf{f}))$ denote the probability of $\mathbf{x}$ being categorized as positive class by G_i, where σ is the Softmax operator. The loss function for training G_i's is as follows:

$$\mathcal{L}_{bce} = \frac{1}{n_l} \sum_{(\mathbf{x}_i, y_i) \in \mathcal{D}_L} -\log(p^{y_i}) - \min_{j \neq y_i} \log(1 - p^j), \quad (1)$$

Algorithm 1: Our Proposed Algorithm for Open-set AL

Input:
  Labeled data $\mathcal{D}_L$, unlabeled data $\mathcal{D}_U$, number of AL cycles R, known categories K, per-cycle budget b, models F, C, T, and $\{G_i\}_{i=1}^K$
Process:
  $\mathcal{D}_{AU} \leftarrow \emptyset$ *# Initial active unknowns*
  Update models F, C, and $\{G_i\}_{i=1}^K$ by minimizing $\mathcal{L}_{total}$ in Eq. 9
  for $c = 0, 1, ..., R-1$ **do**
    $\forall x \in D_U, S_d(x) \leftarrow 0$ *# Initialization*
    if $\mathcal{D}_{AU} \neq \emptyset$ **do**
      Cluster the features of $\mathcal{D}_{AU}$ into K clusters
      For cluster i, compute the center $\mathbf{c}_i$ using Eq. 5
      $\forall x \in D_U$, compute $S_d(x)$ via Eq. 6
    end if
    $\forall x \in D_U$, compute $S_c(x)$ via Eq. 2
    Cluster the features of $\mathcal{D}_U$ into K clusters, $\{C_1, C_2, ..., C_K\}$ *# Diversity*
    for $j = 1, 2, ..., K$ **do**
      $\mathcal{S}_j \leftarrow \{S_c(x) - S_d(x) | \forall x \in C_j\}$ *# Uncertainty*
      $X_j \leftarrow$ select the $\frac{b}{K}$ samples with the lowest values from $\mathcal{S}_j$, and annotate them
    end
    # All queries for the current cycle:
    $X^{active} \leftarrow X_1 \cup X_2 \cup ... \cup X_K$
    # Knowns and active unknowns:
    Obtain X^k and X^u
    $\mathcal{D}_L \leftarrow \mathcal{D}_L \cup X^k$, $\mathcal{D}_{AU} \leftarrow \mathcal{D}_{AU} \cup X^u$,
    $\mathcal{D}_U \leftarrow \mathcal{D}_U / X^{active}$ *# Update datasets*
    Update the target model T via minimizing the cross-entropy loss on D_L
  end

where n_l is the number of known samples in the batch. This loss is a modified version of binary cross-entropy (BCE) (Saito and Saenko 2021) that only updates the positive and the nearest negative decision boundaries for each sample. This is to mitigate the bias of a BC towards the negative class which includes lots of samples.

Closed-set Entropy. For a given sample $\mathbf{x}$, we define closed-set entropy score as follows:

$$S_c(\mathbf{x}) = \frac{1}{K \cdot \log(2)} \sum_{i=1}^{K} H_i(\mathbf{x}), \tag{2}$$

where $H_i(\mathbf{x})$ denotes the entropy of G_i given as:

$$H_i(\mathbf{x}) = -p^i \cdot \log(p^i) - \left(1 - p^i\right) \cdot \log(1 - p^i). \tag{3}$$

S_c measures the average normalized entropy of BC's.

High S_c on Unknown Samples. While training a BC via Eq. (1) obviously ensures its low entropy for known samples, it does not necessarily guarantee a high entropy for all unknown samples. To ensure high S_c for unknown samples, we minimize the following objective on $\mathcal{D}_{AU}$:

$$\mathcal{L}_{em} = \frac{1}{K \cdot n_{au}} \sum_{\mathbf{x} \in \mathcal{D}_{AU}} \sum_{i=1}^{K} -\frac{1}{2} \log(p^i) - \frac{1}{2} \log(1 - p^i), \tag{4}$$

where n_{au} is the number of active unknown samples in the batch. This loss encourages uniform probability outputs $p = [\frac{1}{2}, \frac{1}{2}]$ for unknown samples.

Property 1. *Minimizing L_{em} is equivalent to maximizing the entropy of each G_i.*

Proof. We have $p^i + (1 - p^i) = 1$, and $p^i \in (0, 1)$. By applying Jensen's inequality for concave functions, we obtain $H_i(\mathbf{x}) \leq \log(2)$ and $\mathcal{L}_{em} \geq \log(2)$, where the equality happens *iff* $p^i = (1 - p^i) = \frac{1}{2}$.

Training for Distance-based Entropy Scoring

S_c alone is insufficient for open-set active sampling, as it can be misled by unknown samples in close proximity to a known category. Hence, we employ a distance-based entropy score S_d to achieve higher precision in selecting known samples. Typically, an unknown sample lies near the distribution of its ground-truth category while being distant from other categories, and hence it exhibits a low S_d. Conversely, a known sample remains distant from all unknown categories, resulting in a high S_d.

Distance-based Entropy. We leverage $\mathcal{D}_{AU}$ samples for computing S_d. Having no access to their precise category labels, we first cluster these samples using the FINCH clustering algorithm (Sarfraz, Sharma, and Stiefelhagen 2019). We fix the number of clusters to K. Denoting the obtained cluster labels for $\mathcal{D}_{AU}$ samples as $\{\hat{y}_i\}_{i=1}^{N_{AU}}$, we then compute the center of each cluster as follows:

$$\mathbf{c}_i = \frac{\sum_{(\mathbf{x}, \hat{y}) \in \mathcal{D}_{AU}} \mathbb{I}\{\hat{y} = i\} \cdot F(\mathbf{x})}{\sum_{(\mathbf{x}, \hat{y}) \in \mathcal{D}_{AU}} \mathbb{I}\{\hat{y} = i\}}, \tag{5}$$

where $\mathbb{I}\{\cdot\}$ is the indicator function, and $N_{AU} = |\mathcal{D}_{AU}|$. Finally, S_d is defined as:

$$\begin{aligned} S_d(\mathbf{x}) &= \frac{-1}{\log(K)} \sum_{i=1}^{K} q_i(\mathbf{x}) \cdot \log(q_i(\mathbf{x})), \\ q_i(\mathbf{x}) &= \frac{e^{-\|F(\mathbf{x}) - \mathbf{c}_i\|/T}}{\sum_{j=1}^{K} e^{-\|F(\mathbf{x}) - \mathbf{c}_j\|/T}}, \end{aligned} \tag{6}$$

where $q_i(\mathbf{x})$ is the probability of a sample $\mathbf{x}$ belonging to the i-th cluster, and T is a temperature. Basically, we calculate the distances of a sample from cluster centers, form the probability vector $[q_1(\mathbf{x}), q_2(\mathbf{x}), ..., q_K(\mathbf{x})]$, and compute its normalized entropy. We chose FINCH over K-means (MacQueen et al. 1967) due to its superior speed and efficiency.

Low S_d on Unknown Samples. Using cross-entropy loss for training $F(\cdot)$ generally ensures high distance of known samples from $\mathbf{c}_i$'s in the feature space, resulting in high S_d values. However, low S_d values for unknown samples cannot be guaranteed since cross-entropy loss treats all active unknowns as one open-set class. Hence, we further regularize the features to impose more compactness within each

cluster while maintaining a larger margin between different clusters. Specifically, for a sample $\mathbf{x} \in \mathcal{D}_{AU}$ with cluster label $\hat{y}$, we utilize Tuplet loss (Sohn 2016; Miller et al. 2021) to enforce a low distance between $F(\mathbf{x})$ and the cluster center $\mathbf{c}_{\hat{y}}$, and a large margin between the distance to $\mathbf{c}_{\hat{y}}$ and the distance to $\mathbf{c}_{j \neq \hat{y}}$. The loss can be written as follows:

$$\mathcal{L}_t = \frac{1}{n_{au}} \sum_{\mathbf{x} \in \mathcal{D}_{AU}} \log \left(1 + \sum_{j \neq \hat{y}}^{K} e^{D_{\hat{y}} - D_j}\right) + \beta D_{\hat{y}}, \quad (7)$$

where $D_i = \|F(\mathbf{x}) - \mathbf{c}_i\|$.

Query Strategy

As described in previous sections, we formulate our query strategy by combining Eq. 2 and Eq. 6 as

$$S(\mathbf{x}) = S_c(\mathbf{x}) - S_d(\mathbf{x}). \quad (8)$$

$S(\mathbf{x})$ score estimates the uncertainty of a sample with respect to the data distributions of both known (S_c) and unknown (S_d) categories. A known sample remains near its corresponding known category ($S_c \downarrow$) and distinct from all unknown categories ($S_d \uparrow$), while the opposite holds for an unknown sample ($S_d \downarrow$ and $S_c \uparrow$). As a result, $S(\mathbf{x})$ can effectively separate known samples from unknown ones.

Furthermore, to select informative samples, we propose to query from different regions in the feature space to minimize the redundancy of the selected samples. To be specific, our query strategy is as follows. We first consider the unlabeled samples which are classified as one of the known categories by $C(\cdot)$ and cluster them into K clusters using the FINCH algorithm. Then, within each cluster, we sort the samples based on the $S(\mathbf{x})$ score and select the first $\frac{b}{K}$ samples with the lowest scores for annotation, where b denotes the per-cycle annotation budget (see Algorithm 1). The importance of each component of our query strategy is further studied in ablation studies.

Overall Loss

Our proposed method is trained in an end-to-end manner by minimizing the following total objective:

$$\mathcal{L}_{total} = \begin{cases} \mathcal{L}_{ce} + \mathcal{L}_{bce} & \text{if } \mathcal{D}_{AU} = \emptyset \\ \mathcal{L}_{ce} + \mathcal{L}_{bce} + \mathcal{L}_{em} + \lambda \mathcal{L}_t & \text{if } \mathcal{D}_{AU} \neq \emptyset \end{cases}, \quad (9)$$

where we minimize $\mathcal{L}_{bce}$ on $\mathcal{D}_L$, $\mathcal{L}_{em}$ and $\mathcal{L}_t$ on $\mathcal{D}_{AU}$, and $\mathcal{L}_{ce}$ on $\mathcal{D}_L \cup \mathcal{D}_{AU}$. Note we do not consider $\mathcal{L}_{em}$ and $\mathcal{L}_t$ in the total objective before the first AL cycle since $\mathcal{D}_{AU} = \emptyset$.

Training the Target Model

After querying the samples at each AL cycle, a target model is trained on the updated $\mathcal{D}_L$ dataset, using the standard cross-entropy loss. The performance of this model in K-way classification is utilized for our evaluations.

Experiments

We perform extensive experiments on the CIFAR-10, CIFAR-100 (Krizhevsky, Hinton et al. 2009), and TinyImageNet (Yao and Miller 2015) datasets to demonstrate the effectiveness of our approach.

Datasets. The CIFAR-10 and CIFAR-100 datasets each contain 50000 images for training and 10000 images for testing, while they consist of 10 and 100 categories, respectively. TinyImageNet is a large-scale dataset containing 100000 training images and 20000 testing images in 200 categories.

Experimental Setting. For each dataset, we consider some randomly chosen classes to be the *knowns* and the remaining classes to be the *unknowns* using a mismatch ratio. The mismatch ratio is defined as $\frac{|\mathcal{K}|}{|\mathcal{K}|+|\mathcal{U}|}$, where $|\mathcal{K}|$ is the number of known classes and $|\mathcal{U}|$ is the number of unknown classes. For CIFAR-10, CIFAR-100, and TinyImageNet, we initialize the labeled dataset by randomly sampling 1%, 8%, and 8% of the samples from known classes, respectively. In all of our experiments, we perform 10 cycles of active sampling, and 1500 samples are queried for annotation in each cycle. For fair experimental results, each experiment is conducted four times with varying known/unknown class splits across all the compared methods. The average results from these runs are then reported.

Implementation Details. In all experiments, we train a ResNet18 (He et al. 2016) as our backbone network and one-layer fully-connected networks as binary classifiers. In each AL cycle, we train models for 300 epochs via SGD optimizer (Ruder 2016) with an initial learning rate of 0.01, a momentum of 0.9, and a weight decay of 0.005. The learning rate is decayed by 0.5 every 60 epochs. The batch size is set to 128 for all experiments. We generally set the values of both β and λ to 0.1. We utilize PyTorch (Paszke et al. 2019) to implement our method and an NVIDIA A5000 GPU to run each experiment. We do not use any pre-trained model.

Baselines. We compare our method with the following AL and open-set AL approaches, namely Random, Entropy (Wang and Shang 2014), Certainty (Luo, Schwing, and Urtasun 2013), BALD (Tran et al. 2019), Coreset (Sener and Savarese 2017), BADGE (Ash et al. 2019), MQNet (Park et al. 2022), and LfOSA (Ning et al. 2022), from which LfOSA and MQNet are the SOTA methods for open-set AL.

Main Results

Classification Results. Fig. 3, 4, and 5 show the classification results corresponding to various methods on CIFAR-10, CIFAR-100, and TinyImageNet, respectively. It can be seen that the proposed approach outperforms other baselines nearly across all datasets and mismatch ratios. Our method can effectively identify known samples within unlabeled data by utilizing the proposed entropy scores in the query strategy. As a result, it shows excellent performance in challenging scenarios with high unknown ratios. Specifically, our method outperforms recent open-set AL methods MQNet and LfOSA by margins of 3.88% and 6.00%, respectively, on CIFAR-100 with the mismatch ratio of 20%. As the mismatch ratio increases, we observe a drop in the performance gap between LfOSA and standard AL methods. This is because LfOSA mainly relies on the purity of selected samples, which becomes less effective in high mismatch ratios, where there is an abundance of unlabeled known data. In contrast, our approach maintains a large per-

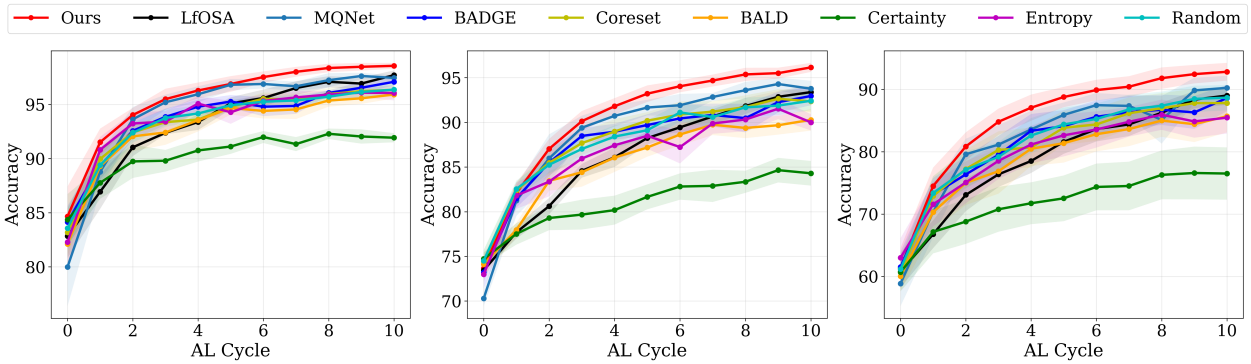

Figure 3: Classification accuracy comparison on CIFAR-10 (mismatch ratios from left to right: 20%, 30%, and 40%).

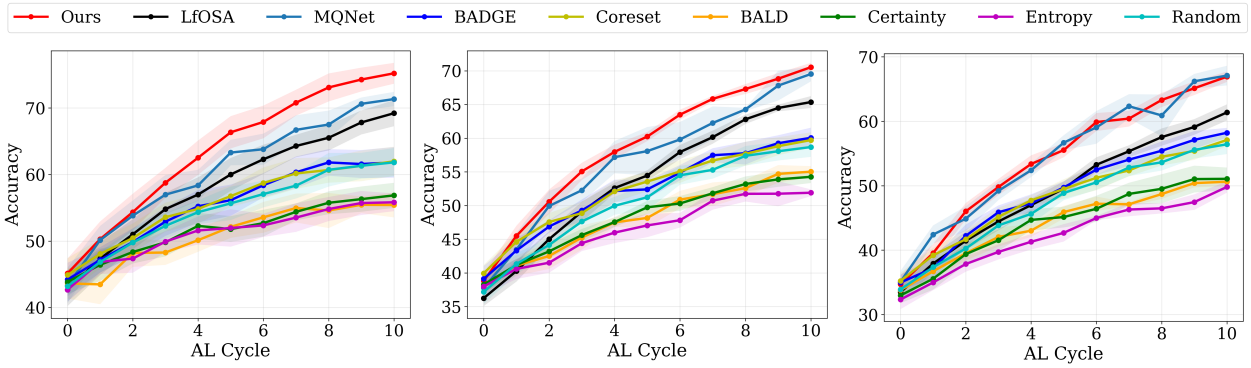

Figure 4: Classification accuracy comparison on CIFAR-100 (mismatch ratios from left to right: 20%, 30%, and 40%).

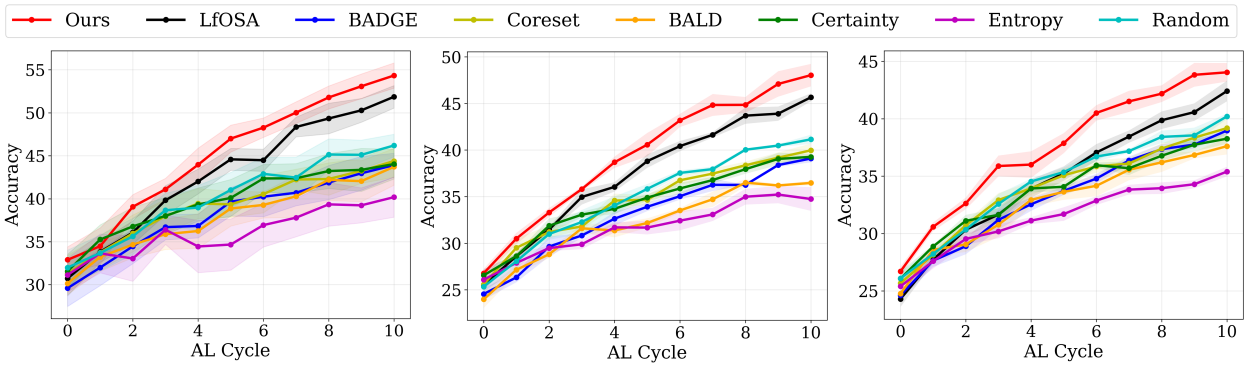

Figure 5: Classification accuracy comparison on TinyImageNet (mismatch ratios from left to right: 20%, 30%, and 40%).

formance margin compared to the standard AL methods by sampling from clusters of unlabeled data to ensure diversity.

Precision Results. The precision of different AL methods in selecting known samples is shown in Fig. 6. As shown in this figure, LfOSA maintains a high precision by focusing on selecting as many known samples as possible which can lead to sampling uninformative samples in low unknown ratio scenarios. Conversely, the precision of MQNet declines rapidly after the first few cycles which does not yield optimal results when the unknown ratio is high due to prioritizing informativeness over purity. However, our approach strikes a balance between these two methods. It maintains high precision across AL cycles by employing two distinct entropy scores and simultaneously selects diverse samples through sampling from clusters. This shows the effectiveness of our method in both high and low unknown ratio settings.

Ablation Study

In this section, we conduct the ablation study on CIFAR-100 with the mismatch ratio of 20% to show the effectiveness of

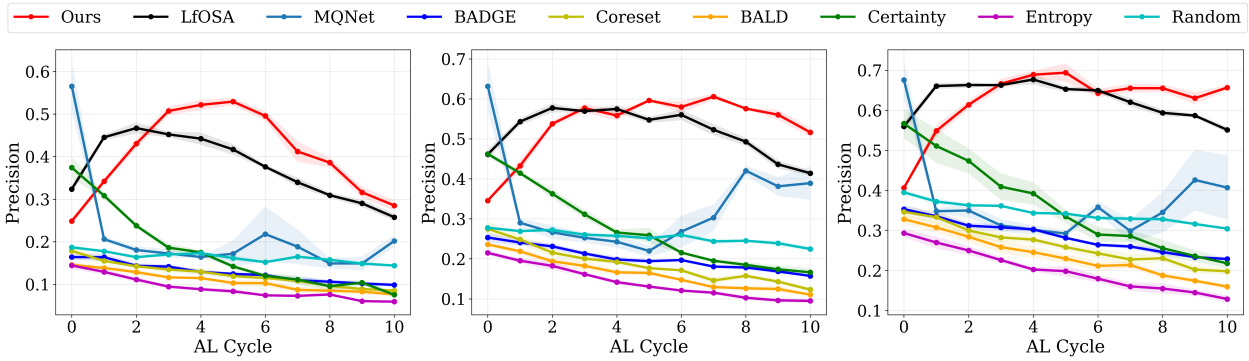

Figure 6: Precision results on CIFAR-100 (mismatch ratios from left to right: 20%, 30%, and 40%).

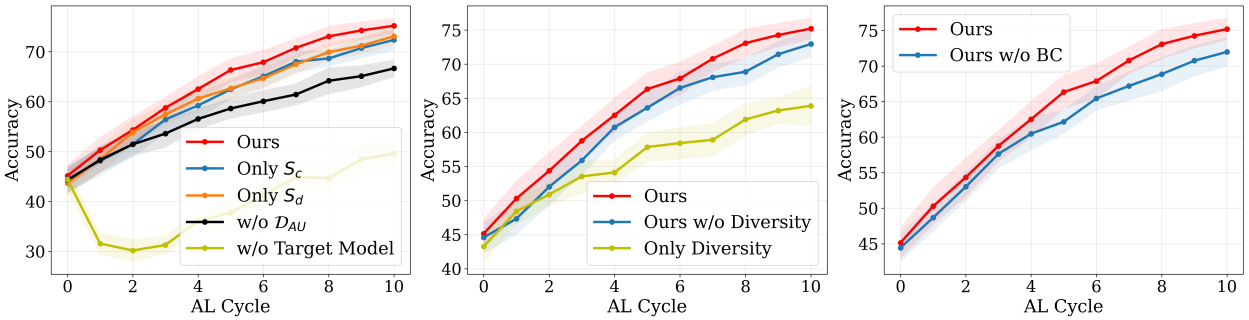

Figure 7: Ablation study on CIFAR-100 with the mismatch ratio of 20%

each component within our framework. Fig. 7 (left) studies the following cases:

Only S_c. It indicates using only the S_c in the query strategy. Accordingly, we do not utilize $\mathcal{L}_t$ in this setting.

Only S_d. It uses only S_d in the query strategy. Accordingly, we do not leverage $\mathcal{L}_{em}$ in this setting. It can be observed that removing each entropy score leads to reduced accuracy performance across all AL cycles. This shows the effectiveness of combining these two entropy scores for the query strategy in our framework.

w/o $\mathcal{D}_{AU}$. It denotes we do not utilize $\mathcal{D}_{AU}$ in any part of our method training. We can observe the importance of training with active unknown samples to achieve satisfactory open-set AL performance.

w/o Target Model. It indicates the utilization of the trained feature extractor F and the classifier C for the final evaluation on the testing data, as opposed to training a separate target model on $\mathcal{D}_L$. The observed performance drop emphasizes the need for a separate target model for evaluations.

In Fig. 7 (middle), we evaluate the importance of diversity in our proposed query strategy as follows:

w/o Diversity. In this experiment, clustering is not utilized in the query strategy. Instead, we choose the samples with the lowest $S(\mathbf{x})$ scores from all unlabeled data globally, rather than from each cluster. We see the performance of the last AL cycle drops by 2.25% compared to our approach.

Only Diversity. In this experiment, we randomly select samples from each cluster, rather than sorting them by the $S(\mathbf{x})$ score. The performance decreases by 11.33% indicating that diversity sampling alone is not effective for selecting informative and valid samples.

In Fig. 7 (right) we study the role of binary classifiers $\{G_i\}_{i=1}^K$ in our framework:

w/o BC. In this experiment, we remove the BC block from our framework. Not using $\{G_i\}_{i=1}^K$ to form the S_c score, we utilize the first K logit outputs of C to calculate the closed-set entropy. The accuracy declines by a margin of 3.21% in this setting which shows the effectiveness of utilizing BC's in our framework.

Conclusion

In this paper, we propose a novel framework for addressing the problem of open-set active learning where we leverage both known and unknown class distribution. Specifically, our approach includes a closed-set entropy score that quantifies the uncertainty of a sample with respect to distributions of known categories and a distance-based entropy that measures uncertainty regarding distributions of unknown categories. By utilizing these entropy scores, we effectively separate the known and unknown samples, and followed by clustering, we select the most informative samples. We conducted extensive experiments on CIFAR-10, CIFAR-100, and TinyImageNet, showing our proposed approach's effectiveness in both high and low open-set noise ratio scenarios.

Acknowledgements

Research was sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-23-2-0008. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

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

Balcan, M.-F.; Broder, A.; and Zhang, T. 2007. Margin based active learning. In *Learning Theory: 20th Annual Conference on Learning Theory, COLT 2007, San Diego, CA, USA; June 13-15, 2007. Proceedings 20*, 35–50. Springer.

Bendale, A.; and Boult, T. E. 2016. Towards open set deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1563–1572.

Chen, G.; Qiao, L.; Shi, Y.; Peng, P.; Li, J.; Huang, T.; Pu, S.; and Tian, Y. 2020. Learning open set network with discriminative reciprocal points. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, 507–522. Springer.

Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2017. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4): 834–848.

Cheng, Z.; Zhang, X.-Y.; and Liu, C.-L. 2023. Unified Classification and Rejection: A One-versus-All Framework. *arXiv preprint arXiv:2311.13355*.

Du, P.; Zhao, S.; Chen, H.; Chai, S.; Chen, H.; and Li, C. 2021. Contrastive coding for active learning under class distribution mismatch. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8927–8936.

Ge, Z.; Demyanov, S.; Chen, Z.; and Garnavi, R. 2017. Generative openmax for multi-class open set classification. *arXiv preprint arXiv:1707.07418*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hino, H.; and Eguchi, S. 2022. Active learning by query by committee with robust divergences. *Information Geometry*, 1–26.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. *arXiv preprint arXiv:2304.02643*.

Kirsch, A.; Van Amersfoort, J.; and Gal, Y. 2019. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. *Advances in neural information processing systems*, 32.

Kothawade, S.; Beck, N.; Killamsetty, K.; and Iyer, R. 2021. Similar: Submodular information measures based active learning in realistic scenarios. *Advances in Neural Information Processing Systems*, 34: 18685–18697.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images. *Technical Report TR-2009, University of Toronto*.

Lu, J.; Xu, Y.; Li, H.; Cheng, Z.; and Niu, Y. 2022. Pmal: Open set recognition via robust prototype mining. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 1872–1880.

Luo, W.; Schwing, A.; and Urtasun, R. 2013. Latent Structured Active Learning. In Burges, C.; Bottou, L.; Welling, M.; Ghahramani, Z.; and Weinberger, K., eds., *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

MacQueen, J.; et al. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, 281–297. Oakland, CA, USA.

Miller, D.; Sunderhauf, N.; Milford, M.; and Dayoub, F. 2021. Class anchor clustering: A loss for distance-based open set recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 3570–3578.

Moon, W.; Park, J.; Seong, H. S.; Cho, C.-H.; and Heo, J.-P. 2022. Difficulty-aware simulator for open set recognition. In *European Conference on Computer Vision*, 365–381. Springer.

Neal, L.; Olson, M.; Fern, X.; Wong, W.-K.; and Li, F. 2018. Open Set Learning with Counterfactual Images. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Nguyen, H. T.; and Smeulders, A. 2004. Active learning using pre-clustering. In *Proceedings of the twenty-first international conference on Machine learning*, 79.

Ning, K.-P.; Zhao, X.; Li, Y.; and Huang, S.-J. 2022. Active learning for open-set annotation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 41–49.

Oza, P.; and Patel, V. M. 2019. C2ae: Class conditioned auto-encoder for open-set recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2307–2316.

Park, D.; Shin, Y.; Bang, J.; Lee, Y.; Song, H.; and Lee, J.-G. 2022. Meta-Query-Net: Resolving Purity-Informativeness Dilemma in Open-set Active Learning. *Advances in Neural Information Processing Systems*, 35: 31416–31429.

Parvaneh, A.; Abbasnejad, E.; Teney, D.; Haffari, G. R.; Van Den Hengel, A.; and Shi, J. Q. 2022. Active learning by feature mixing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12237–12246.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Redmon, J.; Divvala, S.; Girshick, R.; and Farhadi, A. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 779–788.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.

Ruder, S. 2016. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*.

Safaei, B.; Vibashan, V.; de Melo, C. M.; Hu, S.; and Patel, V. M. 2023. Open-Set Automatic Target Recognition. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Saito, K.; and Saenko, K. 2021. Ovanet: One-vs-all network for universal domain adaptation. In *Proceedings of the ieee/cvf international conference on computer vision*, 9000–9009.

Sarfraz, S.; Sharma, V.; and Stiefelhagen, R. 2019. Efficient parameter-free clustering using first neighbor relations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8934–8943.

Scheirer, W. J.; de Rezende Rocha, A.; Sapkota, A.; and Boult, T. E. 2012. Toward open set recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(7): 1757–1772.

Sener, O.; and Savarese, S. 2017. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*.

Seung, H. S.; Opper, M.; and Sompolinsky, H. 1992. Query by committee. In *Proceedings of the fifth annual workshop on Computational learning theory*, 287–294.

Shu, Y.; Shi, Y.; Wang, Y.; Huang, T.; and Tian, Y. 2020. P-odn: Prototype-based open deep network for open set recognition. *Scientific reports*, 10(1): 7146.

Sohn, K. 2016. Improved deep metric learning with multi-class n-pair loss objective. *Advances in neural information processing systems*, 29.

Sun, X.; Yang, Z.; Zhang, C.; Ling, K.-V.; and Peng, G. 2020. Conditional gaussian distribution learning for open set recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13480–13489.

Tran, T.; Do, T.-T.; Reid, I.; and Carneiro, G. 2019. Bayesian generative active deep learning. In *International Conference on Machine Learning*, 6295–6304. PMLR.

VS, V.; Oza, P.; and Patel, V. M. 2023. Towards online domain adaptive object detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 478–488.

Vs, V.; Oza, P.; Sindagi, V. A.; and Patel, V. M. 2022. Mixture of Teacher Experts for Source-Free Domain Adaptive Object Detection. In *2022 IEEE International Conference on Image Processing (ICIP)*, 3606–3610. IEEE.

VS, V.; Yu, N.; Xing, C.; Qin, C.; Gao, M.; Niebles, J. C.; Patel, V. M.; and Xu, R. 2023. Mask-free OVIS: Open-Vocabulary Instance Segmentation without Manual Mask Annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23539–23549.

Wang, D.; and Shang, Y. 2014. A new active labeling method for deep learning. *2014 International Joint Conference on Neural Networks (IJCNN)*, 112–119.

Wei, K.; Iyer, R.; and Bilmes, J. 2015. Submodularity in data subset selection and active learning. In *International conference on machine learning*, 1954–1963. PMLR.

Xu, Z.; Yu, K.; Tresp, V.; Xu, X.; and Wang, J. 2003. Representative sampling for text classification using support vector machines. In *Advances in Information Retrieval: 25th European Conference on IR Research, ECIR 2003, Pisa, Italy, April 14–16, 2003. Proceedings 25*, 393–407. Springer.

Yang, H.-M.; Zhang, X.-Y.; Yin, F.; Yang, Q.; and Liu, C.-L. 2020. Convolutional prototype network for open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5): 2358–2370.

Yao, L.; and Miller, J. 2015. Tiny imagenet classification with convolutional neural networks. *CS 231N*, 2(5): 8.

Yoshihashi, R.; Shao, W.; Kawakami, R.; You, S.; Iida, M.; and Naemura, T. 2019. Classification-reconstruction learning for open-set recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4016–4025.