

# Training Ensembles with Inliers and Outliers for Semi-supervised Active Learning

Vladan Stojnić

Zakaria Laskar

Giorgos Toliás

Visual Recognition Group, Faculty of Electrical Engineering, Czech Technical University in Prague

stojnvla,laskazak,toliageo@fel.cvut.cz

## Abstract

Deep active learning in the presence of outlier examples poses a realistic yet challenging scenario. Acquiring unlabeled data for annotation requires a delicate balance between avoiding outliers to conserve the annotation budget and prioritizing useful inlier examples for effective training. In this work, we present an approach that leverages three highly synergistic components, which are identified as key ingredients: joint classifier training with inliers and outliers, semi-supervised learning through pseudo-labeling, and model ensembling. Our work demonstrates that ensembling significantly enhances the accuracy of pseudo-labeling and improves the quality of data acquisition. By enabling semi-supervision through the joint training process, where outliers are properly handled, we observe a substantial boost in classifier accuracy through the use of all available unlabeled examples. Notably, we reveal that the integration of joint training renders explicit outlier detection unnecessary; a conventional component for acquisition in prior work. The three key components align seamlessly with numerous existing approaches. Through empirical evaluations, we showcase that their combined use leads to a performance increase. Remarkably, despite its simplicity, our proposed approach outperforms all other methods in terms of performance. Code: <https://github.com/vladan-stojnic/active-outliers>

## 1. Introduction

Deep learning achieves considerable results on a variety of tasks but is data-hungry, while data annotation has a high cost and is a tedious process. Using the annotation budget wisely is important, which is the focus of active learning [37]. Given a small labeled and a large unlabeled set, the goal is to acquire a subset of the latter to annotate via a labeling oracle. The acquisition function is responsible for selecting examples that will benefit the classifier training the most compared to the given labeled set. Acquisition, annotation, and model training are three steps that are typi-

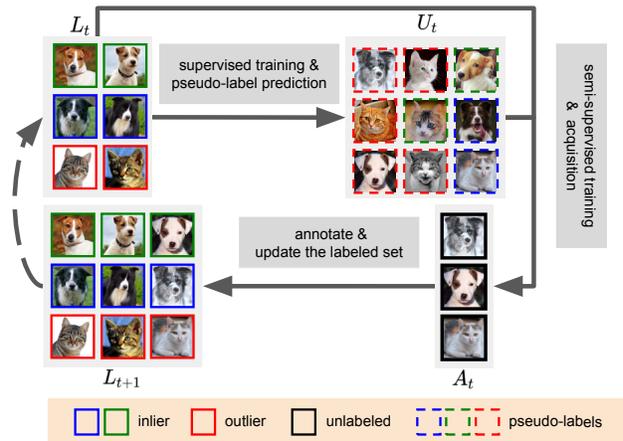


Figure 1. Overview of an active learning round with semi-supervision.  $L_t$ ,  $U_t$ , and  $A_t$  denote labeled, unlabeled, and acquired sets at acquisition round  $t$ , respectively.

cally interleaved over a sequence of consecutive rounds.

Early research is solely focusing on unlabeled sets that are outlier free, *i.e.* all examples come from the categories of interest. Information-theoretic criteria [13, 43] per example are often used to promote the most uncertain predictions for annotation, while other methods process all examples jointly and focus on diversity by pairwise comparisons [24] or large coverage [42]. Surprisingly, random selection achieves good results [32, 44] in deep active learning. This is pronounced even more with the use of semi-supervision, which is deemed more important than the acquisition itself.

However, the outlier-free setup is not realistic. Nevertheless, the presence of outliers is a setup that has attracted less attention. Avoiding outliers is essential so as not to waste the annotation budget. Therefore, existing work [9, 24, 34, 35] explicitly or implicitly performs outlier detection and filtering. Our proposed framework optionally includes outlier filtering, but we question its necessity and investigate the conditions under which it may be useful. Additionally, standard acquisition functions are previously shown to fail [9, 24, 34, 35], but we discover that simple ingredients are missing to make them competitive.

Acquiring outliers together with inliers is unavoidable, especially when their presence is extensive. Therefore, we opt to take advantage of them by training a joint classifier for the inlier classes and the outlier class. The joint classifier obtains increased inlier-class accuracy after each acquisition round and improved ability to perform acquisition. Despite the fact that the joint training enables outlier detection and filtering, we show that, given an appropriate acquisition function, outlier filtering may not be necessary. The performance gap between the use of filtering and no filtering decreases for better acquisition functions. The joint classifier has an additional benefit. It provides pseudo-labels for both types of examples allowing us to exploit semi-supervision and benefit by the unlabeled set which is inherently available according to the active learning task definition. Finally, we rely on the power of ensembles to improve pseudo-labeling and to equip the acquisition with a measure of statistical dispersion. We do not use ensembles during testing, a choice that does not compromise the test-time complexity. An overview of the proposed approach is shown in Figure 1.

In summary, the proposed method consists of several components whose synergy is a key ingredient in surpassing all existing methods, while some of these components are applied for the first time to this specific setup. Our findings are also useful in combination with other approaches and are expected to be useful for future approaches. We perform experiments for a varying amount of outliers on benchmarks created from ImageNet, CIFAR100, and Tiny-ImageNet datasets. The contributions on active learning of this work are summarized as follows:

- We demonstrate the effectiveness of joint training with inliers and outliers, enabling the use of standard acquisition functions that were previously deemed ineffective in the presence of outliers.
- We introduce the use of semi-supervision as a key ingredient for achieving high performance, which is the first application of this technique in the context of outlier-inclusive scenarios.
- Our approach incorporates ensembles during training only, resulting in a significant performance boost without compromising test-time complexity.
- The key components of this work are theoretically compatible with existing approaches; the practical performance benefits of their combination are empirically demonstrated in our experiments.
- We conduct an extensive evaluation across a wide range of outlier percentages, from 0% to 90%, using non-tiny resolution images. Furthermore, we commit to sharing our code and experimental protocol publicly, aiming to enhance consistency in experimental setups across future studies.

## 2. Related work

We review the related work on different setups of active learning, on the related task of semi-supervised learning with outliers, and on outlier detection in an open-world setting.

### 2.1. Active learning

**Acquisition in outlier-free active learning:** The two main families of scoring function are uncertainty-based and diversity-based. Uncertain examples are assumed informative, while uncertainty is measured in different ways, such as entropy [43], confidence [49], margin [11, 38], or mutual information between model parameters and model predictions [13]. Improved uncertainty prediction is obtained through model ensembles [2] or multiple input augmentations [20]. Other definitions of uncertainty use prediction inconsistency over input augmentations [14] or feature perturbations [36]. The second family includes methods that use the diversity of examples in the acquired set. CoreSet [42] and Cluster-Margin [7] select diverse examples that well approximate the whole unlabeled set, while other works [1, 8] combine both notions of diversity and uncertainty. Hacoen *et al.* [17] propose to annotate diverse but certain examples in low-budget regimes but diverse and uncertain examples in the high-budget regime. All these methods are developed for and evaluated in outlier-free setups but are known to fail with outliers [9, 24]. In this work, we show under which conditions such approaches become effective again.

**Unlabeled examples in active learning:** Recent work [14, 30, 32, 44, 46] demonstrates that the way unlabeled examples are used in learning is much more important than the selection process itself. This is the case for semi-supervision [32, 44] and for self-supervision in the pre-training stage [44]. Different selection strategies make little or no difference in these setups, with random selection remaining a good enough choice, whose performance is often under-reported in the fully supervised setup [33]. Improper classifier and hyperparameter tuning lead to unfair method comparisons, requiring proper benchmarking [30, 33]. The aforementioned semi-supervised methods are not directly applicable to our setup due to the presence of outliers, which is one of the issues we handle in our work. Other examples include the popular consistency criterion [3, 48] that is used to perform the acquisition by Gao *et al.* [15], acquisition via classifiers trained to distinguish between labeled and unlabeled examples [23, 45], and synthesizing examples with GANs [10].

**Outliers in active learning** are inherently present in real-world cases. Two recent methods, namely CCAL [9] and SIMILAR [24], propose to counterbalance informativeness and diversity with inlier confidence. CCAL [9] uses self-supervision in an innovative way to improve acquisition

but discards it for classifier training, while we show that it is very beneficial for improving classification accuracy. SIMILAR is the top-performing competitor, whose acquisition is elegant and principled but suffers from low scalability due to the costly optimization of each round. LfOSA [34] filters out unlabelled examples based on outlier-class confidence and then performs the selection based on the maximum confidence. MQNet [35] presents acquisition as a purity-informativeness dilemma, meaning that a good acquisition function should balance purity, *i.e.* the proportion of inlier samples in the acquired set, and informativeness. To construct such a function, they train an MLP on top of standard measures from the literature. Assuming an unknown outlier percentage, these approaches should also work in the outlier-free setup. Nevertheless, we compare and show that this is not always the case. Our method is simpler and scalable, demonstrates synergy with existing scoring functions, and enjoys the benefits of semi-supervision, which is a key ingredient for boosting performance.

## 2.2. Outliers in semi-supervised learning

Similarly to active learning with outliers, the goal is to minimize outlier influence, and all methods rely on different kinds of outlier detectors. MTC [52], D3SL [16], and UASD [6] rely on the Otsu threshold, SSL, and ensembles, respectively, to improve detection. RETRIEVE [22] additionally proposes sub-modular functions to select a subset of good coverage. OpenMatch [40] uses one one-vs-all outlier detector per inlier class to overcome the lack of labeled outliers, which is a major difference from our setup, where an increasing amount of labeled outliers becomes available over consecutive rounds.

## 2.3. Outlier detection

Outlier detection or anomaly detection [21, 41, 47, 51] is a relevant task that aims to solve a binary classification problem of correctly detecting outlier examples. Post-hoc outlier detection methods propose a scoring mechanism on top of an already trained feature backbone [28] to detect outlier examples. Some scoring measures include distance in the feature space [28], pseudo-label confidence [29], or entropy [21]. Other approaches [4, 19, 27, 31, 41, 47] adjust the training to maximize the test-time separability of inliers and outliers. The main challenge is the unavailability of outliers during training. This is based on the assumption that outliers can come from any distribution in an open-set setup, and thus the objective is to learn an unbiased detector. This is not true in our active learning setup, where we have access to both unlabeled and labeled outliers; the latter typically become available after the first acquisition round.

## 3. Method

We define the task of active learning with multiple acquisition rounds and present the proposed approach.

### 3.1. Task formulation

We consider active learning for the classification of object categories  $C$ , with  $K = |C|$ . We consider an additional class, called outlier class  $C_o$ , meant for examples that are not from inlier classes  $C$ . Examples from  $C$  and  $C_o$  are called inliers and outliers, respectively. Initially, we are given a labeled set  $L_0$  and an unlabeled set  $U_0$ , which consist of inliers only, and both inliers and outliers, respectively. Active learning consists of sequential rounds that include the *acquisition* of a subset of the unlabeled set and the *annotation* of the acquired subset. The acquisition should satisfy two objectives that are challenging to balance. Firstly, acquired sets should be as outlier-free as possible. Secondly, the newly acquired and annotated examples should contribute the most to improving the  $K$ -way classifier compared to the current labeled set.

At round  $t$ , the acquisition process makes use of the available examples in  $L_t$  and  $U_t$  to select set  $A_t \subset U_t$  whose labels are assigned by an annotation oracle. Acquired set  $A_t$  may include outliers too, which are labeled by class  $C_o$ . Then, the two sets are updated, *i.e.*  $L_{t+1} = L_t \cup A_t$  and  $U_{t+1} = U_t \setminus A_t$ . The size of  $A_t$  is fixed and equal to the annotation budget per round  $B = |A_t|$ , which is a parameter of the task.

### 3.2. Overview

We start by training a backbone network on all examples in  $L_0 \cup U_0$  by Self-Supervised Learning (SSL), which is known to be beneficial for active learning [44]. At the beginning of round  $t$  for  $t > 0$ ,  $M$  deep network classifiers are initialized by the result of SSL and trained for  $K + 1$  classes with examples in  $L_t$ . The independently trained classifiers are ensembled to perform pseudo-labeling of  $U_t$ , which are used to continue training the  $M$  classifiers in a semi-supervised way on the union of  $L_t$  and  $U_t$ . Then, the ensemble of classifiers, trained with semi-supervision, is used to equip the acquisition process, which optionally includes explicit outlier filtering along with a measure of example uncertainty and/or diversity. Acquired examples are finally annotated by a labeling oracle to obtain  $L_{t+1}$  and  $U_{t+1}$ . Round 0 is a special case that we discuss towards the end of this section.

During test time we do not use any ensembles, which are only used during internal processes to improve pseudo-labeling and acquisition. The test accuracy is evaluated with a single network which is the result of semi-supervised learning. The overall process is summarized in Figure 2, depicting all stages of a single round.

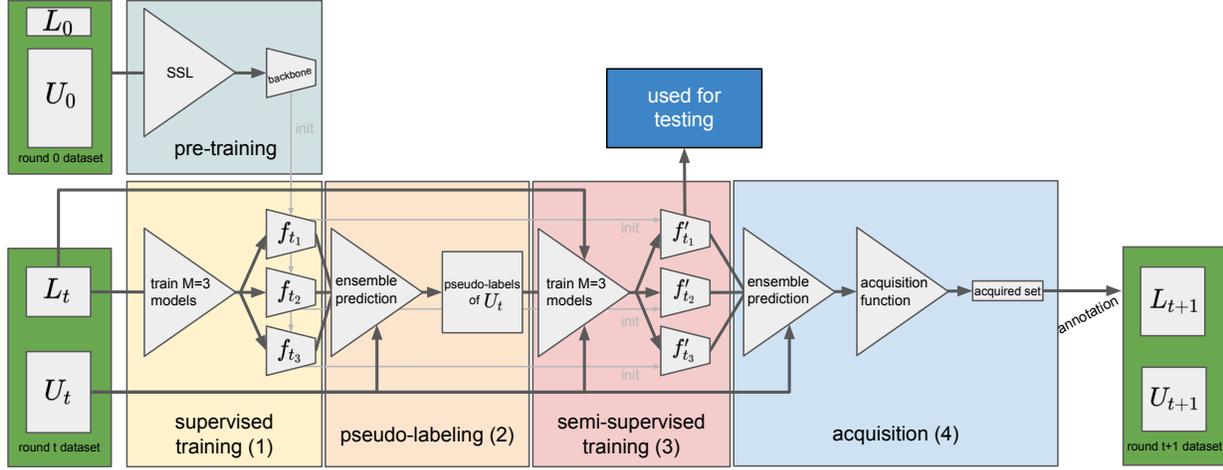


Figure 2. Overview of all active learning stages for the proposed approach during round  $t$ , for  $t > 0$ . It includes independently training  $M$  networks for  $K + 1$ -way classification, acquisition, and annotation. Acquisition exploits the ensemble classifier predictions on the unlabeled set and optionally includes outlier filtering. During round 0, SSL is employed to train the backbone, which is used as initialization for rounds  $t > 0$ , and acquisition with random sampling is performed (not shown in the figure). Testing is performed for  $K$ -way classification on inliers only with a single network; ensembles are only used for internal processes, *i.e.* pseudo-labeling and acquisition.

### 3.3. Training

The network classifier is a function  $f : \mathcal{X} \rightarrow \mathbb{R}^{K+1}$ , where  $\mathcal{X}$  is the space of all examples, and the output space consists of all inlier classes and the outlier one. We consider  $M$  different networks, and the predicted probability distribution for example  $x \in \mathcal{X}$  at round  $t$  by network  $i \in [1, \dots, M]$  is denoted by  $f_{t_i}(x)$ . Network ensembling is performed by averaging the  $M$  output probabilities and is denoted by  $F_t(x)$ , and the probability of the  $j$ -th class is given by  $F_t(x)_j$ . The different networks of the ensemble are trained in the same way, but with a different seed. The seed only affects dataset shuffling and the random augmentations per image. We experimented with the most promising approaches from the work of Wortsman *et al.* [50] to enforce network diversity but did not observe any improvements.

**SSL pre-training:** Before the first round, SSL is performed by instance discrimination, where a positive pair is formed by two different augmentations of the same example, and a negative example is formed by simply picking a different example. This step uses all examples in  $L_0 \cup U_0$  without any labels. SimCLR [5] is the method we choose, following the work of Du *et al.* [9].

**Supervised training:** At the beginning of round  $t$ , each of the  $M$  classifiers is trained by minimizing empirical loss

$$\mathcal{L}(L_t) = \frac{1}{|L_t|} \sum_{x \in L_t} \ell(f_{t_i}(x), y(x)), \quad (1)$$

where  $y(x) \in [1, \dots, K + 1]$  is the label of  $x$ , and  $\ell(\cdot)$  is the cross-entropy loss.

**Semi-supervised training:** We use the unlabeled examples, but only after we first train in a fully supervised way with (1). Then, we generate pseudo-labels  $\hat{y}_t(x) = \arg \max_j F_t(x)_j$  for all examples in  $U_t$ . Each pseudo-label is assigned a weight according to the certainty of the prediction given by

$$w_t(x) = 1 - \frac{H(F_t(x))}{\log(K + 1)}, \quad (2)$$

which is inversely proportional to the normalized entropy and bounded in  $[0, 1]$ , with entropy given by function  $H$ . The role of normalization is to ensure that the range of values is comparable to the weights of labeled examples, which we set equal to 1. We initialize  $M$  classifiers with the result obtained by (1), but use both labeled and unlabeled examples with weighted loss terms given by

$$\mathcal{L}_{\text{semi}}(L_t, U_t) = \frac{1}{N} \sum_{x \in L_t \cup U_t} w_t(x) \ell(f'_{t_i}(x), \hat{y}_t(x)), \quad (3)$$

where  $\hat{y}_t(x) = y(x)$  and  $w_t(x) = 1$  for the labeled examples, and  $N = |L_t| + |U_t|$ . We use  $f'_{t_i}$  and  $F'_t$  for the networks obtained with this semi-supervised way to differentiate from the ones of the previous stage. To evaluate the classification accuracy of round  $t$ , one of the  $M$  networks is randomly picked and used.

**Round 0:** Before any acquisition, at round  $t = 0$ , there are no labeled outliers; therefore, training the  $K + 1$ -way classifier is not possible. We simply perform random acquisition at this stage. In summary, we train the backbone via SSL and then perform random acquisition and annotation that results in  $L_1$  and  $U_1$ .

### 3.4. Acquisition

To acquire examples during round  $t$ , we exploit prediction  $F'_t(x)$  for  $x \in U_t$  to assign acquisition value  $a_t(x)$ . This value is composed of a measure of example uncertainty or diversity, obtained via function  $\tilde{a}_t(x)$ , which is optionally combined with outlier filtering via the ensemble predictions. In particular, the final acquisition value is  $a_t(x) = \tilde{a}_t(x) \mathbb{1}_{\hat{y}_t(x) \neq C_o}$  to include outlier filtering and assign 0 value to examples predicted as outliers, or just  $a_t(x) = \tilde{a}_t(x)$  without filtering. One of the standard choices is to rely on the ensemble of classifiers, measure statistical dispersion among them, and choose examples with large disagreement. In particular, we estimate the Variation-Ratio (VR) [12] given by

$$\tilde{a}_t(x) = 1 - \frac{|\{i : \hat{y}'_{t_i}(x) = \hat{y}'_t(x)\}|}{M}, \quad (4)$$

where  $\hat{y}'_{t_i} = \arg \max_j f'_{t_i}(x)_j$  is the pseudo-label of the  $i$ -th classifier. VR measures the proportion of pseudo-labels from a single classifier that disagree with the pseudo-label from the ensemble classifier. Examples with large disagreement get assigned large scores. At the end, we sort examples in descending order based on  $a_t(x)$  and acquire the first  $B$  examples. Other candidate functions are entropy  $\tilde{a}_t(x) = H(F_t(x))$ , uniform random score generator  $\tilde{a}_t(x) \sim \mathcal{U}_{[0,1]}$ , CoreSet, maximum confidence  $\tilde{a}_t(x) = 1 - \max_j F_t(x)_j$ , and more. In our experiments, we identify cases where outlier filtering is needed or not, and where simple measures become effective despite the presence of outliers.

## 4. Experiments

We discuss training details, datasets, experimental protocol, competing methods, and present the results.

**Datasets and experimental setup:** Most existing methods do not share experimental setup details [9], release source code [34], evaluate on a wide range of outlier ratios [24], or evaluate on non-tiny resolution images [9, 24, 34]. To address this, we use ImageNet ILSVRC2012 [39], TinyImageNet [26], and CIFAR100 [25], to generate benchmarks for active learning with outliers.

The amount of outlier examples in the initial unlabeled set is quantified by the *outlier ratio*, *i.e.* the ratio of the number of outliers over the number of all examples. We consider 25 ImageNet classes that correspond to dog breeds as inlier classes and examples from 700 different classes as outliers. We use the training split to create  $L_0$  and  $U_0$ . In particular,  $L_0$  is generated by 20 randomly chosen examples per inlier class, and  $U_0$  by 500 randomly chosen examples per inlier class and randomly chosen outlier examples so that the outlier ratio is equivalent to 0, 0.05, 0.2, 0.5, 0.8, 0.9 for six different benchmarks. The outlier examples of a particular outlier ratio are a subset of the outlier examples for a larger

outlier ratio. The test set is formed by all examples of inlier classes in the original validation split of ImageNet and contains 1,250 examples in total. In a similar way, we generate benchmarks from TinyImageNet and CIFAR100. We report average classification accuracy over 5 seeds defining different selections for  $L_0$ , but always the same  $U_0$ .

Accuracy reported for round  $t$  is with the classifier trained after  $t$  acquisition rounds, *i.e.* one for round 0 and  $t - 1$  for the follow-up rounds, but the acquisition of round  $t$  is not included. As a reference comparison, we train classifiers with (1) at  $t = 0$  and report the achieved classification accuracy. We report the acquisition *inlier rate*, which is the percentage of inliers among the acquired set per round, defined for round 0 with random acquisition and for all other rounds except the last one, where testing is performed before acquisition. The annotation budget is  $B = 500$  for ImageNet and  $B = 100$  for TinyImageNet and CIFAR100. We use ResNet18 [18] as the backbone. Details regarding CIFAR100 and TinyImageNet benchmarks, detailed reporting of average performance and standard deviation in table format for our main experiments, and additional implementation details can be found in the supplementary material. Additionally, we evaluate our method on the experimental setup followed by MQNet [35] and report results in the supplementary material.

**Baselines and other methods:** The main variant of **our approach** is with  $M = 5$ , includes outlier filtering, and uses VR as a scoring function, unless otherwise stated. We often refer to the numbered steps, as shown in Figure 2, to clarify particular design choices in ablations. The following acquisition functions are used within our approach: **Random** selection, **Entropy**-based selection [43], **VR** [12], and **CoreSet** [42]. We compare with **BADGE** [1] as a recent, well-performing, and representative method developed for the outlier-free setup. Additionally, we compare with **CCAL** [9], **SIMILAR** [24], **LfOSA** [34], and **MQNet** [35] as approaches that perform in the presence of outliers. We observe that SSL-based network initialization is beneficial for all these methods; therefore, we use it to evaluate them. Due to this choice, the reported performance for these methods is noticeably higher than their off-the-shelf application. Note that CCAL, in the original work, uses SSL for the backbones used in the acquisition process but not as classifier initialization. We run CoreSet, BADGE, CCAL, SIMILAR, and MQNet using the provided implementations, after integrating them into our implementation framework. We implement LfOSA by ourselves.

**Comparison with other methods:** We perform extensive experiments on ImageNet, which we consider as the most realistic setup due to the normal-sized images and the larger number of categories in the outlier class. In Figure 3 and Figure 4, we present classification accuracy and inlier rate, respectively, for baselines, SoA approaches, and our

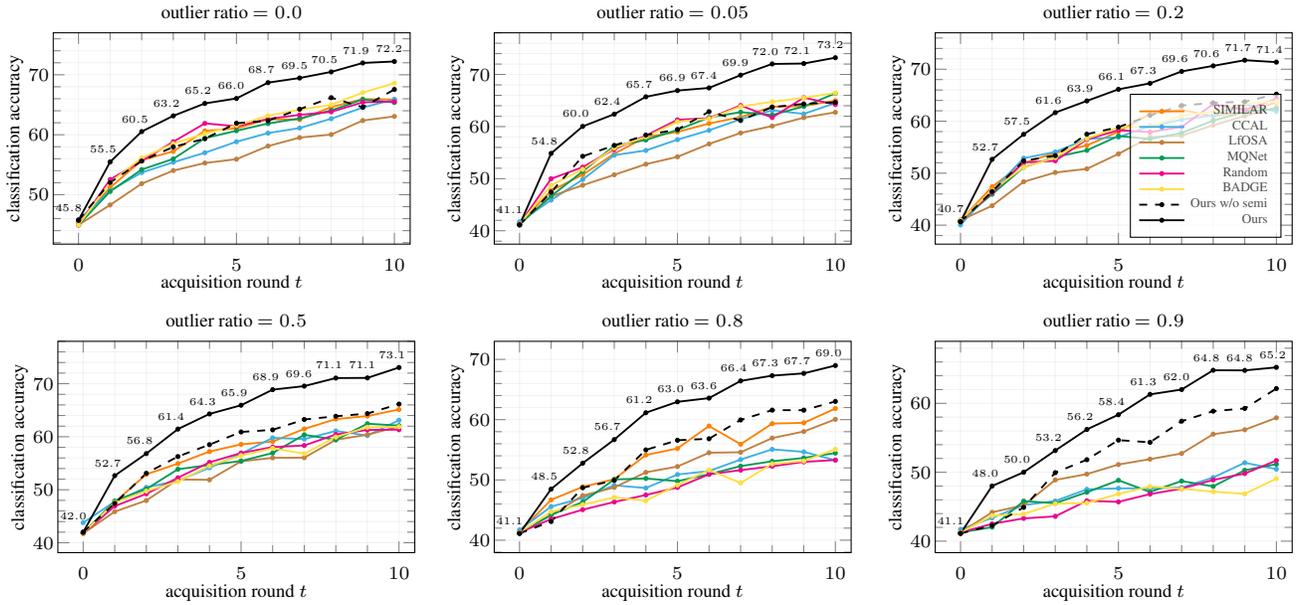


Figure 3. Comparison of classification accuracy over multiple active learning rounds for varying outlier ratios on ImageNet. SIMILAR is excluded for 0.9 outlier ratio since we were not able to run it even on a machine with 800GB of RAM.

method. Our approach achieves the best results among all setups by a large margin. Semi-supervision is a key ingredient, as shown in the work of Simeoni *et al.* [44] too; their case was only the 0.0 outlier ratio. Even without semi-supervision (steps 2 and 3 in Figure 2 are skipped), our

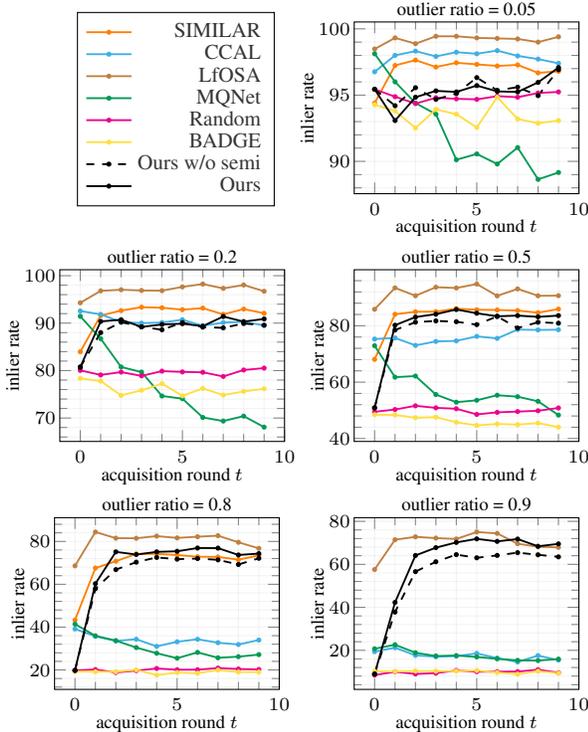


Figure 4. Comparison of inlier rate over multiple active learning rounds for varying outlier ratios on ImageNet. Round 0 is with random selection for our approach and with their specific choice for each method. Reporting round 10 is not included because testing and evaluation are performed before acquisition.

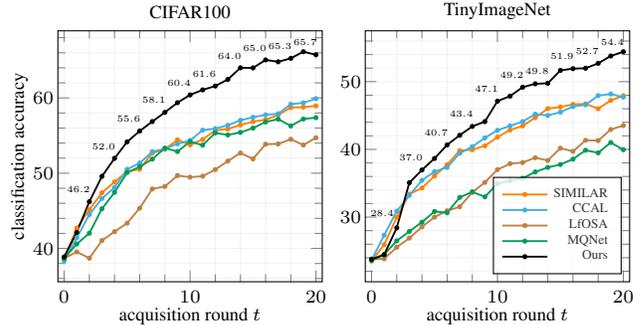


Figure 5. Comparison of classification accuracy over active learning rounds on CIFAR100 and TinyImageNet for 0.8 outlier ratio.

method is either on par or better than others, especially in the presence of more outliers. This happens without achieving the higher inlier rate. Choosing mostly inliers does not mean that the acquired examples are useful for training. For instance, LfOSA achieves the highest inlier rate but performs poorly in the low presence of outliers. As shown in the supplementary material, this is due to the fact that the selected examples are not diverse enough and, therefore, are less useful for training.

Other methods (SIMILAR, CCAL) achieve moderate inlier rates, as ours, but CCAL performs poorly, while SIMILAR is the top-performing competitor. Nevertheless, it is costly to run and not scalable (we could not run it for the 0.9 outlier ratio). MQNet performs well on low outlier ratios but fails in the large presence of outliers. BADGE and Random perform well only without any or with few outliers, as expected. Overall, note that performance differences get larger for more outliers, in the challenging setups, while many methods are well-performing for the low presence of outliers.

We report the same comparison with the top competitors on CIFAR100 and TinyImageNet in Figure 5 to confirm the good performance of the proposed method in two additional datasets. Note that both of these are with tiny resolution images, which forms the most standard setup in prior work for active learning with outliers. CCAL performs on par with SIMILAR, presumably because using two SSL-based networks is a choice tailored to these benchmarks.

**Impact of training with outliers:** First, we validate the importance of outlier filtering for different acquisition functions. To perform that, we perform acquisition with and without outlier filtering, *i.e.* setting predicted outliers to zero score or not. Results are presented in Figure 6 (left). Simple scoring functions such as Random and Entropy benefit a lot from outlier filtering, which increases their performance and inlier rate by a large margin. This is much less the case for CoreSet, and nearly not at all for VR. Moreover, with filtering, even simple Entropy is top performing.

We go one step further and investigate performance without filtering for the standard choice of jointly training a  $K + 1$ -way classifier and compare it with the case of using a  $K$ -way classifier for the acquisition part. That means that during step 3 in semi-supervised learning, we only train with labeled and pseudo-labeled inliers. Results are presented in Figure 6 on the right. It turns out that there is a large performance drop if outliers are not used in the training; the joint training makes outlier filtering less necessary.

To take a closer look, we present an analysis of the VR values in Figure 7. Without filtering, VR with  $K$ -way acquires a large number of outliers (left), while it acquires noticeably fewer outliers with the joint  $K + 1$ -way training despite no filtering (middle). We explain this by the fact that inlier examples from different classes are more similar to each other than to outliers examples. Therefore, inlier examples more often result in disagreements. This may be seen as an outcome of the chosen inlier classes, which nevertheless follows prior work [9] and imitates a realistic scenario. The outlier filtering (right) does not significantly

improve acquisition because it operates better on the low VR regime, where more outliers than inliers are rejected. In the high VR regime, where acquired examples belong, an equal amount of inliers is filtered out too.

**Impact of ensembles:** In Figure 8, we show the impact of the ensemble size on classification accuracy. Using a network ensemble has a significant impact on performance already from  $M = 3$ . The increased pseudo-label accuracy is the main source of the improvement. Training 20 networks have good benefits in early rounds. Nevertheless, we opt for  $M = 5$  as the standard setup to lower training complexity and experimentation time. Detailed time reporting and comparison is included in the supplementary material.

In case we opt for using the ensemble predictor during inference too, the achieved performance on ImageNet for outlier ratio = 0.5,  $M = 5$ , at round 10 is equal to 75.81%.

**Our components improve other methods:** We investigate whether our key components, *i.e.* joint training, ensembles, and semi-supervision, are beneficial to other methods too. In particular, we implement the combination for CCAL, SIMILAR, and CoreSet. We add an outlier class to the classifier for CoreSet and CCAL, while SIMILAR already includes it. This allows us to perform pseudo-labeling and semi-supervision. For CoreSet and SIMILAR to benefit from model ensembling, we use average features and similarity matrices, respectively. CCAL performs acquisition based on a fixed network obtained during pre-training; therefore, semi-supervision or ensembling do not affect its acquisition phase. Results in Figure 9 show that all methods benefit from our key ingredients. However, compared to much more complicated methods, our simple and intuitive approach is top-performing. Note that the top inlier rate does not necessarily result in top performance. Additionally, we see that classical active learning methods (CoreSet) become competitive again.

**Impact of semi-supervision on the acquisition:** In Figure 10 (left), we present the impact of performing acquisition before or after (default) semi-supervision. In the case

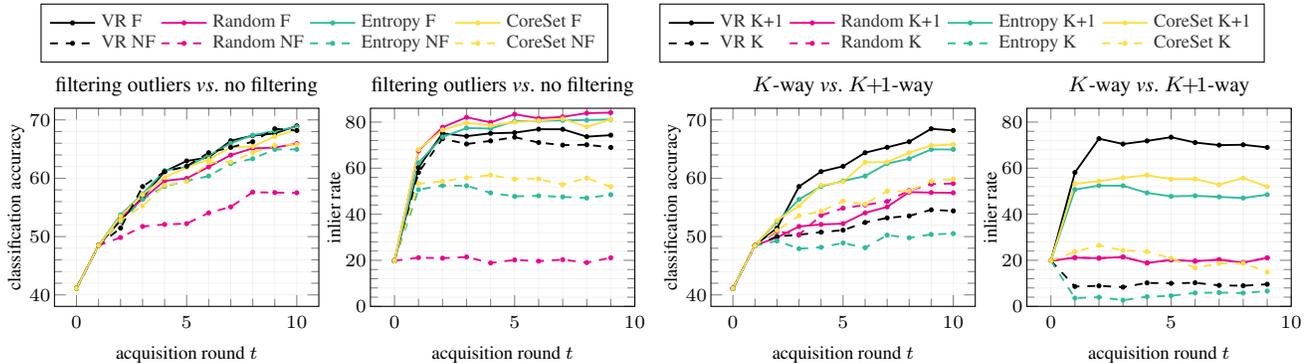


Figure 6. Performance and inlier rate comparison for the proposed approach with different scoring functions for two different experiments. Left: with or without outlier filtering. Right: with a  $K$ -way or a  $K + 1$ -way classifier during semi-supervised learning (step 3), without outlier filtering in both cases. NF: no outlier filtering. F: outlier filtering is used. Experiments on ImageNet with 0.8 outlier ratio.

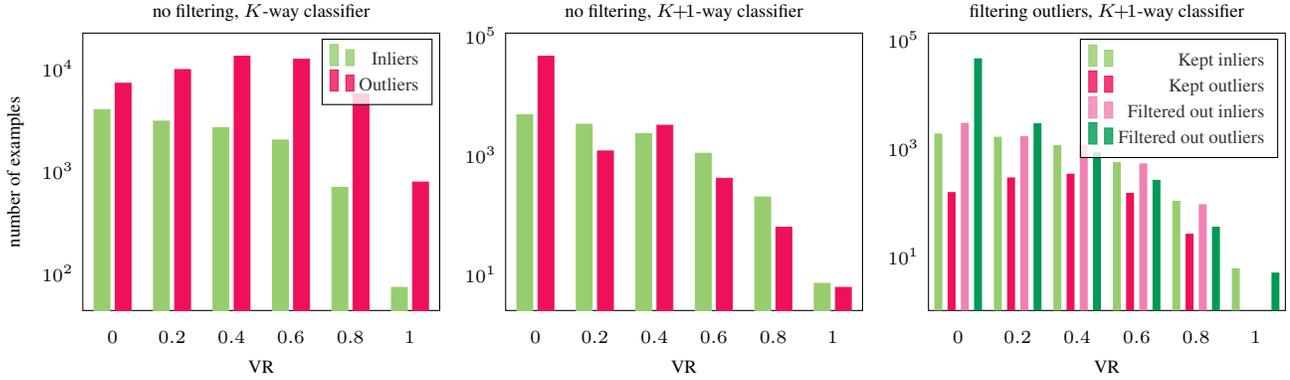


Figure 7. Histograms of VR values for unlabeled examples with our method during round 3 and an outlier ratio of 0.8 on ImageNet. Histograms are created separately for inliers and outliers, and for filtered out or kept inliers/outliers in the case outlier filtering is used.

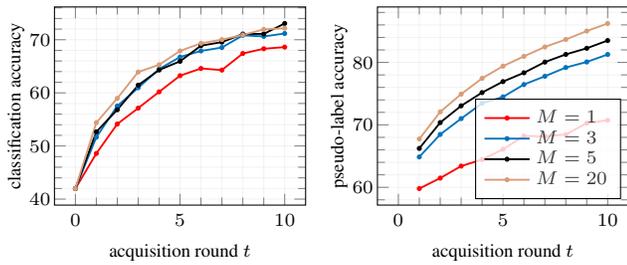


Figure 8. Classification accuracy (left) and accuracy of the predicted pseudo-labels (right) for varying ensemble size on ImageNet for 0.5 outlier ratio. The scoring function is VR for all cases with  $M > 1$  and random for  $M = 1$ , where VR is not defined.

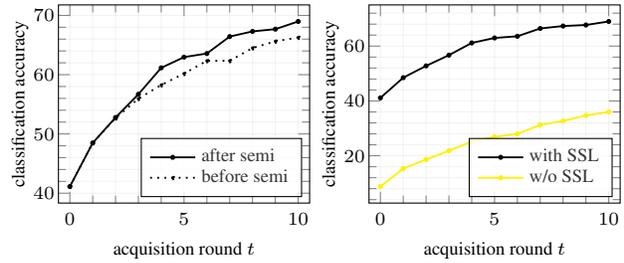


Figure 10. Left: classification accuracy when the acquisition is performed before and after (default) training with semi-supervision. Right: impact of self-supervised pre-training on classification performance. Experiments on ImageNet with 0.8 outlier ratio.

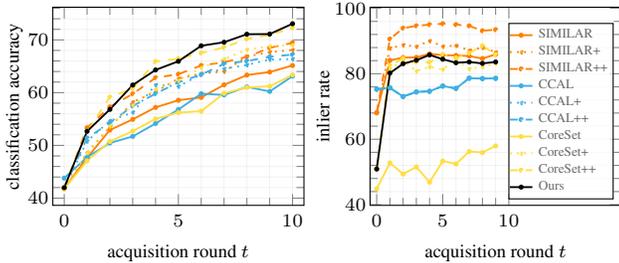


Figure 9. Impact of adding our key ingredients to other methods. Performance and inlier rate comparison when joint training and semi-supervision (denoted by +) and ensembles (all three ingredients together, denoted by ++)

are combined with other methods. Experiments on ImageNet with 0.5 outlier ratio. of performing acquisition before semi-supervision, we still use semi-supervision to train the network used for evaluation. Results show that semi-supervision improves acquisition, although the difference is small in the early rounds.

**Impact of self-supervised pre-training:** We present the impact of self-supervised pre-training on classification accuracy in Figure 10 (right). Results show that it provides a significant benefit of 30% in round 0. This performance gap is retained across all active learning rounds too. This result confirms the findings of earlier work [44].

## 5. Conclusions

We improve the state-of-the-art performance on active learning with outliers by a large margin. This is achieved by three ingredients: joint training with inliers and outliers, semi-supervision via pseudo-labeling, and network ensembles that are used in a way not to increase the test-time complexity. Some of our findings are shown to be compatible with existing acquisition functions, and their applicability goes beyond existing approaches due to the universality of the proposed framework. Simple acquisition functions that were thought to fail in this setup are able to reach state-of-the-art performance within our framework. We will publicly release the source code and datasets of our extensive evaluation for reproducibility and to improve the setup discrepancy from which the current literature suffers.

## Acknowledgement

This work was supported by Junior Star GACR GM 21-28830M, by Czech Technical University in Prague grant No. SGS23/173/OHK3/3T/13, and by Programme Johannes Amos Comenius CZ.02.01.01/00/22\_010/0003405.

## References

- [1] Jordan T. Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. In *ICLR*, 2020. 2, 5
- [2] William H Beluch, Tim Genewein, Andreas Nürnberger, and Jan M Köhler. The power of ensembles for active learning in image classification. In *CVPR*, pages 9368–9377, 2018. 2
- [3] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *NeurIPS*, 2019. 2
- [4] Petra Bevandić, Ivan Krešo, Marin Oršić, and Siniša Šegvić. Discriminative out-of-distribution detection for semantic segmentation. *arXiv preprint arXiv:1808.07703*, 2018. 3
- [5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 4
- [6] Yanbei Chen, Xiatian Zhu, Wei Li, and Shaogang Gong. Semi-supervised learning under class distribution mismatch. In *AAAI*, 2020. 3
- [7] Gui Citovsky, Giulia DeSalvo, Claudio Gentile, Lazaros Karydas, Anand Rajagopalan, Afshin Rostamizadeh, and Sanjiv Kumar. Batch active learning at scale. 2021. 2
- [8] Gui Citovsky, Giulia DeSalvo, Claudio Gentile, Lazaros Karydas, Anand Rajagopalan, Afshin Rostamizadeh, and Sanjiv Kumar. Batch active learning at scale. In *NeurIPS*, 2021. 2
- [9] Pan Du, Suyun Zhao, Hui Chen, Shuwen Chai, Hong Chen, and Cuiping Li. Contrastive coding for active learning under class distribution mismatch. In *ICCV*, 2021. 1, 2, 4, 5, 7
- [10] Sayna Ebrahimi, Will Gan, Kamyar Salahi, and Trevor Darrell. Minimax active learning. In *arXiv*, 2020. 2
- [11] Zeyad Ali Sami Emam, Hong-Min Chu, Ping-Yeh Chiang, Wojciech Czaja, Richard Leapman, Micah Goldblum, and Tom Goldstein. Active learning at the imagenet scale. In *arXiv*, 2021. 2
- [12] Linton C Freeman. *Elementary applied statistics: for students in behavioral science*. 1965. 5
- [13] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *ICML*, 2017. 1, 2
- [14] Mingfei Gao, Zizhao Zhang, Guo Yu, Sercan Ö Arık, Larry S Davis, and Tomas Pfister. Consistency-based semi-supervised active learning: Towards minimizing labeling cost. In *ECCV*, 2020. 2
- [15] Mingfei Gao, Zizhao Zhang, Guo Yu, Sercan Ö Arık, Larry S Davis, and Tomas Pfister. Consistency-based semi-supervised active learning: Towards minimizing labeling cost. In *ECCV*, 2020. 2
- [16] Lan-Zhe Guo, Zhen-Yu Zhang, Yuan Jiang, Yu-Feng Li, and Zhi-Hua Zhou. Safe deep semi-supervised learning for unseen-class unlabeled data. In *ICML*, 2020. 3
- [17] Guy Hacohen, Avihu Dekel, and Daphna Weinshall. Active learning on a budget: Opposite strategies suit high and low budgets. *arXiv preprint arXiv:2202.02794*, 2022. 2
- [18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5
- [19] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *ICLR*, 2018. 3
- [20] SeulGi Hong, Heonjin Ha, Junmo Kim, and Min-Kook Choi. Deep active learning with augmentation-based consistency estimation. In *arXiv*, 2020. 2
- [21] Rui Huang, Andrew Geng, and Yixuan Li. On the importance of gradients for detecting distributional shifts in the wild. In *NIPS*, 2021. 3
- [22] Krishnateja Killamsetty, Xujiang Zhao, Feng Chen, and Rishabh Iyer. Retrieve: Coreset selection for efficient and robust semi-supervised learning. *NIPS*, 2021. 3
- [23] Kwanyoung Kim, Dongwon Park, Kwang In Kim, and Se Young Chun. Task-aware variational adversarial active learning. In *CVPR*, 2021. 2
- [24] Suraj Kothawade, Nathan Beck, Krishnateja Killamsetty, and Rishabh Iyer. Similar: Submodular information measures based active learning in realistic scenarios. In *NeurIPS*, 2021. 1, 2, 5
- [25] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. 5
- [26] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. Technical report, Stanford University, 2015. 5
- [27] Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. In *ICLR*, 2017. 3
- [28] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *NIPS*, 2018. 3
- [29] Shiyu Liang, Yixuan Li, and R Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *ICLR*, 2018. 3
- [30] Carsten T Lüth, Till J Bungert, Lukas Klein, and Paul F Jaeger. Toward realistic evaluation of deep active learning algorithms in image classification. *arXiv preprint arXiv:2301.10625*, 2023. 2
- [31] Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. In *NIPS*, 2018. 3
- [32] Sudhanshu Mittal, Maxim Tatarchenko, Özgün Çiçek, and Thomas Brox. Parting with illusions about deep active learning. In *arXiv*, 2019. 1, 2
- [33] Prateek Munjal, Nasir Hayat, Munawar Hayat, Jamshid Sourati, and Shadab Khan. Towards robust and reproducible active learning using neural networks. In *CVPR*, 2022. 2
- [34] Kun-Peng Ning, Xun Zhao, Yu Li, and Sheng-Jun Huang. Active learning for open-set annotation. In *CVPR*, 2022. 1, 3, 5
- [35] Dongmin Park, Yooju Shin, Jihwan Bang, Youngjun Lee, Hwanjun Song, and Jae-Gil Lee. Meta-query-net: Resolving purity-informativeness dilemma in open-set active learning. In *NeurIPS*, 2022. 1, 3, 5
- [36] Amin Parvaneh, Ehsan Abbasnejad, Damien Teney, Gholamreza Reza Haffari, Anton Van Den Hengel, and Javen Qin-

- feng Shi. Active learning by feature mixing. In *CVPR*, 2022. 2
- [37] Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zihui Li, Brij B Gupta, Xiaojiang Chen, and Xin Wang. A survey of deep active learning. *ACM Computing Surveys*, 54(9):1–40, 2021. 1
- [38] Dan Roth and Kevin Small. Margin-based active learning for structured output spaces. In *Machine Learning: ECML*, 2006. 2
- [39] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 2015. 5
- [40] Kuniaki Saito, Donghyun Kim, and Kate Saenko. Open-match: Open-set semi-supervised learning with open-set consistency regularization. *NIPS*, 2021. 3
- [41] Vikash Sehwal, Mung Chiang, and Prateek Mittal. Ssd: A unified framework for self-supervised outlier detection. In *ICLR*, 2021. 3
- [42] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *ICLR*, 2018. 1, 2, 5
- [43] Burr Settles. Active learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences, 2009. 1, 2, 5
- [44] Oriane Siméoni, Mateusz Budnik, Yannis Avrithis, and Guillaume Gravier. Rethinking deep active learning: Using unlabeled data at model training. In *ICPR*, 2020. 1, 2, 3, 6, 8
- [45] Samarth Sinha, Sayna Ebrahimi, and Trevor Darrell. Variational adversarial active learning. In *ICCV*, 2019. 2
- [46] Shuang Song, David Berthelot, and Afshin Rostamizadeh. Combining mixmatch and active learning for better accuracy with fewer labels. *arXiv preprint arXiv:1912.00594*, 2019. 2
- [47] Jihoon Tack, Sangwoo Mo, Jongheon Jeong, and Jinwoo Shin. Csi: Novelty detection via contrastive learning on distributionally shifted instances. In *NIPS*, 2020. 3
- [48] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIPS*, 2017. 2
- [49] Dan Wang and Yi Shang. A new active labeling method for deep learning. 2014. 2
- [50] Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, Rebecca Roelofs, Raphael Gontijo Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *ICML*, 2022. 4
- [51] Jingkan Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *arXiv preprint arXiv:2110.11334*, 2021. 3
- [52] Qing Yu, Daiki Ikami, Go Irie, and Kiyoharu Aizawa. Multi-task curriculum framework for open-set semi-supervised learning. In *ECCV*, 2020. 3