Assessing and enhancing robustness of active learning strategies to spurious bias

Anonymous authors
Paper under double-blind review

Abstract

In the presence of spurious correlation, the deep neural network (DNN) trained using empirical risk minimization (ERM) tends to rely on spurious features during predictions, particularly when the target label exhibits spurious correlations with certain attributes in the training set. Prior works have proposed methods to mitigate bias caused by spurious correlations in passive learning scenarios. In this work, we focus on investigating the performance of common active learning (AL) algorithms under spurious bias and designing an AL algorithm that is robust to spurious bias. AL is a framework that iteratively acquires new samples to progressively improve the classifier. In AL loops, sample acquisition is directed by the informativeness criteria, such as uncertainty and representativeness. The concept behind these criteria shares similarities with approaches to addressing spurious correlations in passive settings (i.e., underrepresented samples are deemed informative and thus given higher value during training). In fact, the previous research has demonstrated the potential of AL in addressing out-of-distribution problems. Hence, with an appropriately defined acquisition function, a sample-efficient framework can be established to effectively handle spurious correlations. Inspired by recent works on simplicity bias, we propose **D**omain-**I**nvariant **A**ctive **L**earning (DIAL) which leverages the disparity in training dynamics between overrepresented and underrepresented samples, selecting samples that exhibit "slow" training dynamics. DIAL involves no excessively resource-intensive computations beyond the standard training process and feedforward inference, making it more scalable for addressing real-world problems with AL. Empirical results demonstrates that DIAL not only outperforms baselines in achieving robustness performance under the spurious correlation scenarios but also on the standard ML datasets.

1 Introduction

Deep neural networks (DNNs) commonly demand a substantial number of training samples for the optimization of their parameters. The necessary size of the training dataset to prevent overfitting tends to grow proportionally with the model's size. Annotation of samples is often a resource-intensive and time-consuming process. Active learning (AL) emerges as a specialized approach designed to tackle this challenge. In contrast to supervised learning, AL algorithms dynamically recommend informative samples for annotation by humans. The objective of AL is to achieve optimal performance with least number of annotated samples.

Spurious correlations refer to the scenario where certain (potentially simpler) task-irrelevant attributes in the training set are highly correlated with the target labels. Consider the scenario where a DNN is trained to distinguish between images containing cars and bicycles. In the training dataset, an unintended sampling bias might emerge, leading to a situation where the majority of car images happen to be predominantly of a particular color, say blue, while the majority of bicycle images tend to have a different color, like red. This sampling bias inadvertently introduces a spurious correlation between the object category and the color attribute. As a result, the trained DNN may mistakenly learn to associate the presence of a certain color with a particular object class, leading to erroneous predictions when faced with images featuring cars or bicycles of different colours. For example, during deployment, the model might classify any blue object as a bicycle and any red object as a car, regardless of other features. This issue can have adverse effects in real-world

applications – relying on these false associations can result in flawed predictions, inaccurate analyses, and misguided actions, particularly in critical domains such as healthcare (Oakden-Rayner et al., 2020) and social sciences (Angwin et al., 2016).

One common category of existing approaches for learning robust DNNs (Sagawa et al., 2020; Liu et al., 2021; Sohoni et al., 2020; Nam et al., 2020) rely on identification of underrepresented samples, often achieved through explicit labels or estimation (more detailed discussions are provided in Section 2). This process resembles the acquisition step in the AL loop where the informativeness is based on the representativeness of samples in the labelled pool. The key distinctions between standard AL strategies and existing approaches to spurious correlation are as follows: (1) the former assesses informativeness solely based on sample features, while the latter evaluates not only based on features but also considers label information (in our illustrative scenario, class label – cars and bicycles – and/or attribute label – red and blue); (2) after evaluation, the former appends newly labelled samples to existing labelled dataset while the latter assigns importance weightings for subsets of samples.

Those distinctions, though, are blur. The phase when passive approaches to spurious correlation identify underrepresented samples essentially corresponds to a single-step active acquisition but without explicit labels. Furthermore, the act of appending samples to an existing dataset is equivalent to increasing the importance of specific populations in the dataset (resampling versus reweighing).

Tamkin et al. (2022) demonstrated that uncertainty-based AL methods can inherently address various out-of-distribution problems. Their research showed that AL methods improve both overall performance and the performance of underrepresented subgroups compared to random sampling. This improvement arises from the increased of labelled disambiguating samples – those whose spurious attributes do not align with their labels in the labelled pool – from the unlabelled pool by AL methods. However, uncertainty-based acquisition can sometimes miss these disambiguating samples if their predictions are overly confident due to reliance on spurious features, resulting in low uncertainty and causing these samples to be neglected in the AL loop. Consequently, the robustness improvement is limited. Therefore, this work aims to design a deep learning AL acquisition method that can accurately identify underrepresented/disambiguating samples.

Our motivation stems from recent research on *simplicity bias*, where DNNs trained with stochastic gradient descent (SGD) tend to prioritize learning simple features over complex ones (Shah et al., 2020; Nakkiran et al., 2019). Consequently, the DNN will exhibit invariance to complex features, leading to potential detrimental effects on generalization, especially when the simple features are completely irrelevant to the task. The lack of robustness under spurious correlation arises from the bias towards much simpler (and thus, easier-to-learn) features over task-intrinsic ones (Teney et al., 2022; Vasudeva et al., 2023; Bell & Sagun, 2023). In essence, such classifiers adopt simplistic decision rules heavily influenced by the statistical bias present in the training samples, e.g., relying on the color instead of other features. Several works in passive settings have leveraged this phenomenon to design methods that demonstrate strong effectiveness Liu et al. (2021); Murali et al. (2023); Yang et al. (2024). Moreover, from the perspective of training loss profiles, Nam et al. (2020) showed a discrepancy in terms of training dynamics, where the loss of samples aligned with the bias converges faster than those unaligned.

Inspired by this observation, we introduce **D**omain-**I**nvariant **A**ctive **L**earning (DIAL), which utilizes training dynamics as a proxy to assess the informativeness of unlabelled samples. Our experiments demonstrate that DIAL effectively selects underrepresented samples from the labelled pool, thereby enhancing the robustness of the classifier. Furthermore, although the concept of training dynamics is often associated with gradients, DIAL does not require gradient operations but instead relies solely on feedforward inference during sample acquisition, making it efficient in the context of deep learning.

Our contributions in this work can be summarized as follows:

- We show the potential of addressing spurious correlations in the sample-efficient manner.
- Drawing inspiration from simplicity bias, we proposed DIAL, a deep AL strategy designed to mitigate spurious bias. DIAL is easy to implement and highly compatible in deep learning.
- We perform extensive experiments on real-world datasets, demonstrating that our proposed method, DIAL, achieves better robustness than existing AL methods.

2 Related work

Active learning. The primary goal of AL algorithms is to query labels for the most informative samples. These criteria for informativeness are broadly categorized into two groups (Settles, 2012): *uncertainty-based* and *representation-based* sampling. One classic notion of uncertainty-based sampling involves assessing the distance of the sample to the decision boundary, including the margin (Roth & Small, 2006), the entropy of softmax outputs (Wang & Shang, 2014), and the maximum class predictive score (Wang & Shang, 2014). In the context of probabilistic formulation, another notion is derived from the information-theoretic perspective where Roy & McCallum (2001) uses the expected information gain, and Houlsby et al. (2011) measures the mutual information between the model posterior and the samples. Gal et al. (2017) incorporates Monte Carlo dropout Gal & Ghahramani (2016) to overcome the mutual information of (Houlsby et al., 2011) for DNNs and Kirsch et al. (2019) further extends the approach to consider the dependencies between samples in the batch. Representation-based strategies, on the other hand, adopt the notion of representativeness as the measure for informativeness. Settles et al. (2007) evaluates representativeness in the gradient space which selects samples with maximum norm of gradient in order to maximize the change of model's parameter. Sener & Savarese (2018) and Geifman & El-Yaniv (2017) label samples that exhibit diversity w.r.t. the current labelled pool with the goal of constructing a proxy for the complete dataset. Similarly, Ash et al. (2020) seeks to promote diversity using gradient embeddings. A closely related study to our proposed method is presented in Jung et al. (2022), where they proposed utilizing snapshot-ensemble (Huang et al., 2016) instead of commonly used methods such as deep ensemble (Lakshminarayanan et al., 2017) during the acquisition phase. However, in Jung et al. (2022), a large learning rate or a learning rate scheduler is necessary to ensure that each snapshot converges to local minimals. In contrast, our proposed method employs the standard training procedure (i.e., a typically small constant learning rate) to capture the dynamics of the learning.

Active learning with neural tangent kernel. Recently, there has been a line of works grounded in the neural tangent kernel (NTK) framework. Mohamadi et al. (2022) introduced a look-ahead strategy, utilizing the NTK to approximate the DNN if the label of an unlabelled sample were acquired. Subsequently, it selects samples that maximize the expected output w.r.t. the current trained DNN. In a related study, Wang et al. (2022) explores a different approach, selecting samples based on the criterion of maximizing the change in training dynamics. Furthermore, Kong et al. (2022) uses the eigenvalue of the Gram matrix generated by the NTK to determine the diversity of queries.

Active learning for fairness Conventional AL aims to enhance model performance by selectively querying the most informative samples from the unlabelled pool. However, this often overlooks the fairness aspect, potentially perpetuating or even exacerbating biases present in the data. To address this, several works have proposed fairness-aware AL strategies. Similarly, Anahideh et al. (2022); Sharaf et al. (2022); Fajri et al. (2024) incorporate fairness constraints directly into the sample selection process, aiming to achieve trade-off between fairness and model performance. However, these approaches necessitate explicit access to bias attributes in the labelled or even unlabelled pool. This requirement can limit their practicality in real-world scenarios where access to bias attributes might be restricted due to privacy regulations. In this work, we aim to eliminate this requirement, aligning with the current research trends in passive settings.

Subgroup robustness. Numerous methods are available to tackle bias caused by spurious correlations in the dataset. These methods share a common objective: to mitigate bias by improving the *worst-group performance* so that the disparities between overall performances is minimized. Examples include subgroup robust approaches (Sagawa et al., 2020; Nam et al., 2020; Liu et al., 2021; Yao et al., 2022; Sohoni et al., 2020) and representation learning approaches Arjovsky et al. (2020); Zemel et al. (2013). Some of these methods require explicit spurious attribute information (e.g., (Sagawa et al., 2020; Yao et al., 2022)), thereby limiting their use cases. In contrast, our proposed algorithm requires no additional information beyond the class label. Specifically, upon the AL algorithm's selection of samples, the expert is only required to annotate them with class labels.

3 Background

Notation. We consider classification problems. Given a set of samples $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n) \mid x \in \mathcal{X}, y \in \mathcal{Y}\}$ and a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, the task is to learn a DNN $f_\theta : \mathcal{X} \to \mathcal{Y}$ parameterized by θ via empirical risk minimization (ERM):

$$\theta^* = \arg\min_{\theta \in \Theta} \sum_{i=1}^{n} \ell\left(f_\theta(x_i), y_i\right) \tag{1}$$

We omit the symbol θ when the context is clear. Furthermore, we express the DNN obtained at time t as f_{θ_t}.

Active learning. In pool-based AL scenario, starting with a labelled pool $\mathcal{D}_L$ the algorithm sequentially queries an oracle for annotations of some unlabelled samples from $\mathcal{D}_U = \{x_1, \ldots, x_m\}$. We overload the notation by using $f_{\mathcal{D}_L}$ to denote the DNN obtained on the labelled pool $\mathcal{D}_L$. Formally, at each AL iteration, the active learner chooses some samples from the unlabelled pool $\mathcal{Q} \in \mathcal{D}_U$ (typically conditioned on $f_{\mathcal{D}_L}$). Following this selection, the oracle (i.e., human annotator) assigns labels for $\mathcal{Q}$. Subsequently, the pool gets an update: $\mathcal{D}_L \leftarrow \mathcal{D}_L \cup Q$ and $\mathcal{D}_U \leftarrow \mathcal{D}_U \setminus \mathcal{Q}$. This process continues until a stopping criterion is met, e.g., when a labelling budget is exhausted.

Spurious bias. Typically, one would anticipate learning a classifier that makes predictions based on semantic features (i.e., features that are relevant to the prediction task). However, in many instances, the feature domain comprises other meaningful attributes that are irrelevant to the task such as image background (Sagawa et al., 2020), and demographic identities (Liu et al., 2015; Borkan et al., 2019). We refer to these attributes as *spurious attributes*, denoted by $\mathcal{S}$, and we define subgroups as $g \in \mathcal{G} := \mathcal{Y} \times \mathcal{S}$. Formally, spurious correlation refers to the scenario where a statistical dependency between the task and the spurious attributes is observed solely within the training set (Yang et al., 2023):

$$\mathbb{P}_{\text{tr}}(X, Y, S) \propto \mathbb{P}_{\text{tr}}(Y \mid S)\, \mathbb{P}_{\text{tr}}(S). \tag{2}$$

This correlation, however, is absent in the test set:

$$\mathbb{P}_{\text{te}}(X, Y, S) \propto \mathbb{P}_{\text{te}}(Y)\, \mathbb{P}_{\text{te}}(S). \tag{3}$$

More specifically, in the training set, we have $\mathbb{P}_{\text{tr}}(Y \mid S) \neq \mathbb{P}_{\text{tr}}(Y)$. Learning becomes challenging in the presence of spurious correlation because the classifier may shift its reliance from semantic features to spurious attributes, leading to errors during deployment.

To assess robustness with respect to spurious bias, we adopt two key performance metrics in our analysis. The first metric is the *average accuracy*: $\mathbb{E}_{(x,y)}\left[\mathbb{1}\left[f(x) = y\right]\right]$, which provides an assessment of the overall model performance. The second metric is the *worst-group accuracy* (Sagawa et al., 2020) defined as the "accuracy" of the worst-performing subgroup: $\min_{g' \in \mathcal{G}} \mathbb{E}_{(x,y)|g=g'}\left[\mathbb{1}\left[f(x) = y\right]\right]$. More formally, the worst-group accuracy is known as the *worst true positive rate of one group versus the other groups*. This is a commonly used evaluation measure in the field of subgroup robustness (Idrissi et al., 2022; Yang et al., 2023). In line with Nam et al. (2020), we term samples correctly classified by the bias attribute based on the biased training set as *bias-aligned* (BA) samples, and those misclassified as *bias-confliting* (BC) samples. Throughout the paper, for more broader context, we use the terms *underrepresented* and *overrepresented* to refer to BC and BA samples, respectively.

AL for spurious bias. The primary reason for poor robustness under spurious correlations is due to the underrepresentation of certain subgroups in the training set. Consequently, many passive approaches address this issue directly by augmenting the representation of underrepresented samples through upsampling. In this study, we adopt a similar approach in the active way, i.e., improving the representation of underrepresented samples through the acquisition of new samples. In passive frameworks, the criteria for upsampling typically rely on some proxy measure, commonly the predictive error Sagawa et al. (2020); Liu et al. (2021); Duchi & Namkoong (2021). However, this approach is unfeasible in AL, as neither the target labels nor the bias labels are observable before acquisitions. This raises the question: how can we identify underrepresented samples from the unlabelled pool without any labeling information? The following section describes our motivation and principle behind our proposed method.

4 Limitation of uncertainty-based active learning in enhancing robustness

The goal of uncertainty-based AL methods is to minimize model uncertainty, thus enhancing overall performance (Settles, 2012; Huang et al., 2014). This is achieved by selecting samples that the current model finds less certain about. While there are various metrics to quantify uncertainty, they all share a common principle: samples close to the decision boundary tend to be more informative.

Figure 1 illustrates a 2D binary classification task where the features are generated by Gaussian distributions $x \sim \mathcal{N}([\mu_1, \mu_2], \Sigma)$, with identical diagonal covariance matrix $\Sigma = \sigma^2 I$ for every subgroup. The class label Y is determined by the 1st feature coordinate x_1 (x-axis), and the spurious attribute is defined by the 2nd feature coordinate x_2 (y-axis). All subgroups are linearly separable along the central axis. Due to spurious correlations, samples with $y = \square$ are more frequently observed in red while samples with $y = \circ$ are more frequently observed in blue, making other combinations rare. We then train a linear classifier for the task. To emulate the simplicity bias phenomenon during training, we incorporate simplicity regularization.[1]

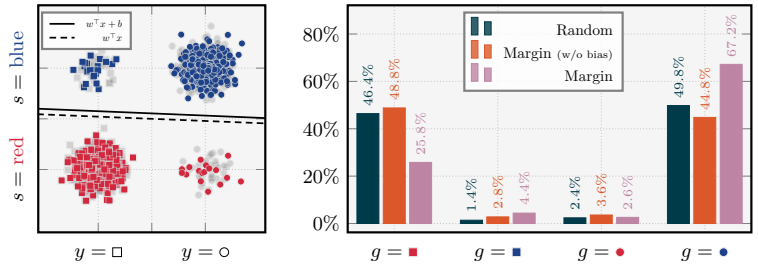


Figure 1: **(left)** The shapes and colours represent the class and spurious attributes, respectively, with grey indicating unlabelled points. The BA (or BC) subgroups are on the off-diagonal (or diagonal). **(right)** The distribution (over subgroups) of the queried batch.

One potential limitation of uncertainty-based AL methods is their reliance on predictive uncertainty as the criterion (Roth & Small, 2006; Wang & Shang, 2014), which can be biased if the predictive model itself is biased. For instance, in Figure 1, the uncertainty measure primarily reflects the classifier's uncertainty about the spurious attribute (color) rather than the target (shape). Consequently, a BA sample might exhibit low uncertainty even if it is misclassified. In this illustrative example, we consider two linear classifiers: one with additive bias and one without.

When employing the unbiased classifier, the mean of the BC samples in the unlabelled pool is closer to the decision boundary compared to the BA samples. However, due to their disproportion, there are significantly more BA samples with similar levels of ambiguity as the BC samples. Hence, the AL agent queries a small number of BC samples alongside many BA samples.

Conversely, the biased classifier shifts the decision boundary towards certain subgroups. As shown in the bar chart in Figure 1, uncertainty sampling enhances the representation of BC subgroups compared to random sampling. However, this behaviour is substantially influenced by the classifier's characteristics. For example, a biased classifier makes the blue subgroups more uncertain than the red ones, resulting in uncertainty sampling favouring these subgroups and leading to the acquisition of a higher proportion of BA samples (blue rectangles) and a smaller proportion of BC samples (red circles). We observed similar behaviour on the real dataset (see Section 5.3).

Despite this limitation, uncertainty-based approaches still offer advantages over random sampling. However, they might exhibit slow progress in improving robustness due to sampling bias. To address this challenge, we require an approach capable of identifying and isolating BC/underrepresented samples. By prioritizing these samples, we can expedite the development of a more robust classifier through AL.

5 Motivation

5.1 Bias-aligned samples are learned faster

[1]We introduce a regularization term $|w_1|$ to the loss function, simulating that learning x_1 is more challenging than x_2. Note that this simplicity bias is the key cause of poor robustness under spurious correlations, as evidenced by recent studies. It can be shown that enforcing a complexity bias through a $|w_1|$ term would yield an unbiased classifier, even in the presence of data bias.

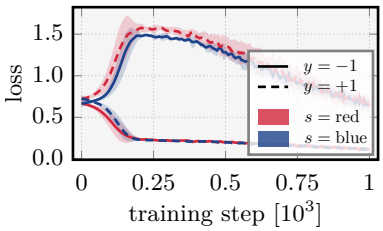
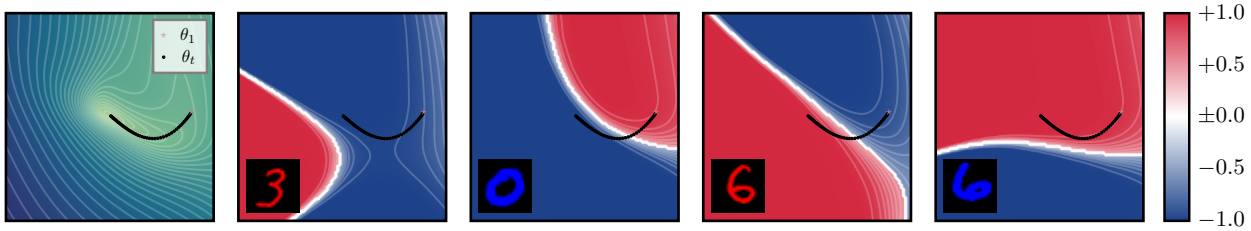


Figure 2: Training loss of aggregated over different subgroups. The colour represents the bias (digit colour).

When training a DNN with the ERM objective on a biased dataset, a notable phenomenon arises: the loss of the BA groups decreases much more rapidly than that of the BC groups, as discussed in Nam et al. (2020); Qiu et al. (2023). This discrepancy arises because the model tends to learn a decision rule that heavily relies on the bias in the early training stage. As a result, the loss for BA subgroups rapidly approaches zero, while the loss for BC subgroups remains high. To demonstrate this phenomenon, we consider a binary classification task involving a binary spurious attribute. We construct a toy example using a binarized MNIST dataset, where digits less than five are labelled as the negative class, while those above are labelled as the positive class. To simulate spurious bias, we assign colours ($\mathcal{S} = \{\text{red}, \text{blue}\}$) to digits (e.g., $\{$ 0 1 4 3 $\}^{y=-1}_{s=\text{red}}$, $\{$ 3 4 2 1 $\}^{y=-1}_{s=\text{blue}}$, $\{$ 6 7 5 8 $\}^{y=+1}_{s=\text{red}}$, $\{$ 6 8 6 7 $\}^{y=+1}_{s=\text{blue}}$), with the training set primarily featuring *negative-class* digits in red and *positive-class* digits in blue which establishes a statistical relationship between colour and target label. As depicted in Figure 2, the loss of BA subgroups shows notably quicker convergence compared to the BC subgroups. This suggests that BC or underrepresented samples can be readily identified by their loss with a model from early training stages (Liu et al., 2021). However, this methodology becomes impractical in AL scenarios where labels are unavailable. Moreover, pseudo-labels, often utilized as substitutes for ground truth labels in AL algorithms (Ash et al., 2020; Wang et al., 2022), may not be suitable, as the model could confidently make incorrect predictions based on incorrect contexts, making the samples distinguishable by loss (further discussion is provided in Section 5.2). Nevertheless, this discrepancy in learning dynamics serves as a useful indicator to discern sample membership, which is the primary motivation behind our approach. This is explored in the next section.

5.2 Inference of learning dynamics

The training dynamics can be further interpreted from an inference perspective. Specifically, predicting an unseen BA sample is anticipated to exhibit sustained low variance throughout the entire training process, while the prediction of a BC sample gradually converges towards the ground truth over time.



(a) training loss (b) $(y = -1, \text{red})$ (c) $(y = -1, \text{blue})$ (d) $(y = +1, \text{red})$ (e) $(y = +1, \text{blue})$

Figure 3: Visualization of multiple metrics in the parameter space. (a) illustrates the training loss surface. The predicted probability of test samples (displayed in the corners) from different subgroup is depicted in (b) and (e) for BA subgroups, and (c) and (d) for BC subgroups, where the sign indicates the predicted label (+ for positive and − for negative classes). The learning trajectory is represented by • with ⋆ denoting the initial point. The model traverses through a significantly greater number of levels in the probability map for BC subgroups compared to BA subgroups, indicating more pronounced predictive variations in the learning trajectory for BC samples.

Figure 3 presents the view of the loss landscape alongside the predicted probability of unseen samples from multiple subgroups in the parameter space. Distinct behaviors emerge for for BC samples ($\{(y = -1, \text{blue}), (y = +1, \text{red})\}$) compared to BA samples ($\{(y = -1, \text{red}), (y = +1, \text{blue})\}$) as the model converges to the local minimum. In the case of BA samples, the model consistently delivers correct predictions with minor variations in confidence. Conversely, for BC samples, the model initially tends to make incorrect predictions, and significant fluctuations in predictions occur until it eventually converges to the ground truth. This exploration underscores the intricate dynamics involved in handling underrepresented samples offering

an insight into designing an acquisition function based on the learning dynamics. Furthermore, to accurately capture the dynamics, one approach is to measure the rate of change of the output of the model w.r.t. training time, denoted as $\partial f(x)/\partial t$ for a given input x. This quantity can be computed analytically using the NTK framework (Lee et al., 2019; Jacot et al., 2018). Additionally, Wang et al. (2022) also incorporated the notion of learning dynamics into their acquisition explicitly computed using NTKs. However, this analytical approach faces practical challenges, particularly in the context of large-scale DNNs with numerous parameters. The computational complexity escalates significantly, particularly in AL scenarios where the computation must scale with the number of AL loops. This computational burden highlights the necessity of exploring alternative methodologies. Thus, in this paper we aimed to propose an alternative measure that offers scalability and computational efficiency.

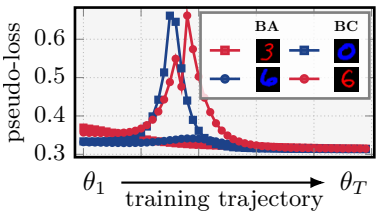


Figure 4: Pseudo-loss progression along the training trajectory from initial (left) to end (right).

Pseudo-labels can be misleading. Moreover, we explore the feasibility of leveraging pseudo-loss for identification of underrepresented samples. Figure 4 visualises the pseudo-loss $\ell(f_\theta(x), \bar{y})$ across different training steps, where $\bar{y} = \arg\max f_\theta(x)$ denotes the pseudo-label assigned to individual sample x. In the initial training phase, the differences are primarily observable among the different spurious attributes $\mathcal{S} = \{\text{red}, \text{blue}\}$. A clearer differentiation emerges between BA and BC subgroups, ultimately reaching disparate levels, as training progresses. Eventually, with more training iterations, the pseudo-loss tends to converge to a similar level. Although the pseudo-loss provides a precise indication of a sample's representativeness concerning the training set, obtaining accurate pseudo labels can be challenging, requiring careful selection of hyperparameters such as the number of training steps, early stopping, and etc. Otherwise, inaccurate pseudo-labels can mislead the selection of unseen samples in AL.

5.3 Domain-Invariant Active Learning

Algorithm 1 Domain-Invariant Active Learning

Input: number of gradient update steps T, number of acquisitions k

for $n \in [N]$ **do**

  ▷ *Train f_θ and collect checkpoints* ◁

  $\mathcal{M} \leftarrow \{\}$

  for $t \in [T]$ **do**

    $\theta_{t+1} \leftarrow \theta_t + \eta \nabla_{\theta_t} \ell(\mathcal{D}_L)$

    $\mathcal{M} \leftarrow \mathcal{M} \cup \{\theta_{t+1}\}$

  ▷ *acquire new samples* ◁

  if $n < N$ **then**

    compute $d_{\mathcal{M}}(x)$ for all $x \in \mathcal{D}_U$

    acquire $Q = \{\text{top-k } d_{\mathcal{M}}\}$

    $\mathcal{D}_L \leftarrow \mathcal{D}_L \cup Q$; $\mathcal{D}_U \leftarrow \mathcal{D}_U \setminus Q$

output $f_{\mathcal{D}_L}$

Tracking tranining trajectory. Based on the above observations, we proposed an active learning strategy DIAL, outlined in Algorithm 1. Unlike other AL acquisition functions, which typically rely on the inference from a single trained model $f_{\mathcal{D}_L}$, DIAL leverages the complete training trajectory in the acquisition process. Prior to the acquisition step, DIAL tracks the trajectory by collecting checkpoints $\mathcal{M} = \{f_{\theta_1}, \ldots, f_{\theta_T} \mid \mathcal{D}_L\}$ throughout the training. This can be viewed as discretizing the evolution of the model over time $f_{\theta(t)}$, akin to approximating the quantity $\partial f(x)/\partial t$. In practical implementation, DIAL does not track every single step but instead samples a fixed number of checkpoints evenly across the training process. This approach mirrors that of Jung et al. (2022), with the main difference being that DIAL does not incorporate varying learning rates; rather, it employs standard stochastic gradient descent with a fixed learning rate. This enables us to capture the differences in learning dynamics between overrepresented and underrepresented samples, a characteristic that naturally emerges in the standard training procedure. Our empirical findings validate the efficacy of this approach, particularly under biased scenarios.

Acquisition function. After collecting the trajectory of training $\mathcal{M}$, DIAL computes the informativeness of $x \in \mathcal{D}_U$ using $d_{\mathcal{M}}$. Essentially, $d_{\mathcal{M}}$ computes the pairwise disparity of predictions between each $\theta \in \mathcal{M}$

$$d_{\mathcal{M}}(x) = \sum_{\theta' \in \mathcal{M}} \sum_{\theta \in \mathcal{M}} \phi(f_\theta(x), f_{\theta'}(x)) \tag{4}$$

where ϕ is the distance function measuring the discrepancy between the outputs from two different models. A high value of $d_{\mathcal{M}}(x)$ indicates a slow evolution of predictions for x during training, suggesting that x is either a challenging sample or belongs to a underrepresented group. We also explored some existing ensemble-based acquisition functions in our experiments (refer to Section 7) since the set of trajectories $\mathcal{M}$ can be viewed as an ensemble of the model.

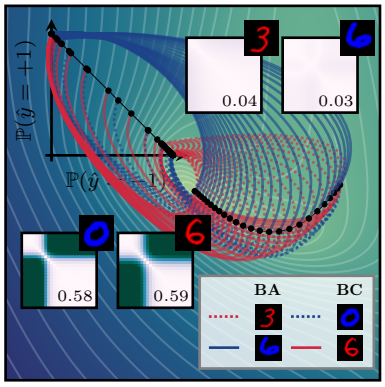


Figure 5: Illustration of DIAL alongside pairwise distance matrices for samples from each subgroup. Average values of distance matrices are shown in the corner of the matrices.

Working principle. The working principle of DIAL is visually detailed in Figure 5. Predictions for individual samples across various subgroups by different checkpoints in the trajectory $\mathcal{M} = \{\theta_1, \ldots, \theta_T\}$ are mapped onto the probability simplex located at the top left of Figure 5. Here, it can be observed that the predictions associated with BA samples typically tend to aggregate towards the vertices of the simplex. In contrast, predictions of BC samples exhibit a more dispersed pattern, spanning more uniformly across the simplex. This variation in distribution across the simplex indicates a substantial difference in prediction consistency between checkpoints, especially noticeable among the BC samples. This phenomenon is further quantified through the pairwise distances matrix, where distances are calculated with metric $\phi(f_\theta(x), f_{\theta'}(x))$. The distance matrices reveal greater discrepancies in the predictions over successive training checkpoints for BC samples compared to BA samples, suggesting less consistency in the model's performance across different training stages for BC samples. In light of this observations, it becomes a useful strategy to enhance the balancedness of $\mathcal{D}_L$. By selecting samples $x \in \mathcal{D}_U$ that demonstrate high values of $d_{\mathcal{M}}$, it is likely to include more samples from underrepresented subgroups into $\mathcal{D}_L$. Such targeted acquisition is instrumental in mitigating spurious correlations present in the existing $\mathcal{D}_L$. Ultimately, this approach leads to the development of a model that is more robust and better attuned to variations w.r.t. the spurious attribute.

DIAL acquires more balanced batches. In Figure 6, we can see that the distributions for the BC subgroups are long-tailed, skewing towards higher informativeness. This pattern is especially pronounced for DIAL, where the distributions for both BC subgroups are more symmetric. Conversely, other methods like Margin and Entropy, while capturing the informativeness of the BC subgroups, produce distributions that are less symmetric. For instance, the BC subgroup ($y = +1, s =$ red) demonstrates a higher density towards the upper score range compared to its counterpart ($y = -1, s =$ blue). This disparity in distribution suggests bias in the acquisition process, where one underrepresented subgroup will be disproportionately represented over the other. Given this context, when adopting a greedy approach to sample selection (i.e., maximize the informativeness of each batch of samples), DIAL demonstrates a superior capability to ensure a more balanced inclusion of samples from underrepresented subgroups. Moreover, this implies that DIAL can achieve better group-wise performance, such as improved worst-group accuracy.

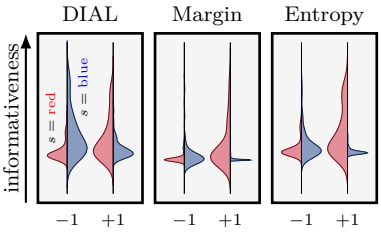


Figure 6: Distributions of informativeness across subgroups with acquisition functions.

6 Experiments

In this section, we provide empirical evaluations on various benchmarks along with several baseline methods. The results are averaged across five trials conducted with different random seeds. Mean and standard deviation are presented in all figures and tables. All experimental details can be found in Appendix A.

Datasets and DNNs. We assess the performance of all AL algorithms using two categories of benchmark datasets. The first category is datasets with spurious correlations for evaluating subgroup robustness, including Waterbirds (Sagawa et al., 2020), CelebA (Liu et al., 2015) and Corrupted CIFAR-10 (Hendrycks & Dietterich, 2019). Waterbirds is the classification task for landbird vs. waterbird where samples are

associated with different backgrounds (spurious attribute): land or water; in the training set landbirds (or waterbirds) are more likely to appear on land (or on water). In the CelebA dataset, the task is to determine if the person in the image has black hair, where gender is considered the spurious attribute. The dataset predominantly consists of females without black hair. Corrupted CIFAR-10 are construed using the standard CIFAR-10 dataset with several types of textures applied onto the images where some classes are entangled with some specific textures in the training set. The second category is the common AL benchmark datasets: CIFAR-10 (Krizhevsky & Hinton, 2009) and SVHN Netzer et al. (2011) for assessing the performance of DIAL in the unbiased environment. Further details regarding the datasets are provided in Appendix A.1. In terms of network architectures, we employ ResNet50 and ResNet18 (He et al., 2016). Comprehensive information such as training hyperparameters can be found in Appendix A.2.

Baselines. In addition to Random sampling, which selects samples randomly to simulate passive learning, we include the following baseline methods: BADGE (Ash et al., 2020), CoreSet (Sener & Savarese, 2018), Margin (Roth & Small, 2006), Entropy (Wang & Shang, 2014), and Certainty (Wang & Shang, 2014). Further details on these baseline methods can be found in Appendix A.3.

Metrics. Regarding performance evaluation, we use worst-group accuracy as the primary metric for Waterbirds and CelebA, and average accuracy for Corrupted CIFAR-10 (as suggested by Nam et al. (2020), given the unbiased test split). While for all common AL datasets, we rely on the average accuracy. Additionally, we also incorporate the area under learning curve (ALC) metric, which quantifies the area under the respective metric curve (y-axis) across the number of queried batches (x-axis).

6.1 Results

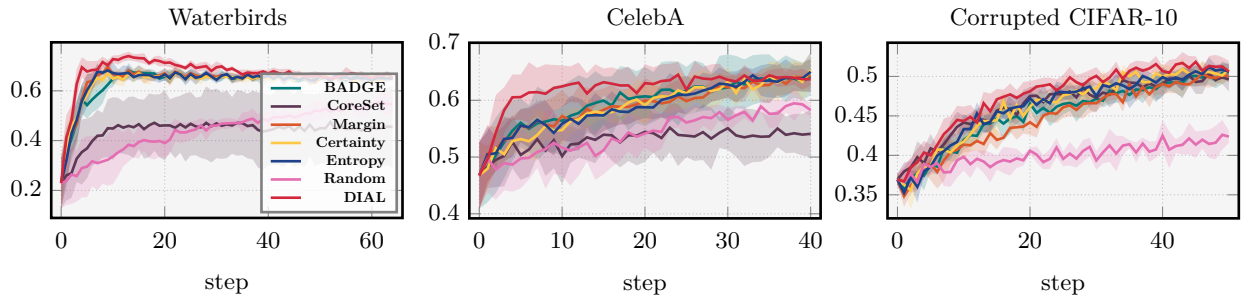


Figure 7: Performance vs. number of AL steps. The figures display worst-group accuracy for Waterbirds and CelebA, alongside average accuracy for Corrupted CIFAR-10.

Does AL improve subgroup performance? Figure 7 presents the performance achieved by various AL algorithms. With the exception of CoreSet on Waterbirds, all baseline methods consistently outperform passive learning (Random) throughout the run. This observation indicates that AL strategies inherently possess the capability to enhance subgroup robustness in a sample-efficient manner, validating our research hypothesis. Particularly noteworthy is the performance of DIAL, which consistently outperforms all baseline methods, especially on Waterbirds. At its peak performance, DIAL achieves approximately 74% accuracy with only around 13% of the total training samples, surpassing state-of-the-art algorithms such as those by Sagawa et al. (2020) (73.1%) and Liu et al. (2021) (71.2%)[2].

While uncertainty methods demonstrate competitive performance, diversity-based approaches (BADGE and CoreSet) exhibit less potency. This discrepancy can be attributed to the enforced diversity criteria, which result in a certain proportion of queried samples being drawn from overrepresented subgroups, thereby failing to adequately address spurious correlations in the labelled pool. It is important to note the decreasing trend observed for DIAL after reaching peak performance on Waterbirds, which is attributed to the insufficient underrepresented samples in $\mathcal{D}_U$ (see Figure 9a), exacerbating spurious correlations upon continued sample

[2]The numerical values are directly taken from Yang et al. (2023)

acquisition. Consequently, we use relatively small query batch sizes to prevent excessive consumption of underrepresented samples, ensuring optimal performance (further analysis is provided in Section 7).

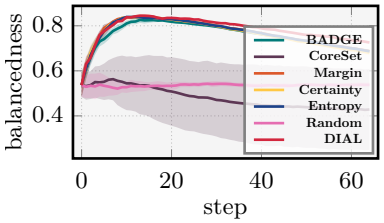
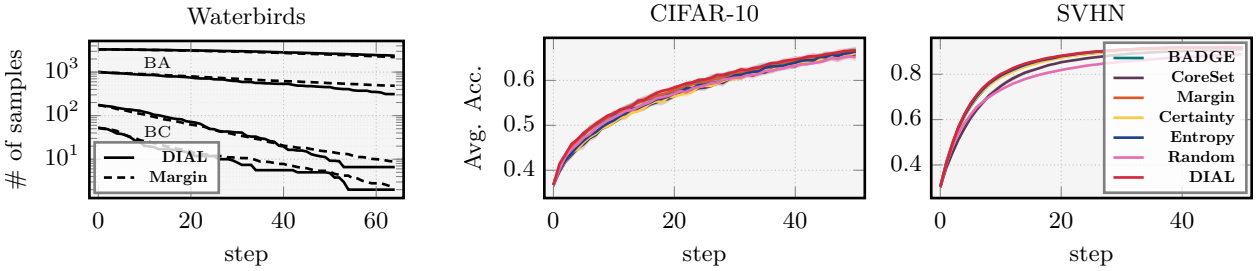


Figure 8: Balancedness[3] of $\mathcal{D}_L$ on Waterbirds throughout the run. Higher implies that $\mathcal{D}_L$ is more balanced, i.e., a value of 1 means that the numbers of samples from every subgroups $g \in \mathcal{G}$ are equal.

Additionally, Figure 8 illustrates the balancedness of $\mathcal{D}_L$, quantified by the normalized entropy of subgroup proportions: $\sum_{g\in\mathcal{G}} -p_g \log p_g / \log |\mathcal{G}|$, where p_g represents the proportion of subgroup g. Due to space limit, the balancedness for other datasets are presented in Appendix B. The analysis reveals a correlation between balancedness and subgroup performance, as observed in Figure 7: higher balancedness in $\mathcal{D}_L$ corresponds to improved robustness. The performance trend of Random remains consistent throughout, indicating that performance gains over AL steps stem primarily from the overall growth of the labelled set rather than the acquisition of underrepresented samples. During the early steps, when the balancedness achieved by DIAL is comparable to that of other methods, the subgroup performance of DIAL aligns closely with that of its counterparts. However, as DIAL approaches peak balancedness, subgroup performance reaches its zenith. This observation reinforces our core objective of designing a subgroup-robust AL strategy: acquiring new samples to balance the training pool, thereby achieving optimal robustness.



(a) subgroup sample size in $\mathcal{D}_U$

(b) performance on common datasets

Figure 9: (a) Evolution of $\mathcal{D}_U$ on Waterbirds. Each line represents an individual subgroup. Due to the bias, there are substantially more BA subgroup samples initially. Observably, DIAL tends to "consume" a greater proportion of BC samples, extending to even the secondary subgroup within BA. (b) Performance vs. number of AL steps for common datasets. The y-axis represents the average accuracy.

Standard AL setting. Figure 9b shows the performance curves for the standard datasets. DIAL exhibits superior performance on both datasets, with its advantage being particularly pronounced on CIFAR-10. This observation underscores the versatility of DIAL: not only does it excel in imbalanced settings, but it also proves effective in conventional settings.

	First 10 AL steps						Full run					
Algorithm	**Waterbirds**	**CelebA**	**Co. CIFAR-10**	**SVHN**	**CIFAR-10**	**Avg.**	**Waterbirds**	**CelebA**	**Co. CIFAR-10**	**SVHN**	**CIFAR-10**	**Avg.**
Random	$2.89_{\pm 1.0}$	$4.99_{\pm 0.5}$	$3.81_{\pm 0.1}$	$5.89_{\pm 0.1}$	$4.64_{\pm 0.1}$	4.56	$28.36_{\pm 2.6}$	$21.66_{\pm 1.1}$	$20.00_{\pm 0.4}$	$39.58_{\pm 0.2}$	$28.62_{\pm 0.1}$	29.42
Margin	**$5.54_{\pm 0.4}$**	**$5.39_{\pm 0.3}$**	$3.83_{\pm 0.1}$	**$6.30_{\pm 0.1}$**	**$4.70_{\pm 0.1}$**	**5.30**	**$41.02_{\pm 0.6}$**	$23.58_{\pm 1.0}$	$22.38_{\pm 0.2}$	**$42.00_{\pm 0.2}$**	$29.06_{\pm 0.3}$	33.81
Certainty	$5.47_{\pm 0.3}$	$5.23_{\pm 0.2}$	$3.89_{\pm 0.1}$	$6.21_{\pm 0.1}$	$4.56_{\pm 0.0}$	5.15	$40.72_{\pm 0.4}$	$23.50_{\pm 0.7}$	$23.10_{\pm 0.3}$	$41.71_{\pm 0.2}$	$28.48_{\pm 0.1}$	33.38
Entropy	$5.51_{\pm 0.5}$	$5.36_{\pm 0.3}$	$3.89_{\pm 0.1}$	$6.26_{\pm 0.1}$	$4.58_{\pm 0.0}$	5.29	$40.92_{\pm 0.6}$	$23.57_{\pm 0.9}$	$23.10_{\pm 0.2}$	$41.97_{\pm 0.2}$	$28.70_{\pm 0.2}$	**33.82**
CoreSet	$3.61_{\pm 1.0}$	$5.08_{\pm 0.4}$	**$4.05_{\pm 0.1}$**	$5.79_{\pm 0.1}$	$4.53_{\pm 0.1}$	4.84	$28.20_{\pm 8.1}$	$21.20_{\pm 1.5}$	**$23.17_{\pm 0.3}$**	$40.64_{\pm 0.2}$	$28.45_{\pm 0.2}$	31.07
BADGE	$5.16_{\pm 0.5}$	$5.37_{\pm 0.5}$	$3.87_{\pm 0.1}$	$6.22_{\pm 0.1}$	$4.70_{\pm 0.1}$	5.23	$40.56_{\pm 0.5}$	**$23.83_{\pm 1.9}$**	$22.68_{\pm 0.3}$	$41.82_{\pm 0.2}$	**$29.06_{\pm 0.3}$**	33.69
DIAL	**$6.16_{\pm 0.4}$**	**$5.84_{\pm 0.5}$**	**$4.07_{\pm 0.1}$**	**$6.30_{\pm 0.1}$**	**$4.74_{\pm 0.0}$**	**5.57**	**$43.10_{\pm 0.6}$**	**$24.83_{\pm 1.0}$**	**$23.79_{\pm 0.3}$**	**$42.09_{\pm 0.1}$**	**$29.25_{\pm 0.1}$**	**34.61**

Table 1: The ALC for all datasets with their corresponding metrics. The left section of the table shows the ALC over the initial 10 AL steps, while the right section displays the ALC for the entire run. The highest and second-highest values are highlighted in **bold red** and **bold blue**, respectively.

Table 1 provides a comprehensive numerical overview of the performance of all methods. Throughout the early stages and the entire run, DIAL consistently outperforms all baseline methods across all datasets. Following

[3]The curves of Margin and Entropy overlap.

closely, uncertainty-based methods exhibit the second-best performance on all datasets except for Corrupted CIFAR-10 (and CIFAR-10 in full run). Margin and Entropy methods, on average, demonstrate competitive performance. Furthermore, while DIAL may not exhibit notably superior performance on SVHN, it remains comparable to baseline methods, indicating its applicability to problems where issues of imbalance or spurious correlations are unaware, without compromising performance. Refer to Appendix B for comprehensive summaries of other metrics per dataset, similar to those presented in Yang et al. (2023).

7 Ablation study

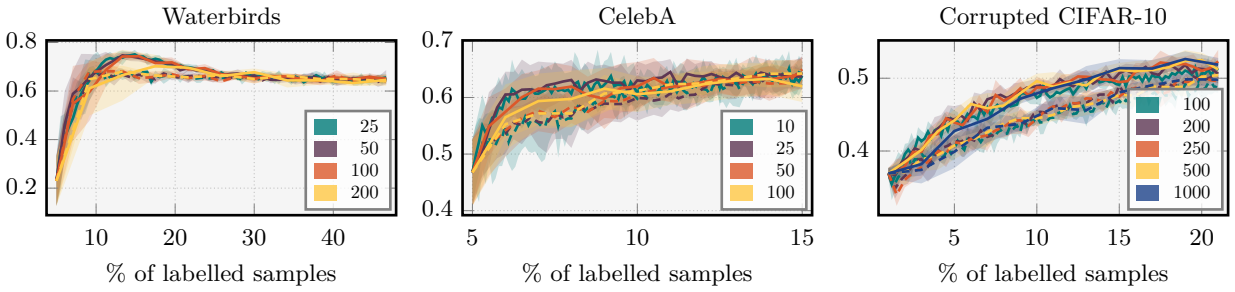


Figure 10: Learning curve (worst-group accuracy for Waterbirds and CelebA, and average accuracy for Corrupted CIFAR-10) across different query batch sizes. The solid line and dashed line correspond DIAL and Margin, respectively.

Effect of query batch size Figure 10 presents the subgroup robustness performance w.r.t. the query batch size k. Similar to standard AL outcomes, the subgroup performance tends to decline as k increases. This decline occurs because larger values of k compel the querying of uninformative samples, which, in this context, correspond to overrepresented or BC samples. Remarkably, with a batch size of 25, DIAL achieved 76.1% on Waterbirds, outperforming most passive baselines as reported in Yang et al. (2023). Thus, when dealing with imbalance in AL, it is recommanded to use smaller query batch sizes. Additionally, it is noteworthy that DIAL consistently outperforms Margin, which is the best-performing baseline overall, across all query batch sizes.

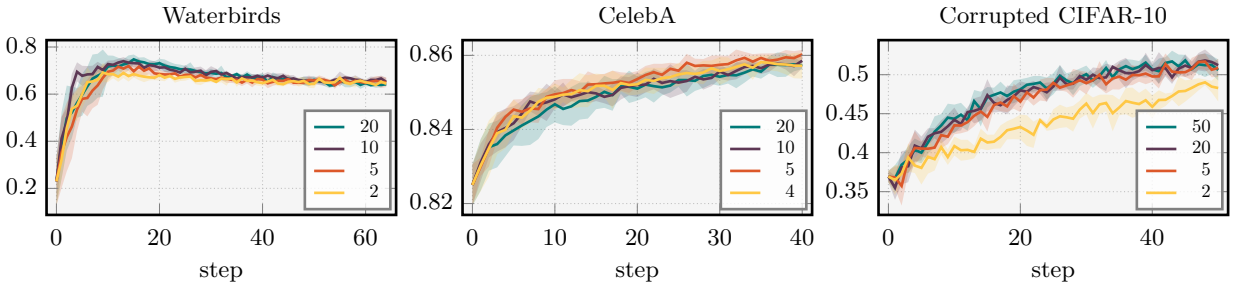
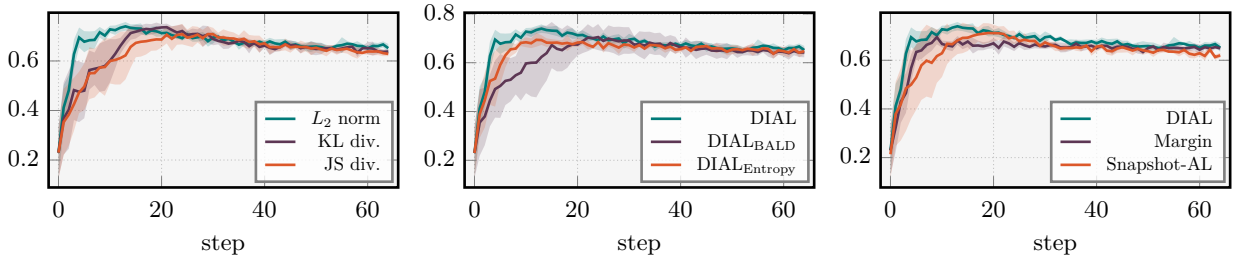


Figure 11: Learning curve across different size of checkpoints $|\mathcal{M}|$.

Effect of $|\mathcal{M}|$ Figure 11 depicts the learning curve across different densities of learning dynamics for DIAL. We uniformly sample model checkpoints throughout the training process. We observed that a larger size of $\mathcal{M}$ (denser sampling) generally leads to higher performance, as it provides a more accurate estimation of the learning dynamics. In the case of Waterbirds, the slow increase in performance in the beginning when using $|\mathcal{M}| = 20$ could be attributed to noise caused training instability. However, despite this, it achieved higher peak performance compared to other configurations. Overall, it's crucial to strike a balance in the sampling rate. Sampling rates should not be excessively high (resulting in larger $\mathcal{M}$), as this may capture training noise. Conversely, they should not be too low (resulting in smaller $\mathcal{M}$), as this could lead to missing information of learning dynamics.



(a) different ϕ (b) different $d_{\mathcal{M}}$ (c) compared with Jung et al. (2022)

Figure 12: Performance of DIAL across different configurations evaluated on Waterbirds, with worst-group accuracy plotted on the y-axis. (a) Evaluation of various distance functions ϕ. (b) Assessment of different common acquisition criteria on training dynamics $d_{\mathcal{M}}$. (c) Assessment of the approach proposed by Jung et al. (2022).

Effect of distance functions ϕ. We conducted a comparison of distance functions utilized to capture predictive discrepancies among checkpoints. The discrepancy is evaluated using the softmax output and we found that this approach resulted in much better performance, even with L_2 norm. The comparison in Figure 3a shows that both the L_2 norm and Kullback-Leibler (KL) divergence demonstrate similar peak performance. However, the L_2 norm achieves this with significantly fewer acquisition steps. Jensen-Shannon (JS) divergence, a symmetric version of the KL divergence, initially behaves similarly to the KL divergence but diverges after approximately 10 acquisition steps, eventually converging to a comparable level. Overall, the L_2 norm proves adept at accurately detecting predictive discrepancies, particularly on underrepresented samples, leading to rapid improvements in robustness. The diminished performance with divergence metrics may be attributed to the non-linear operation, which tends to neglect very small discrepancies.

Effect of discrepancy criteria $d_{\mathcal{M}}$. In addition to Equation (4), we also explored alternative criteria for assessing informativeness with training dynamics $\mathcal{M}$. One such criteria is Entropy, defined as $d_{\mathcal{M}}(x) = \mathrm{H}(f_{\mathcal{M}}(x))$, where $f_{\mathcal{M}}(x) = \frac{1}{|\mathcal{M}|}\sum_{\theta\in\mathcal{M}} f_\theta(x)$ and H is the entropy measure which evaluates the overall predictive entropy of a sample x throughout the training. Additionally, we also incorporate BALD (Houlsby et al., 2011) $d_{\mathcal{M}}(x) = \mathrm{H}(f_{\mathcal{M}}(x)) + \frac{1}{|\mathcal{M}|}\sum_{\theta\in\mathcal{M}} \mathrm{H}(f_\theta(x))$ to measure the predictive disagreement among models throughout the training. Figure 3b shows that both variants ($\text{DIAL}_{\text{Entropy}}$ and $\text{DIAL}_{\text{BALD}}$) do not yield performance gain compared to the proposed criteria (in Equation (4)).

Comparison with Snapshot-AL We conducted a comparison between DIAL and Snapshot-AL as proposed by Jung et al. (2022). To ensure a fair comparison, we obtained an equal number of checkpoints for both DIAL and Snapshot-AL, and employed the same disagreement criteria as outlined in Equation (4). The primary distinction lies in the application of the cosine annealing learning rate scheduler (Loshchilov & Hutter, 2017) during training with Snapshot-AL. As illustrated in Figure 3c, Snapshot-AL, which utilizes diverse checkpoints generated with the learning rate scheduler, consistently underperforms DIAL throughout run. This further corroborates our assertion that the diversity of ensembles or checkpoints does not effectively facilitate the identification of underrepresented samples. Rather, tracking the learning dynamics provides a more accurate means of identification of such samples. Nevertheless, Snapshot-AL demonstrates superior peak performance compared to Margin, despite its relatively lower performance in the early rounds.

8 Conclusion

In this work, we investigate spurious correlations within in the AL settings. We demonstrate a potential failure mode of uncertainty-based AL approaches in handling spurious correlations. To address this, we propose a sample AL acquisition approach that effectively identifies underrepresented samples from the unlabelled pool. Our experiments on real-world datasets show that our proposed method mitigates bias in the labelled dataset and outperforms baseline methods. Additionally, we provide a detailed analysis highlighting the effectiveness of our approach with respect to various AL hyperparameters.

References

Hadis Anahideh, Abolfazl Asudeh, and Saravanan Thirumuruganathan. Fair active learning. *Expert Systems with Applications*, 199:116981, August 2022. ISSN 09574174. doi: 10.1016/j.eswa.2022.116981.

Julia Angwin, Jeff Larson, Mattu Surya, and Lauren Kirchner. How We Analyzed the COMPAS Recidivism Algorithm. *ProPublica (5 2016)*, 9, 2016.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant Risk Minimization, March 2020.

Jordan T. Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep Batch Active Learning by Diverse, Uncertain Gradient Lower Bounds. In *International Conference on Learning Representations*, 2020.

Parmida Atighehchian, Frederic Branchaud-Charron, Jan Freyberg, Rafael Pardinas, Lorne Schell, and George Pearse. Baal, a bayesian active learning library, 2022.

Samuel James Bell and Levent Sagun. Simplicity Bias Leads to Amplified Performance Disparities. In *2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 355–369, Chicago IL USA, June 2023. ACM. ISBN 9798400701924. doi: 10.1145/3593013.3594003.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced Metrics for Measuring Unintended Bias with Real Data for Text Classification. In *Companion Proceedings of The 2019 World Wide Web Conference*, pp. 491–500, San Francisco USA, May 2019. ACM. ISBN 978-1-4503-6675-5. doi: 10.1145/3308560.3317593.

John C. Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3), June 2021. ISSN 0090-5364. doi: 10.1214/20-AOS2004.

Ricky Maulana Fajri, Akrati Saxena, Yulong Pei, and Mykola Pechenizkiy. FAL-CUR: Fair Active Learning using Uncertainty and Representativeness on Fair Clustering. *Expert Systems with Applications*, 242:122842, May 2024. ISSN 09574174. doi: 10.1016/j.eswa.2023.122842.

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1050–1059, New York, New York, USA, June 2016. PMLR.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep Bayesian Active Learning with Image Data. In *International Conference on Machine Learning*, pp. 1183–1192. PMLR, July 2017.

Yonatan Geifman and Ran El-Yaniv. Deep Active Learning over the Long Tail, November 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Dan Hendrycks and Thomas Dietterich. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. In *Proceedings of the International Conference on Learning Representations*, 2019. doi: 10.48550/arXiv.1903.12261.

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian Active Learning for Classification and Preference Learning. December 2011.

Gao Huang, Yixuan Li, Geoff Pleiss, Zhuang Liu, John E. Hopcroft, and Kilian Q. Weinberger. Snapshot Ensembles: Train 1, Get M for Free. In *International Conference on Learning Representations*, November 2016.

Sheng-Jun Huang, Rong Jin, and Zhi-Hua Zhou. Active Learning by Querying Informative and Representative Examples. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(10):1936–1949, October 2014. ISSN 0162-8828, 2160-9292. doi: 10.1109/TPAMI.2014.2307881.

Badr Youbi Idrissi, Martin Arjovsky, Mohammad Pezeshki, and David Lopez-Paz. Simple data balancing achieves competitive worst-group-accuracy. In *Proceedings of the First Conference on Causal Learning and Reasoning*, volume 177 of *Proceedings of Machine Learning Research*, pp. 336–351. PMLR, April 2022.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural Tangent Kernel: Convergence and Generalization in Neural Networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pp. 8580–8589, Red Hook, NY, USA, 2018. Curran Associates Inc.

Seohyeon Jung, Sanghyun Kim, and Juho Lee. A Simple Yet Powerful Deep Active Learning With Snapshots Ensembles. In *International Conference on Learning Representations*, September 2022.

Andreas Kirsch, Joost R. van Amersfoort, and Yarin Gal. BatchBALD: Efficient and Diverse Batch Acquisition for Deep Bayesian Active Learning. In *Advances in Neural Information Processing Systems*, pp. 7026–7037, 2019.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton A. Earnshaw, Imran S. Haque, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A Benchmark of in-the-Wild Distribution Shifts. In *International Conference on Machine Learning (ICML)*, 2021.

Seo Taek Kong, Soomin Jeon, Dongbin Na, Jaewon Lee, Hong-Seok Lee, and Kyu-Hwan Jung. A Neural Pre-Conditioning Active Learning Algorithm to Reduce Label Complexity. In *Advances in Neural Information Processing Systems*, May 2022.

A. Krizhevsky and G. Hinton. Learning Multiple Layers of Features from Tiny Images. *Master's thesis, Department of Computer Science, University of Toronto*, 2009.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 6405–6416, Red Hook, NY, USA, December 2017. Curran Associates Inc. ISBN 978-1-5108-6096-4.

Jaehoon Lee, Lechao Xiao, Samuel S. Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide Neural Networks of Any Depth Evolve as Linear Models under Gradient Descent. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2019. Curran Associates Inc.

Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just Train Twice: Improving Group Robustness without Training Group Information. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 6781–6792. PMLR, July 2021.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep Learning Face Attributes in the Wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Ilya Loshchilov and Frank Hutter. SGDR: Stochastic Gradient Descent with Warm Restarts. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

Mohamad Amin Mohamadi, Wonho Bae, and Danica J. Sutherland. Making Look-Ahead Active Learning Strategies Feasible with Neural Tangent Kernels. In *Advances in Neural Information Processing Systems*, 2022.

Nihal Murali, Aahlad Manas Puli, Ke Yu, Rajesh Ranganath, and Kayhan Batmanghelich. Shortcut Learning Through the Lens of Early Training Dynamics. In *ICML 2023 Workshop on Spurious Correlations, Invariance, and Stability*, February 2023. doi: 10.48550/arXiv.2302.09344.

Preetum Nakkiran, Gal Kaplun, Dimitris Kalimeris, Tristan Yang, Benjamin L. Edelman, Fred Zhang, and Boaz Barak. SGD on Neural Networks Learns Functions of Increasing Complexity. In *Advances in Neural Information Processing Systems*, 2019. doi: 10.48550/arXiv.1905.11604.

Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from Failure: De-biasing Classifier from Biased Classifier. In *Advances in Neural Information Processing Systems*, volume 33, pp. 20673–20684, 2020.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading Digits in Natural Images with Unsupervised Feature Learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011*, 2011.

Luke Oakden-Rayner, Jared Dunnmon, Gustavo Carneiro, and Christopher Re. Hidden Stratification Causes Clinically Meaningful Failures in Machine Learning for Medical Imaging. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, pp. 151–159, Toronto Ontario Canada, April 2020. ACM. ISBN 978-1-4503-7046-2. doi: 10.1145/3368555.3384468.

GuanWen Qiu, Da Kuang, and Surbhi Goel. Complexity Matters: Dynamics of Feature Learning in the Presence of Spurious Correlations. In *Mathematics of Modern Machine Learning Workshop at NeurIPS 2023*, 2023.

Dan Roth and Kevin Small. Margin-Based Active Learning for Structured Output Spaces. In *Machine Learning: ECML 2006*, pp. 413–424, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg. ISBN 978-3-540-46056-5.

Nicholas Roy and Andrew McCallum. Toward Optimal Active Learning through Sampling Estimation of Error Reduction. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pp. 441–448, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. ISBN 1-55860-778-1.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization. In *International Conference on Learning Representations*, 2020.

Ozan Sener and Silvio Savarese. Active Learning for Convolutional Neural Networks: A Core-Set Approach. In *International Conference on Learning Representations*, 2018.

Burr Settles. *Active Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Springer International Publishing, Cham, 2012. ISBN 978-3-031-00432-2 978-3-031-01560-1. doi: 10.1007/978-3-031-01560-1.

Burr Settles, Mark Craven, and Soumya Ray. Multiple-Instance Active Learning. In *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007.

Harshay Shah, Kaustav Tamuly, Aditi Raghunathan, Prateek Jain, and Praneeth Netrapalli. The Pitfalls of Simplicity Bias in Neural Networks. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Amr Sharaf, Hal Daume Iii, and Renkun Ni. Promoting Fairness in Learned Models by Learning to Active Learn under Parity Constraints. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 2149–2156, Seoul Republic of Korea, June 2022. ACM. ISBN 978-1-4503-9352-2. doi: 10.1145/3531146.3534632.

Nimit Sohoni, Jared Dunnmon, Geoffrey Angus, Albert Gu, and Christopher Ré. No Subclass Left Behind: Fine-Grained Robustness in Coarse-Grained Classification Problems. In *Advances in Neural Information Processing Systems*, volume 33, pp. 19339–19352, 2020.

Alex Tamkin, Dat Nguyen, Salil Deshpande, Jesse Mu, and Noah Goodman. Active Learning Helps Pretrained Models Learn the Intended Task. In *Advances in Neural Information Processing Systems*, April 2022.

Damien Teney, Ehsan Abbasnejad, Simon Lucey, and Anton Van Den Hengel. Evading the Simplicity Bias: Training a Diverse Set of Models Discovers Solutions with Superior OOD Generalization. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16740–16751, New Orleans, LA, USA, June 2022. IEEE. ISBN 978-1-66546-946-3. doi: 10.1109/CVPR52688.2022.01626.

Bhavya Vasudeva, Kameron Shahabi, and Vatsal Sharan. Mitigating Simplicity Bias in Deep Learning for Improved OOD Generalization and Robustness. In *ICML 2023 Workshop on Spurious Correlations, Invariance, and Stability*, October 2023.

C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Datase. Technical report, 2011.

Dan Wang and Yi Shang. A New Active Labeling Method for Deep Learning. In *2014 International Joint Conference on Neural Networks (IJCNN)*, pp. 112–119, Beijing, China, July 2014. IEEE. ISBN 978-1-4799-1484-5 978-1-4799-6627-1. doi: 10.1109/IJCNN.2014.6889457.

Haonan Wang, Wei Huang, Ziwei Wu, Hanghang Tong, Andrew J. Margenot, and Jingrui He. Deep Active Learning by Leveraging Training Dynamics. In *Advances in Neural Information Processing Systems*, volume 35, pp. 25171–25184, 2022.

Yu Yang, Eric Gan, Gintare Karolina Dziugaite, and Baharan Mirzasoleiman. Identifying Spurious Biases Early in Training through the Lens of Simplicity Bias. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2024. doi: 10.48550/arXiv.2305.18761.

Yuzhe Yang, Haoran Zhang, Dina Katabi, and Marzyeh Ghassemi. Change is Hard: A Closer Look at Subpopulation Shift. In *International Conference on Machine Learning*, February 2023.

Huaxiu Yao, Yu Wang, Sai Li, Linjun Zhang, Weixin Liang, James Zou, and Chelsea Finn. Improving Out-of-Distribution Robustness via Selective Augmentation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 25407–25437. PMLR, July 2022.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning Fair Representations. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 325–333, Atlanta, Georgia, USA, June 2013. PMLR.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 Million Image Database for Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6): 1452–1464, June 2018. ISSN 0162-8828, 2160-9292, 1939-3539. doi: 10.1109/TPAMI.2017.2723009.

A Experimental setup

A.1 Datasets

In all experiments, we use the training split as the unlabelled pool and evaluate the performance using the test split. At the start of the AL loops, a subset of samples is randomly labelled to form the initial labelled pool. Dataset statistics are provided in Figure A.1.

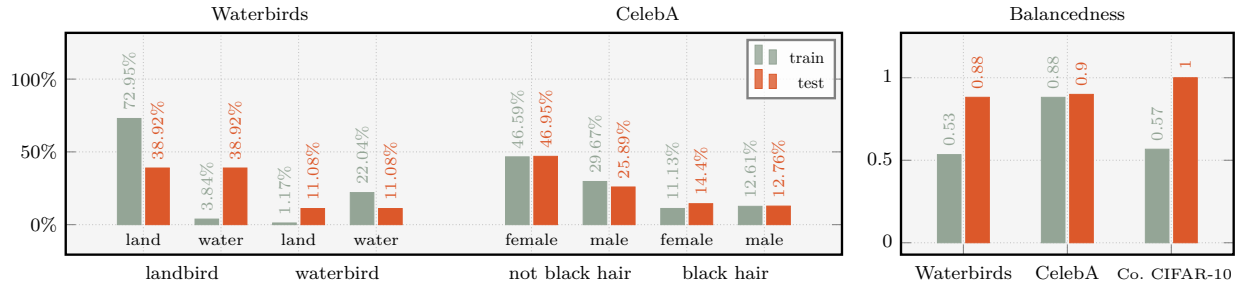
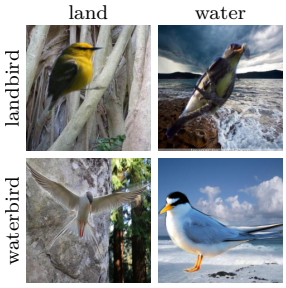


Figure A.1: **(left)** The subgroup distribution of the training and the test splits over all subgroups. **(right)** The balancedness of the dataset.



Waterbirds. The Waterbirds dataset is a synthetic dataset created by overlaying bird images from Wah et al. (2011) onto various scene images from Zhou et al. (2018). The target labels are *landbird* and *waterbird*, while the spurious attribute is the background: *land* or *water*. In the training set, there is a spurious correlation between the bird class and the background, with the majority subgroups (or BA subgroups) being *landbirds on land* and *waterbirds on water*. Both the target and spurious labels are binary: $\mathcal{Y} = \text{landbird}, \text{waterbird}$ and $\mathcal{S} = \text{land}, \text{water}$, resulting in four subgroups ($|\mathcal{G}| = 4$). The train and test splits consist of 4,795 and 5,794 samples, respectively.

CelebA. The CelebA dataset (Liu et al., 2015) comprises over 200,000 celebrity images annotated with 40 different attributes. In our experiment, we use the `Black_Hair` attribute as the target and the `Male` attribute as the spurious attribute. The dataset exhibits a spurious correlation between hair color and gender; specifically, with most females having non-black hair. Both target and spurious labels are binary: $\mathcal{Y} = \text{not black hair}, \text{black hair}$ and $\mathcal{S} = \text{female}, \text{male}$, resulting in four subgroups ($|\mathcal{G}| = 4$). Due to computational constraints, we randomly subsampled 10,000 samples from the training split and used the original test split, which contains 19,962 samples.

Corrupted CIFAR-10. The Corrupted CIFAR-10 dataset is generated from the standard CIFAR-10 (Krizhevsky & Hinton, 2009) by applying corruptions as proposed by Hendrycks & Dietterich (2019). The target labels are the original CIFAR-10 labels, while the spurious labels correspond to the types of corruption: (1) Gaussian Noise, (2) Shot Noise, (3) Impulse Noise, (4) Speckle Noise, (5) Gaussian Blur, (6) Defocus Blur, (7) Glass Blur, (8) Motion Blur, (9) Zoom Blur, (10) Original. Following the procedure defined by Nam et al. (2020), we generate the dataset with a corruption severity of 1 and a ratio of 95% BA samples to create the spurious correlation. Subgroups where the numeric index of target label and that of spurious label coincide ($y = s$) are considered BA subgroups, while others are BC subgroups. Both target and spurious labels have a cardinality of 10, resulting in 100 subgroups ($|\mathcal{G}| = 100$). The subgroups are uniformly distributed, with BA and BC subgroups having proportions of 95%/10 and 5%/90, respectively. The train and test splits consist of 50,000 and 10,000 samples, respectively, with the test split being unbiased (uniformly distributed across all subgroups).

A.2 Implementation details and hyperparameter settings

Our code for the experiments is based on the Baal AL framework (Atighehchian et al., 2022). We use Baal implementations for Random, Margin, and Entropy, while BADGE and Coreset are sourced from the repository by Ash et al. (2020). We construct the Waterbirds and CelebA datasets using the WILDS package (Koh et al., 2021) and the Corrupted CIFAR-10 dataset using the repository by Nam et al. (2020). For evaluation, we use the code by Yang et al. (2023). Additionally, for visualizing the training trajectories in Section 5, we deploy the loss-landscape-anim repository. We also followed some hyperparameters setting and data preprocessing procedure specified in Yang et al. (2023) and Nam et al. (2020).

Dataset	# of initial samples	query batch size	architecture	pre-trained weights	optimizer	learning rate	weight decay	momentum	batch size
Waterbirds	5% of total	0.625% of total	ResNet50	ImageNet-1K	SGD	1×10^{-3}	1×10^{-4}	0.9	108
CelebA	5% of total	0.25% of total	ResNet50	ImageNet-1K	SGD	1×10^{-3}	1×10^{-4}	0.9	108
Co. CIFAR-10	500	200	ResNet18	ImageNet-1K	Adam	1×10^{-3}	1×10^{-4}	–	128
CIFAR-10	500	500	ResNet18	–	SGD	1×10^{-3}	5×10^{-4}	0.9	128
SVHN	500	500	ResNet18	ImageNet-1K	SGD	1×10^{-3}	1×10^{-4}	0.9	128

A.3 Baselines

Random — Uniformly label k samples from the unlabelled pool, resembling the passive setting.

Margin — An uncertainty-based AL method that selects samples based on the difference between the top two predictive class probabilities: $\arg\max_{x \in \mathcal{D}_U} f(x)^0 - f(x)^1$, where $f(x)^0$ and $f(x)^1$ are the largest and second largest softmax entries of $f(x)$ (Roth & Small, 2006).

Entropy — An uncertainty-based AL method that selects the most uncertain samples based on predictive entropy: $\arg\max_{x \in \mathcal{D}_U} H(f(x))$, where H is the entropy defined as $H(p) = \sum_{i=1}^{C} -p_i \log(1/p_i)$ (Wang & Shang, 2014).

Cetainty — An uncertainty-based AL method that selects samples with the smallest predictive probability: $\arg\min_{x \in \mathcal{D}_U} \max_{i \in [C]} f(x)_i$ (Wang & Shang, 2014).

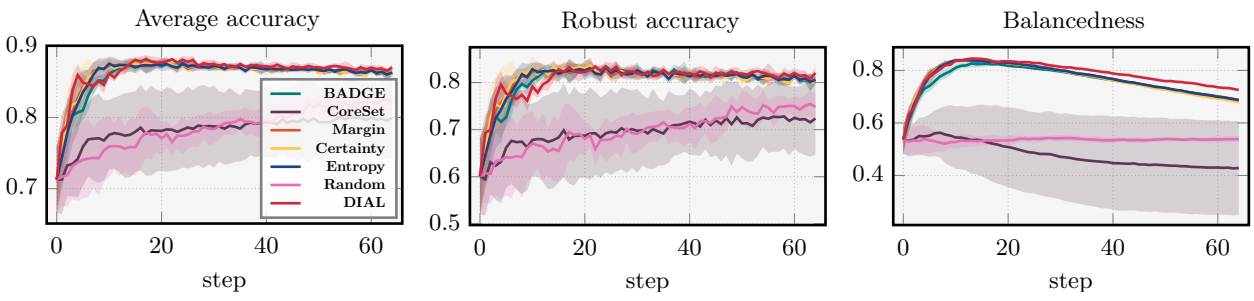
BADGE — A diversity-based AL method. For each unlabelled sample, it computes the gradient embeddings (the first derivative of the pseudo-loss with respect to the output of the penultimate layer). It then selects a subset of k samples using the k-MEANS++ seeding algorithm on the gradient embeddings (Ash et al., 2020).

Coreset — A diversity-based AL method that selects a subset of k samples by solving the coreset problem on the feature embeddings (the output of the penultimate layer) (Sener & Savarese, 2018).

B Additional results

In this section, we present additional results and analysis on datasets with spurious correlations. We include the ALC over evaluation metrics as presented in Yang et al. (2023) (refer to Section B.3 in Yang et al. (2023) for details). Additionally, we report the *robust accuracy*, defined as the *worst-case expected accuracy over attribute* (Sohoni et al., 2020): $\min_{s' \in \mathcal{S}} \mathbb{E}_{(x,y)|s=s'} [\mathbb{1}[f(x) = y]]$.

B.1 Waterbirds



B.1.1 ALC over evaluation metrics across different AL methods

Algorithm	Avg Acc.	Worst Acc.	Avg Prec.	Worst Prec.	Avg F1.	Worst F1.	Adjusted Acc.	Balanced Acc.	AUROC	ECE
Random	$50.2_{\pm 1.4}$	$28.4_{\pm 2.6}$	$45.3_{\pm 1.4}$	$33.0_{\pm 2.3}$	$46.2_{\pm 1.5}$	$37.7_{\pm 2.0}$	$48.0_{\pm 1.4}$	$48.0_{\pm 1.4}$	$54.5_{\pm 1.2}$	$9.7_{\pm 1.4}$
Margin	$55.2_{\pm 0.3}$	$41.0_{\pm 0.6}$	$51.1_{\pm 0.4}$	$42.2_{\pm 0.8}$	$52.1_{\pm 0.4}$	$46.1_{\pm 0.5}$	$53.6_{\pm 0.3}$	$53.6_{\pm 0.3}$	$58.9_{\pm 0.2}$	$6.1_{\pm 0.3}$
Certainty	$55.2_{\pm 0.4}$	$40.7_{\pm 0.4}$	$51.1_{\pm 0.4}$	$42.3_{\pm 0.8}$	$52.1_{\pm 0.4}$	$46.0_{\pm 0.5}$	$53.6_{\pm 0.3}$	$53.6_{\pm 0.3}$	$58.9_{\pm 0.2}$	$6.1_{\pm 0.3}$
Entropy	$55.2_{\pm 0.4}$	$40.9_{\pm 0.6}$	$51.0_{\pm 0.5}$	$42.1_{\pm 0.8}$	$52.1_{\pm 0.4}$	$46.0_{\pm 0.6}$	$53.6_{\pm 0.3}$	$53.6_{\pm 0.3}$	$58.9_{\pm 0.2}$	$6.1_{\pm 0.3}$
CoreSet	$50.1_{\pm 3.4}$	$28.2_{\pm 8.1}$	$45.3_{\pm 3.5}$	$33.0_{\pm 5.6}$	$46.0_{\pm 3.8}$	$37.5_{\pm 5.2}$	$47.8_{\pm 3.6}$	$47.8_{\pm 3.6}$	$54.2_{\pm 3.0}$	$9.4_{\pm 2.5}$
BADGE	$55.1_{\pm 0.2}$	$40.6_{\pm 0.5}$	$50.9_{\pm 0.2}$	$41.9_{\pm 0.4}$	$51.9_{\pm 0.2}$	$45.8_{\pm 0.3}$	$53.4_{\pm 0.2}$	$53.4_{\pm 0.2}$	$58.8_{\pm 0.1}$	$6.2_{\pm 0.2}$
DIAL	$55.4_{\pm 0.3}$	$43.1_{\pm 0.6}$	$51.4_{\pm 0.3}$	$42.4_{\pm 0.6}$	$52.5_{\pm 0.3}$	$46.7_{\pm 0.4}$	$54.2_{\pm 0.3}$	$54.2_{\pm 0.3}$	$59.2_{\pm 0.2}$	$6.0_{\pm 0.2}$

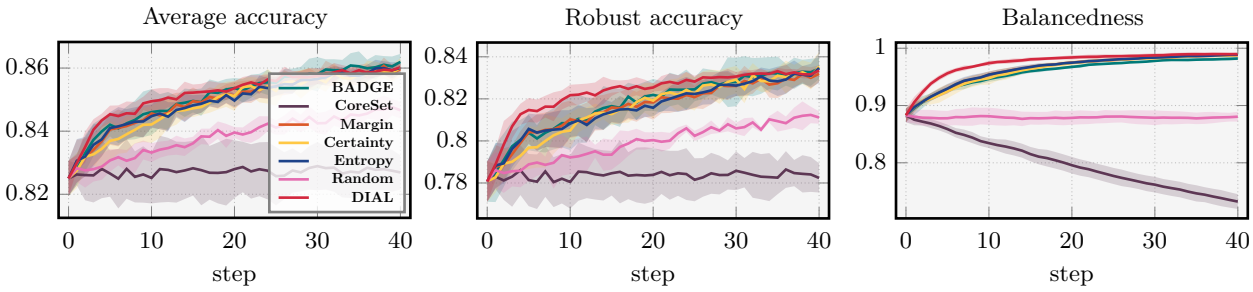
B.1.2 ALC over evaluation metrics across different number of checkpoints

$\lvert\mathcal{M}\rvert$	Avg Acc.	Worst Acc.	Avg Prec.	Worst Prec.	Avg F1.	Worst F1.	Adjusted Acc.	Balanced Acc.	AUROC	ECE
2	$55.3_{\pm 0.3}$	$41.3_{\pm 0.5}$	$51.2_{\pm 0.3}$	$42.4_{\pm 0.6}$	$52.2_{\pm 0.3}$	$46.2_{\pm 0.4}$	$53.7_{\pm 0.3}$	$53.7_{\pm 0.3}$	$59.0_{\pm 0.2}$	$6.0_{\pm 0.2}$
5	$55.2_{\pm 0.4}$	$41.4_{\pm 0.8}$	$51.1_{\pm 0.5}$	$42.2_{\pm 0.9}$	$52.2_{\pm 0.4}$	$46.2_{\pm 0.6}$	$53.7_{\pm 0.4}$	$53.7_{\pm 0.4}$	$58.9_{\pm 0.3}$	$6.1_{\pm 0.4}$
10	$55.4_{\pm 0.3}$	$43.1_{\pm 0.6}$	$51.4_{\pm 0.3}$	$42.4_{\pm 0.6}$	$52.5_{\pm 0.3}$	$46.7_{\pm 0.4}$	$54.2_{\pm 0.3}$	$54.2_{\pm 0.3}$	$59.2_{\pm 0.2}$	$6.0_{\pm 0.2}$
20	$55.1_{\pm 0.4}$	$42.5_{\pm 1.2}$	$51.0_{\pm 0.4}$	$41.9_{\pm 0.7}$	$52.1_{\pm 0.4}$	$46.2_{\pm 0.6}$	$53.9_{\pm 0.4}$	$53.9_{\pm 0.4}$	$59.0_{\pm 0.4}$	$6.3_{\pm 0.3}$

B.1.3 ALC over evaluation metrics across different query batch size and AL methods

Algorithm	k	Avg Acc.	Worst Acc.	Avg Prec.	Worst Prec.	Avg F1.	Worst F1.	Adjusted Acc.	Balanced Acc.	AUROC	ECE
DIAL	25	$69.0_{\pm 0.5}$	$52.8_{\pm 1.1}$	$63.9_{\pm 0.6}$	$52.5_{\pm 0.9}$	$65.2_{\pm 0.6}$	$57.8_{\pm 0.8}$	$67.3_{\pm 0.5}$	$67.3_{\pm 0.5}$	$73.8_{\pm 0.4}$	$7.7_{\pm 0.4}$
Margin		$69.0_{\pm 0.4}$	$51.3_{\pm 0.4}$	$63.8_{\pm 0.5}$	$52.7_{\pm 1.0}$	$65.1_{\pm 0.5}$	$57.6_{\pm 0.6}$	$67.0_{\pm 0.3}$	$67.0_{\pm 0.3}$	$73.7_{\pm 0.2}$	$7.7_{\pm 0.4}$
DIAL	50	$34.6_{\pm 0.2}$	$26.5_{\pm 0.6}$	$32.0_{\pm 0.2}$	$26.4_{\pm 0.4}$	$32.7_{\pm 0.2}$	$29.0_{\pm 0.3}$	$33.7_{\pm 0.2}$	$33.7_{\pm 0.2}$	$36.9_{\pm 0.2}$	$3.8_{\pm 0.2}$
Margin		$34.5_{\pm 0.2}$	$25.6_{\pm 0.4}$	$31.9_{\pm 0.3}$	$26.4_{\pm 0.5}$	$32.6_{\pm 0.3}$	$28.8_{\pm 0.4}$	$33.5_{\pm 0.2}$	$33.5_{\pm 0.2}$	$36.8_{\pm 0.1}$	$3.8_{\pm 0.2}$
DIAL	100	$17.2_{\pm 0.1}$	$13.1_{\pm 0.4}$	$16.0_{\pm 0.2}$	$13.1_{\pm 0.3}$	$16.3_{\pm 0.2}$	$14.4_{\pm 0.2}$	$16.8_{\pm 0.2}$	$16.8_{\pm 0.2}$	$18.4_{\pm 0.1}$	$1.9_{\pm 0.1}$
Margin		$17.2_{\pm 0.1}$	$12.9_{\pm 0.1}$	$15.9_{\pm 0.1}$	$13.1_{\pm 0.2}$	$16.2_{\pm 0.1}$	$14.3_{\pm 0.1}$	$16.7_{\pm 0.1}$	$16.7_{\pm 0.1}$	$18.4_{\pm 0.1}$	$2.0_{\pm 0.1}$
DIAL	200	$8.6_{\pm 0.1}$	$6.4_{\pm 0.3}$	$7.9_{\pm 0.1}$	$6.5_{\pm 0.1}$	$8.1_{\pm 0.1}$	$7.1_{\pm 0.1}$	$8.3_{\pm 0.1}$	$8.3_{\pm 0.1}$	$9.2_{\pm 0.1}$	$1.0_{\pm 0.1}$
Margin		$8.6_{\pm 0.1}$	$6.4_{\pm 0.1}$	$7.9_{\pm 0.1}$	$6.5_{\pm 0.1}$	$8.1_{\pm 0.1}$	$7.1_{\pm 0.1}$	$8.3_{\pm 0.0}$	$8.3_{\pm 0.0}$	$9.2_{\pm 0.0}$	$1.0_{\pm 0.0}$

B.2 CelebA



B.2.1 The ALC over evaluation metrics across different AL methods

Algorithm	Avg Acc.	Worst Acc.	Avg Prec.	Worst Prec.	Avg F1.	Worst F1.	Adjusted Acc.	Balanced Acc.	AUROC	ECE
Random	$33.5_{\pm 0.1}$	$21.7_{\pm 1.1}$	$32.4_{\pm 0.2}$	$30.5_{\pm 0.7}$	$31.1_{\pm 0.1}$	$26.5_{\pm 0.3}$	$30.3_{\pm 0.3}$	$30.4_{\pm 0.3}$	$36.0_{\pm 0.2}$	$2.9_{\pm 0.1}$
Margin	$34.0_{\pm 0.1}$	$23.6_{\pm 1.0}$	$32.9_{\pm 0.1}$	$30.9_{\pm 0.2}$	$31.9_{\pm 0.2}$	$27.8_{\pm 0.3}$	$31.2_{\pm 0.2}$	$31.3_{\pm 0.2}$	$36.5_{\pm 0.2}$	$2.3_{\pm 0.2}$
Certainty	$34.0_{\pm 0.1}$	$23.5_{\pm 0.7}$	$32.9_{\pm 0.1}$	$31.0_{\pm 0.2}$	$31.8_{\pm 0.3}$	$27.6_{\pm 0.5}$	$31.1_{\pm 0.4}$	$31.1_{\pm 0.3}$	$36.4_{\pm 0.1}$	$2.4_{\pm 0.2}$
Entropy	$34.0_{\pm 0.1}$	$23.6_{\pm 0.9}$	$32.9_{\pm 0.1}$	$30.8_{\pm 0.1}$	$31.9_{\pm 0.2}$	$27.8_{\pm 0.3}$	$31.2_{\pm 0.3}$	$31.3_{\pm 0.2}$	$36.4_{\pm 0.2}$	$2.3_{\pm 0.2}$
CoreSet	$33.1_{\pm 0.3}$	$21.2_{\pm 1.5}$	$31.7_{\pm 0.5}$	$29.4_{\pm 1.0}$	$30.6_{\pm 0.4}$	$25.7_{\pm 0.7}$	$29.8_{\pm 0.5}$	$29.9_{\pm 0.5}$	$35.5_{\pm 0.4}$	$2.6_{\pm 0.3}$
BADGE	$34.0_{\pm 0.1}$	$23.8_{\pm 1.9}$	$33.0_{\pm 0.1}$	$31.0_{\pm 0.4}$	$32.0_{\pm 0.3}$	$27.9_{\pm 0.6}$	$31.3_{\pm 0.5}$	$31.4_{\pm 0.5}$	$36.5_{\pm 0.1}$	$2.4_{\pm 0.3}$
DIAL	$34.1_{\pm 0.1}$	$24.8_{\pm 1.0}$	$33.1_{\pm 0.2}$	$31.2_{\pm 0.5}$	$32.0_{\pm 0.2}$	$28.0_{\pm 0.4}$	$31.3_{\pm 0.3}$	$31.4_{\pm 0.3}$	$36.6_{\pm 0.1}$	$2.6_{\pm 0.2}$

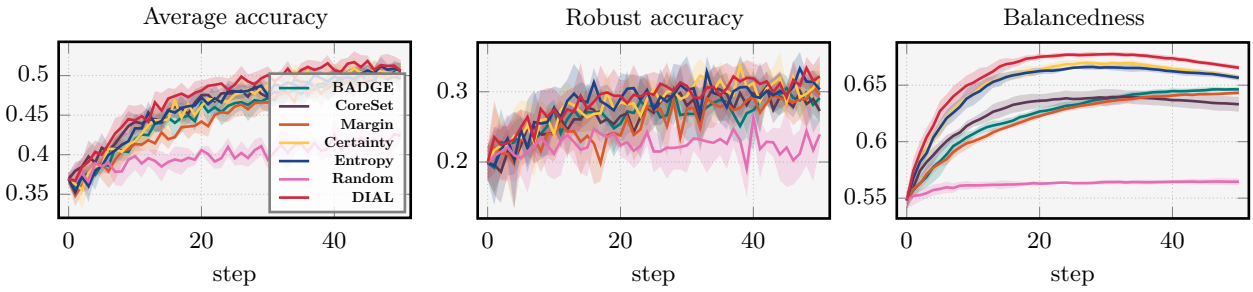
B.2.2 ALC over evaluation metrics across different number of checkpoints

$\|\mathcal{M}\|$	Avg Acc.	Worst Acc.	Avg Prec.	Worst Prec.	Avg F1.	Worst F1.	Adjusted Acc.	Balanced Acc.	AUROC	ECE
2	34.1 $\pm$0.1	25.2 $\pm$1.4	32.9 $\pm$0.1	30.6 $\pm$0.5	32.1 $\pm$0.2	28.2 $\pm$0.5	31.5 $\pm$0.4	31.6 $\pm$0.4	36.5 $\pm$0.1	2.3 $\pm$0.2
4	34.0 $\pm$0.1	24.6 $\pm$1.6	33.0 $\pm$0.2	31.1 $\pm$0.7	32.0 $\pm$0.3	27.9 $\pm$0.5	31.2 $\pm$0.4	31.3 $\pm$0.4	36.5 $\pm$0.1	2.6 $\pm$0.3
5	34.1 $\pm$0.1	24.8 $\pm$1.0	33.1 $\pm$0.2	31.2 $\pm$0.5	32.0 $\pm$0.2	28.0 $\pm$0.4	31.3 $\pm$0.3	31.4 $\pm$0.3	36.6 $\pm$0.1	2.6 $\pm$0.2
10	34.0 $\pm$0.1	24.3 $\pm$1.0	33.0 $\pm$0.1	31.2 $\pm$0.5	31.9 $\pm$0.2	27.7 $\pm$0.4	31.1 $\pm$0.3	31.2 $\pm$0.3	36.5 $\pm$0.1	2.7 $\pm$0.2
20	34.0 $\pm$0.1	24.2 $\pm$1.4	33.0 $\pm$0.2	31.4 $\pm$0.6	31.8 $\pm$0.4	27.5 $\pm$0.7	31.0 $\pm$0.5	31.0 $\pm$0.5	36.5 $\pm$0.1	2.8 $\pm$0.3

B.2.3 ALC over evaluation metrics across different query batch size and AL methods

Algorithm	k	Avg Acc.	Worst Acc.	Avg Prec.	Worst Prec.	Avg F1.	Worst F1.	Adjusted Acc.	Balanced Acc.	AUROC	ECE
DIAL	10	85.1 $\pm$0.2	61.4 $\pm$2.8	82.5 $\pm$0.5	77.8 $\pm$1.5	79.9 $\pm$0.6	69.6 $\pm$1.1	78.1 $\pm$0.9	78.2 $\pm$0.9	91.3 $\pm$0.3	6.6 $\pm$0.5
Margin		85.0 $\pm$0.3	59.2 $\pm$3.0	82.2 $\pm$0.3	77.0 $\pm$0.9	79.8 $\pm$0.6	69.6 $\pm$1.1	78.1 $\pm$0.9	78.3 $\pm$0.8	91.2 $\pm$0.3	5.8 $\pm$0.6
DIAL	20	42.6 $\pm$0.1	30.8 $\pm$1.4	41.3 $\pm$0.2	39.1 $\pm$0.7	40.0 $\pm$0.3	34.8 $\pm$0.5	39.0 $\pm$0.4	39.1 $\pm$0.4	45.7 $\pm$0.1	3.3 $\pm$0.2
Margin		42.5 $\pm$0.1	29.7 $\pm$1.1	41.1 $\pm$0.3	38.6 $\pm$0.7	39.9 $\pm$0.2	34.8 $\pm$0.3	39.0 $\pm$0.2	39.1 $\pm$0.2	45.6 $\pm$0.2	2.8 $\pm$0.2
DIAL	25	34.1 $\pm$0.1	24.8 $\pm$1.0	33.1 $\pm$0.2	31.2 $\pm$0.5	32.0 $\pm$0.2	28.0 $\pm$0.4	31.3 $\pm$0.3	31.4 $\pm$0.3	36.6 $\pm$0.1	2.6 $\pm$0.2
Margin		34.0 $\pm$0.1	23.6 $\pm$1.0	32.9 $\pm$0.1	30.9 $\pm$0.2	31.9 $\pm$0.2	27.8 $\pm$0.3	31.2 $\pm$0.2	31.3 $\pm$0.2	36.5 $\pm$0.2	2.3 $\pm$0.2
DIAL	50	17.0 $\pm$0.1	12.2 $\pm$0.7	16.5 $\pm$0.1	15.6 $\pm$0.3	16.0 $\pm$0.2	13.9 $\pm$0.3	15.6 $\pm$0.2	15.6 $\pm$0.2	18.3 $\pm$0.0	1.4 $\pm$0.2
Margin		17.0 $\pm$0.1	11.9 $\pm$0.4	16.4 $\pm$0.1	15.4 $\pm$0.3	16.0 $\pm$0.1	13.9 $\pm$0.1	15.6 $\pm$0.1	15.7 $\pm$0.1	18.2 $\pm$0.1	1.2 $\pm$0.1
DIAL	100	8.5 $\pm$0.0	6.0 $\pm$0.4	8.3 $\pm$0.0	7.8 $\pm$0.1	8.0 $\pm$0.1	6.9 $\pm$0.2	7.8 $\pm$0.1	7.8 $\pm$0.1	9.1 $\pm$0.0	0.7 $\pm$0.1
Margin		8.5 $\pm$0.0	6.0 $\pm$0.1	8.2 $\pm$0.0	7.7 $\pm$0.1	8.0 $\pm$0.0	7.0 $\pm$0.0	7.8 $\pm$0.0	7.8 $\pm$0.0	9.1 $\pm$0.0	0.6 $\pm$0.0

B.3 Corrupted CIFAR-10



B.3.1 ALC over evaluation metrics across different AL methods

Algorithm	Avg Acc.	Worst Acc.	Avg Prec.	Worst Prec.	Avg F1.	Worst F1.	Adjusted Acc.	Balanced Acc.	AUROC	ECE
Random	20.0 $\pm$0.4	0.2 $\pm$0.0	20.8 $\pm$0.4	11.2 $\pm$0.4	20.1 $\pm$0.4	11.4 $\pm$0.5	20.0 $\pm$0.4	20.0 $\pm$0.4	39.3 $\pm$0.2	24.7 $\pm$0.4
Margin	22.4 $\pm$0.2	0.6 $\pm$0.1	23.0 $\pm$0.2	12.7 $\pm$0.1	22.5 $\pm$0.2	13.1 $\pm$0.4	22.4 $\pm$0.2	22.4 $\pm$0.2	41.0 $\pm$0.1	22.4 $\pm$0.2
Certainty	23.1 $\pm$0.3	0.9 $\pm$0.1	23.7 $\pm$0.2	13.5 $\pm$0.2	23.2 $\pm$0.3	14.1 $\pm$0.3	23.1 $\pm$0.3	23.1 $\pm$0.3	41.6 $\pm$0.1	21.8 $\pm$0.3
Entropy	23.1 $\pm$0.2	0.9 $\pm$0.1	23.7 $\pm$0.2	13.5 $\pm$0.1	23.2 $\pm$0.2	14.2 $\pm$0.2	23.1 $\pm$0.2	23.1 $\pm$0.2	41.6 $\pm$0.1	21.8 $\pm$0.2
CoreSet	23.2 $\pm$0.3	0.7 $\pm$0.1	23.7 $\pm$0.3	13.1 $\pm$0.4	23.3 $\pm$0.3	13.6 $\pm$0.5	23.2 $\pm$0.3	23.2 $\pm$0.3	41.5 $\pm$0.2	21.3 $\pm$0.3
BADGE	22.7 $\pm$0.3	0.8 $\pm$0.1	23.3 $\pm$0.2	13.1 $\pm$0.2	22.8 $\pm$0.3	13.5 $\pm$0.2	22.7 $\pm$0.3	22.7 $\pm$0.3	41.3 $\pm$0.2	22.0 $\pm$0.3
DIAL	23.8 $\pm$0.3	1.0 $\pm$0.2	24.3 $\pm$0.3	13.7 $\pm$0.2	23.9 $\pm$0.3	14.4 $\pm$0.3	23.8 $\pm$0.3	23.8 $\pm$0.3	42.1 $\pm$0.2	20.8 $\pm$0.3

B.3.2 ALC over evaluation metrics across different number of checkpoints

$\|\mathcal{M}\|$	Avg Acc.	Worst Acc.	Avg Prec.	Worst Prec.	Avg F1.	Worst F1.	Adjusted Acc.	Balanced Acc.	AUROC	ECE
2	21.8 $\pm$0.3	0.5 $\pm$0.1	22.4 $\pm$0.3	12.3 $\pm$0.3	21.9 $\pm$0.3	12.6 $\pm$0.3	21.8 $\pm$0.3	21.8 $\pm$0.3	40.6 $\pm$0.2	22.6 $\pm$0.3
5	23.4 $\pm$0.3	1.2 $\pm$0.2	23.9 $\pm$0.2	13.6 $\pm$0.2	23.5 $\pm$0.2	14.3 $\pm$0.3	23.4 $\pm$0.3	23.4 $\pm$0.3	41.8 $\pm$0.1	21.3 $\pm$0.2
20	23.6 $\pm$0.2	1.2 $\pm$0.3	24.2 $\pm$0.2	13.6 $\pm$0.3	23.7 $\pm$0.2	14.4 $\pm$0.3	23.6 $\pm$0.2	23.6 $\pm$0.2	42.0 $\pm$0.1	21.0 $\pm$0.2
50	23.8 $\pm$0.3	1.0 $\pm$0.2	24.3 $\pm$0.3	13.7 $\pm$0.2	23.9 $\pm$0.3	14.4 $\pm$0.3	23.8 $\pm$0.3	23.8 $\pm$0.3	42.1 $\pm$0.2	20.8 $\pm$0.3

B.3.3 ALC over evaluation metrics across different query batch size and AL methods

Algorithm	k	Avg Acc.	Worst Acc.	Avg Prec.	Worst Prec.	Avg F1.	Worst F1.	Adjusted Acc.	Balanced Acc.	AUROC	ECE
Margin	100	$44.1_{\pm 0.4}$	$1.1_{\pm 0.1}$	$45.3_{\pm 0.3}$	$25.0_{\pm 0.6}$	$44.3_{\pm 0.4}$	$25.9_{\pm 0.8}$	$44.1_{\pm 0.4}$	$44.1_{\pm 0.4}$	$81.6_{\pm 0.3}$	$46.0_{\pm 0.5}$
DIAL		$46.9_{\pm 0.3}$	$1.8_{\pm 0.1}$	$47.9_{\pm 0.3}$	$27.0_{\pm 0.5}$	$47.1_{\pm 0.3}$	$28.2_{\pm 0.3}$	$46.9_{\pm 0.3}$	$46.9_{\pm 0.3}$	$83.8_{\pm 0.2}$	$43.0_{\pm 0.4}$
Margin	200	$22.4_{\pm 0.2}$	$0.6_{\pm 0.1}$	$23.0_{\pm 0.2}$	$12.7_{\pm 0.1}$	$22.5_{\pm 0.2}$	$13.1_{\pm 0.4}$	$22.4_{\pm 0.2}$	$22.4_{\pm 0.2}$	$41.0_{\pm 0.1}$	$22.4_{\pm 0.2}$
DIAL		$23.8_{\pm 0.3}$	$1.0_{\pm 0.2}$	$24.3_{\pm 0.3}$	$13.7_{\pm 0.2}$	$23.9_{\pm 0.3}$	$14.4_{\pm 0.3}$	$23.8_{\pm 0.3}$	$23.8_{\pm 0.3}$	$42.1_{\pm 0.2}$	$20.8_{\pm 0.3}$
Margin	250	$17.8_{\pm 0.2}$	$0.6_{\pm 0.2}$	$18.4_{\pm 0.2}$	$10.2_{\pm 0.2}$	$17.9_{\pm 0.2}$	$10.6_{\pm 0.2}$	$17.8_{\pm 0.2}$	$17.8_{\pm 0.2}$	$32.8_{\pm 0.1}$	$17.8_{\pm 0.2}$
DIAL		$19.0_{\pm 0.1}$	$0.9_{\pm 0.1}$	$19.4_{\pm 0.1}$	$11.1_{\pm 0.1}$	$19.1_{\pm 0.1}$	$11.5_{\pm 0.1}$	$19.0_{\pm 0.1}$	$19.0_{\pm 0.1}$	$33.7_{\pm 0.1}$	$16.6_{\pm 0.1}$
Margin	500	$8.9_{\pm 0.1}$	$0.3_{\pm 0.1}$	$9.2_{\pm 0.1}$	$5.2_{\pm 0.1}$	$9.0_{\pm 0.1}$	$5.3_{\pm 0.2}$	$8.9_{\pm 0.1}$	$8.9_{\pm 0.1}$	$16.4_{\pm 0.1}$	$8.7_{\pm 0.1}$
DIAL		$9.5_{\pm 0.1}$	$0.4_{\pm 0.0}$	$9.7_{\pm 0.1}$	$5.6_{\pm 0.1}$	$9.5_{\pm 0.1}$	$5.7_{\pm 0.2}$	$9.5_{\pm 0.1}$	$9.5_{\pm 0.1}$	$16.9_{\pm 0.0}$	$8.1_{\pm 0.1}$
Margin	1000	$4.5_{\pm 0.0}$	$0.1_{\pm 0.0}$	$4.6_{\pm 0.0}$	$2.6_{\pm 0.1}$	$4.5_{\pm 0.0}$	$2.7_{\pm 0.1}$	$4.5_{\pm 0.0}$	$4.5_{\pm 0.0}$	$8.2_{\pm 0.0}$	$4.2_{\pm 0.0}$
DIAL		$4.7_{\pm 0.0}$	$0.2_{\pm 0.0}$	$4.8_{\pm 0.0}$	$2.8_{\pm 0.1}$	$4.7_{\pm 0.0}$	$2.9_{\pm 0.1}$	$4.7_{\pm 0.0}$	$4.7_{\pm 0.0}$	$8.4_{\pm 0.0}$	$4.0_{\pm 0.0}$

B.4 Some analysis

B.4.1 DIAL upweights the minority subgroups

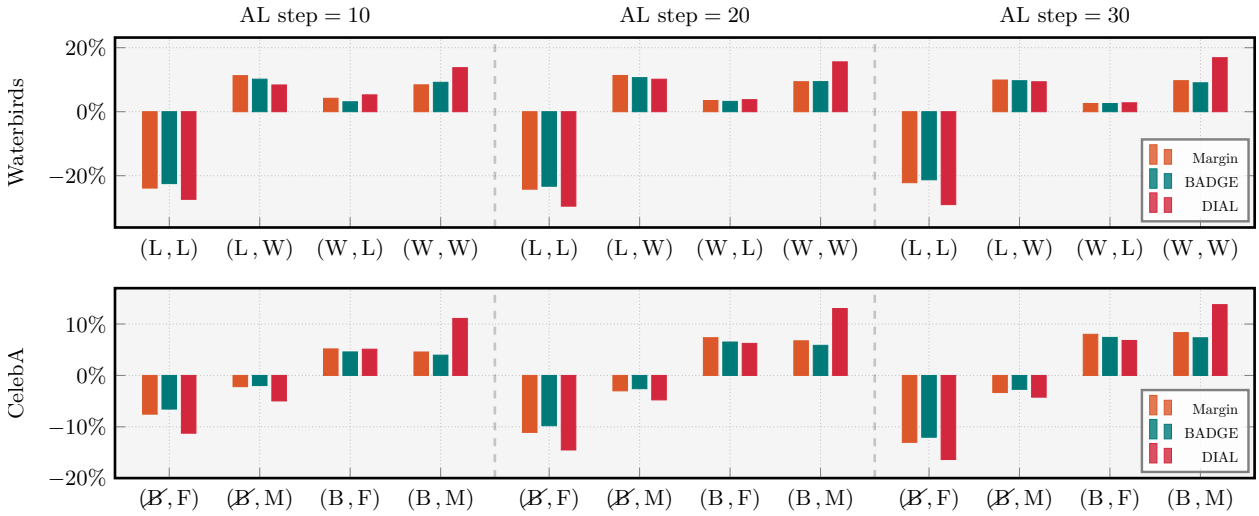


Figure B.2: The change of the distribution of $\mathcal{D}_L$ over (Y, S) relative to the initial labelled pool across different AL methods. The subgroups (L, W) and (W, L) correspond *landbird in water* and *waterbird in land* respectively. Similarly, the subgroups (B́, F) and (B, M) correspond to *female with non-black hair* and *male with black hair*, respectively.

Discussion. Figure B.2 shows that AL effectively enhance the representation of underrepresented subgroups. As more samples are queried, the proportion of these underrepresented (or overrepresented) subgroups shifts: underrepresented subgroups increase while overrepresented ones decrease. Notably, DIAL demonstrates a low preference for querying the overrepresented subgroups, such as "*landbird in land*" for the Waterbirds dataset and "*non-black hair female*" for CelebA, surpassing other AL methods in this aspect.

B.4.2 Comparison of performance with passive setting

Discussion. Figure B.3 shows the performance comparison between active and passive settings. Each performance value for the AL methods is the peak achieved throughout the run. The results indicate that most AL methods consistently outperform the ERM baseline with significantly fewer samples across various performance metrics, except for Corrupted CIFAR-10. Especially, DIAL stands out as the most sample-efficient AL method among the baselines (i.e., located at the top-left corner of the plot, indicating high performance with the small number of samples), particularly for the worst-group accuracy metric.

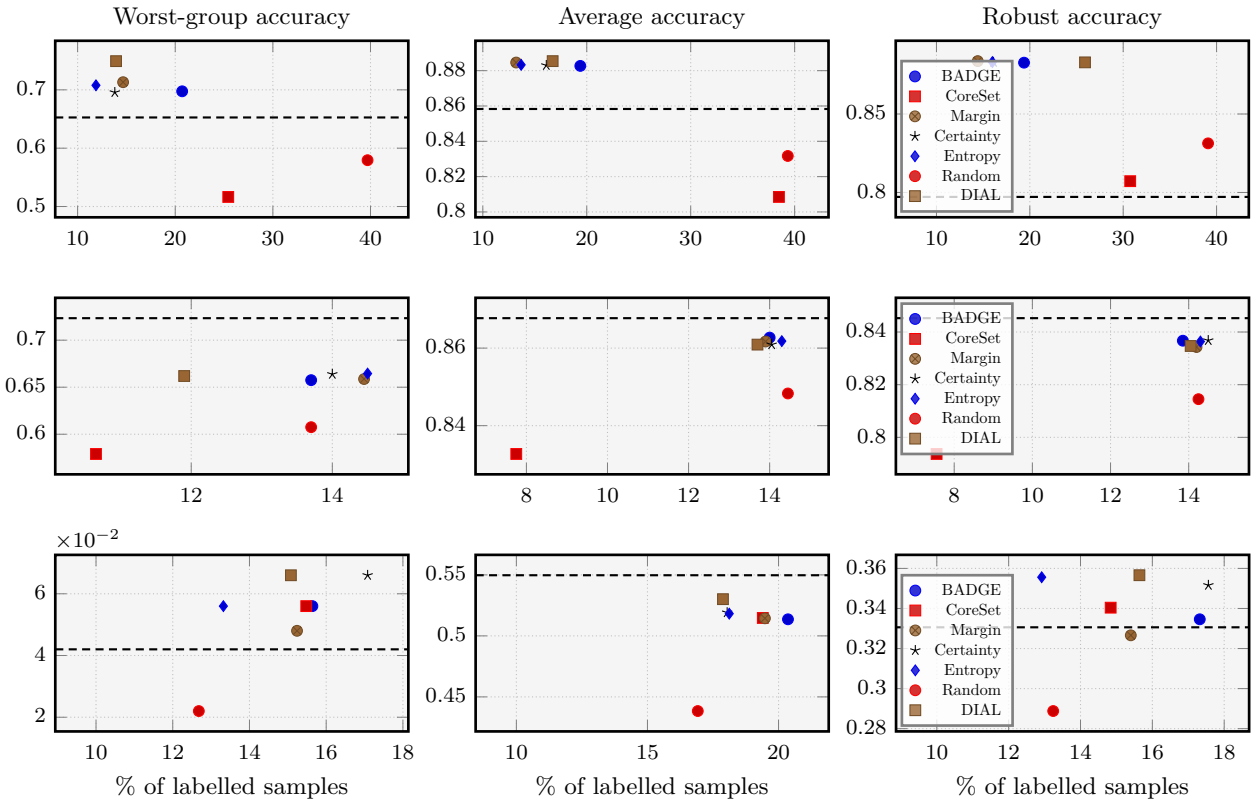


Figure B.3: **(top)** Waterbirds **(middle)** CelebA **(bottom)** Corrupted CIFAR-10. The horizontal dashed line indicates the passive performance (ERM on the compete labelled training set). The y-axis and x-axis represent the performance and the portion of labelled pool, respectively.

C Early stopping active learning

The stopping criteria in AL is used to terminate the AL loops, balancing the performance and labeling cost. Since this is not main scope in this work, we only present the early stopped results retrospectively based on some metrics proposed by Yang et al. (2023): (1) test worst-group accuracy (C.1): oracle criteria based on the *worst-group accuracy* on the test split. (2) validation worst-group accuracy (C.2): criteria based on the *worst-group accuracy* on the validation split, assuming the spurious attributes are known. (3) validation average accuracy (C.3): criteria based on the *average accuracy* on the validation split, assuming the spurious attributes are unknown.

C.1 Test worst-group accuracy

C.1.1 Waterbirds

Algorithm	% of samples	Avg Acc.	Worst Acc.	Avg Prec.	Worst Prec.	Avg F1.	Worst F1.	Adjusted Acc.	Balanced Acc.	AUROC	ECE
Random	$42.5_{\pm 0.0}$	$82.4_{\pm 1.3}$	$55.4_{\pm 4.4}$	$75.1_{\pm 1.5}$	$58.1_{\pm 2.5}$	$76.8_{\pm 1.5}$	$65.3_{\pm 2.2}$	$79.6_{\pm 1.5}$	$79.6_{\pm 1.5}$	$89.0_{\pm 1.1}$	$11.9_{\pm 1.1}$
Margin	$10.4_{\pm 0.0}$	$86.9_{\pm 1.1}$	$69.5_{\pm 2.9}$	$80.6_{\pm 1.5}$	$66.8_{\pm 2.8}$	$82.4_{\pm 1.2}$	$73.6_{\pm 1.7}$	$85.1_{\pm 0.9}$	$85.1_{\pm 0.9}$	$92.8_{\pm 0.5}$	$8.6_{\pm 0.9}$
Certainty	$10.4_{\pm 0.0}$	$87.1_{\pm 0.6}$	$67.1_{\pm 3.4}$	$80.9_{\pm 0.8}$	$67.6_{\pm 1.9}$	$82.5_{\pm 0.7}$	$73.5_{\pm 0.9}$	$84.7_{\pm 0.7}$	$84.7_{\pm 0.7}$	$92.6_{\pm 0.3}$	$8.2_{\pm 0.6}$
Entropy	$10.4_{\pm 0.0}$	$86.8_{\pm 1.5}$	$68.2_{\pm 3.3}$	$80.5_{\pm 1.9}$	$66.8_{\pm 3.8}$	$82.2_{\pm 1.6}$	$73.2_{\pm 2.3}$	$84.8_{\pm 1.1}$	$84.8_{\pm 1.1}$	$92.6_{\pm 0.7}$	$8.5_{\pm 1.4}$
CoreSet	$36.4_{\pm 0.0}$	$79.3_{\pm 5.4}$	$47.5_{\pm 14.4}$	$71.7_{\pm 5.9}$	$53.0_{\pm 9.2}$	$73.1_{\pm 6.3}$	$60.2_{\pm 8.8}$	$76.0_{\pm 6.2}$	$76.0_{\pm 6.2}$	$85.7_{\pm 5.1}$	$13.9_{\pm 3.4}$
BADGE	$12.8_{\pm 0.0}$	$87.3_{\pm 1.0}$	$68.1_{\pm 3.1}$	$81.1_{\pm 1.3}$	$67.8_{\pm 2.5}$	$82.7_{\pm 1.1}$	$73.9_{\pm 1.6}$	$85.1_{\pm 0.9}$	$85.1_{\pm 0.9}$	$92.9_{\pm 0.5}$	$8.4_{\pm 0.9}$
DIAL	$12.8_{\pm 0.0}$	$87.0_{\pm 1.0}$	$74.0_{\pm 1.7}$	$80.7_{\pm 1.1}$	$66.0_{\pm 2.3}$	$82.8_{\pm 1.0}$	$74.4_{\pm 1.3}$	$86.4_{\pm 0.5}$	$86.4_{\pm 0.5}$	$93.5_{\pm 0.4}$	$9.4_{\pm 1.0}$

C.1.2 CelebA

Algorithm	% of samples	Avg Acc.	Worst Acc.	Avg Prec.	Worst Prec.	Avg F1.	Worst F1.	Adjusted Acc.	Balanced Acc.	AUROC	ECE
Random	$14.5_{\pm 0.0}$	$84.7_{\pm 0.4}$	$59.5_{\pm 3.1}$	$81.7_{\pm 0.4}$	$76.3_{\pm 1.2}$	$79.5_{\pm 1.0}$	$69.2_{\pm 1.8}$	$77.8_{\pm 1.4}$	$78.0_{\pm 1.3}$	$91.0_{\pm 0.4}$	$6.8_{\pm 0.9}$
Margin	$15.0_{\pm 0.0}$	$86.0_{\pm 0.2}$	$64.3_{\pm 3.0}$	$83.3_{\pm 0.3}$	$78.4_{\pm 1.2}$	$81.4_{\pm 0.5}$	$72.1_{\pm 0.9}$	$79.8_{\pm 0.8}$	$80.0_{\pm 0.8}$	$92.2_{\pm 0.2}$	$4.7_{\pm 1.0}$
Certainty	$15.0_{\pm 0.0}$	$86.0_{\pm 0.3}$	$65.3_{\pm 3.1}$	$83.2_{\pm 0.4}$	$78.0_{\pm 0.7}$	$81.5_{\pm 0.5}$	$72.4_{\pm 0.8}$	$80.1_{\pm 0.6}$	$80.2_{\pm 0.6}$	$92.2_{\pm 0.3}$	$5.0_{\pm 0.7}$
Entropy	$15.0_{\pm 0.0}$	$86.0_{\pm 0.2}$	$65.0_{\pm 3.5}$	$83.2_{\pm 0.2}$	$78.0_{\pm 1.3}$	$81.5_{\pm 0.6}$	$72.4_{\pm 1.1}$	$80.1_{\pm 1.0}$	$80.2_{\pm 1.0}$	$92.2_{\pm 0.1}$	$4.8_{\pm 0.9}$
CoreSet	$10.8_{\pm 0.0}$	$82.8_{\pm 0.9}$	$55.2_{\pm 4.3}$	$79.1_{\pm 1.2}$	$72.4_{\pm 2.1}$	$76.9_{\pm 1.3}$	$65.2_{\pm 2.2}$	$75.1_{\pm 1.5}$	$75.4_{\pm 1.5}$	$88.8_{\pm 1.0}$	$5.9_{\pm 1.0}$
BADGE	$15.0_{\pm 0.0}$	$86.2_{\pm 0.3}$	$64.1_{\pm 3.2}$	$83.7_{\pm 0.2}$	$79.0_{\pm 0.7}$	$81.6_{\pm 0.6}$	$72.5_{\pm 1.0}$	$80.1_{\pm 0.8}$	$80.2_{\pm 0.8}$	$92.4_{\pm 0.2}$	$5.4_{\pm 0.5}$
DIAL	$11.0_{\pm 0.0}$	$85.6_{\pm 0.2}$	$64.5_{\pm 2.3}$	$82.9_{\pm 0.2}$	$78.0_{\pm 0.5}$	$80.9_{\pm 0.3}$	$71.3_{\pm 0.6}$	$79.2_{\pm 0.5}$	$79.4_{\pm 0.5}$	$91.8_{\pm 0.1}$	$6.2_{\pm 0.2}$

C.2 Validation worst-group accuracy

C.2.1 Waterbirds

Algorithm	% of samples	Avg Acc.	Worst Acc.	Avg Prec.	Worst Prec.	Avg F1.	Worst F1.	Adjusted Acc.	Balanced Acc.	AUROC	ECE
Random	$41.9_{\pm 0.0}$	$82.0_{\pm 1.4}$	$55.2_{\pm 2.9}$	$74.7_{\pm 1.5}$	$57.2_{\pm 2.6}$	$76.3_{\pm 1.5}$	$64.8_{\pm 2.0}$	$79.4_{\pm 1.3}$	$79.4_{\pm 1.3}$	$88.8_{\pm 1.1}$	$12.3_{\pm 1.3}$
Margin	$10.4_{\pm 0.0}$	$86.9_{\pm 1.1}$	$69.5_{\pm 2.9}$	$80.6_{\pm 1.5}$	$66.8_{\pm 2.8}$	$82.4_{\pm 1.2}$	$73.6_{\pm 1.7}$	$85.1_{\pm 0.9}$	$85.1_{\pm 0.9}$	$92.8_{\pm 0.5}$	$8.6_{\pm 0.9}$
Certainty	$10.4_{\pm 0.0}$	$87.1_{\pm 0.6}$	$67.1_{\pm 3.4}$	$80.9_{\pm 0.8}$	$67.6_{\pm 1.9}$	$82.5_{\pm 0.7}$	$73.5_{\pm 0.9}$	$84.7_{\pm 0.7}$	$84.7_{\pm 0.7}$	$92.6_{\pm 0.3}$	$8.2_{\pm 0.6}$
Entropy	$10.4_{\pm 0.0}$	$86.8_{\pm 1.5}$	$68.2_{\pm 3.3}$	$80.5_{\pm 1.9}$	$66.8_{\pm 3.8}$	$82.2_{\pm 1.6}$	$73.2_{\pm 2.3}$	$84.8_{\pm 1.1}$	$84.8_{\pm 1.1}$	$92.6_{\pm 0.7}$	$8.5_{\pm 1.4}$
CoreSet	$36.4_{\pm 0.0}$	$79.3_{\pm 5.4}$	$47.5_{\pm 14.4}$	$71.7_{\pm 5.9}$	$53.0_{\pm 9.2}$	$73.1_{\pm 6.3}$	$60.2_{\pm 8.8}$	$76.0_{\pm 6.2}$	$76.0_{\pm 6.2}$	$85.7_{\pm 5.1}$	$13.9_{\pm 3.4}$
BADGE	$12.8_{\pm 0.0}$	$87.3_{\pm 1.0}$	$68.1_{\pm 3.1}$	$81.1_{\pm 1.3}$	$67.8_{\pm 2.5}$	$82.7_{\pm 1.1}$	$73.9_{\pm 1.6}$	$85.1_{\pm 0.9}$	$85.1_{\pm 0.9}$	$92.9_{\pm 0.5}$	$8.4_{\pm 0.9}$
DIAL	$12.8_{\pm 0.0}$	$87.0_{\pm 1.0}$	$74.0_{\pm 1.7}$	$80.7_{\pm 1.1}$	$66.0_{\pm 2.3}$	$82.8_{\pm 1.0}$	$74.4_{\pm 1.3}$	$86.4_{\pm 0.5}$	$86.4_{\pm 0.5}$	$93.5_{\pm 0.4}$	$9.4_{\pm 1.0}$

C.2.2 CelebA

Algorithm	% of samples	Avg Acc.	Worst Acc.	Avg Prec.	Worst Prec.	Avg F1.	Worst F1.	Adjusted Acc.	Balanced Acc.	AUROC	ECE
Random	$14.5_{\pm 0.0}$	$84.7_{\pm 0.4}$	$59.5_{\pm 3.1}$	$81.7_{\pm 0.4}$	$76.3_{\pm 1.2}$	$79.5_{\pm 1.0}$	$69.2_{\pm 1.8}$	$77.8_{\pm 1.4}$	$78.0_{\pm 1.3}$	$91.0_{\pm 0.4}$	$6.8_{\pm 0.9}$
Margin	$14.5_{\pm 0.0}$	$86.1_{\pm 0.3}$	$64.1_{\pm 2.5}$	$83.4_{\pm 0.2}$	$78.3_{\pm 0.6}$	$81.6_{\pm 0.5}$	$72.4_{\pm 1.0}$	$80.1_{\pm 0.8}$	$80.2_{\pm 0.8}$	$92.2_{\pm 0.2}$	$4.8_{\pm 0.5}$
Certainty	$15.0_{\pm 0.0}$	$86.0_{\pm 0.3}$	$65.3_{\pm 3.1}$	$83.2_{\pm 0.4}$	$78.0_{\pm 0.7}$	$81.5_{\pm 0.5}$	$72.4_{\pm 0.8}$	$80.1_{\pm 0.6}$	$80.2_{\pm 0.6}$	$92.2_{\pm 0.3}$	$5.0_{\pm 0.7}$
Entropy	$13.3_{\pm 0.0}$	$85.9_{\pm 0.1}$	$64.4_{\pm 1.1}$	$82.9_{\pm 0.3}$	$77.3_{\pm 0.6}$	$81.4_{\pm 0.2}$	$72.3_{\pm 0.3}$	$80.1_{\pm 0.2}$	$80.3_{\pm 0.2}$	$92.0_{\pm 0.2}$	$4.9_{\pm 0.5}$
CoreSet	$10.8_{\pm 0.0}$	$82.8_{\pm 0.9}$	$55.2_{\pm 4.3}$	$79.1_{\pm 1.2}$	$72.4_{\pm 2.1}$	$76.9_{\pm 1.3}$	$65.2_{\pm 2.2}$	$75.1_{\pm 1.5}$	$75.4_{\pm 1.5}$	$88.8_{\pm 1.0}$	$5.9_{\pm 1.0}$
BADGE	$14.5_{\pm 0.0}$	$86.1_{\pm 0.3}$	$63.8_{\pm 3.4}$	$83.5_{\pm 0.2}$	$78.6_{\pm 0.5}$	$81.5_{\pm 0.6}$	$72.4_{\pm 1.0}$	$80.0_{\pm 0.8}$	$80.1_{\pm 0.8}$	$92.4_{\pm 0.1}$	$5.4_{\pm 0.6}$
DIAL	$11.0_{\pm 0.0}$	$85.6_{\pm 0.2}$	$64.5_{\pm 2.3}$	$82.9_{\pm 0.2}$	$78.0_{\pm 0.5}$	$80.9_{\pm 0.3}$	$71.3_{\pm 0.6}$	$79.2_{\pm 0.5}$	$79.4_{\pm 0.5}$	$91.8_{\pm 0.1}$	$6.2_{\pm 0.2}$

C.3 Validation average accuracy

C.3.1 Waterbirds

Algorithm	% of samples	Avg Acc.	Worst Acc.	Avg Prec.	Worst Prec.	Avg F1.	Worst F1.	Adjusted Acc.	Balanced Acc.	AUROC	ECE
Random	$42.5_{\pm 0.0}$	$82.4_{\pm 1.3}$	$55.4_{\pm 4.4}$	$75.1_{\pm 1.5}$	$58.1_{\pm 2.5}$	$76.8_{\pm 1.5}$	$65.3_{\pm 2.2}$	$79.6_{\pm 1.5}$	$79.6_{\pm 1.5}$	$89.0_{\pm 1.1}$	$11.9_{\pm 1.1}$
Margin	$17.7_{\pm 0.0}$	$87.9_{\pm 0.7}$	$65.1_{\pm 2.8}$	$82.0_{\pm 1.0}$	$70.0_{\pm 2.2}$	$83.3_{\pm 0.8}$	$74.5_{\pm 1.0}$	$85.0_{\pm 0.6}$	$85.0_{\pm 0.6}$	$93.0_{\pm 0.3}$	$8.3_{\pm 0.7}$
Certainty	$13.5_{\pm 0.0}$	$87.6_{\pm 0.8}$	$65.4_{\pm 2.3}$	$81.6_{\pm 1.0}$	$69.1_{\pm 2.0}$	$83.0_{\pm 0.9}$	$74.1_{\pm 1.2}$	$84.9_{\pm 0.7}$	$84.9_{\pm 0.7}$	$92.9_{\pm 0.3}$	$7.9_{\pm 0.7}$
Entropy	$11.0_{\pm 0.0}$	$87.6_{\pm 1.0}$	$67.1_{\pm 4.4}$	$81.5_{\pm 1.3}$	$68.7_{\pm 2.0}$	$83.0_{\pm 1.3}$	$74.2_{\pm 1.9}$	$85.1_{\pm 1.4}$	$85.1_{\pm 1.4}$	$92.8_{\pm 0.6}$	$7.8_{\pm 0.6}$
CoreSet	$38.9_{\pm 0.0}$	$80.0_{\pm 5.2}$	$44.7_{\pm 13.8}$	$72.3_{\pm 5.9}$	$54.4_{\pm 9.3}$	$73.6_{\pm 6.2}$	$60.6_{\pm 8.7}$	$75.9_{\pm 6.0}$	$75.9_{\pm 6.0}$	$85.9_{\pm 4.8}$	$13.2_{\pm 3.5}$
BADGE	$16.5_{\pm 0.0}$	$87.6_{\pm 0.5}$	$64.9_{\pm 1.1}$	$81.5_{\pm 0.8}$	$69.0_{\pm 1.4}$	$82.9_{\pm 0.6}$	$74.0_{\pm 0.9}$	$84.7_{\pm 0.5}$	$84.7_{\pm 0.5}$	$92.8_{\pm 0.1}$	$8.3_{\pm 0.4}$
DIAL	$20.7_{\pm 0.0}$	$88.1_{\pm 0.5}$	$69.2_{\pm 1.5}$	$82.2_{\pm 0.8}$	$69.7_{\pm 1.7}$	$83.7_{\pm 0.6}$	$75.3_{\pm 0.7}$	$85.9_{\pm 0.3}$	$85.9_{\pm 0.3}$	$93.5_{\pm 0.2}$	$8.2_{\pm 0.5}$

C.3.2 CelebA

Algorithm	% of samples	Avg Acc.	Worst Acc.	Avg Prec.	Worst Prec.	Avg F1.	Worst F1.	Adjusted Acc.	Balanced Acc.	AUROC	ECE
Random	$14.8_{\pm 0.0}$	$84.8_{\pm 0.3}$	$59.3_{\pm 1.1}$	$81.8_{\pm 0.5}$	$76.4_{\pm 1.1}$	$79.6_{\pm 0.4}$	$69.4_{\pm 0.6}$	$77.9_{\pm 0.5}$	$78.1_{\pm 0.5}$	$91.0_{\pm 0.3}$	$6.6_{\pm 0.5}$
Margin	$14.5_{\pm 0.0}$	$86.1_{\pm 0.3}$	$64.1_{\pm 2.5}$	$83.4_{\pm 0.2}$	$78.3_{\pm 0.6}$	$81.6_{\pm 0.5}$	$72.4_{\pm 1.0}$	$80.1_{\pm 0.8}$	$80.2_{\pm 0.8}$	$92.2_{\pm 0.2}$	$4.8_{\pm 0.5}$
Certainty	$14.2_{\pm 0.0}$	$85.9_{\pm 0.3}$	$63.9_{\pm 2.5}$	$83.3_{\pm 0.2}$	$78.4_{\pm 0.6}$	$81.2_{\pm 0.6}$	$71.9_{\pm 1.1}$	$79.7_{\pm 0.9}$	$79.8_{\pm 0.9}$	$92.1_{\pm 0.3}$	$5.3_{\pm 0.5}$
Entropy	$14.5_{\pm 0.0}$	$86.0_{\pm 0.2}$	$63.9_{\pm 2.2}$	$83.3_{\pm 0.2}$	$78.1_{\pm 0.7}$	$81.5_{\pm 0.4}$	$72.3_{\pm 0.8}$	$80.0_{\pm 0.6}$	$80.2_{\pm 0.6}$	$92.1_{\pm 0.1}$	$4.7_{\pm 0.5}$
CoreSet	$7.8_{\pm 0.0}$	$82.8_{\pm 1.1}$	$52.2_{\pm 3.3}$	$79.6_{\pm 1.7}$	$74.2_{\pm 3.0}$	$76.4_{\pm 1.4}$	$64.0_{\pm 2.1}$	$74.3_{\pm 1.4}$	$74.5_{\pm 1.3}$	$88.8_{\pm 1.4}$	$6.5_{\pm 0.9}$
BADGE	$14.5_{\pm 0.0}$	$86.1_{\pm 0.3}$	$63.8_{\pm 3.4}$	$83.5_{\pm 0.2}$	$78.6_{\pm 0.5}$	$81.5_{\pm 0.6}$	$72.4_{\pm 1.0}$	$80.0_{\pm 0.8}$	$80.1_{\pm 0.8}$	$92.4_{\pm 0.1}$	$5.4_{\pm 0.6}$
DIAL	$14.8_{\pm 0.0}$	$85.9_{\pm 0.2}$	$63.7_{\pm 1.8}$	$83.6_{\pm 0.2}$	$79.4_{\pm 0.6}$	$81.1_{\pm 0.3}$	$71.5_{\pm 0.6}$	$79.2_{\pm 0.5}$	$79.4_{\pm 0.5}$	$92.2_{\pm 0.1}$	$6.5_{\pm 0.5}$