Navigating the Pitfalls of Active Learning Evaluation: A Systematic Framework for Meaningful Performance Assessment

Conference Paper

Author(s):
Lüth, Carsten; Bungert, Till; Klein, Lukas; Jäger, Paul F.

Publication date:
2023

Permanent link:
https://doi.org/10.3929/ethz-b-000652157

Navigating the Pitfalls of Active Learning Evaluation: A Systematic Framework for Meaningful Performance Assessment

Carsten T. Lüth[1,2] **Till J. Bungert**[1,2] **Lukas Klein**[1,2,3] **Paul F. Jaeger**[1,2]

[1]German Cancer Research Center (DKFZ) Heidelberg, Interactive Machine Learning Group, Germany
[2]Helmholtz Imaging, German Cancer Research Center (DKFZ) Heidelberg, Heidelberg, Germany
[3]Institute for Machine Learning, ETH Zürich, Switzerland

carsten.lueth@dkfz-heidelberg.de

Abstract

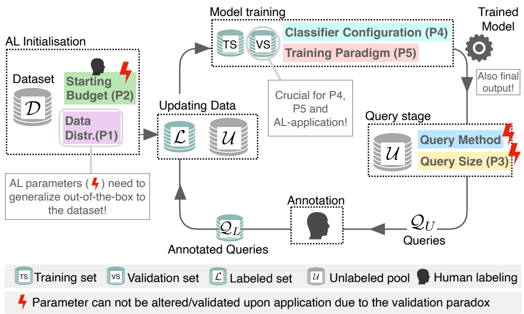
Active Learning (AL) aims to reduce the labeling burden by interactively selecting the most informative samples from a pool of unlabeled data. While there has been extensive research on improving AL query methods in recent years, some studies have questioned the effectiveness of AL compared to emerging paradigms such as semi-supervised (Semi-SL) and self-supervised learning (Self-SL), or a simple optimization of classifier configurations. Thus, today's AL literature presents an inconsistent and contradictory landscape, leaving practitioners uncertain about whether and how to use AL in their tasks. In this work, we make the case that this inconsistency arises from a lack of systematic and realistic evaluation of AL methods. Specifically, we identify five key pitfalls in the current literature that reflect the delicate considerations required for AL evaluation. Further, we present an evaluation framework that overcomes these pitfalls and thus enables meaningful statements about the performance of AL methods. To demonstrate the relevance of our protocol, we present a large-scale empirical study and benchmark for image classification spanning various data sets, query methods, AL settings, and training paradigms. Our findings clarify the inconsistent picture in the literature and enable us to give hands-on recommendations for practitioners. The benchmark is hosted at https://github.com/IML-DKFZ/realistic-al.

1 Introduction

Active Learning (AL) is a popular approach to efficiently label large pools of data by interactively querying the most informative samples for training a classification system. While some parts of the AL community actively propose new query methods (QMs) to advance the field [14, 30], others report AL to be generally outperformed by alternative training paradigms to standard training (ST) such as Semi-Supervised Learning (Semi-SL) [20, 43] and Self-Supervised Learning (Self-SL) [6], or even by well-configured standard baselines [44]. To add to the confusion, further studies have shown that AL can even decrease classification performance in certain settings, a phenomenon referred to as the "cold start problem" [6, 20, 43].

This heterogeneous state of research in AL poses a significant challenge for anyone seeking to efficiently annotate their dataset and facing the questions: *On which tasks and datasets is AL beneficial? And how to best employ AL on my dataset?*

37th Conference on Neural Information Processing Systems (NeurIPS 2023).



(a)

Pitfall	P1	P2	P3	P4		P5	
Compared Aspect	Data	Starting	Query	Perform.	HP Optim.	Self	Semi
Related Work	Distr.	Budget	Size	Baselines	& Val Split	SL	SL
Munjal et al. [44]	(✔)		✔	✔	✔		
Mittal et al. [43]		(✔)		(✔)			✔
Bengar et al. [6]		✔	(✔)			✔	
Gao et al. [20]		✔	(✔)				✔
Yi et al. [54]	✔					(✔)	
Krishnan et al. [35]	(✔)						
Kim et al. [30]	(✔)	(✔)	(✔)				
Beck et al. [4]		(✔)	(✔)	(✔)			
Zhan et al. [58]	✔		✔				
Chan et al. [11]				✔		✔	✔
Ours	✔	✔	✔	✔	✔	✔	✔

(b)

Figure 1: **(a)**: The five pitfalls (P1-P5) for meaningful evaluation in the context of the Active Learning loop. Detailed information is provided in Sec. 2. **(b)**: The five pitfalls are highly prevalent in the current literature (green ticks denote successful avoidance of the respective pitfall). A detailed correspondance between individual studies and pitfalls is provided in Appendix B. Our study is the first to avoid all pitfalls and enable trustworthy performance assessment of AL methods.

We argue that the inconsistency arises from a lack of systematic and realistic evaluation of AL methods. AL inherently comes with specific requirements on how methods will be applied in practice. First and foremost, the stated purpose of "reducing the labeling effort on a task" implies that AL methods need to be rolled out to unseen datasets different from the labeled development data. This inherent requirement of cross-task generalization needs to be reflected in method evaluation, posing a need to test methods under diverse settings to identify robust and trustworthy configurations for the subsequent "blind" application on real-life tasks (see "validation paradox", Sec. 2.2). However, such considerations are generally neglected in AL research, as identified in our work by means of five key pitfalls in the current literature, spanning from a lack of tested AL settings and tasks to a lack of appropriate baselines (see Figure 1 and P1-P5 in Sec. 2.3).

To this end, we present an evaluation framework for deep active classification that overcomes the five pitfalls and demonstrate the relevance of this contribution by means of a large-scale empirical study spanning various datasets, QMs, AL settings, and training paradigms. To give one concrete example for the disruptive impact of our work: We address the widespread pitfall of neglecting classifier configuration (see P4 in Figure 1b) by introducing a light-weight protocol for high-quality configuration. The results demonstrate how this simple protocol lets even our random-query baseline exceed the originally reported performance of recent AL methods [1, 30, 35] on the widely-used CIFAR-10/100 datasets, while drastically reducing computational effort compared to the configuration protocol of Munjal et al. [44].

This example showcases, that the novelty of our work does not lie in presenting entirely novel results and insights, but in the fact that our comprehensive and systematic approach is the first to address all five key pitfalls and thus to provide *trustworthy* insights into the real-life capabilities of current AL methods. By relating these insights to recent studies that have been subject to flawed evaluation, we are able to resolve existing inconsistencies in the field and provide robust guidelines for when and how to apply AL on a given task.

2 Realistic Evaluation in Active Learning

2.1 Active Learning Task Formulation

As depicted in Figure 1a, AL describes a classification task, where a dataset $\mathcal{D}$ is given that is divided into a labeled set $\mathcal{L}$ and an unlabeled pool $\mathcal{U}$. Initially, only a fraction of the data is labelled ("starting budget"), which is ideally split further into a training set and a validation set for hyperparameter (HP) tuning and performance monitoring. After initial training, the QM is used to generate queries $\mathcal{Q}_{\mathcal{U}}$ of a certain amount ("query size") that represent the most informative samples from $\mathcal{U}$ based on the current classifier predictions. Subsequently, queried samples are labeled ($\mathcal{Q}_{\mathcal{L}}$), moved from $\mathcal{U}$ to $\mathcal{L}$, and the classifier is re-trained on $\mathcal{L}$. This process is repeated until classifier performance is satisfying, as monitored on the validation set.

2.2 Overview over critical concepts in Active Learning evaluation

Evaluating an AL algorithm typically means testing how much classification performance is gained by data samples queried over several training iterations. The QM selecting those samples is considered useful if the performance gains exceed the gains of randomly queried samples. While this process is well-established, it is prone to neglecting critical concepts for evaluation and thus to over-simplification of how AL algorithms are applied in practice.

AL validation paradox. When applying AL on a new, mostly unlabelled dataset, one can not directly validate which QM type or configuration performs best, as this would require labeling individual AL trajectories through the dataset for each setting. This excessive labeling would directly contradict the goal of AL to reduce the labeling effort, a predicament which we refer to as the AL validation paradox. Notably, this is in contrast to standard ML, where fixed labeled training and validation splits generally suffices for model selection and configuration.

Special requirements on evaluation. As the described paradox impedes on-the-spot validation, it forces one to, instead, estimate how well certain QMs will perform on the given task based solely on prior knowledge. The quality of this prior knowledge depends on how extensively the respective QM has been validated prior to application, i.e. on the development data. This implies several critical requirements on evaluation:

- A generalization gap between development and application arises, because the latter typically comes with practical constraints such as on the size of the starting budget, the query size, or a given class imbalance. To avoid AL failure, such constraints need to be anticipated by *evaluating QMs on a realistic range of the corresponding design choices*.
- Given these inherent generalization requirements in AL, it is crucial to simulate the generalization process by evaluating QMs on *roll-out datasets*, i.e. tasks that were not part of the configuration process on the *development datasets*.
- Evaluation needs to acknowledge that AL is an elaborate and costly training paradigm, and thus ensure that methods are *tested against simpler and cheaper alternatives*. For instance, if the gain of a QM over random sampling vanishes by simply tuning the learning rate of the classifier, or by a single self-supervised pretraining, the QM, and AL in general, provides no practical value.

To effectively describe the current oversight of these requirements in the field, we identified five concrete evaluation pitfalls (P1-P5), which are detailed in Section 2.3.

Cold start problem. Neglecting these pitfalls could render AL ineffective, or even counterproductive, for specific tasks, as it might result in QMs underperforming relative to a random-sampling baseline. Such failures occur predominantly in few-label settings and are a recognized challenge in the research community, often termed as the cold start problem [20].

2.3 Current pitfalls of Active Learning evaluation

The current AL literature features an inconsistent landscape of evaluation protocols, but, as Figure 1b shows, none of them adhere to all requirements for evaluation described above. To study the current oversight of these requirements, we identify five key *pitfalls* (P1-P5) that need to be overcome for meaningful evaluation of QMs. For a visual overview of the pitfalls and how they integrate into the AL setting see Figure 1a.

P1: Lack of evaluated data distribution settings. To ensure that QMs work out-of-the-box in real-life settings, they need to be evaluated on a broad data distribution. Relevant aspects of a distribution in the AL-context go beyond the data domain and include class distribution, the relative difficulty of separation across classes, as well as a potential mismatch between the frequency and importance of classes. All of these aspects directly affect the functionality of a QM and may lead to real-life failure of AL when not considered in the evaluation.

Current practice: Most current work is limited to evaluating QMs on balanced datasets from one specific domain (e.g. CIFAR-10/100) and under the assumption of equal class importance. To our knowledge, testing the generalizability of a fixed QM setting to new datasets ("roll-out") has not been performed before. There are some experiments conducted on an artificially imbalanced dataset

(CIFAR-LT) [35, 44] suggesting good AL performance. Further, Gal et al. [19] study AL on the ISIC-2016 dataset, but obtain volatile results due to the small dataset size. Atighehchian et al. [3] study AL on the MIO-TCD dataset and reported performance improvements for underrepresented classes. In the large study Beck et al. [4] perform experiments on five mostly balanced datasets datasets and [58] use 13 datasets for mostly standard AL experiments.

→ *Proposed solution:* We argue that the underrepresentation of class-imbalanced datasets in the field is one reason for current doubts regarding the general functionality of AL. Real-life settings will most likely not be class balanced providing a natural advantage of AL over random sampling. We propose to consider diverse datasets with real class imbalances as an essential part of AL evaluation and advocate for the inclusion of "roll-out" datasets, as a real-life test of selected and fixed AL settings.

P2: Lack of evaluated starting budgets. There are two reasons for why this parameter is an essential aspect of AL evaluation: 1) Upon application, the budget might be fixed and the QM is required to generalize out-of-the-box to this setting. 2) We are interested in the minimal budget at which the QM works since a too large budget implies inefficient labeling (equivalent to random queries) and a too small budget is likely to cause AL failure (cold start problem). This search needs to be performed prior to an AL application due to the validation paradox.

Current practice: Most recent studies evaluate AL on a single starting budget made of thousands of samples on datasets such as CIFAR-10 and CIFAR-100[30, 44, 54, 55]. Information-theoretic publications commonly use a smaller starting budget [19, 32], but typically on even simpler datasets such as MNIST [38]. Beck et al. [4] compare two starting budgets on MNIST reporting no performance drop. On the other hand, some studies benchmarking AL against Semi-SL and Self-SL compare two [6] or three [20, 43] starting budgets often with the conclusion that smaller starting budgets lead to AL failure. Bengar et al. [6] report that there exists a relationship between the number of classes in a task and the optimal starting budget (the intuition being that class number is a proxy for task complexity).

→ *Proposed solution:* To overcome this pitfall and resolve the current contradictions, we evaluate all QMs for three different starting budgets on all datasets. We refer to these settings as the *low-, mid-, and high-label regime*. Extending on the findings of Bengar et al. [6], adequate budget sizes are determined using heuristics based on the number of classes per task.

P3: Lack of evaluated query sizes. The number of samples queried for labelling in each AL iteration is an essential aspect of QM evaluation. This is because, upon application, this parameter might be predefined by the compute-versus-label cost ratio of the respective task (a smaller query size amounts to higher computational efforts but might enable more informed queries and thus less labeling). Since query size cannot be validated on the task at hand due to the validation paradox, the generalizability of QMs to various settings of this parameter needs to be evaluated beforehand.

Current practice: In current literature, there is a concerning disconnect between theoretical and practical papers regarding what constitutes a reasonable query size. Information-theoretical papers typically select the smallest query size possible and QMs such as BatchBALD [32] are specifically designed to simulate reduced query sizes [19, 46]. In contrast, practically-oriented papers usually employ larger query sizes [30, 43, 44, 51, 55], but only in combination with large starting budgets (P2), where cold start problems generally do not occur. Only a few studies perform limited evaluations of varying query sizes. Beck et al. [4] and Zhan et al. [58] report a negligible effect of varying query sizes, but only evaluate in combination with large starting budgets (1000 samples). In line with this, Munjal et al. [44] conclude that the choice of query size does not matter, but only compared two large values (2500 versus 5000 samples) on a fixed large starting budget (5000 samples). Atighehchian et al. [3] come to a similar conclusion, but also only considered a relatively large starting budget (500) for ImageNet-pretrained models on CIFAR-10, where, again, no cold start problem occurs. Bengar et al. [6] employ varying query sizes without further analysis of the parameter.

→ *Proposed solution:* To overcome this pitfall and reliably study the effect of query sizes also in low-label, i.e. high-risk, settings, we evaluate all QMs for three different query sizes in combination with varying starting budgets on all datasets (i.e. as part of the low-, mid-, and high-label regimes). For a specific focus on the effect of query size in the low-label settings, we perform an additional ablation with varying query sizes on a small fixed starting budget.

P4: Neglection of the classifier configuration. As stated in Sec. 2.2, when aiming to draw conclusions about the performance or usefulness of a QM, it is critical that this evaluation be based on well-configured classifiers. Otherwise, performance gains might be attributed to AL that could have been achieved by simple hyperparameter (HP) modifications. Separating a validation split from the training data is a crucial requirement for sound HP tuning.

Current practice: Most studies in AL literature do not report how HPs are obtained and do not mention the use of validation splits [30, 35, 43, 51, 55]. Typically, reported settings are copied from fully labeled data scenarios. In some cases, even the proposed QMs feature delicate HPs without reporting how they were optimized raising the question of whether these settings generalize to new data [30, 51, 55]. Munjal et al. [44] demonstrate how adequate HP tuning on a validation set allows a random query baseline to outperform current QMs under their originally proposed HP settings. However, they run a full grid search for every QM and AL training iteration, which might not be feasible in practice.

$\rightarrow$ *Proposed solution:* To overcome this pitfall and enable meaningful performance assessment of AL methods, we define a validation dataset of a size deducted heuristically from the starting budget. Based on this data, a small selection of HPs (learning rate, weight decay and data augmentations [17]) is tuned only once per AL experiment while training on the starting budget. The limited search space and discarding of multiple tuning iterations result in a lightweight and practically feasible protocol for classifier configuration.

P5: Neglection of alternative training paradigms. Analogously to arguments made in P4, meaningful evaluation of AL requires comparison against alternative approaches that address the same problem. Specifically, the training paradigms Self-SL [13, 25] and Semi-SL [39, 52] have shown strong potential to make efficient use of an unlabeled data pool in a classification task thus alleviating the labeling burden. Additionally to benchmarking AL against Self-SL and Semi-SL, the question arises of whether AL can yield performance gains when combined with these paradigms.

Current practice: While most AL studies do not consider Self-SL and Semi-SL, there are a few recent exceptions: Bengar et al. [6] benchmark AL in combination with Self-SL and conclude that AL only yields gains under sufficiently high starting budgets. However, these results suffer from inadequate classifier configuration (P4). Yi et al. [54] propose a QM in combination with Self-SL, but the employed Self-SL strategy is limited by compatibility with the proposed QM. Further, Gao et al. [20] combine Semi-SL with AL and report superior performance compared to ST for CIFAR-10/100 and ImageNet [48], i.e. the datasets on which Semi-SL methods have been developed. Similarly, Mittal et al. [43] evaluate the combination of Semi-SL and AL on CIFAR-10/100, reporting strong improvements compared to ST and find that AL decreases performance for small starting budgets.

$\rightarrow$ *Proposed solution:* To overcome this pitfall and resolve current inconsistencies, we benchmark all QMs against both Self-SL and Semi-SL, and evaluate the combination of AL with these paradigms. Crucially, we are the first to study these relations as part of a reliable evaluation, i.e. while avoiding all other key pitfalls (P1-P4).

3 Experimental Setup

This section describes the design of our empirical study in light of the proposed improvements for AL evaluation (detailed experimental settings can be found in Appendix D). We first address P1 by extending our evaluation to 5 different datasets, containing different label distributions. Specifically, these datasets include CIFAR-10, CIFAR-100, CIFAR-10 LT, ISIC-2019 and MIO-TCD, where the first three are developmental datasets and the latter two are used exclusively for the proposed roll-out evaluation. Further, we address P2 and P3 by defining three different *label regimes* which we refer to as "low-label", "mid-label" and "high-label" regimes. Starting budgets and query sizes are both set to $5 \times C$, $25 \times C$ and $100 \times C$ for the three label regimes, where C denotes the number of classes. [1] To address P4, we configure our classifiers for all three label regimes based on a validation set five

[1]These deviate for CIFAR-100 [$5 \times C$ (low-), $10 \times C$ (mid-), $50 \times C$ (high-label)] due to a smaller ratio of dataset size to number of classes.

times the size of the starting budget. Further, addressing P4 and P5 we use a ResNet-18 [24] as the backbone in all experiments and optimize the essential HPs for each respective training paradigm. At last, we address P5 by comparing randomly initialized models (standard training) against Self-SL pre-trained models and Semi-SL models.

Compared query methods. In this section, we describe the applied QM in more general terms and refer to Appendix A for details. We focus exclusively on QMs which do not alter the classifier and require no configuration of additional HPs except for bayesian QMs which are based on dropout. Generally, QMs can be divided into two categories as they are either based on uncertainty estimation or enforcing exploration. **Random:** The baseline all QMs are compared against which randomly draws samples from the pool $\mathcal{U}$. **Core-Set:** This explorative QM aims to find the core-set of a convolutional neural network [49] by means of a K-Center Greedy approximation on intermediate representations. **Entropy:** This uncertainty-based QM selects the samples with the highest entropy across predicted class scores [50]. **BALD:** This uncertainty-based QM uses the mutual information between the class label and the model parameters with regard to each sample for greedy selection [27], it was introduced with dropout for deep bayesian active learning[19]. **BADGE:** This QM performs a clustering based on per-sample gradient vectors obtained via proxy labels. This enables a selection that is both diverse and guided by uncertainty [2].

Datasets. The initial datasets for our experiments are **CIFAR-10/100** [36]. For further analysis we added **CIFAR-10 LT** [10], an artificially created dataset built upon CIFAR-10 with a long-tail distribution of classes following an exponential decay for the training split. The imbalance factor ρ was selected to be 50, following [35]. On all three of these datasets, we use accuracy on the class-balanced test set as the primary performance metric. Finally, we selected **ISIC-2019**[15, 16, 53] and **MIO-TCD**[42] as roll-out datasets to verify the generalizability of AL methods to new settings. Both of these datasets feature inherently imbalanced class distributions, and are more likely to be subject to label noise compared to the CIFAR datasets. As such, we deem the two roll-out datasets an essential step toward a realistic evaluation of AL methods. For both MIO-TCD and ISIC-2019, we use balanced accuracy as the primary performance measure (Appendix D).

Active learning setup. We report performance measures for each dataset on identical test splits based on three experiments using different seeded models and different train and validation splits to reduce the possible influence of these parameters on our results. Further, we train the models from scratch on every training step to avoid correlated queries [32].

Training paradigms. Randomly initialized and supervised-trained models are referred to as ST models. Further, we use the popular contrastive SimCLR [13] training as a basis for Self-SL pre-training. These models are fine-tuned and are referred to as Self-SL models. Self-SL models have a two-layer MLP as a classification head to make better use of the representations (ablation in Appendix E). For ST and Self-SL models, we obtain bayesian models by adding dropout to the final representations before the classification head following [19]. As a Semi-SL method, we use FixMatch [52] which combines the principles of consistency and uncertainty reduction. Due to the long training times (factor 80) compared to ST training, we only ran experiments in the low- and mid-label regime while increasing the query size by a factor of three to reduce training costs.

Hyperparameter selection. The HPs for our models are selected for each label regime before the AL loop is started using the corresponding validation set. For Self-SL and ST models we use a fixed training recipe (Scheduler, Optimizer, Batch Size, Epochs) and only optimize learning rate, weight decay and data augmentations. The data augmentations used are standard augmentations and Randaugment which uses stronger augmentations acting as a regularization [17]. For Semi-SL methods we fix all HPs with the exception of learning rate and weight decay following Sohn et al. [52]. Model selection for ST and Self-SL models is based on the best validation set epoch, while for Semi-SL models the final checkpoint is used. For imbalanced datasets, we use oversampling for ST and Self-SL pre-trained models Buda et al. [8] and use weighted cross-entropy-loss and distribution alignment for FixMatch.

Low-Label query size ablation. To investigate the effect of query size in the low-label regime, we conduct an ablation with Self-SL pre-trained models on CIFAR-100 and ISIC-2019. For CIFAR-100 the query sizes are 50, 500 and 2000, while for ISIC-2019 they are 10, 40 and 160.

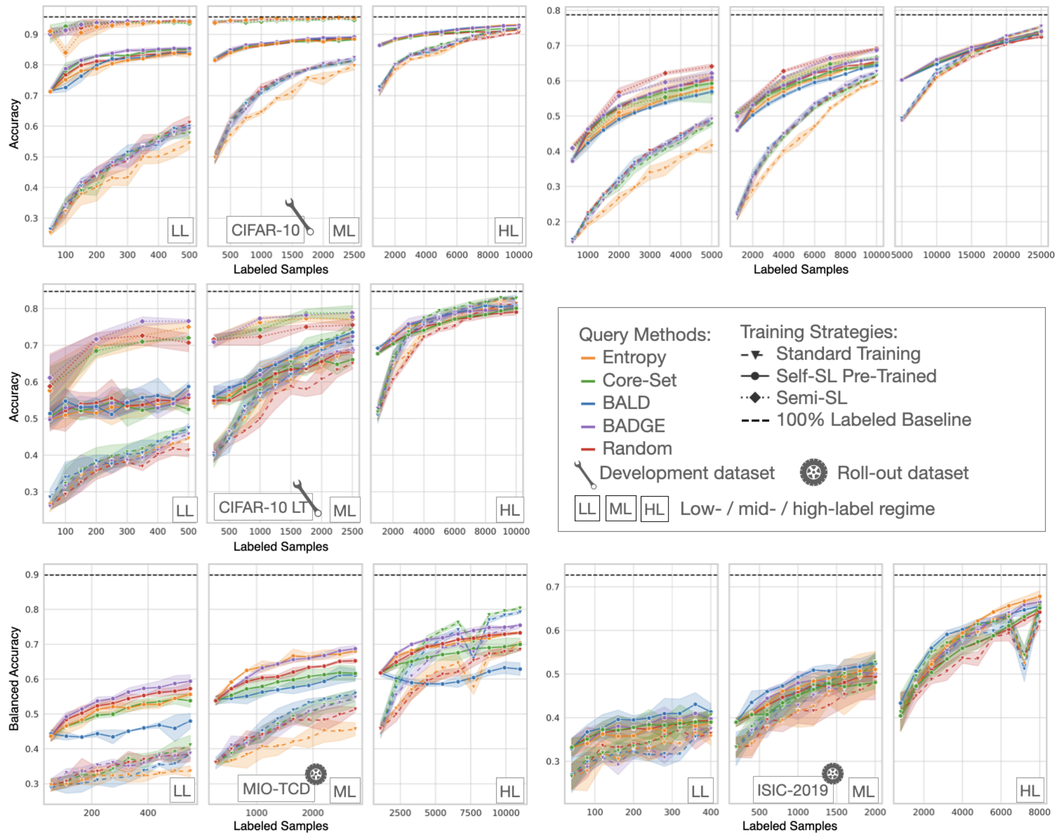


Figure 2: Results obtained with our proposed evaluation protocol over five different datasets and the three label regimes. These experiments are to our knowledge the largest conducted study for AL and reveal insights along the lines of the five key parameters as discussed in Sec. 2. The strong performance dip on MIO-TCD and ISIC-2019 is discussed in Sec. 4.

4 Results & Discussion

The results of our empirical study are shown in Figure 2. An in-depth analysis of results for individual datasets and analysis based on the pairwise penalty matrix [2] and area under the budget curve [57] can be found in Appendix F. Here, we will discuss the main findings along the lines of our five identified pitfalls of evaluation (P1-P5). Our findings demonstrate the relevance of the proposed protocol for realistic evaluation and its potential to generate trustworthy insights on when and how AL works. If not otherwise mentioned, all references in this chapter refer to AL studies.

P1 Data distribution. The proposed evaluation over a diverse selection of dataset distributions including specific roll-out datasets proved essential for realistic evaluation of QMs as well as the different training strategies. One main insight is the fact that class distribution is a crucial predictor for the potential performance gains of AL on a dataset: Performance gains of AL are generally higher on imbalanced datasets and occur consistently even for ST models with a small starting budget, which are typically prone to experience cold start problems. This observation is consistent with a few previous studies [30, 35, 54]. Further, our results underpin the importance of the roll-out datasets e.g. when looking at the sub-random performance of BALD (with Self-SL) and Entropy (with ST) on MIO-TCD. Such worst-case failures of AL application (increased compute and labeling effort due to AL) could not have been predicted based on development data where all AL-parameters are optimized. Another example is the lack of generalizability of Semi-SL, where performance in relation to other Self-SL and ST decreases gradually with data complexity (going from CIFAR-10 to CIFAR-100/-10 LT to the two roll-out datasets MIO-TCD and ISIC-2019.)

P2 Starting budget. The comprehensive study of various starting budgets on all datasets reveals that AL methods are more robust with regard to small starting budgets than previously reported

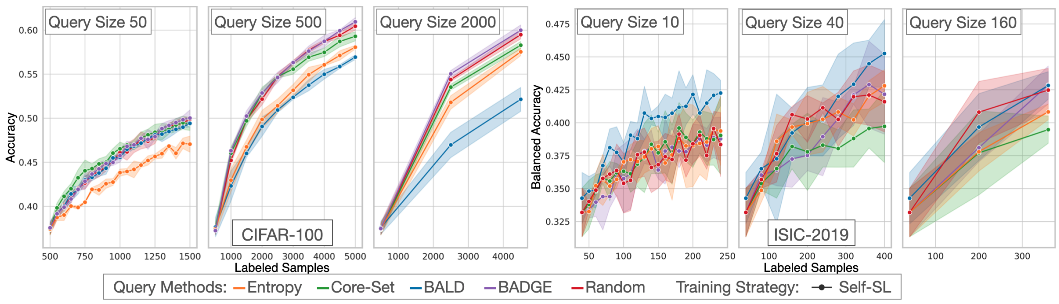


Figure 3: Low-label query size ablation on CIFAR-100 and ISIC-2019. On CIFAR-100, reducing the query size resolves the observed failure mode of BALD. However, no improvement is observed on ISIC-2019, presumably because BALD already shows the best performance. Further, BADGE performs consistently well across all query sizes without failure modes, revealing its robustness also for low-label settings.

[6, 20, 43]. With the exception of Entropy we did not observe cold start problems even for any QM even in combination with notoriously prone ST models. The described robustness is presumably enabled by our thorough classifier configuration (P4) and heuristically adapted query sizes (P3). This finding has great impact potential suggesting that AL can be applied at earlier points in the annotation process thereby further reducing the labeling cost, especially in combination with BADGE which performed consistently better or on par with Random. Similarly for Self-SL models, AL performs well on small starting budgets with the exceptions of CIFAR-100 (BALD, Entropy) and MIO-TCD (BALD, Core-Set, Entropy).

P3 Query size. Based on our evaluation of the query size we can empirically confirm its importance with regard to 1) general AL performance and 2) counteracting the cold start problem. The are, however, surprising findings indicating that the exact interaction between query size and performance remains an open research question. For instance, we observe the cold start problem for BALD with Self-SL on CIFAR-100 ($\sim 50\%$ accuracy at 2k labeled samples for query sizes of 500 (low-label regime) and 1k (mid-label regime). On the other hand, in the high label regime (budget of 5k and query size of 5k) ST and Self-SL models with similar accuracies of 50% and 60%, respectively, benefit from BALD. Since cold start problems are commonly associated with large query sizes, this finding seems counter-intuitive, but has been reported before although without further investigations [6, 20, 43]. To gain a better understanding of this phenomenon and how QMs interact with query size in the low-label regime, we performed a dedicated experiment series for Self-SL training on CIFAR-100 and ISIC-2019 (see Figure 3 and Appendix F.2). Due to a significant improvement of BALD on CIFAR-100 for even smaller query sizes from sub-Random to Random performance and no worsening on ISIC-2019, we can conclude that small query sizes represent an effective countermeasure against the cold-start problem for BALD contrary to the findings of [44, 58] and currently not considered as a solution [6, 20, 43]. This potentially explains the gap between these findings and more theoretical works which advertise using the smallest query size possible [19, 33]. Importantly, while smaller query sizes appear as a potential solution for the observed instabilities of BALD, they seem to have no considerable effect on the other QMs. Moreover, our ablation reveals that even in the low-label settings, BADGE is the overall most reliable of all compared QMs and exhibits no sub-Random performance. This adds to the existing reports of the robustness of BADGE for higher label regimes [4, 58].

P4 Classifer configuration. Our results show that method configuration on a properly sized validation set is essential for realistic evaluation in AL. For instance, our method configuration has the same effect on the classifier performance as increasing the number of labeled training samples by a factor of ~ 5 (e.g. our ST model reach approximately the same accuracy of $\sim 44\%$ trained on 200 samples compared to models in the AL study by Bengar et al. [6] trained on 1k samples). The effectiveness of our proposed lightweight HP selection on the starting budget including only three parameters (Sec. 3) is demonstrated by the fact that all our ST models substantially outperform respective models found in relevant literature [6, 20, 30, 35, 43, 54, 55] where HP optimization is generally neglected. Details for this comparison are provided in Appendix H. This raises the question of to which extent reported AL advantages could have been achieved by simple classifier configurations. Further, our models also generally outperform expensively configured models by Munjal et al. [44]. Thus, we conclude that manually constraining the search space renders HP optimization feasible in practice without decreasing performance and ensures performance gains by Active Learning are not overstated. The importance of the proposed strategy to optimize HPs on the starting budget for each new dataset is supported by the fact that the resulting configurations change across datasets.

P5 Alternative training paradigms. Based on our study benchmarking AL in the context of both Self-SL and Semi-SL, we see that while Self-SL generally leads to improvements across all experiments, Semi-SL only leads to considerably improved performance on the simpler datasets CIFAR-10/100, on which Semi-SL methods are typically developed. Generally, models trained with either of the two training paradigms receive a lower performance gain from AL (over random querying) compared to ST. Crucially, Self-SL models converge around 2.5 times faster than ST models, while the training time of Semi-SL models is around 80 times longer than ST and often yields only small benefits over Self-SL or ST models. The fact that AL entails multiple training iterations amplifies the computational burden of Semi-SL, rendering their combination prohibitively expensive in most practical scenarios. Further, the fact that our Semi-SL models based on Fixmatch do not seem to generalize to more complex datasets in our setting stands in stark contrast to conclusions drawn by [20, 43] as to which the emergence of Semi-SL renders AL redundant. Interestingly, the exact settings where Semi-SL does not provide benefits in our study are the ones where AL proved advantageous. The described contradiction with the literature underlines the importance of our proposed protocol testing for a method's generalizability to unseen datasets. This is especially critical for Semi-SL, which is known for unstable performance on noisy and class imbalanced datasets as noted by studies focusing on Semi-SL [7, 22, 23, 29, 45, 56].

Limitations. We propose a light-weight and practically feasible strategy for HP optimization and made other design choices (e.g. ResNet-18 classifier), thus we can not guarantee that our configurations are optimal for all compared training paradigms and would like to provide a critical discussion: **1)** The ResNet-18 in combination with our shortened training times, might hinder Semi-SL performance more than other training paradigms. This setting is necessary to be able to cope with the computational cost of Semi-SL (factor ~ 400 training time compared with ST). **2)** The validation set size of $5\times$ starting budget size (i.e. training set) could be considered as larger than practically desirable, where most data would be used for training. This design decision follows the study of Oliver et al. [45], showing that an adequately sized validation set is necessary for proper HP selection (especially for Semi-SL). **3)** We observe a performance dip of ST models on MIO-TCD and ISIC-2019 at $\sim$7k samples, which we attribute to our HP selection scheme. This indicates that HPs might need to be re-selected occasionally at certain training iterations. However, such cases are immediately detected in practice allowing for correction where necessary by re-optimizing the HPs (see Appendix I.1). A more extensive discussion of limitations can be found in Appendix I.

5 Conclusion & Take-Aways

Our experiments provide strong empirical evidence that the current evaluation protocols do not sufficiently answer the key question every potential AL practitioner faces: Should I employ AL on my dataset? Answering this question entails estimating whether an AL algorithm will provide performance gains over random queries and thus whether the expected reduction in labeling cost outweighs both the additional computational and engineering cost attributed to AL. We argue that our proposed protocol for realistic evaluation represents a cornerstone towards enabling informed decisions in this context. This is made possible by focusing on evaluating the generalizability of AL to new settings under real-world conditions. This perspective manifests itself in the form of five key pitfalls in the current AL literature regarding data distribution, starting budget, query size, classifier configuration, and alternative training paradigms (see Sec. 2.3). Even though the thorough evaluation increases the computational cost once during method development, it will lead to a net cost reduction in the long-term by enabling an informed selection of AL methods in practice, thereby effectively reducing annotation cost upon application.

We hope that the proposed evaluation protocol in combination with the publicly-available benchmark will help to push active learning towards robust and wide-spread real-world application.

Main empirical insights revealed by our study:

- Assessment of AL methods in the literature is substantially hampered by subpar classifier configurations, but meaningful assessment can be restored by our protocol for lightweight hyperparameter tuning on the starting budget.
- AL generally provides substantial gains in class-imbalanced settings.
- BADGE is the best-performing QM across a realistic range of datasets, starting budgets, and query sizes and exhibits no failure modes (i.e. sub-Random performance).
- Combining AL with Self-SL considerably improves performance, shortens the training time, and stabilizes optimization, especially in low-label settings.
- FixMatch results indicate that Semi-SL methods perform well on datasets where they have been developed, but may struggle to generalize to more realistic scenarios such as class imbalanced data. Combining the paradigm with AL additionally suffers from extensive training times.
- BALD with Self-SL pre-trained models benefits from smaller query sizes on small starting budgets possibly circumventing the "cold start problem".

Take-aways for developing and proposing new Active Learningl algorithms:

AL methods should be tested for generalizability on roll-out datasets and come with a clear recipe for real-world application including how to adapt all design choices to new settings. Since the expected benefit of AL on a new setting increases with lower application costs, we believe there is a high potential for wide-spread real-world use of AL by reducing the two prevalent cost factors: **1)** Engineering costs, which are reduced by building easy-to-use AL tools. **2)** Computational costs, which are reduced by explicitly including methods that shorten the training time in AL.

Take-aways for the cost-benefit analysis of deploying Active Learning:

1. Since we identify BADGE as a robust query method exhibiting no sub-random-sampling performance across all tested settings, the potential harm of AL is minimized allowing to base decisions around deploying AL on the described cost-benefit analysis.
2. The expected benefit is high in settings, where there is a mismatch between the task-specific importance of individual classes and their frequency in the dataset (e.g. class-imbalanced datasets in combination with tasks requiring a balanced classifier).
3. The expected benefit further increases with the labeling cost that will be reduced by AL. AL is thus likely to yield a net benefit in settings with high label cost and low computational and engineering costs.

Future research:

Future research may investigate to what extend AL benefits from foundation models such as CLIP [47], which have been shown to reach strong performance in low-label settings through finetuning or knowledge extraction. Additionally, their rapid re-finetuning during AL iterations may enable real-time interactive labelling. A further important aspect might be to study the effect of AL training on the classifiers' ability to detect outliers [18] or catch failures under distributions shifts [28].

Acknowledgements

This work was funded by Helmholtz Imaging (HI), a platform of the Helmholtz Incubator on Information and Data Science. This work is supported by the Helmholtz Association Initiative and Networking Fund under the Helmholtz AI platform grant (ALEGRA (ZT-I-PF-5-121)). We thank Maximilian Zenk, Sebastian Ziegler and Fabian Isensee for insightful discussions and feedback.

References

[1] Sharat Agarwal, Himanshu Arora, Saket Anand, and Chetan Arora. Contextual diversity for active learning. In *European Conference on Computer Vision*, pages 137–153. Springer, 2020.

[2] Jordan T. Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep Batch Active Learning by Diverse, Uncertain Gradient Lower Bounds. *arXiv:1906.03671 [cs, stat]*, February 2020.

[3] Parmida Atighehchian, Frédéric Branchaud-Charron, and Alexandre Lacoste. Bayesian active learning for production, a systematic study and a reusable library. *arXiv:2006.09916 [cs, stat]*, June 2020.

[4] Nathan Beck, Durga Sivasubramanian, Apurva Dani, Ganesh Ramakrishnan, and Rishabh Iyer. Effective evaluation of deep active learning on image classification tasks. *arXiv preprint arXiv:2106.15324*, 2021.

[5] William H. Beluch, Tim Genewein, Andreas Nurnberger, and Jan M. Kohler. The Power of Ensembles for Active Learning in Image Classification. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9368–9377, Salt Lake City, UT, June 2018. IEEE. ISBN 978-1-5386-6420-9. doi: 10.1109/CVPR.2018.00976.

[6] Javad Zolfaghari Bengar, Joost van de Weijer, Bartlomiej Twardowski, and Bogdan Raducanu. Reducing label effort: Self-supervised meets active learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 1631–1639, October 2021.

[7] Sayedali Shetab Boushehri, Ahmad Bin Qasim, Dominik Waibel, Fabian Schmich, and Carsten Marr. Systematic comparison of incomplete-supervision approaches for biomedical imaging classification. Preprint, Bioinformatics, December 2020.

[8] Mateusz Buda, Atsuto Maki, and Maciej A. Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259, 2018. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2018.07.011. URL `https://www.sciencedirect.com/science/article/pii/S0893608018302107`.

[9] Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122, 2013.

[10] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32, 2019.

[11] Yao-Chun Chan, Mingchen Li, and Samet Oymak. On the Marginal Benefit of Active Learning: Does Self-Supervision Eat its Cake? In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3455–3459, Toronto, ON, Canada, June 2021. IEEE. ISBN 978-1-72817-605-5. doi: 10.1109/ICASSP39728.2021.9414665.

[12] Mark Chen, Alec Radford, Rewon Child, Jeff Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative Pretraining from Pixels. *ICML*, page 13, 2020.

[13] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR, July 2020.

[14] Gui Citovsky, Giulia DeSalvo, Claudio Gentile, Lazaros Karydas, Anand Rajagopalan, Afshin Rostamizadeh, and Sanjiv Kumar. Batch active learning at scale. *Advances in Neural Information Processing Systems*, 34:11933–11944, 2021.

[15] Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pages 168–172. IEEE, 2018.

[16] Marc Combalia, Noel CF Codella, Veronica Rotemberg, Brian Helba, Veronica Vilaplana, Ofer Reiter, Cristina Carrera, Alicia Barreiro, Allan C Halpern, Susana Puig, et al. Bcn20000: Dermoscopic lesions in the wild. *arXiv preprint arXiv:1908.02288*, 2019.

[17] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703, 2020.

[18] Arnav Das, Gantavya Bhatt, Megh Bhalerao, Vianne Gao, Rui Yang, and Jeff Bilmes. Accelerating batch active learning using continual learning techniques. *arXiv preprint arXiv:2305.06408*, 2023.

[19] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *International Conference on Machine Learning*, pages 1183–1192. PMLR, 2017.

[20] Mingfei Gao, Zizhao Zhang, Guo Yu, Sercan Ö. Arık, Larry S. Davis, and Tomas Pfister. Consistency-Based Semi-supervised Active Learning: Towards Minimizing Labeling Cost. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, volume 12355, pages 510–526. Springer International Publishing, Cham, 2020. ISBN 978-3-030-58606-5 978-3-030-58607-2. doi: 10.1007/978-3-030-58607-2_30.

[21] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised Representation Learning by Predicting Image Rotations. *arXiv:1803.07728 [cs]*, March 2018.

[22] Lan-Zhe Guo, Zhen-Yu Zhang, Yuan Jiang, Yu-Feng Li, and Zhi-Hua Zhou. Safe deep semi-supervised learning for unseen-class unlabeled data. In *International Conference on Machine Learning*, pages 3897–3906. PMLR, 2020.

[23] Lan-Zhe Guo, Zhi Zhou, and Yu-Feng Li. Robust deep semi-supervised learning: A brief introduction. *arXiv preprint arXiv:2202.05975*, 2022.

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[25] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum Contrast for Unsupervised Visual Representation Learning. *arXiv:1911.05722 [cs]*, March 2020.

[26] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. In *The collected works of Wassily Hoeffding*, pages 409–426. Springer, 1994.

[27] Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian Active Learning for Classification and Preference Learning. *arXiv:1112.5745 [cs, stat]*, December 2011.

[28] Paul F Jaeger, Carsten Tim Lüth, Lukas Klein, and Till J Bungert. A call to reflect on evaluation practices for failure detection in image classification. In *The Eleventh International Conference on Learning Representations*, 2023.

[29] Jaehyung Kim, Youngbum Hur, Sejun Park, Eunho Yang, Sung Ju Hwang, and Jinwoo Shin. Distribution aligning refinery of pseudo-label for imbalanced semi-supervised learning. *Advances in neural information processing systems*, 33:14567–14579, 2020.

[30] Kwanyoung Kim, Dongwon Park, Kwang In Kim, and Se Young Chun. Task-Aware Variational Adversarial Active Learning. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8162–8171, Nashville, TN, USA, June 2021. IEEE. ISBN 978-1-66544-509-2. doi: 10.1109/CVPR46437.2021.00807.

[31] Andreas Kirsch and Yarin Gal. A Practical & Unified Notation for Information-Theoretic Quantities in ML. *arXiv:2106.12062 [cs, stat]*, December 2021.

[32] Andreas Kirsch, Joost van Amersfoort, and Yarin Gal. BatchBALD: Efficient and diverse batch acquisition for deep bayesian active learning. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[33] Andreas Kirsch, Sebastian Farquhar, Parmida Atighehchian, Andrew Jesson, Frederic Branchaud-Charron, and Yarin Gal. Stochastic Batch Acquisition for Deep Active Learning. *arXiv:2106.12059 [cs, stat]*, January 2022.

[34] Jannik Kossen, Sebastian Farquhar, Yarin Gal, and Tom Rainforth. Active Testing: Sample-Efficient Model Evaluation. *arXiv:2103.05331 [cs, stat]*, June 2021.

[35] Ranganath Krishnan, Nilesh Ahuja, Alok Sinha, Mahesh Subedar, Omesh Tickoo, and Ravi Iyer. Improving Robustness and Efficiency in Active Learning with Contrastive Loss. *arXiv:2109.06873 [cs]*, September 2021.

[36] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[37] Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-Tuning can Distort Pretrained Features and Underperform Out-of-Distribution. *arXiv:2202.10054 [cs]*, February 2022.

[38] Yann LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

[39] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on Challenges in Representation Learning, ICML*, volume 3, page 896, 2013.

[40] Peng Liu, Lizhe Wang, Guojin He, and Lei Zhao. A Survey on Active Deep Learning: From Model-driven to Data-driven, February 2022.

[41] Xiao Liu, Fanjin Zhang, Zhenyu Hou, Li Mian, Zhaoyu Wang, Jing Zhang, and Jie Tang. Self-supervised learning: Generative or contrastive. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–1, 2021. doi: 10.1109/TKDE.2021.3090866.

[42] Zhiming Luo, Frédéric Branchaud-Charron, Carl Lemaire, Janusz Konrad, Shaozi Li, Akshaya Mishra, Andrew Achkar, Justin Eichel, and Pierre-Marc Jodoin. Mio-tcd: A new benchmark dataset for vehicle classification and localization. *IEEE Transactions on Image Processing*, 27 (10):5129–5141, 2018. doi: 10.1109/TIP.2018.2848705.

[43] Sudhanshu Mittal, Maxim Tatarchenko, Özgün Çiçek, and Thomas Brox. Parting with Illusions about Deep Active Learning. *arXiv:1912.05361 [cs]*, December 2019.

[44] Prateek Munjal, Nasir Hayat, Munawar Hayat, Jamshid Sourati, and Shadab Khan. Towards robust and reproducible active learning using neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 223–232, 2022.

[45] Avital Oliver, Augustus Odena, Colin Raffel, Ekin D. Cubuk, and Ian J. Goodfellow. Realistic Evaluation of Deep Semi-Supervised Learning Algorithms. *arXiv:1804.09170 [cs, stat]*, June 2019.

[46] Robert Pinsler, Jonathan Gordon, Eric Nalisnick, and José Miguel Hernández-Lobato. Bayesian Batch Active Learning as Sparse Subset Approximation. *arXiv:1908.02144 [cs, stat]*, February 2021.

[47] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021. URL `https://proceedings.mlr.press/v139/radford21a.html`.

[48] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

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

[50] Burr Settles. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009.

[51] Samarth Sinha, Sayna Ebrahimi, and Trevor Darrell. Variational adversarial active learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[52] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. FixMatch: Simplifying semi-supervised learning with consistency and confidence. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 596–608. Curran Associates, Inc., 2020.

[53] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1): 1–9, 2018.

[54] John Seon Keun Yi, Minseok Seo, Jongchan Park, and Dong-Geol Choi. Using Self-Supervised Pretext Tasks for Active Learning. *arXiv:2201.07459 [cs]*, January 2022.

[55] Donggeun Yoo and In So Kweon. Learning loss for active learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[56] Maximilian Zenk, David Zimmerer, Fabian Isensee, Paul F Jäger, Jakob Wasserthal, and Klaus Maier-Hein. Realistic evaluation of fixmatch on imbalanced medical image classification tasks. In *Bildverarbeitung für die Medizin 2022*, pages 291–296. Springer, 2022.

[57] Xueying Zhan, Huan Liu, Qing Li, and Antoni B Chan. A comparative survey: Benchmarking for pool-based active learning. In *IJCAI*, pages 4679–4686, 2021.

[58] Xueying Zhan, Qingzhong Wang, Kuan-hao Huang, Haoyi Xiong, Dejing Dou, and Antoni B Chan. A comparative survey of deep active learning. *arXiv preprint arXiv:2203.13450*, 2022.