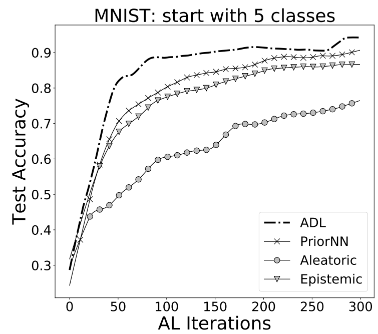
We thank all reviewers for their valuable feedback and constructive suggestions. Major comments are addressed below.

Reply to Reviewer 1

Q1: Need for an evidence-based subjective logic (SL) framework. The SL framework provides the theoretical underpinning to perform a principled, fine-grained analysis between the 1st-order uncertainty (i.e., predictive entropy as the total uncertainty) and 2nd-order uncertainty (vacuity + dissonance), where evidence plays a central role to unveil the underlying (dynamic) relationship among different uncertainties. Understanding this dynamic relationship is essential to derive a theoretically sound data sampling process in AL. In particular, Theorem 1 shows the total uncertainty dynamically shifts between high vacuity and high dissonance as more evidence is collected. Guided by this theory, AL can be regarded as an evidence collection process. The evidence-based uncertainties (i.e., vacuity + dissonance) derived under SL, offer a natural way to determine the sources of uncertainty during AL, which starts by focusing on vacuity in early stage when evidence is limited and then gradually shifts to dissonance. Since vacuity and dissonance both depend on evidence, using evidence provides a principled way to trace the dynamic shift between these two sources of uncertainty to best guide data sampling in AL. This is the **key advantage** over other types of uncertainty, such as epistemic and aleatoric, which only focus a certain aspect of uncertainty, and their (dynamic) relationship is hard to be precisely quantified as in vacuity and dissonance. Thus, using these uncertainty measures lacks the capability to dynamically adjust the sampling process as the nature and focus of uncertainty change when more data samples are labeled. In sum, while there are various forms of uncertainty measures, the evidence-based (2nd-order) uncertainty, i.e., vacuity + dissonance, offers the most suitable way for active sampling, as justified by our theory and empirical evaluation. The right figure includes additional comparison with other uncertainties: epistemic, aleatoric (Kendall & Gal, 2017), and distributional uncertainty of prior networks (Malinin & Gales, 2018). As can be seen, ADL converges much faster than other uncertainty based sampling functions, which empirically confirms its effectiveness in AL. We will report the comparison results on all other datasets in the revised paper.



Q2: Choice of architecture. We choose a relatively simple architecture to demonstrate that the good AL performance is due to the sampling function instead of a strong classifier. Several works (eg, [7] and [11]) follow a similar rationale.

Q3: Experiments of a greater scale. We thank the reviewer for suggesting these large-scale image datasets. Limited by time, we were not able to conduct the experiments on these large datasets and collect the active learning results. An interesting future direction is to combine the architectures suggested by the reviewer and our sampling function and apply to these large datasets.

Reply to Reviewer 3

Q1: What "evidence-based entropy" is when claiming entropy can be decomposed into vacuity and dissonance. Thank you for the suggestion. Entropy decomposition means a high entropy dynamically shifts between a high vacuity and a high dissonance as more evidence is collected, instead of a simple sum of these two uncertainties. We will make this clear in the revised paper.

Q2: Relation to prior work. The prior networks model (Malinin & Gales, 2018) proposes distributional uncertainty (DU) for OOD detection. While DU can be regarded as a type of epistemic uncertainty that can be used for data sampling in AL, the prior network needs to be properly trained as its parameter must encapsulate knowledge of both in-domain distribution and the decision boundary, making it not very suitable for AL. This is also evidenced by our additional comparison result in Fig. (a). The Noise-Contrastive Priors (Hafner et al. 2018) can also be used for OOD detection as it encourages high uncertainty near the boundary of the training data. However, in the initial phase of AL when the training data is very limited, this measure can be insufficient to explore data samples faraway from the training data. R1 (Q1) offers a deeper discussion on why vacuity + dissonance is more effective for AL than other uncertainties.

Reply to Reviewer 4

Q1: Compare to prior networks...ability to capture vacuity. For comparison with prior networks, please refer to Fig. (a) and R3 (Q2). As the model continues to be trained with AL, it remains uncertainty aware but the total uncertainty (especially vacuity) will decrease. Fig.3 (page 7) shows how vacuity changes along with AL.

Q2: Parameter update and complexity. ADL is retrained once a labeled sample is added. Time complexity is $O(training\_size \times epoch)$, which is efficient given the small training size in AL.

Q3: Minor note. DU should be EU; $W = K$ is commonly used for a non-informative prior; it should be $C \to \infty$; β should be non-negative, which is ensured in our experiments. We will fix these typos.