A Detailed settings for the AL experiments on MoBYv2AL

Datasets. For the quantitative evaluation, we put forward four well-known image classification datasets: CIFAR-10, CIFAR-100 [24], SVHN[14] and FashionMNIST[40]. CIFAR-10 and CIFAR-100 contain the same 50000 training examples but with different labelling systems (10 and 100 classes). SVHN and FashionMNIST are separated into ten classes each as CIFAR-10. However, both datasets are larger, with 73257 coloured street numbers and 60000 grayscale images for FashionMNIST. Although CIFAR-10/100 and FashionMNIST have class-balanced data, this is not the case for SVHN. From another perspective, deploying grayscale images from FashionMNIST challenges our contrastive learning approach, previously customised to RGB data.

Models. We mentioned in the Methodology that we use different CNNs for feature encoders. To show that MoBYv2 is robust to architectural changes, we opt for VGG-16 [31] in the CIFAR-10/100 quantitative experiments and for ResNet-18 [18] in SVHN and FashionMNIST. Moreover, for the SSL comparison with CSAL we align the encoder with WideResNet-28[43].

AL settings. Under the exploration-exploitation trade-off, we characterise the budget to select as an exploiting factor while the exploration is captured in the number selection cycles. The initial random-sampled labelled dataset varies between the CIFAR-10/100 experiments and SVHN/FashionMNIST. For CIFAR-10/100, we consider 10% (5000) of the entire training set as labelled and the rest as unlabelled data. The budget is limited to 5% (2500) samples for selection, and we repeat this cycle seven times. In the second set of experiments, we test our method in a more restrictive environment with a starting set of 1000 labelled and a similar fixed budget. Despite this, we expanded the exploration to 10 cycles reaching 10000 labelled data. As a performance measurement, we evaluate the average of 5 trials testing accuracy in the AL framework.

B Selection function analysis

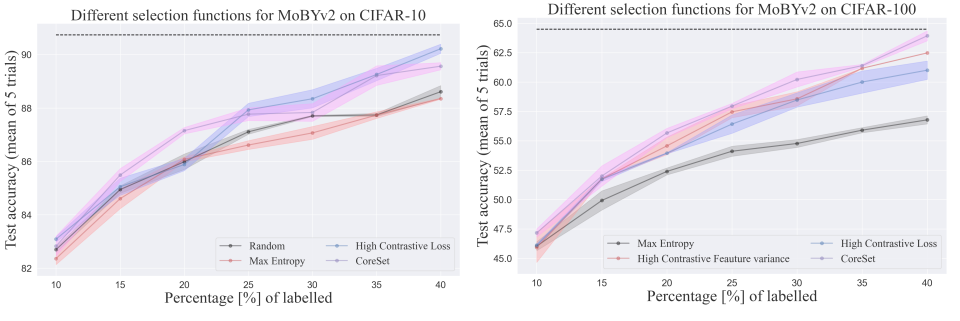

Figure B.1: Quantitative evaluation with different selection functions for CIFAR-10 (**left**), CIFAR-100 (**right**) [Zoom in for better view]

Our proposed pipeline, MoBYv2AL, can easily adapt to multiple selection methods. Here, we quantitatively motivate the choice of CoreSet from section 3. Therefore, we re-

evaluate MoBYv2AL on CIFAR-10/100 benchmarks in Figure B.1. We vary the selection of the new budget between random, maximum class entropy and CoreSet. Intuitively, we also analyse the effect of selecting unlabelled examples with high contrastive loss.

In both benchmarks, sampling with random or max entropy benefits the less MoBYv2AL pipeline. On the other hand, a representativeness-oriented method like CoreSet suits our hypothesis better. When sampling with high contrastive loss, we detected repetitive examples from some specific classes. This can be explained by higher contextual variance in that category. Specifically, on CIFAR-10, animal classes (cat, deer, dog), with stronger patterns, were more preferred than the vehicle ones (car, truck, ship).

For a better visual analysis, we have simulated a toy-set experiment with the first five classes from SVHN. Here, we take t-SNE[36] representations of the MoBYv2AL query encoder outputs of unlabelled data. In Figure B.2, the samples marked with crosses construct the new labelled set. The selection behaviour of the Max Entropy and CoreSet can be inter-

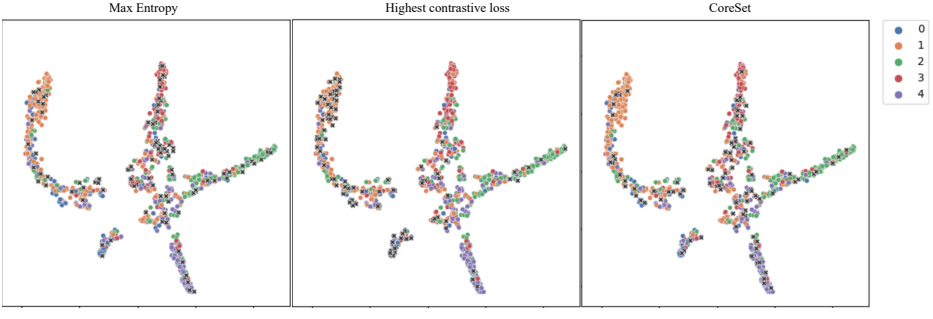

Figure B.2: Qualitative AL selection analysis on MoBYv2. t-SNE representations at the first selection stage for 5 classes of SVHN. [Zoom in for better view]

preted as expected: on the left side, the uncertainty-based technique tracks the most class-variant images; CoreSet, on the right side, samples both in and out-of-distribution according to the Euclidean space.