International Journal of Computer Vision (2025) 133:4294–4316
https://doi.org/10.1007/s11263-025-02372-z

# Realistic Evaluation of Deep Active Learning for Image Classification and Semantic Segmentation

**Sudhanshu Mittal[1] · Joshua Niemeijer[2,3] · Özgün Çiçek[4] · Maxim Tatarchenko[4] · Jan Ehrhardt[3] · Jörg P. Schäfer[2] · Heinz Handels[3] · Thomas Brox[1]**

Received: 27 May 2024 / Accepted: 4 February 2025 / Published online: 28 February 2025


**Abstract**

Active learning aims to reduce the high labeling cost involved in training machine learning models on large datasets by efficiently labeling only the most informative samples. Recently, deep active learning has shown success on various tasks. However, the conventional evaluation schemes are either incomplete or below par. This study critically assesses various active learning approaches, identifying key factors essential for choosing the most effective active learning method. It includes a comprehensive guide to obtain the best performance for each case, in image classification and semantic segmentation. For image classification, the AL methods improve by a large-margin when integrated with data augmentation and semi-supervised learning, but barely perform better than the random baseline. In this work, we evaluate them under more realistic settings and propose a more suitable evaluation protocol. For semantic segmentation, previous academic studies focused on diverse datasets with substantial annotation resources. In contrast, data collected in many driving scenarios is highly redundant, and most medical applications are subject to very constrained annotation budgets. The study evaluates active learning techniques under various conditions including data redundancy, the use of semi-supervised learning, and differing annotation budgets. As an outcome of our study, we provide a comprehensive usage guide to obtain the best performance for each case.

**Keywords** Active learning · Semi supervised learning · Classification · Segmentation

## 1 Introduction

In *Active learning* (AL), the objective is the reduction of annotation cost by selecting those samples for annotation, which are expected to yield the largest increase in the model's performance. Active learning is based on the attractive idea that some samples are more valuable for learning than others - by identifying those in the pool of unlabeled data, we can use an annotator's time more efficiently. It assumes that raw data can be collected in abundance for most large-scale data applications, but annotation limits the use of this data.

Various previous works have proposed solutions to the challenge of AL, which is ubiquitous in most machine learning applications. Yet there exists a skepticism about its benefits over random or some manual prior. This doubt largely arises from inconsistency in method performances

Communicated by Ullrich Köthe.

Sudhanshu Mittal and Joshua Niemeijer have contributed equally to this work.

Özgün Çiçek and Maxim Tatarchenko have done work at the University of Freiburg.

✉ Sudhanshu Mittal
mittal@cs.uni-freiburg.de

✉ Joshua Niemeijer
Joshua.Niemeijer@dlr.de

Jan Ehrhardt
jan.ehrhardt@uni-luebeck.de

Jörg P. Schäfer
Joerg.Schaefer@dlr.de

Heinz Handels
heinz.handels@uni-luebeck.de

Thomas Brox
brox@cs.uni-freiburg.de

[1] University of Freiburg, Freiburg, Germany

[2] German Aerospace Center (DLR), Braunschweig, Germany

[3] University of Luebeck, Lübeck, Germany

[4] Robert Bosch GmbH, Gerlingen, Germany

across studies, which differ in architectures, augmentation strategies, and optimization techniques. Also the effectiveness of the AL acquisition methods w.r.t task difficulty and data distributions remains unclear. Semi-supervised learning, besides active learning, is a way to deal with this situation of high annotation cost. Semi-supervised learning (SSL) and AL share a common objective of obtaining maximum performance from minimum supervision. Therefore, it is sensible to integrate both ideas, yet the combination of active learning with semi-supervised learning is understudied. We aim to study this combination and provide clarity on the abovementioned inconsistencies.

In this work, we systematically assess the state of the field and challenge the principal hypothesis behind active learning: *active selection of the samples to be labeled leads to a significant reduction in the annotation effort compared to random selection.* We systematically study the behavior of active learning methods under different training conditions in order to present a realistic perspective. Our study identifies that existing works are effective, but only under certain training conditions. They are not consistent across different model variables like data distribution, annotation budget, supervision type, and regularization. This work provides an extensive analysis of existing active learning methods under these diverse variables.

We conduct a detailed two-part analysis of active learning (AL) methods for image classification and semantic segmentation. For image classification, we evaluate the consistency of AL methods across similar datasets like CIFAR-10 and CIFAR-100, then examine their performance under the influence of strong augmentations, semi-supervised learning objectives, and varying annotation budgets. For the dense semantic segmentation task, data is often collected as video streams, especially for navigation applications such as autonomous driving. Such video stream data is very different from previously tested benchmarks in active learning literature; it is highly redundant. We assess AL methods for segmentation with varying levels of data redundancy, across different annotation budgets, and in conjunction with semi-supervised learning, to understand their behavior under these conditions.

### 1.1 Active Learning for Image Classification

Our first study seeks answers to the following scientific questions about active learning for image classification:

1. Since a widely accepted evaluation protocol is missing, methods are often tested under incompatible circumstances. We evaluate the effect of compatible experimental settings on the ranking of methods. In particular, **do AL methods work consistently well in conjunction with data augmentation?**
2. Contemporary papers on active learning largely ignore the progress of the closely related field of semi-supervised learning, where approaches effectively operate under the same assumptions with regard to the used data. **What is the effect as concepts from semi-supervised learning are integrated into active learning?**
3. Existing methods are typically not evaluated in a low-budget setting - a mode crucially important to kick-start network training on a new dataset. **How do active learning concepts work in such a low-budget regime?**

We conducted an extensive comparison of existing active learning approaches for image classification, revealing that recent progress is negligible under realistic conditions, see Fig. 1a. Integration of modern semi-supervised learning into active learning gives a significant boost to the acquisition functions. However, the difference w.r.t. random baseline with SSL becomes negligible.

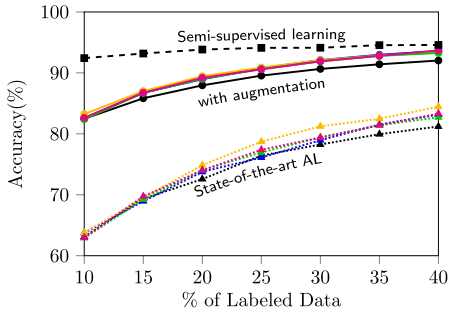
### 1.2 Active Learning for Semantic Segmentation

In the second part of the study, we conduct analysis for semantic segmentation. As a result, we show that the findings for image classification only hold under certain conditions for semantic segmentation. We noticed that the state-of-the-art AL methods for segmentation had been evaluated only in a particular experimental setup— highly diverse benchmark datasets with a comparatively large annotation budget; see Table 1b. Its applicability in other settings with different data distribution and annotation budgets is highly relevant but an unstudied topic. Additionally, we do not know how active learning methods integrate with semi-supervised learning. In this work, we also seek answers to specific missing questions.

**1. How do different active learning methods perform when the dataset has many redundant samples?** Samples with highly overlapping information are referred to as redundant samples, for example, the consecutive frames of a video. Many commonly used segmentation datasets were originally collected as videos for practical reasons, e.g., Cityscapes, CamVid, BDD100k (Yu et al., 2018). Since active learning methods were only tested on filtered versions of these datasets, their applicability on redundant datasets is open and highly relevant.

**2. What happens when the initial unlabeled pool is also used for training along with annotated samples using semi-supervised learning (SSL)?** For image classification, few have shown that integration of SSL into AL is advantageous for different tasks. However, for semantic segmentation, this combination is not well studied.

**3. What happens when the annotation budget is low? Which methods scale best in such low-budget settings?** Semantic segmentation annotations can be expensive for specific applications, especially in the medical domain.




(a) Image Classification

| Dataset↓ | Annotation Budget | | | |
|---|---|---|---|---|
| | Low | | High | |
| Supervision | AL | SSL-AL | AL | SSL-AL |
| Diverse | ✓ | ✓ | ✓ | ✓ |
| Redundant | ✓ | ✓ | ✓ | ✓ |

(b) Semantic Segmentation

**Fig. 1** **a** State-of-the-art active learning methods for image classification do not consistently use modern data augmentation techniques or advances in the closely related field of semi-supervised learning, which leads to the wrong impression about the current state of the field. The figure shows the results for image classification using CIFAR 10. It shows that AL methods *with data augmentation* significantly outperform state-of-the-art AL methods usually tested in the literature. Random sampling with *semi-supervised learning*, which is overlooked in AL evaluations even outperforms AL methods trained with data augmentation. **b** We study current active learning (AL) methods for semantic segmentation over 3 dimensions - dataset distribution, annotation budget, and integration of semi-supervised learning (SSL-AL). Green cells denote newly studied settings in this work. Previous AL works correspond to the grey cells. This work provides a guide to use AL under all the above conditions

Therefore, it is critical to understand the behavior of the various active learning methods in low-budget settings.

In this work, we report the results of an empirical study designed to find answers to the above-raised questions. The outcome of this study yields new insights and provides, as the major contribution of this work, a guideline for the best selection of available techniques under the various tested conditions. We also propose a new exemplary evaluation task (A2D2-3K) for driving scenarios based on the highly redundant A2D2 dataset, which is closer to the raw data collection scheme in a driving case.

## 2 Related Work

Many previous works provide a comprehensive survey, which usually includes definitions, taxonomy, dataset benchmarks and methodologies. Our work is different from prior surveys on deep active learning in numerous ways. While prior surveys (Li et al., 2024; Ren et al., 2021) focus on summarizing the existing active learning methods and highlighting their strengths and weaknesses, our work focuses on providing a quantitative comparison between these prior methods. Some works (Zhan et al., 2022) also include comparative analysis, however they are limited to a few standard settings where datasets have an i.i.d. and diverse data distribution. In this work, we not only aim to provide a fair comparison across existing methods, but also show that active learning methods show different behaviors based on experiment properties like distributional property of the datasets, nature of the AL method, different levels of task difficulty and different training objectives. Below, we categorize prior work based on these important properties that play a significant role on the performance of AL methods.

### 2.1 Acquisition Objective: Uncertainty vs. Representation

Here, we discuss the works divided by the measure of the value of a sample.

*Uncertainty-based methods* try to find the samples which are hard to learn. In these methods, samples with the most predictive uncertainty are considered as most informative for labeling purposes. Several methods have been proposed to estimate uncertainty for neural networks using Bayesian (Blundell et al., 2015; Gal & Ghahramani, 2016a, b; Kendall & Gal, 2017) and non-Bayesian approaches (Lakshminarayanan et al., 2017; Osband et al., 2016). Gal et al. (2017) proposed to estimate posterior uncertainty using Monte Carlo dropout for active learning. Wang et al. (2017) used the entropy of the softmax output in a neural network as a proxy uncertainty measure. Beluch et al. (2018) used the ensemble method to estimate prediction uncertainty and select new samples based on a statistical measure of committee disagreement called variation ratio (Johnson, 1966). They show this method outperforms all other uncertainty-based methods.

*Representation-based methods* (Sener & Savarese, 2017; Yang et al., 2017), also referred to as density-based methods, try to find a diverse set of samples that optimally represents the complete dataset distribution. Sener and Savarese (2017) formulated the active learning problem as core-set selection and showed effectiveness for CNNs. This method utilizes the geometry of data points using Euclidean distances and selects

samples that maximize the coverage of all samples. *Learning-based approaches* (Sinha et al., 2019; Yoo & Kweon, 2019) use an auxiliary network module and loss function to learn a measure of information gain from new samples. Yoo and Kweon (2019) proposed to learn a loss prediction module to predict target losses of unlabeled samples and select samples with the highest predicted loss. It can also be considered a pseudo-uncertainty heuristic. Sinha et al. (2019) proposed a semi-supervised active learning approach that learns a VAE-GAN hybrid network to select unlabeled samples that are not well represented in the labeled set. It can also be considered a representation-type method.

### 2.2 Acquisition Type: Single-sample vs. Batch Acquisition

The acquisition methods can be categorized into single-sample-based and batch-based approaches. They assess the value of new samples for selecting individually and collectively as a batch, respectively.

*Single sample acquisition* takes the top $b$ samples according to the score of the acquisition function to select a batch of size $b$. Several methods follow this selection scheme based on either epistemic uncertainty or representation score. For example, uncertainty-based methods try to select the most uncertain samples to acquire a batch. Many methods, such as EqualAL (Golestaneh & Kitani, 2020) Ensemble+AT (Lakshminarayanan et al., 2017), and CEAL (Wang et al., 2017), estimate uncertainty based on the output probabilities. Epistemic uncertainty, estimated using Entropy (Shannon, 1948), is often used as a strong baseline in several active learning works (Golestaneh & Kitani, 2020; Shin et al., 2021; Rangnekar et al., 2022). Some methods, namely BALD (Houlsby et al., 2011) and DBAL (Gal et al., 2017) employed a Bayesian approach using Monte Carlo Dropout (Gal & Ghahramani, 2016b) to measure the epistemic uncertainty. Representation-based methods aim to select the most representative samples of the dataset that are not yet covered by the labeled samples. Numerous adversarial learning-based methods utilize an auxiliary network to score samples based on this measure, including DAAL (Wang et al., 2020), VAAL (Sinha et al., 2019), and WAAL (Shui et al., 2020). For our study, we employ Entropy, EqualAL, and BALD to represent single-sample acquisition methods due to their direct applicability to segmentation tasks. We did not include deep ensemble-based methods due to their limited scalability and adversarial methods due to their hyperparameter sensitivity. In general, single-sample acquisition approaches select individually very informative samples but do not optimize the joint improvement obtained with the whole batch.

*Batch-based acquisition* methods acquire the whole batch of size $b$ to maximize cumulative information gain. Sener and Savarese (2017) formulated the acquisition function as a core-set selection approach based on the feature representations. It is a representation-based approach that selects the batch of samples jointly to represent the whole data distribution. BatchBALD (Kirsch et al., 2019) is a greedy algorithm that selects a batch of points by estimating the joint mutual information between the whole batch and the model parameters. This method was also proposed to remedy the mode collapse issue, where the acquisition function collapses into selecting only similar samples (see Sect. 4.3.1 for details). However, it is limited to simple image classification datasets like MNIST (Deng, 2012) since its computation complexity grows exponentially with the batch size. Some more recent batch-based methods include k-MEANS++ (Zhdanov, 2019), GLISTER (Killamsetty et al., 2021), ADS (Jia et al., 2019), but these methods only evaluate on image classification tasks. For the study, we selected the Coreset method (Sener & Savarese, 2017) to represent batch-based methods due to its effectiveness, simplicity, and easy scalability to the segmentation task.

### 2.3 Active Learning for Semantic Segmentation

Along with the task of active learning for image classification, we also focus on semantic segmentation. Suggestive Annotation (Yang et al., 2017), Cereals (Mackowiak et al., 2018), and VAAL (Sinha et al., 2019) are a few previous works that have shown the applicability of deep active learning for semantic segmentation.

When applied to semantic segmentation, active learning methods must choose which area of the image is to be considered for the acquisition: the full image (Sinha et al., 2019), superpixels (Cai et al., 2021), polygons (Golestaneh & Kitani, 2020), or each pixel (Shin et al., 2021). There is no common understanding so far of which approach is cheaper and more effective. Thus, our study uses the straightforward image-wise selection and annotation procedure. Most existing methods for segmentation are based on the model's uncertainty for the input image, where the average score over all pixels in the image is used to select top-k images. Entropy (Shannon, 1948) (estimated uncertainty) is a widely used active learning baseline for selection. This function computes per-pixel entropy for the predicted output and uses the averaged entropy as the final score. EqualAL (Golestaneh & Kitani, 2020) determines the uncertainty based on the consistency of the prediction on the original image and its horizontally flipped version. The average value over all the pixels is used as the final score. BALD (Houlsby et al., 2011) is often used as a baseline in existing works. It is employed for segmentation by adding dropout layers in the decoder module of the segmentation model and then computing the pixel-wise mutual information using multiple forward passes. Coreset (Sener & Savarese, 2017) is a batch-based

approach that was initially proposed for image classification, but it can be easily modified for segmentation. For e.g., the pooled output of the ASPP (Chen et al., 2018) module in the DeepLabv3+ (Chen et al., 2018) model can be used as the feature representation for computing distance between the samples. Our study includes Entropy, EqualAL, BALD, and Coreset approaches for the analysis, along with the random sampling baseline. Most AL methods for semantic segmentation use single-sample acquisition and show superior performance over batch acquisition function Coreset. This chapter also studies the integration of these methods with semi-supervised learning.

### 2.4 Semi-supervised Active Learning

Most representation-based AL methods use unlabeled samples to learn the underlying distribution, but only a few methods use semi-supervised learning to improve their selection criteria (Sinha et al., 2019; Sener & Savarese, 2017; Ravanbakhsh et al., 2019; Wang et al., 2022). Sinha et al. (2019) used an unlabeled pool to learn its distribution against the distribution of labeled samples. Still, they did not take advantage to improve the feature representation of the target model itself. Sener and Savarese (2017) have also previously shown the advantage of using the unlabeled pool for learning the target model. Wang et al. (2022) also explored the usage of the most-certain samples from the un-labeled pool using pseudo-labeling, but the pseudo-labeling process can easily propagate erroneous labels if not tuned properly. Ravanbakhsh et al. (2019) proposed a GAN-based approach to use the unlabelled pool and utilize the discriminator score to query low-confident samples for active learning. Two concurrent open-source works (Gao et al., 2020; Anonymous, 2020) have also shown some similar findings to our work. However, they are restricted to only image classification.

Some recent works have also studied active learning methods with the integration of SSL for segmentation, but their scope is limited only to special cases like subsampled driving datasets (Rangnekar et al., 2022) or low labeling budget, both cases with only single-sample acquisition methods. Our work provides an overview of the integration of SSL and active learning for the image classification and semantic segmentation task.

### 2.5 Current Benchmarks

Current AL methods for image classification are mostly tested on CIFAR-10 and CIFAR-100 datasets, which are perfectly balanced. Some recent works (Kim et al., 2021) have also tested on higher resolution datasets like Caltech101, which is naturally imbalanced. In this work, we also use CIFAR-10 and CIFAR-100 for our study on image classification. However, we evaluate the AL methods with various new settings like strong augmentation, integration of SSL, and low-annotation budget. Current AL methods for semantic segmentation are usually evaluated on driving datasets due to the industrial focus on autonomous driving. These datasets include Cityscapes (Cordts et al., 2016), BDD100K (Yu et al., 2018) and CamVid (Brostow et al., 2008). Some works evaluate more generic datasets like PASCAL-VOC (Everingham et al., 2010). Medical datasets (Zhang et al., 2016; Codella et al., 2018; Simpson et al., 2019) are also common for the AL studies due to extremely high annotation cost. In this work, we focus on driving datasets and introduce a more realistic driving AL task. We also provide a case-study for medical data domain fitting to one the studied cases.

## 3 Active Learning for Image Classification

In this section, we assess the performance of state-of-the-art AL methods for image classification and compare them with the integration of data regularization and a state-of-the-art semi-supervised learning approach. We also challenge the previously proposed methods under a low annotation budget where the initial model is trained with fewer labeled samples, followed by a few new sample selections at each AL cycle.

### 3.1 Integration of AL with Label-efficient Learning

*Active Learning with Data Augmentation.* Recently, various regularization techniques have been proposed to improve model generalization with minimal labeled data. Although these methods show consistent success in various applications, they have been ignored by the works on active learning. In this section, we study the effect of one such regularization: *data augmentation*. Data augmentation is a widely accepted regularization technique, which increases the power of machine learning models, particularly when there is little labeled data. Nevertheless, several latest AL works (Beluch et al., 2018; Sinha et al., 2019) resort to either not using any augmentation during training or only doing simplistic augmentations like horizontal flipping. The behavior of active learning under the influence of strong data augmentation is largely unknown. In the experiments, we apply strong augmentations, including color and geometric augmentations, during the training phase of the model. The acquisition function selects samples based on this model.

*Integration of Active Learning with Semi-supervised Learning.* A largely common practice in the previous works has been to utilize the unlabeled pool only for sampling, although it is available throughout the learning process (otherwise, one could not sample from it) and could be used more rigorously. Using semi-supervised learning, we can utilize this unlabeled pool for training the model itself and thus learn an improved query function using unlabeled samples.

To this end, we employed the UDA (Xie et al., 2019) semi-supervised learning method. UDA applies a consistency loss between differently augmented unlabeled samples to learn from unlabeled samples. We integrated SSL into the AL methods by training the model using the UDA objective and defining the query function based on this model. In each cycle, the target model is trained using UDA instead of the standard supervised training.

*Active Learning under Low-annotation Budget.* We observed in the literature that there is an inconsistency in the methods' behavior when switching from CIFAR-10 to CIFAR-100. This challenges the principal assumption of active learning that a dedicated selection strategy always improves over a random selection of samples. *We ask whether active learning benefits from a low-budget setting, where every sample is particularly crucial.* In certain applications, such as medical image analysis, already 10000 annotated samples can be very costly. Thus, training with only a few labeled samples in the beginning is attractive. We study the behavior of active learning methods where the initial and sampling annotations budget is 10 to 20 times smaller than usually studied in previous works.

### 3.2 Experiment Setup

We evaluate and compare following **baseline methods**:

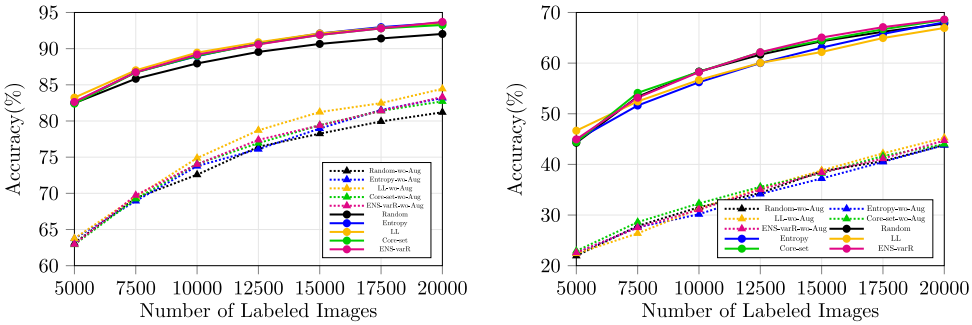
- *Random*: A new set of samples is selected randomly from the unlabeled pool and is added to the labeled pool with annotations.
- *Entropy*: Shannon (1948) is an information-theoretic measure used as an uncertainty metric for sampling. This method naively selects samples for which the pseudo-probabilities predicted by the softmax classifier have the highest entropy. For the entropy method, we use the softmax output of the final fully-connected layer to calculate the entropy of the prediction.
- *Ensemble with Variation Ratio (ENS-varR):* The second method, which selects samples based on an uncertainty criterion, relies on using ensembles. It has been shown to consistently outperform all other uncertainty-based approaches for active learning by Beluch et al. (2018). The core of the method is to calculate the variation ratio (varR) metric given as the proportion of predicted class labels that are not the modal class prediction:

$$var\,R = 1 - \frac{f_m}{T}, \tag{1}$$

where $f_m$ is the frequency of the modal class and $T$ is the number of ensemble members. This heuristic is motivated by the query-by-committee algorithm proposed by Seung et al. (1992). The query function selects the samples with larger varR values. The ensemble is only used for sample querying - the target performance is still reported for a single model. Similar to Beluch et al. (2018), we use an ensemble of 5 models for our experiments.
- *Core-set*: This type of method selects a batch of samples such that the performance of the model trained on the labeled set matches the performance of the model trained on the whole dataset (Paul et al., 2014). The recent core-set approach proposed by Sener and Savarese (2017) casts the core-set selection problem as a k-center problem and proposes a robust k-center approach. The proposed approach chooses a subset such that the largest distance between the chosen point and unlabeled points is minimized in the feature space. For the core-set approach, we make use of the k-center greedy implementation since it is much faster and only performs marginally worse than the robust version.
- *Learning Loss (LL)*: This method (Yoo & Kweon, 2019) proposes a loss prediction module that is attached to the target network to estimate the loss value of the unlabeled samples. The samples with the largest predicted loss are selected for annotation. This auxiliary module is trained to preserve the pairwise ranking of the original loss values, which is imposed using a hinge loss function over random pairs of samples in a minibatch.
- *Unsupervised Data Augmentation (UDA)*: UDA (Xie et al., 2019) is a semi-supervised learning method for image classification. It uses consistency regularization to learn from unlabeled samples along with AutoAugment (Cubuk et al., 2019) and other augmentation techniques to reduce overfitting. We selected this method because: 1) it shows state-of-the-art performance, 2) it is based on a simple idea and is easy to implement. Also, the method performs well even when the number of labeled samples is very small. Our implementation used online data augmentation instead of the offline one in the original work (Xie et al., 2019).

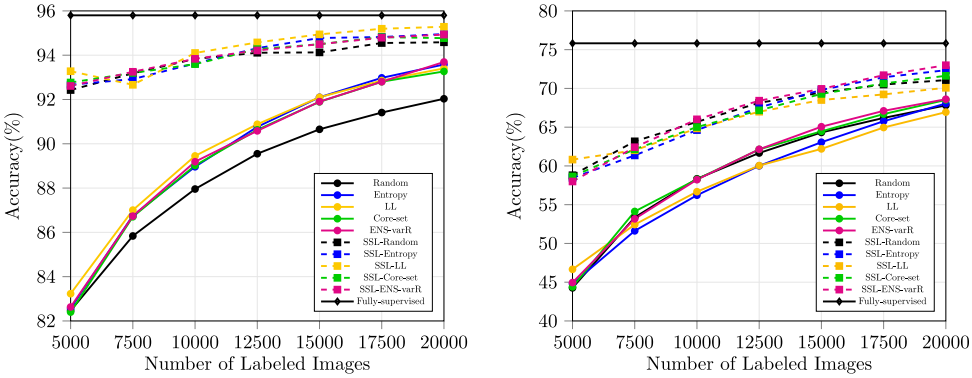
*Datasets.* We evaluate the methods on the CIFAR-10 and CIFAR-100 datasets. Both datasets contain the same set of 60,000 images, assigned to 10 and 100 classes, respectively. The training and test set contains 50,000 and 10,000 images, respectively. CIFAR-10 is the most commonly tested dataset in the field of active learning. CIFAR-100 is an extension with 100 classes, which makes the task more challenging. The initial labeling budget is $\mathcal{B}_i = 5000$, and the sampling budget is $\mathcal{B}_s = 2500$ labels for each cycle. We tested this configuration for 6 sampling cycles (i.e. going from 10% to 40% labeled samples). In the first step, we randomly sampled a class-balanced subset of samples from the unlabeled pool.

*Evaluation metrics.* We evaluate AL methods in different data budget settings, referred to as the $\mathcal{B}_i$-$\mathcal{B}_s$ setting, where $\mathcal{B}_i$ is the initial label budget, $\mathcal{B}_s$ is the sampling-label budget,




(a) CIFAR-10 (b) CIFAR-100

**Fig. 2** Using data augmentation on CIFAR-10 significantly improves the performance of active learning methods and makes the relative difference between them less pronounced. The performance of AL methods on CIFAR-100 improves significantly when using up-to-date image augmentation. Results without augmentation are denoted as 'X-wo-Aug'. These results show that active learning methods when used with latest data augmentation during training lose their significance




(a) CIFAR-10 (b) CIFAR-100

**Fig. 3** Combining AL methods with semi-supervised learning leads to significant performance improvement on CIFAR-10 compared to the raw AL case. Results shown in the large-budget setting with $\mathcal{B}_i = 5000, \mathcal{B}_s = 2500$. Integrating SSL and AL leads to overall performance improvement on CIFAR-100, however, not all combinations consistently outperform random sampling. Results shown in the large-budget setting with $\mathcal{B}_i = 5000, \mathcal{B}_s = 2500$. These results show that using semi-supervised learning during training further reduces the significance of AL methods. Some methods even underperform the random sampling baseline

and $\mathcal{B}_i$,$\mathcal{B}_s$ refer to the number of labeled images. Images are sampled randomly to fulfill the initial label budget. For the subsequent steps, images are sampled using the AL acquisition function with the sampling-label budget. We test these datasets with $5K - 2.5K$, $500 - 500$, and $250 - 250$ settings.

We use final accuracy to evaluate the performance of the model at each AL cycle step. For the evaluation of the active learning method, we use two metrics: Area Under the Budget Curve (AUC@$\mathcal{B}$) and final accuracy(Acc).

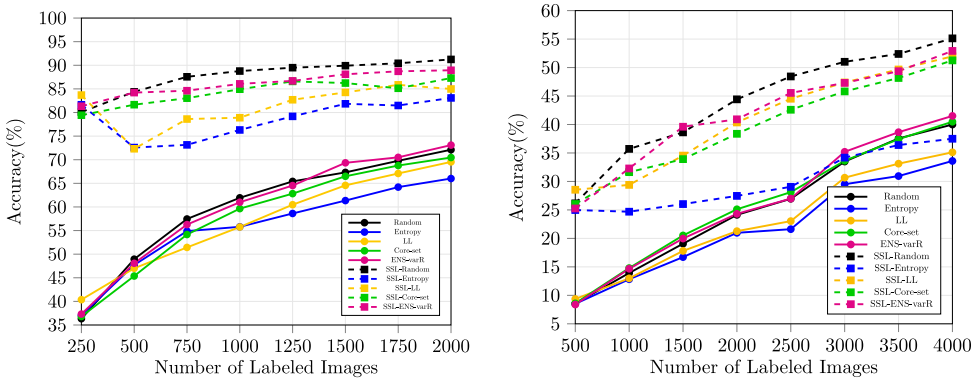
- **AUC@$\mathcal{B}$** is the area under the performance curves, shown in Figs. 2 and 3. It captures a cumulative score of the AL performance curve up to a budget $\mathcal{B}$, where $\mathcal{B}$ is the number of labeled images. We use a total budget of $\mathcal{B}$=20K in the 5K-2.5K setting for CIFAR-10 and CIFAR-100 datasets. We use $\mathcal{B}$=2K for CIFAR-10 in 250-250 setting and $\mathcal{B}$=4K for CIFAR-100 in 500-500 setting.

We use the following formula to compute the Area Under the Budget Curve(AUC) at a total budget $\mathcal{B}$, where $\mathcal{B}$ is the percentage of the labeled dataset:

$$AUC@\mathcal{B} = \sum_{i=1}^{i=N} \frac{(b_{i+1} - b_i)(p_i + p_{i+1})}{2} \tag{2}$$

where N is the number of AL acquisition steps, $b_i$ is the percentage of the labeled dataset from the whole dataset at step $i$, and $p_i$ is the performance of the model in mIoU(%) at step $i$.

- **Acc** reports the Accuracy of the model after using the total labeling budget $\mathcal{B}$. We report performance at an intermediate labeling budget to clearly see the ranking of the AL methods.




(a) CIFAR-10 (b) CIFAR-100

**Fig. 4** When evaluated in the low-budget regime ($\mathcal{B}_i = \mathcal{B}_s = 250$) on CIFAR-10, integrated SSL-AL methods are still better than their raw counterparts, however, SSL with random sampling shows the best performance. When evaluated in the low-budget regime ($\mathcal{B}_i = \mathcal{B}_s = 500$) on CIFAR-100, most integrated SSL-AL methods are still better than their raw counterparts but nothing beats SSL with random sampling

### 3.3 Results

*Integration with Data Augmentation* In this experiment, we validated the importance of elaborate up-to-date image augmentation for the performance of AL methods. We first evaluated all methods without any augmentation. Subsequently, we evaluated the same methods with augmentation, which includes using the AutoAugment policies found by Cubuk et al. (2019), cutout (DeVries & Taylor, 2017), horizontal random flipping, and random cropping. Figure 2 shows that without using any augmentation, all AL methods clearly perform better than the random baseline. The LL method shows distinct improvement over other methods (matching the results from Yoo and Kweon (2019)) and an overall improvement of 3.2% over the random baseline on the CIFAR-10 dataset. When the same experiment is performed with augmentation, all the methods improve drastically in absolute performance. However, the relative effect of using different AL methods becomes far less pronounced: all the AL methods show similar performance within a range of 0.4%. In conclusion, AL works well with data augmentation, but data augmentation blurs the differences between AL strategies: they all perform largely the same.

For completion, we further validate the importance of using up-to-date augmentation for AL methods on the CIFAR-100 dataset. We evaluate all methods with and without augmentation, similar to the CIFAR-10 experiment. The overall conclusion is also very similar: Without augmentation, the LL method shows a distinct improvement of 1.4% over the random baseline; with augmentation, all the methods improve by a large margin in absolute performance, but the relative difference between different methods becomes insignificant and the relative ranking of different methods changes. Performance curves are shown in Fig. 2b.

*Integration with Semi-supervised Learning* We refer to the integrated methods as SSL-X, where X is the name of the AL method. Figures 3a and b show a remarkably strong performance of the SSL method (SSL-Random) on CIFAR10 and CIFAR100: when using 5K random labeled samples, SSL almost reaches the same performance which AL methods achieved on 20K samples picked by the corresponding query functions. Also, for the remaining data ratios, there is a large performance gap between semi-supervised and active learning, both on CIFAR-10 and CIFAR-100. Clearly, semi-supervised learning makes much better use of the same data than active learning.

SSL and AL can be combined, which yields an improvement over raw SSL on CIFAR-10. The SSL-LL method performs best and shows an improvement over the random baseline by 0.7% after 6 cycles. However, on CIFAR-100, the relative ranking of the AL methods changes completely; SSL-LL performs worse than the other methods and struggles even to compete with the random selection method.

The same is true for raw active learning without SSL: on CIFAR-100, some active learning methods do not reach the performance of randomly drawing the samples to be labeled, shown in Fig. 3b.

*High-budget vs Low-budget.* We explored such low-budget settings with $\mathcal{B}_i$ and $\mathcal{B}_s$ for each cycle set to 250 labels for CIFAR-10 and 500 labels for CIFAR-100. We tested this setting for 7 sampling cycles with a total budget of 2000 and 4000 labels for CIFAR-10 and CIFAR-100, respectively. We kept all the augmentation techniques from the previous experiments.

The results are shown in Figs. 4a and b. None of the active learning methods consistently outperforms the random baseline, neither on CIFAR-10 nor on CIFAR-100. This always holds for the combination of active learning and semi-supervised learning, whereas for raw active learning, only ENS-varR could marginally outperform the random baseline. In fact, some techniques perform considerably worse than the random baseline, especially in conjunction with semi-supervised learning, showing that their selection strategy is counter-productive in the low-budget regime.

### 3.4 Proposed Evaluation Protocol

Based on our observations, we formulate a more appropriate evaluation protocol and recommend using it for benchmarking future active learning methods for image classification.

1. AL methods should be evaluated on a broader range of datasets to assess their general robustness.
2. Evaluating AL methods with up-to-date network architectures and up-to-date augmentation techniques is vital.
3. There should always be a direct comparison between AL methods and SSL methods.
4. With the existing large-budget regime, AL methods should also be evaluated in the low-budget regime.

## 4 Active Learning for Semantic Segmentation

In this section, we assess the performance of state-of-the-art AL methods for semantic segmentation and compare them with the integration of semi-supervised learning. All methods are tested on datasets with different levels of redundancy and various levels of annotation budgets.

### 4.1 Conceptual Considerations

Our experiments show that the presence of redundant samples in the data distribution influences the choice of the acquisition function and the training regime that achieve the best performance. The main cause for this is the mode collapse issue, where the acquisition function collapses into selecting only similar samples. Here, we first discuss why and when this mode collapse occurs and how to remedy this issue. Then, we discuss ideal conditions for the successful integration of semi-supervised learning with active learning acquisition functions.

#### 4.1.1 Redundancy Can Cause Mode Collapse

Mode collapse in active learning refers to the circumstance that acquisition functions tend to select similar (redundant) samples when acquiring batches of data (Kirsch et al., 2019). Since the selected similar samples contain highly redundant information, their annotation does not add much new value to the model performance. Figure 7 illustrates this mode collapse issue for a driving dataset case, where samples are selected from dense local feature space clusters using an epistemic uncertainty-based acquisition function. The mode collapse occurs when the dataset contains redundant samples, and the acquisition function is designed to select single samples based on some independent sample scores. Here, redundant samples tend to get similar scores from the acquisition function due to their similarity, i. e., their large overlap in information. Therefore, if one of those samples is selected due to a high score, other similar samples are also selected. This mode collapse effect occurs especially in Deep Active Learning scenarios since the acquisition of big batches is necessary to reduce the overall number of active learning cycles.

Existing deep active learning methods for semantic segmentation show that epistemic uncertainty is a good heuristic to select samples for annotation in common benchmarks. These strategies utilize single-sample acquisition functions and select the set of most valuable samples from the unlabeled pool based on the sampling budget. Since such methods were only tested on diverse datasets which are already curated for diversity, the mode collapse problem does not have a strong effect on their evaluation. However, this is not the case for many real-world applications. Redundancy occurs when there are repeated recordings of similar scenes or when the data is collected in a video format, like driving scenarios. A good acquisition function for such a redundant dataset must be aware of the batch's diversity to address the mode collapse issue. Intuitively, clustering-type approaches are ideal in redundant datasets since they select one sample from each local cluster avoiding single-sample selection traps like the mode collapse issue.

In this section, we argue that mode collapse is a common issue in real-world datasets and is largely ignored due to poor active learning benchmarks, which only cover diverse datasets. We probe previous AL methods for semantic segmentation over different diverse and redundant datasets. We design various redundant datasets based on the driving video dataset A2D2 (Geyer et al., 2020) to reveal how the behavior of active learning methods changes with the level of redundancy in the dataset.

#### 4.1.2 Requirements for Integration of Semi-supervised Learning and Active Learning

Active learning methods use a pool of unlabeled samples only for selecting new samples for annotation. However, this pool can also be used by semi-supervised learning, where the objective is to learn jointly from labeled and unlabeled samples. In this work, we integrate semi-supervised learning with active learning in the context of semantic segmentation, an idea that was previously proposed for classification (Sener & Savarese, 2017; Gao et al., 2020; Munjal et al., 2022). In particular, we train the model using a semi-supervised learning objective, which impacts the resulting model and hence the acquisition function.

Successful integration can also be conceptually explained based on the underlying assumption of semi-supervised learning and the selection principle of the active learning approach. According to the *clustering assumption* of SSL,

if two points belong to the same cluster, then their outputs are likely to be close and can be connected by a short curve (Chapelle et al., 2006). In this regard, when labeled samples align with the clusters of unlabeled samples, the cluster assumption of SSL is satisfied, resulting in a good performance. Consequently, to maximize semi-supervised learning performance, newly selected samples must cover the unlabeled clusters that are not already covered by labeled samples. Only acquisition functions that foster this coverage requirement have the potential to leverage the additional benefits that arise from the integration of semi-supervised learning. A batch-based method, e.g., Coreset, selects samples for annotations to minimize the distance to the farthest neighbor. By transitivity, such labeled samples would have a higher tendency to propagate the knowledge to neighboring unlabeled samples in the cluster and utilize the knowledge of unlabeled samples using a semi-supervised learning objective and help boost the model performance. Similar behavior can also be attained using other clustering approaches that optimize for coverage.

## 4.2 Experiment Setup

### 4.2.1 Tested Approaches

In our study, we test five active learning acquisition functions, including Random, Entropy, EqualAL, BALD, and Coreset. Here Entropy, EqualAL, and BALD approach represent single-sample, and Coreset represents the batch-based approach. All methods select the whole image for annotation. These methods are further described below, along with the segmentation-specific changes.

- *Random:* The samples are selected randomly for annotation from the unlabeled pool.
- *Entropy* (Shannon, 1948): This acquisition function uses per-pixel entropy as an estimation of the epistemic uncertainty (U) for the predicted output $p$. The entropy is computed over the class predictions $p_{c,j}$, where $c \in C$ are the possible classes.

$$H(p_j) = -\sum_{c=1}^{C} p_{c,j} \log p_{c,j} \tag{3}$$

The final score for selection is the average entropy over the number of pixels $j \in N$.

$$U(p) = \frac{1}{N} \sum_{j=1}^{N} H(p_j) \tag{4}$$

This method selects all top-scoring images.

- *EqualAL* (Golestaneh & Kitani, 2020): The EqualAL approach determines the uncertainty (U) based on the self-consistency between the prediction on the original image $p$ and the prediction on its horizontally flipped version $\tilde{p}$. The average uncertainty value over all the pixels is used as the final score.

$$U(p_j, \tilde{p}_j) = \sum_{c=1}^{C} H(p_{c,j}) + \sum_{c=1}^{C} H(\tilde{p}_{c,j}) \tag{5}$$

We use the EqualAL implementation, which trains using only cross-entropy loss to keep the baselines comparable.
- *BALD:* Houlsby et al. (2011) The BALD approach is based on a Monte Carlo Dropout network to compute the pixel-wise Mutual Information of the classification. In our implementation, we employ dropout layers with a dropout ratio of 10% in the decoder layer and, during inference, compute 10 passes that result in $p^t$ (where $t = 1...10$) predictions per image. The Mutual Information (MI) is then computed as follows:

$$\mathrm{MI}(p^t) = H\left(\mathbb{E}(p^t)\right) - \mathbb{E}\left(H(p^t)\right) \tag{6}$$

where $\mathbb{E}(p^t)$ is the is the expected predicted probability over the $t$ stochastic forward passes and $\mathbb{E}(H(p^t))$ is the expected entropy over the individual forward passes.
- *Coreset:* Sener and Savarese (2017) The Coreset approach selects a batch of samples that cover the whole data distribution. It formulates this batch selection as a robust k-center selection problem. Coreset implements a greedy algorithm that iteratively selects unlabeled samples with maximum distance to the nearest neighbor of the so far selected samples. We utilize the k-center greedy approach since it is much faster and only performs slightly worse than the robust formulation. We use the ASPP module output in the DeepLabv3+ (Chen et al., 2018) model as the feature representation. Formally the selection of a new sample $x_i^*$ from the set of unlabeled images can be defined as:

$$x_i^* = \arg\max_{x_i \in \mathcal{X}_U} \min_{x_s \in \mathcal{X}_S} d(f(x_i), f(x_s)) \tag{7}$$

where $d(f(x_i), f(x_s))$ is the distance between the feature representations $f(x_i)$ and $f(x_s)$. The set $x_s \in \mathcal{X}_S$ is represents the already selected images and $x_u \in \mathcal{X}_U$ the unlabeled images.
- *MCD setting:* Since the BALD method requires the introduction of Dropout layers into the architecture, we segregate the methods into two categories: With Monte Carlo Dropout (MCD) and without Monte Carlo Dropout layers. Random, Entropy, EqualAL, and Coreset are without MCD. BALD and Coreset-MCD are based on

MCD. We compare methods in each category separately due to different architectures. We show fully-supervised performance, referred to as '100%' in the result tables, both with (100% MCD) and without MCD (100%) architectures.

*Semi-supervised Learning* To leverage the unlabeled samples, we use the semi-supervised learning s4GAN method (Mittal et al., 2019). It uses adversarial training to align the labeled and unlabeled data distribution and further uses self-training based on the GAN discriminator score. We pair all the used active learning approaches with SSL using this approach. This is marked by the suffix '-SSL' in the experiments. In particular, we train the model using an SSL objective, which impacts the resulting model and hence the acquisition function.

#### 4.2.2 Datasets

Active learning methods are often evaluated on PASCAL-VOC and Cityscapes datasets, where PASCAL-VOC is naturally diverse while Cityscapes is diversified by subsampling from videos. In this section, we test on an additional driving dataset, A2D2, which is highly redundant. We evaluate the methods on these three datasets. To understand the nature of active learning methods over varying levels of redundancy in the dataset, we curate 5 smaller dataset pools from the large, original A2D2 dataset, described further below as A2D2-Pools.

- **Cityscapes** (Cordts et al., 2016) is a driving dataset used to benchmark semantic segmentation tasks. The dataset was originally collected as videos from 27 cities, where a diverse set of images were selected for annotation. Due to the selection, Cityscapes cannot cover the redundant data scenario in our evaluation, although it was derived from videos. As we will see in the results, the nature of the active learning method changes when considering the raw form of data in a driving scenario, and pre-filtering, as done in Cityscapes, is sub-optimal compared to directly applying active learning on the raw data (see Sect. 4.4).
- **PASCAL-VOC** (Everingham et al., 2010) is another widely used segmentation dataset. We use the extended dataset (Hariharan et al., 2011), which consists of 10582 training and 1449 validation images. It contains a wide spectrum of natural images with mixed categories like vehicles, animals, furniture, etc. It is the most diverse dataset in this study.
- **A2D2** (Geyer et al., 2020) is a large-scale driving dataset consisting of 41277 annotated images with a resolution of $1920 \times 1208$ from 23 sequences. It covers an urban setting from highways, country roads, and three cities. It contains labels for 38 categories. We map them to the 19 classes of Cityscapes for our experiments. A2D2 provides annotations for every $\sim 10^{th}$ frame in the sequence and contains a lot of overlapping information between frames. Some consecutive frames are shown in Fig. 9. We utilize 40135 frames from 22 sequences for creating our training sets and one sequence consisting of 1142 images for validation. The validation sequence '*20180925_112730*' is selected based on the maximum class balance. A2D2 represents the most redundant raw dataset in our study.
- **A2D2 Pools.** To obtain a more continuous spectrum between diverse and redundant datasets, we created five smaller dataset pools by subsampling the large A2D2 datasets. Each pool comprises 2640 images, which is comparable in size to the Cityscapes training set. Four pools are curated by subsampling the original dataset, while the fifth pool is created by augmentation. The first four pools, denoted by Pool-Xf (where X is 0, 5, 11, and 21), were created by randomly selecting samples and X consecutive frames for each randomly selected sample from the original A2D2 dataset. Pool-0f contains only randomly selected images. We assume that the consecutive frames contain highly redundant information. Therefore, the pool with more consecutive frames has higher redundancy and lower diversity. The fifth pool, Pool-Aug, contains augmented duplicates in place of the consecutive frames. We create five duplicates of each randomly selected frame by randomly cropping 85% of the image area and adding color augmentation (see Fig. 10).
- **BCSS** (Amgad et al., 2019) Breast Cancer Semantic Segmentation is a dataset comprised of tissue regions from breast cancer images obtained from The Cancer Genome Atlas (TCGA). For annotation, 151 whole-slide images (WSIs) were used, which were cropped into RGB images with a resolution of 512x512 pixels for our experiments. The resulting training set consists of 6,000 images, and the validation set comprises 2,768 images. The training dataset includes 22 initial classes that we map to 5 classes, a common practice due to the sparse representation of many categories. The final classes are Tumor, Stroma, Inflammatory, Necrosis, and Other. Given that the dataset is derived from only 151 WSI images, the resulting distribution is redundant. This redundancy is a typical scenario in medical datasets, where data pools for annotation are often obtained from only a few patients and generally show little variation in anatomy.

**Which dataset is diverse or redundant?** We would like to clarify how we tag a dataset as diverse or redundant. Extreme cases like PASCAL-VOC can be easily tagged as diverse, and BCSS, A2D2 original and A2D2-Pool-5f/11f/21f can be tagged as redundant. However, it is hard

**Table 1** Active Learning results on Cityscapes and A2D2 Pool-0f

| A | AL Method | SSL | Cityscapes | | A2D2 Pool-0f | |
|---|---|---|---|---|---|---|
| | Metric → | | mIoU | AUC | mIoU | AUC |
| S | Random | ✗ | 58.90 | 23.29 ± 0.07 | 48.48 | 19.20 ± 0.07 |
| S | Entropy | ✗ | 61.83 | 24.25 ± 0.06 | 52.40 | **20.37** ± 0.11 |
| S | EqualAL | ✗ | 62.41 | 24.32 ± 0.12 | **52.50** | 20.35 ± 0.04 |
| B | Coreset | ✗ | 60.89 | 23.89 ± 0.26 | 51.14 | 19.88 ± 0.08 |
| S | Random-SSL | ✓ | 60.72 | 23.85 ± 0.22 | 49.69 | 19.60 ± 0.08 |
| S | Entropy-SSL | ✓ | 60.61 | 23.93 ± 0.09 | 50.80 | 19.90 ± 0.27 |
| S | EqualAL-SSL | ✓ | 60.26 | 23.96 ± 0.27 | 51.08 | 20.02 ± 0.04 |
| B | Coreset-SSL | ✓ | **63.14** | **24.47** ± 0.07 | 51.49 | 20.02 ± 0.15 |
| - | 100% | ✗ | 68.42 | 27.37 ± 0.07 | 56.87 | 22.75 |
| *With MC-Dropout decoder* | | | | | | |
| S | BALD | ✗ | 61.87 | 24.28 ± 0.19 | **52.82** | **20.32** ± 0.17 |
| S | BALD-SSL | ✓ | 61.13 | 23.89 ± 0.16 | 52.29 | 20.14 ± 0.12 |
| B | Coreset-MCD | ✗ | 60.60 | 23.78 ± 0.10 | 49.99 | 19.45 ± 0.17 |
| B | Coreset-MCD-SSL | ✓ | **62.24** | **24.37** ± 0.10 | 51.76 | 19.97 ± 0.10 |
| - | 100%-MCD | ✗ | 67.07 | 26.83 | 56.47 | 22.59 |

AUC@50 and mIoU@30 metrics are reported. A denotes the Acquisition method type. S and B denote the single-sample and batch-based acquisition, respectively. We observe that single-sample based methods work better for diverse datasets. Standard deviation over 3 runs are included with the AUC values

to put a redundant/diverse tag for many datasets in the middle of the spectrum. Cityscapes and A2D2-Pool-0f fall in this spectrum since they are curated by sparsely selecting from large video stream data. We consider them as non-redundant/diverse for our study since they behave more like diverse datasets.

**Evaluation metrics** We use mean Intersection over Union (mIoU) to evaluate the performance of the model at each AL cycle step. For the evaluation of the active learning method, we use two metrics: Area Under the Budget Curve (AUC@$\mathcal{B}$) and mean Intersection over Union at a budget $\mathcal{B}$ (mIoU@$\mathcal{B}$). **AUC@$\mathcal{B}$** is the area under the performance curves, shown in Figs. 5 and 12. It captures a cumulative score of the AL performance curve up to a budget $\mathcal{B}$, *where $\mathcal{B}$ is the percentage of the labeled dataset size.* For the experiments on A2D2 pools, we use a total budget of $\mathcal{B}$=50 in the 10-10 setting. For PASCAL-VOC, we run three experiments with $\mathcal{B}$=10, 25, and 50 in 2-2, 5-5, and 10-10 settings, respectively. For Cityscapes, we experiment with B=50 in the 10-10 setting. **mIoU@$\mathcal{B}$** reports the performance of the model after using a certain labeling budget $\mathcal{B}$. We report performance at an intermediate labeling budget to clearly see the ranking of the AL methods. Dataset-pools / code for all experiments are available here - https://github.com/sud0301/best_practices_ALSS

## 4.3 Results

Here, we answer the three questions raised in the introduction of the chapter concerning the behavior of active learning methods w.r.t data distribution in terms of redundancy, integration of semi-supervised learning, and different labeling budgets. For each experiment, we compare random sampling, single-sample, or batch-based acquisition approaches.

### 4.3.1 Impact of Dataset Redundancy

Table 1 and Fig. 5 show the results on Cityscapes and A2D2 Pool-0f. For both datasets, the single-sample (S) method, EqualAL, performs the best in the supervised-only setting. Table 2 and Fig. 11 shows the results obtained on the PASCAL-VOC dataset in 5-5 and 10-10 settings. Single-sample-based methods perform the best in the 10-10 setting, whereas Coreset performs the best in the 5-5 AL setting by a marginal gap w.r.t. random baseline. Table 3, 4 and Fig. 12 show the results for the redundant datasets. The batch-based Coreset method consistently performs the best in all four datasets in the supervised-only setting.

**Diverse datasets need a single-sample method, and redundant datasets need a batch-based method.** We observe that the order of best-performing models changes based on the level of redundancy in the dataset. Single-sample-based acquisition functions perform best on diverse datasets, whereas batch-based acquisition functions perform best on redundant datasets. We attribute this reversed effect to the mode collapse problem, where, for redundant datasets, single-sample acquisition methods select local clusters of similar samples. Diverse datasets are devoid of this issue as they do not possess local clusters due to high diversity across samples. Therefore, diversity-driven acquisition is not critical for diverse datasets.

**Table 2** Active Learning results on PASCAL-VOC dataset in 5-5 and 10-10 settings

| A | AL Method | SSL | PASCAL: 5-5 | | PASCAL: 10-10 | |
|---|---|---|---|---|---|---|
| | Metric → | | mIoU | AUC | mIoU | AUC |
| S | Random | ✗ | 70.70 | 13.92 ± 0.17 | 72.13 | 28.85 ± 0.21 |
| S | Entropy | ✗ | 70.38 | 13.94 ± 0.17 | 73.72 | 29.17 ± 0.26 |
| S | EqualAL | ✗ | 69.14 | 13.82 ± 0.11 | 73.40 | 29.03 ± 0.09 |
| B | Coreset | ✗ | 70.85 | 13.96 ± 0.05 | 73.63 | 29.06 ± 0.11 |
| S | Random-SSL | ✓ | 72.57 | 14.36 ± 0.07 | 75.33 | 29.87 ± 0.08 |
| S | Entropy-SSL | ✓ | 73.36 | 14.51 ± 0.07 | **76.08** | 30.01 ± 0.11 |
| S | EqualAL-SSL | ✓ | **73.39** | **14.55** ± 0.02 | 75.89 | **30.06** ± 0.05 |
| B | Coreset-SSL | ✓ | 72.88 | 14.46 ± 0.09 | 75.91 | 30.03 ± 0.01 |
| - | 100% | ✗ | 77.00 | 15.40 | 77.00 | 30.80 |

AUC@50 and mIoU@30 metric are reported. S and B denotes the single-sample and batch-based acquisition, respectively. We observe all methods show similar performance. Single-sample methods with SSL show marginally better performance than batch-based methods. Standard deviation over 3 runs are included with the AUC values

**Table 3** Active Learning results on A2D2-Pool5f, A2D2-Pool11f

| A | AL Method | SSL | Pool-5f | | Pool-11f | |
|---|---|---|---|---|---|---|
| | Metric → | | mIoU | AUC | mIoU | AUC |
| S | Random | ✗ | 47.58 | 18.69 ± 0.14 | 44.61 | 17.76 ± 0.07 |
| S | Entropy | ✗ | 49.96 | 19.48 ± 0.04 | 47.43 | 18.52 ± 0.06 |
| S | EqualAL | ✗ | 49.50 | 19.29 ± 0.08 | 47.14 | 18.44 ± 0.20 |
| B | Coreset | ✗ | 50.08 | 19.44 ± 0.21 | 47.72 | 18.69 ± 0.06 |
| S | Random-SSL | ✓ | 47.92 | 19.03 ± 0.21 | 45.25 | 18.02 ± 0.24 |
| S | Entropy-SSL | ✓ | 48.78 | 19.31 ± 0.16 | 47.53 | 18.56 ± 0.06 |
| S | EqualAL-SSL | ✓ | 48.80 | 19.28 ± 0.09 | 46.50 | 18.39 ± 0.16 |
| B | Coreset-SSL | ✓ | **50.44** | **19.69** ± 0.03 | **48.99** | **19.01** ± 0.12 |
| - | 100% | ✗ | 53.25 | 21.30 ± 0.14 | 48.85 | 19.54 |
| *With MC-Dropout decoder* | | | | | | |
| S | BALD | ✗ | **50.40** | 19.29 ± 0.07 | 47.85 | 18.74 ± 0.15 |
| S | BALD-SSL | ✓ | 50.33 | 19.62 ± 0.06 | 47.34 | 18.61 ± 0.12 |
| B | Coreset-MCD | ✗ | **50.40** | 19.49 ± 0.19 | 47.67 | 18.61 ± 0.08 |
| B | Coreset-MCD-SSL | ✓ | 50.28 | **19.65** ± 0.09 | **48.60** | **18.96** ± 0.06 |
| - | 100%-MCD | ✗ | 53.82 | 21.53 | 50.86 | 20.34 |

AUC@50 and mIoU@30 metrics are reported. S and B denotes the single-sample and batch-based acquisition, respectively. We observe that batch-based methods consistently outperform single-sample methods on redundant datasets. Standard deviation over 3 runs are included with the AUC values

This observation is consistent for PASCAL-VOC, where single-sample-based uncertainty-type methods perform better than batch-based and random methods in the high-budget setting. The difference between the methods is only marginal here since most acquired samples add ample new information due to the highly diverse nature of the dataset. This difference further diminishes w.r.t. random baseline with a lower labeling budget (e.g. 5-5) since any learned useful bias also becomes weaker. The observations for the 5-5 setting tend towards a very low-budget setting which is further analyzed later in this section.

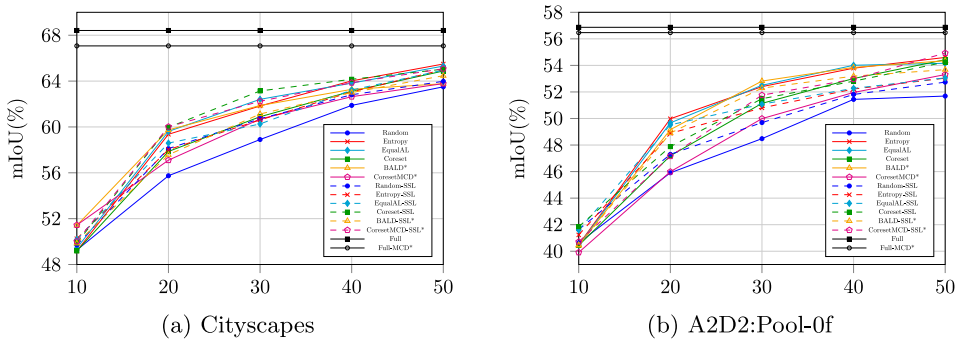
**Mode collapse analysis.** Here, we analyze and visualize the above-mentioned mode collapse issue. We provide a qualitative analysis of the mode collapse issue on the redundant A2D2 Pool-21f. We plot the feature representations using t-SNE to show the selection process for a single-sample-based Entropy function and batch-based Coreset function, shown in Fig. 6. It shows that Entropy acquisition selects many samples within local clusters, which are similar samples with overlapping information. This yields a suboptimal use of the annotation budget. In contrast, Coreset acquisition has a good selection coverage and avoids this mode collapse.

We argue that mode collapse is a common issue in many real-world datasets, containing similar samples. A good

**Table 4** Active Learning results on A2D2-Pool-21f, and A2D2-PoolAug

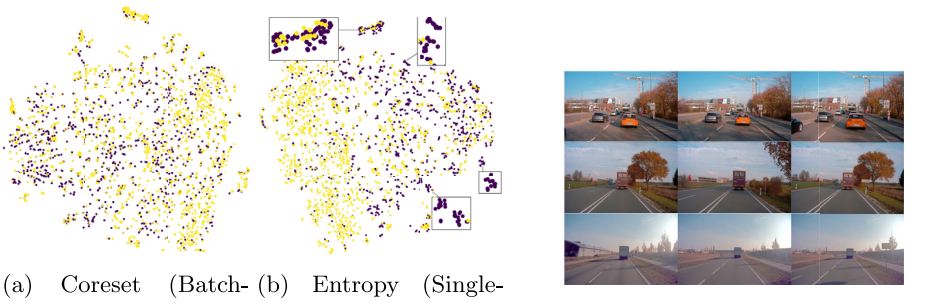
| A | AL Method | SSL | Pool-21f | | Pool-Aug | |
|---|---|---|---|---|---|---|
| | Metric → | | mIoU | AUC | mIoU | AUC |
| S | Random | ✗ | 44.52 | 17.67 ± 0.30 | 43.80 | 17.15 ± 0.15 |
| S | Entropy | ✗ | 46.08 | 18.21 ± 0.12 | 44.51 | 17.33 ± 0.24 |
| S | EqualAL | ✗ | 46.32 | 18.18 ± 0.16 | 44.24 | 17.29 ± 0.20 |
| B | Coreset | ✗ | 46.68 | 18.38 ± 0.11 | 44.70 | 17.54 ± 0.15 |
| S | Random-SSL | ✓ | 46.27 | 18.19 ± 0.18 | 44.17 | 17.29 ± 0.29 |
| S | Entropy-SSL | ✓ | 46.93 | 18.43 ± 0.05 | 44.50 | 17.47 ± 0.09 |
| S | EqualAL-SSL | ✓ | 47.11 | 18.54 ± 0.01 | 44.81 | 17.56 ± 0.13 |
| B | Coreset-SSL | ✓ | **47.62** | **18.69** ± 0.14 | **45.81** | **17.74** ± 0.04 |
| - | 100% | ✗ | 49.23 | 19.69 | 46.03 | 18.41 |
| *With MC-Dropout decoder* | | | | | | |
| S | BALD | ✗ | 46.78 | 18.57 ± 0.13 | **45.53** | **17.80** ± 0.07 |
| S | BALD-SSL | ✓ | 47.06 | 18.57 ± 0.24 | 45.16 | 17.72 ± 0.21 |
| B | Coreset-MCD | ✗ | 46.86 | 18.35 ± 0.09 | 44.74 | 17.50 ± 0.08 |
| B | Coreset-MCD-SSL | ✓ | **47.73** | **18.75** ± 0.06 | 45.37 | 17.75 ± 0.14 |
| - | 100%-MCD | ✗ | 50.43 | 20.17 | 46.62 | 18.65 |

AUC@50 and mIoU@30 metrics are reported. S and B denotes the single-sample and batch-based acquisition, respectively. We observe that batch-based methods consistently outperform single-sample methods on redundant datasets. Standard deviation over 3 runs are included with the AUC values




(a) Cityscapes (b) A2D2:Pool-0f

**Fig. 5** Results on diverse driving datasets. Active learning performance curves on Cityscapes and A2D2:Pool-0f. X-axis shows the percentage of labeled dataset. The methods which utilize MC-dropout in their network architecture are marked with ∗, and are only comparable to other methods with MC-dropout



(a) Coreset (Batch-based AL) (b) Entropy (Single-sample AL) (c) Mode collapse (Single-sample AL)

**Fig. 6** TSNE plots of **a** Coreset and **b** Entropy functions for A2D2 Pool-21f. The yellow points are feature representation from the unlabeled set, the violet points are the acquired points. The batch-based approach has good selection coverage, whereas the single-sample acquisition approach selects similar samples from clusters. Figure **c** shows acquired redundant samples from the violet clusters in (**b**) (Color figure online)

**Table 5** Active Learning results on the PASCAL-VOC and Cityscapes dataset in low-budget 2-2 setting. AUC@10 and mIoU@6 metric are reported.

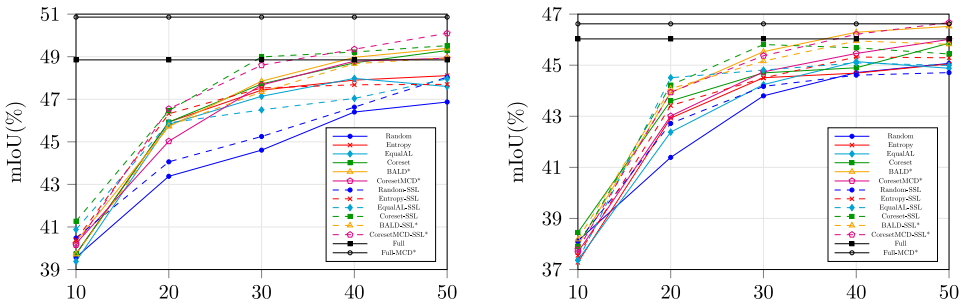
| A | AL Method | SSL | Cityscapes: 2-2 | | PASCAL: 2-2 | |
|---|---|---|---|---|---|---|
| | Metric → | | mIoU@6 | AUC@10 | mIoU@6 | AUC@10 |
| S | Random | ✗ | 46.05 | 3.65 | 66.41 | 5.22 |
| S | Entropy | ✗ | **51.24** | **4.00** | 66.33 | 5.11 |
| B | Coreset | ✗ | 47.26 | 3.74 | 66.24 | 5.19 |
| S | Random-SSL | ✓ | 47.46 | 3.72 | **68.60** | **5.37** |
| S | Entropy-SSL | ✓ | 49.99 | 3.93 | 67.26 | 5.31 |
| B | Coreset-SSL | ✓ | 48.51 | 3.82 | 68.03 | 5.35 |
| - | 100% | ✗ | 68.42 | 5.47 | 77.00 | 6.16 |

A denotes Acquisition method type. S and B denotes the single-sample and batch-based acquisition, respectively

**Table 6** Active Learning results on A2D2 Pool-0f in 2-2 setting and on the BCSS dataset in 1-1 setting

| A | AL Method | SSL | A2D2 Pool-11f 2-2 | | BCSS 1-1 | |
|---|---|---|---|---|---|---|
| | Metric → | | mIoU@6 | AUC@10 | mIoU@3 | AUC@5 |
| S | Random | ✗ | 37.74 | 2.93 | 56.29 | 2.24 |
| S | Entropy | ✗ | 36.37 | 2.92 | 53.67 | 2.15 |
| S | EqualAL | ✗ | 37.28 | 2.97 | 53.72 | 2.15 |
| B | Coreset | ✗ | **39.63** | **3.10** | 56.68 | 2.37 |
| S | Random-SSL | ✓ | 36.46 | 2.90 | 58.93 | 2.34 |
| S | Entropy-SSL | ✓ | 36.70 | 2.93 | 57.96 | 2.28 |
| S | EqualAL-SSL | ✓ | 36.31 | 3.06 | 58.75 | 2.32 |
| B | Coreset-SSL | ✓ | 39.20 | 3.06 | **60.1** | **2.38** |
| - | 100% | ✗ | 48.85 | 3.91 | 65.71 | 2.63 |

AUC@10/AUC@5 and mIoU@6/mIoU@3 metrics are reported. A denotes Acquisition method type. S and B denotes the single-sample and batch-based acquisition, respectively




(a) A2D2:Pool-11f (b) A2D2:Pool-Aug

**Fig. 7** Results on redundant datasets. Active Learning performance curves on A2D2 dataset: Pool-11f and Pool-Aug. The X-axis shows the percentage of labeled datasets. The methods which utilize MC-Dropout in their network architecture are marked with ∗, and are only comparable to other methods with MC-Dropout

acquisition function for such datasets must be aware of the batch's diversity to address the mode collapse issue. It is largely ignored due to the narrow scope of existing AL benchmarks like PASCAL-VOC and Cityscapes, which only cover diverse datasets.

#### 4.3.2 Systematic Integration of SSL

For all redundant datasets, the Coreset-SSL approach consistently performs the best; see results in Table 3, 4 and Figs. 7, 12. For diverse datasets, SSL integration is also helpful, but there is no consistent best approach. For the PASCAL-VOC dataset, single-sample-based methods with SSL show the best performance, shown in Table 2. For Cityscapes, Coreset-

SSL outperforms all other approaches; see Table 1 and Fig. 5. For A2D2-Pool0f, Coreset-SSL improves over Coreset, but the single-sample acquisition method BALD approach shows the best performance.

**Redundant datasets favor the integration of batch-based active learning and semi-supervised learning.** The batch-based acquisition function Coreset always profits from the integration of SSL. Coreset aligns well with the SSL objective since Coreset selects samples from each local cluster, thus covering the whole data distribution. This assists SSL in obtaining maximum information from the unlabeled samples, as discussed in Sect. 4.1. This effect is especially strong in the redundant A2D2 pools, where Coreset-SSL always improves over Coreset and also shows the best performance. In contrast, SSL integration for single-sample methods is either harmful or ineffective, except for the PASCAL-VOC dataset. Interestingly, in Pool-11f, some Coreset-SSL methods even outperform the 100% baseline with less than 30% labeled data. This indicates that some labeled redundant samples can even harm the model (see Fig. 12), possibly due to data imbalance. For Cityscapes, SSL with Coreset yields significant improvement, and SSL even changes the ranking of the methods. We see that EqualAL performs the best in the supervised-only setting, whereas Coreset-SSL surpasses all methods. This slight anomaly in the case of Cityscapes happens because the advantage due to the combination of SSL and batch-based method is greater than the advantage of using single-sample methods in non-redundant datasets. For diverse PASCAL-VOC, all methods align well with SSL. All methods perform well with no clear winner method since all selection criteria select samples that provide good coverage of the data distribution.

#### 4.3.3 Low Annotation Budget

*Active learning is volatile with a low budget.* Experimenting with PASCAL-VOC in the 2-2 budget setting, Random-SSL performs the best, i.e., semi-supervised learning without an active learning component (see Table 5 and Fig. 11). We believe that active learning fails in this setting because it fails to capture any helpful bias for selection in such a low-data regime with diverse samples. Our observations in this low-budget setting confirm and provide stronger empirical support for similar behavior observed in the case of image classification in Sect. 3. For A2D2 Pool-0f and Cityscapes in the 2-2 setting (see Table 5 and 12), the single-sample acquisition performs the best, while its SSL integration is detrimental. These methods possibly learn some useful bias due to the specialized driving domain. For redundant datasets in low-budget settings, the batch-based acquisition is still the most effective way. However, SSL does not contribute any additional improvements for the A2D2 Pool-11f (Table 6) due to insufficient labeled samples to support learning from unlabeled samples. In medical image processing this use case of having a low budget due to the high cost of professionals for annotation and redundant data distributions is especially common. For the medical BCSS dataset the integration of coreset and SSL yields the best results. The two single sample acquisition methods (Entropy and Equal) yield worse results than the random acquisition. This indicates the mode collapse issue caused by the redundant distribution. This results indicates that batch based AL approaches are especially important in the medical domain. Overall, we observe a highly volatile nature of active learning in conjunction with a low budget. The ideal policy transitions from random selection towards batch-based acquisition, as the dataset redundancy goes from low to high.

**Table 7** AL results on the proposed A2D2-3k task

| A | AL Method | SSL | mIoU | AUC |
|---|---|---|---|---|
| B | Uniform | ✗ | 57.75 | – |
| S | Random | ✗ | 56.14 | 5.35 |
| S | Entropy | ✗ | 60.16 | 5.53 |
| B | Coreset | ✗ | 60.30 | 5.55 |
| S | Uniform (@5) + Entropy | ✗ | 60.40 | 5.66 |
| B | Uniform-SSL | ✓ | 58.93 | – |
| S | Random-SSL | ✓ | 57.57 | 5.53 |
| S | Entropy-SSL | ✓ | 59.91 | 5.61 |
| B | Coreset-SSL | ✓ | **61.13** | **5.72** |
| S | Uniform (@5) + Ent-SSL | ✓ | 59.63 | 5.59 |
| - | 100% | ✗ | 66.65 | 6.64 |

mIoU@7.5 and AUC@7.5 are reported. S and B denote the single-sample and batch-based acquisition, respectively. Uniform refers to the temporal subsampling selection process and (@5) means every $5^{th}$ frame. This result shows that batch-based methods with SSL is the best AL strategy for realistic redundant datasets

**Table 8** Overview showing the best performing AL method for each scenario

| Dataset ↓ | Annotation Budget | |
|---|---|---|
| | Low | High |
| Diverse | Random-SSL | Single-SSL |
| Redundant | Batch | Batch-SSL |

Single and Batch refer to single-sample and batch-based method, and Random refers to random selection. Suffix -SSL refers to the usage of semi-supervised learning, otherwise refers that standard supervised learning performed better than SSL based approach

### 4.4 An Exemplar Case Study: A2D2-3K Task

Previous active learning works on semantic segmentation cover only the combination of a diverse dataset and a high annotation budget. In contrast, the collected raw data can be quite redundant, like in video datasets. To study this miss-

ing redundant setting, we propose a new active learning task A2D2-3K for segmentation based on the A2D2 dataset. The aim of the new task is to select 3K images (similar size to Cityscapes) from the original A2D2 dataset (∼40K images) to achieve the best performance. We select 3K images using active learning in 3 cycles with 1K images each. We compare 5 acquisition functions, including Random, Entropy, and Coreset, along with SSL integration. Such video datasets are often manually subsampled based on some prior information like time or location and then used for active learning. Therefore, we also include two such baselines - (a) where 3K samples are uniformly selected based on time information, denoted as Uniform, and (b) where every fifth sample is first selected uniformly to select $\sim 8K$ samples and then applied with Entropy acquisition function, denoted as Uniform(@5)+Entropy. The second approach is closer to previously used active learning benchmarks in the driving context. Results are shown in Table 7. We find that the batch-based Coreset-SSL method performs the best, discussed in Sect. 4.3.2, while the subsampling-based approaches are sub-optimal. This makes an excellent case for active learning in datasets with high redundancy, as active learning filters the data better than time-based subsampling methods.

## 5 Conclusion

In Sect. 3, we only studied active learning models for image classification under the influence of data augmentation, with the integration of semi-supervised learning under different annotation budgets. Our experiments provide strong evidence that the current evaluation protocol used in active learning for image classification is sub-optimal, leading to wrong conclusions about the methods' performance and the state of the field in general. Evaluating CIFAR-100, which is marginally different from CIFAR-10, dramatically changes the ranking of the methods. Applying state-of-the-art data augmentation significantly increases the scores of all methods, making them virtually indistinguishable in terms of the final performance. Modern semi-supervised learning algorithms applied in the conventional active learning setting show a higher relative performance increase than any of the active learning methods proposed in recent years. State-of-the-art active learning approaches often fail to outperform simple random sampling, especially when the labeling budget is small - a setting crucial for many real-world applications.

Our experiments in Sect. 4 shows that active learning is indeed a useful tool for semantic segmentation. However, it is vital to understand the behavior of different active learning methods in various application scenarios. Table 8 provides an overview of the best-performing methods for each scenario for the semantic segmentation task. Our findings indicate that single-sample-based uncertainty is a suitable measure for sample selection in diverse datasets. In contrast, batch-based diversity-driven measures are better suited for datasets with high levels of redundancy. SSL is successfully integrated with batch-based diversity-driven methods. However, it can have a detrimental impact when combined with single-sample-based uncertainty acquisition functions. Active learning with a high annotation budget always performs better than random sampling and is further improved with the integration of semi-supervised learning. The batch-based methods are successful when there is a certain presence of redundancy in the dataset. Active learning with low annotation budgets is highly sensitive to the level of redundancy in the dataset. The optimal active learning policy changes from random selection to single-sample selection and then to batch-based selection based on the level of redundancy in the dataset. These findings have been missing in method development, which is usually optimized only for a few scenarios. The results of this study facilitate a broader view of the task with presumably positive effects in many applications.

## A AL for Image classification

### A.1 Experiment Details

**Training details.** For the network architecture, we consistently use the Wide-Resnet-Network (Zagoruyko & Komodakis, 2016) with depth=28 and width=2 (WRN-28-2). We select WRN due to its efficiency and widespread adoption. WRN-28-2 contains only 1.5M parameters showing close-to-state-of-the-art performance on CIFAR datasets. The WRN-28-2 classification network is optimized using an SGD optimizer with a base learning rate of 3e-2, momentum of 0.9, and weight decay rate of 5e-4. We use a cosine learning rate schedule for training each model. We trained all AL methods (without SSL methods) for 150 epochs per sampling cycle with a batch size of 64. We train the semi-supervised AL methods for 50k iterations per sampling cycle with a batch size of 64 for the labeled loss and a batch size of 320 for the unlabeled loss. We mask out unlabeled examples whose highest probabilities across categories are less than 0.6 and set the softmax-temperature scaling constant to 0.5. Other hyperparameters are used exactly as proposed in Xie et al. (2019). Our implementation is based on the open-source toolbox Pytorch (Paszke et al., 2017).

All results are shown as performance curves. We report the mean performance over 3 trials with different initial labeled sets for all single model-based methods and over 2 trials for ensemble-based methods due to higher computation cost and lower variance.

LL method usually starts with a higher initial performance due to the extra regularization effect from the loss-prediction module. All other methods start from similar initial perfor-

**Table 9** Active Learning results on CIFAR-10 and CIFAR-100 datasets. AUC@20K and Acc@12.5K metrics are reported. The table compares AL results with and without the usage of strong data augmentation during the training of the model

| AL Method | Strong | CIFAR-10 | | CIFAR-100 | |
|---|---|---|---|---|---|
| Metric → | Aug. | Acc | AUC | Acc | AUC |
| Random w/o Aug | ✗ | 76.43 | 44.87 | 34.35 | 20.59 |
| Entropy w/o Aug | ✗ | 76.11 | 45.28 | 34.16 | 20.27 |
| LL w/o Aug | ✗ | 78.71 | 46.09 | 35.28 | 20.77 |
| Coreset w/o Aug | ✗ | 76.94 | 45.39 | 35.57 | 20.98 |
| ENS-varR w/o Aug | ✗ | 77.39 | 45.51 | 35.03 | 20.68 |
| Random w/ Aug | ✓ | 89.55 | 53.27 | 61.67 | 35.99 |
| Entropy w/ Aug | ✓ | 90.76 | 53.96 | 60.01 | 35.31 |
| LL w/ Aug | ✓ | **90.88** | **54.06** | 60.04 | 35.31 |
| Coreset w/ Aug | ✓ | 90.63 | 53.89 | 62.13 | 36.21 |
| ENS-varR w/ Aug | ✓ | 90.58 | 53.94 | **62.15** | **36.25** |
| 100% | ✓ | 95.80 | 57.48 | 75.82 | 45.49 |

**Table 10** Active Learning results on CIFAR-10 and CIFAR-100. AUC@20K and mIoU@12.5K metrics are reported. The table compares AL methods with and without semi-supervised learning

| AL Method | Strong | SSL | CIFAR-10 | | CIFAR-100 | |
|---|---|---|---|---|---|---|
| Metric → | Aug. | | Acc | AUC | Acc | AUC |
| Random w/ Aug | ✓ | ✗ | 89.55 | 53.27 | 61.67 | 35.99 |
| Entropy w/ Aug | ✓ | ✗ | 90.76 | 53.96 | 60.01 | 35.31 |
| LL w/ Aug | ✓ | ✗ | 90.88 | 54.06 | 60.04 | 35.31 |
| Coreset w/ Aug | ✓ | ✗ | 90.63 | 53.89 | 62.13 | 36.21 |
| ENS-varR w/ Aug | ✓ | ✗ | 90.58 | 53.84 | 62.15 | 36.25 |
| Random-SSL | ✓ | ✓ | 94.11 | 56.33 | 68.14 | 40.19 |
| Entropy-SSL | ✓ | ✓ | 94.32 | 56.43 | 67.54 | 40.02 |
| LL-SSL | ✓ | ✓ | **94.58** | **56.58** | 66.99 | 39.70 |
| Coreset-SSL | ✓ | ✓ | 94.27 | 56.41 | 67.17 | 39.94 |
| ENS-varR-SSL | ✓ | ✓ | 94.21 | 56.44 | **68.42** | **40.40** |
| 100% | ✓ | ✗ | 95.80 | 57.48 | 75.82 | 45.49 |

**Table 11** Active Learning results on CIFAR-10 and CIFAR-100. AUC@2K and mIoU@1K metrics are reported. The table compares AL methods with and without semi-supervised learning in a low-annotation budget setting

| AL Method | Strong | SSL | CIFAR-10 | | CIFAR-100 | |
|---|---|---|---|---|---|---|
| Metric → | Aug. | | Acc | AUC | Acc | AUC |
| Random w/ Aug | ✓ | ✗ | 61.93 | 1.91 | 24.13 | 1.68 |
| Entropy w/ Aug | ✓ | ✗ | 55.79 | 1.76 | 20.98 | 1.43 |
| LL w/ Aug | ✓ | ✗ | 55.75 | 1.79 | 21.27 | 1.50 |
| Coreset w/ Aug | ✓ | ✗ | 59.64 | 1.85 | 25.11 | 1.72 |
| ENS-varR w/ Aug | ✓ | ✗ | 61.00 | 1.91 | 24.31 | 1.73 |
| Random-SSL | ✓ | ✓ | **88.78** | **2.67** | **44.42** | **2.80** |
| Entropy-SSL | ✓ | ✓ | 76.30 | 2.35 | 27.46 | 1.84 |
| LL-SSL | ✓ | ✓ | 78.90 | 2.44 | 40.36 | 2.57 |
| Coreset-SSL | ✓ | ✓ | 84.93 | 2.55 | 38.37 | 2.50 |
| ENS-varR-SSL | ✓ | ✓ | 86.07 | 2.60 | 40.91 | 2.65 |

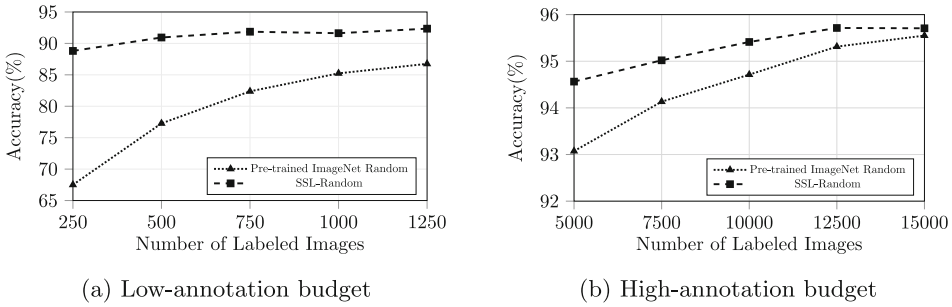



**Fig. 8** The SSL-Random baseline clearly outperforms a fine-tuned network pre-trained on ImageNet in the low-budget setting. Results shown on CIFAR-10. The SSL-Random baseline clearly outperforms a fine-tuned network pre-trained on ImageNet in the large-budget setting. Results shown on CIFAR-10

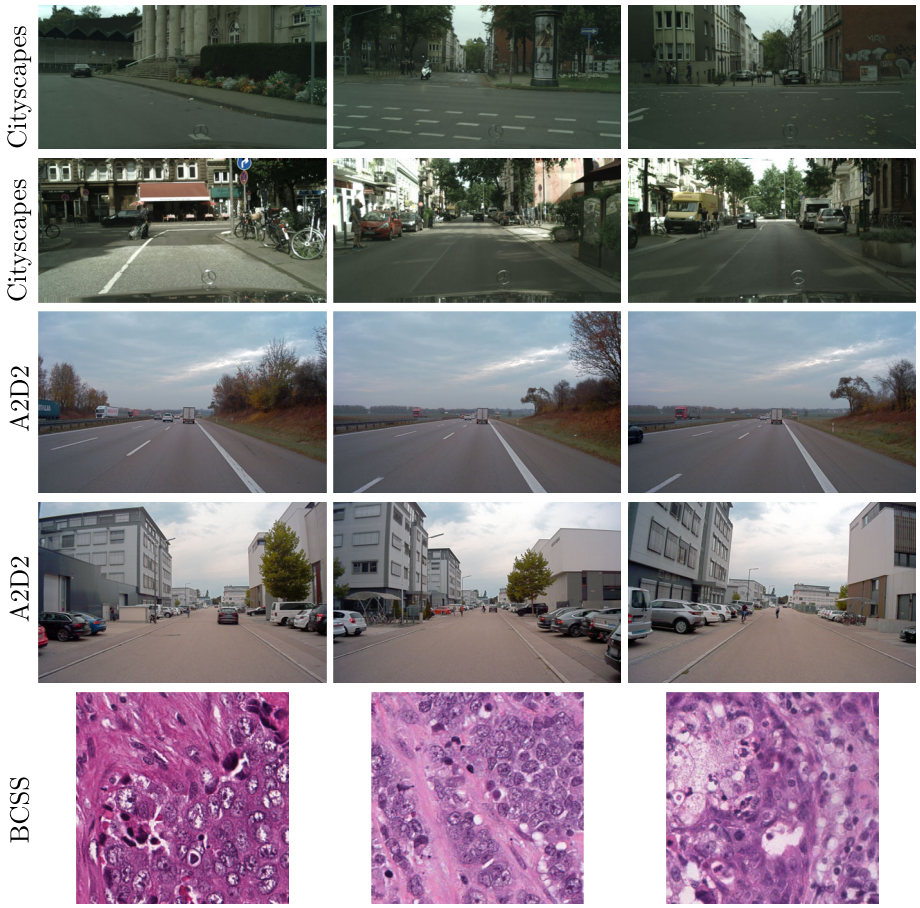



**Fig. 9** Consecutive images from the Cityscapes and A2D2 datasets. This shows even the consecutive images in the Cityscapes dataset are different and diverse, whereas consecutive frames in the A2D2 dataset are very similar, containing redundant information. The bottom row shows examples from the medical BCSS dataset

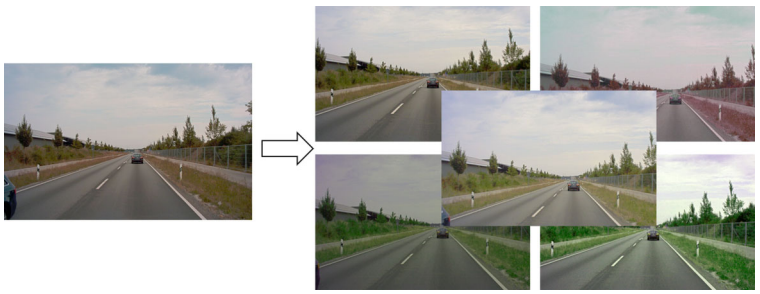
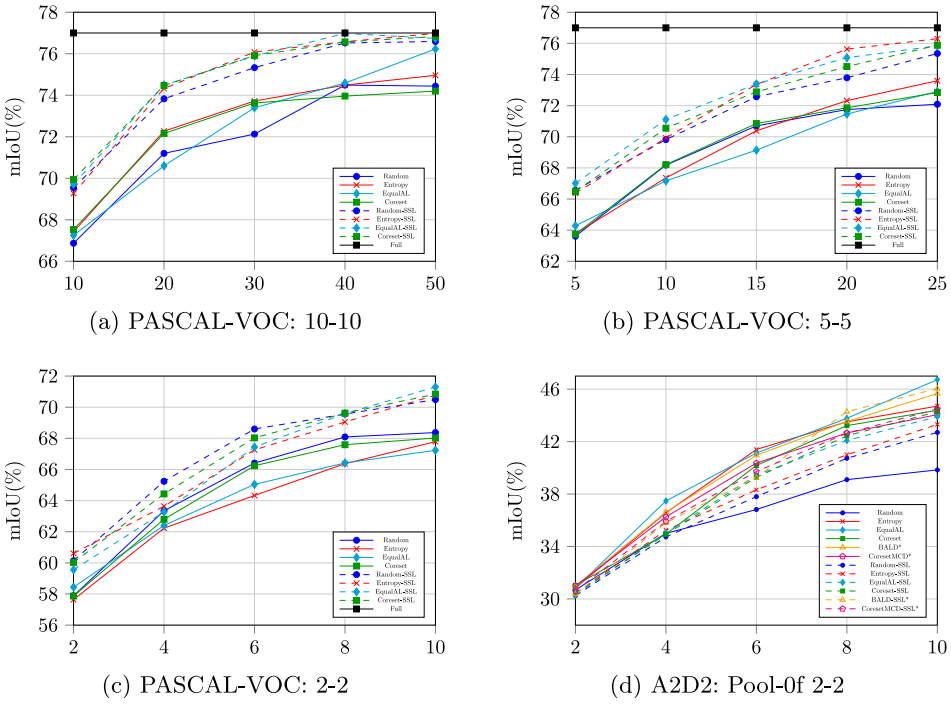


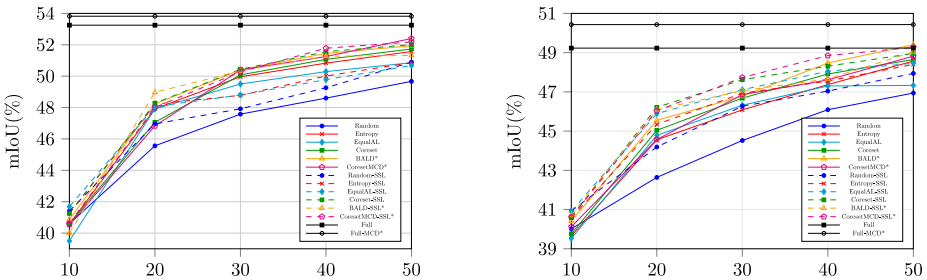
**Fig. 10** A2D2 Pool-Aug. Left: the original image. Right: the duplication through color augmentation and random cropping of the original image




(a) PASCAL-VOC: 10-10 (b) PASCAL-VOC: 5-5

(c) PASCAL-VOC: 2-2 (d) A2D2: Pool-0f 2-2

**Fig. 11** Active learning performance curves on PASCAL-VOC and A2D2:Pool-0f. X-axis shows the percentage of labeled dataset. The methods which utilize MC-Dropout in their network architecture are marked with ∗, and are only comparable to other methods with MC-Dropout




(a) A2D2:Pool-5f (b) A2D2:Pool-21f

**Fig. 12** Results on redundant datasets. Active Learning performance curves on A2D2 dataset: Pool-5f and Pool-21f. The X-axis shows the percentage of labeled datasets. The methods which utilize MC-Dropout in their network architecture are marked with ∗, and are only comparable to other methods with MC-Dropout

mance with a slight difference due to the model variance. This variance is more prominent in the beginning due to the overfitting effect on a small labeled set.

## A.2 Results Tables

***More Results.*** Table 9 shows final Accuracy and AUC for CIFAR-10 and CIFAR-100 datasets for high annotation budget. The table compares the AL results with and without data augmentation. We observer that AL performance improves significantly by using data augmentation for image classification task. Table 10 shows the final accuracy and AUC for CIFAR-10 and CIFAR-100 datasets with and without semi-supervised learning. Adding semi-supervised learning to AL methods shows a significant improvement. However, the gap between random-SSL and other AL-SSL methods is not significant. Table 11 shows final accuracy and AUC for CIFAR-10 and CIFAR-100 datasets for low annotation budget. We observe that Random-SSL performs significantly better than other active learning methods.

***Comparison to Transfer Learning.*** Oliver et al. (2018) argued that transfer learning might be a preferable alternative to semi-supervised learning when a suitable labeled dataset is available for transfer learning. Following the recommendation, we compare the performance of the SSL-Random baseline with a fine-tuned ImageNet pre-trained network on CIFAR-10.

The ImageNet pre-trained network is fine-tuned only on the labeled samples. The experiment was conducted with Resnet-18 due to the availability of pre-trained ImageNet weights. We observe that the SSL-AL method clearly outperforms fine-tuning of a pre-trained ImageNet network in both high- and low-budget settings. We tested both budget setting for 4 sampling cycles. The corresponding results are shown in Fig. 8a and b, respectively. This experiment shows that including an up-to-date semi-supervised learning algo-

**Table 12** Active Learning results on A2D2 Pool-0f in 2-2 setting. AUC@10 and mIoU@6 metrics are reported. A denotes Acquisition method type. S and B denotes the single-sample and batch-based acquisition, respectively

| A | AL Method | SSL | A2D2 Pool-0f 2-2 | |
|---|---|---|---|---|
| | Metric → | | mIoU@6 | AUC@10 |
| S | Random | ✗ | 36.82 | 2.92 |
| S | Entropy | ✗ | **41.40** | 3.18 |
| S | EqualAL | ✗ | 41.13 | **3.22** |
| B | Coreset | ✗ | 40.18 | 3.12 |
| S | Random-SSL | ✓ | 37.80 | 2.99 |
| S | Entropy-SSL | ✓ | 38.32 | 3.03 |
| S | EqualAL-SSL | ✓ | 39.43 | 3.07 |
| B | Coreset-SSL | ✓ | 39.28 | 3.08 |
| - | 100% | ✗ | 56.87 | 4.55 |

rithm in an active learning pipeline makes sense even when large pre-training data is available.

*A2D2-Pool-Aug* The fifth pool, Pool-Aug, contains augmented duplicates in place of the consecutive frames. We create five duplicates of each randomly selected frame by randomly cropping 85% color augmentation (see Fig. 10).

## B AL for Semantic Segmentation

### B.1 Experiment Details

*Datasets visualization* Fig. 9 shows examples of the A2D2 and the Cityscapes dataset. Each row shows three temporally consecutive frames in both labeled datasets. We clearly observe that the images in the A2D2 dataset have high-overlapping information, whereas images in the Cityscapes dataset are quite diverse. Therefore, to create our redundancy experiments, we chose the A2D2 dataset as the base dataset.

*Training details* We used the DeepLabv3+ (Chen et al., 2018) architecture with Wide-ResNet38 (WRN-38) (Wu et al., 2019) backbone for all our experiments. The backbone WRN-38 is pre-trained using ImageNet (Deng et al., 2009). For the supervised learning setting, the model is trained using the SGD optimizer with a base-learning rate of $1e-3$, momentum of 0.9, and a weight decay of $5e-4$. We utilize a polynomial learning rate scheduler with a batch size of 8 and train a model in each AL cycle for 100 epochs. The model is trained with data augmentations, including random cropping and random horizontal flipping. Input image size is $256 \times 512$ for Cityscapes and A2D2 datasets and $321 \times 321$ for the PASCAL-VOC dataset.

We utilize the s4GAN (Mittal et al., 2019) method for semi-supervised learning (SSL). We use the same training setting for the segmentation model as in the supervised learning setting. We use the same hyperparameters as mentioned in Mittal et al. (2019), except for the learning rate of the discriminator, which is set to $2.5e-5$ for Cityscapes and A2D2 experiments. We add 3 dropout layers with a dropout rate of 0.1 in the decoder of the segmentation model for all the MCD-based AL methods.

### B.2 More Results

Figure 12 show the AL-curves of the redundant pools A2D2:Pool-5f and A2D2:Pool-21f. These curves correspond to the results of Table 3 and Table 4 in the main paper. Figure 11 adds the AL-curves for the PASCAL-VOC experiments of table 2 and the low budget experiments on PASCAL-VOC and the A2D2: Pool-0f of Table 5 and Table 12. Table 12 adds further results for the low-budget setting under diverse distributions that have been studied in Table 5 (Fig. 12).

**Funding** Open Access funding enabled and organized by Projekt DEAL. The research leading to the results on the semantic segmentation task is funded by the German Federal Ministry for Economic Affairs and Climate Action within the project "KI Delta Learning" (Forderkennzeichen 19A19013N) and "KI Wissen - Entwicklung von Methoden für die Einbindung von Wissen in maschinelles Lernen". The authors would like to thank the consortium for the successful cooperation. Funded by the Deutsche Forschungsgemeinschaft (DFG) - 401269959, 417962828. This study on the image classification task was supported by the German Federal Ministry of Education and Research via the project Deep-PTL and by the Intel Network of Intelligent Systems.

**Data Availability** Datasets used in this study are publicly available. We use CIFAR-10 and CIFAR-100 datasets for our study on image classification task. CIFAR-10 and CIFAR-100 are available here: https://www.cs.toronto.edu/ kriz/cifar.html We use PASCAL-VOC, Cityscapes, A2D2 and BCSS datasets for our experiments on semantic segmentation. All datasets are publicly available. Below are the links to access these datasets

- PASCAL-VOC - https://host.robots.ox.ac.uk/pascal/VOC/
- Cityscapes - https://www.cityscapes-dataset.com/
- A2D2 - https://www.a2d2.audi/a2d2/en.html
- BCSS - https://bcsegmentation.grand-challenge.org/

## References

Amgad, M., Elfandy, H., Hussein, H., Atteya, L. A., Elsebaie, M. A. T., Abo Elnasr, L. S., Sakr, R. A., Salem, H. S. E., Ismail, A. F., Saad, A. M., Ahmed, J., Elsebaie, M. A. T., Rahman, M., Ruhban, I. A., Elgazar, N. M., Alagha, Y., Osman, M. H., Alhusseiny, A. M., Khalaf, M. M., & Cooper, L. A. D. (2019). Structured crowdsourcing enables convolutional segmentation of histology images. *Bioinformatics, 35*(18), 3461–3467. https://doi.org/10.1093/bioinformatics/btz083

Anonymous: Rethinking deep active learning: Using unlabeled data at model training. *In Submitted to ICLR (2020)*. Under review. https://openreview.net/forum?id=rJehllrtDS

Beluch, W.H., Genewein, T., Nürnberger, A., & Köhler, J.M. (2018). The power of ensembles for active learning in image classification. *In proceedings of the IEEE conference on computer vision and pattern recognition*

Blundell, C., Cornebise, J., Kavukcuoglu, K., & Wierstra, D. (2015). Weight uncertainty in neural networks. *International Conference on Machine Learning, 37*, 1613–1622.

Brostow, G. J., Fauqueur, J., & Cipolla, R. (2008). Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters, 30*, 88.

Cai, L., Xu, X., Liew, J.H., & Foo, C.S. (2021). Revisiting superpixels for active learning in semantic segmentation with realistic annotation costs. *In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*.

Chapelle, O., Schölkopf, B., & Zien, A. (2006). *Semi-supervised learning*. The MIT Press.

Chen, L., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. *CoRR arXiv,1802*, 02611.

Chen, L., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2018). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence, 40*, 834.

Codella, N. C. F., Gutman, D., Celebi, M. E., Helba, B., Marchetti, M. A., Dusza, S. W., Kalloo, A., Liopyris, K., Mishra, N., Kittler, H., & Halpern, A. (2018). Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi). *In IEEE 15th international symposium on biomedical imaging (ISBI)*. https://doi.org/10.1109/ISBI.2018.8363547

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., & Schiele, B. (2016). The cityscapes dataset for semantic urban scene understanding. *In proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*

Cubuk, E. D., Zoph, B., Mane, D., Vasudevan, V., & Le, Q. V. (2019). Autoaugment: Learning augmentation strategies from data. *In IEEE conference on computer vision and pattern recognition*

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. *In IEEE conference on computer vision and pattern recognition*

Deng, L. (2012). The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine, 29*, 141–142.

DeVries, T., & Taylor, G.W. (2017). Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552

Everingham, M., van Gool, L., Williams, C. K. I., Winn, J., & Zisserman, A. (2010). The pascal visual object classes (voc) challenge. *International Journal of Computer Vision, 88*, 303.

Gal, Y., & Ghahramani, Z. (2016a). Bayesian convolutional neural networks with Bernoulli approximate variational inference. *In ICLR-workshop Track*

Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. *International Conference on Machine Learning, 48*, 1050.

Gal, Y., Islam, R., & Ghahramani, Z. (2017). Deep Bayesian active learning with image data. *International Conference on Machine Learning, 70*, 1183.

Gao, M., Zhang, Z., Yu, G., Arık, S.Ö., Davis, L.S., & Pfister, T. (2020). Consistency-based semi-supervised active learning: Towards minimizing labeling cost. *In ECCV*.

Geyer, J., Kassahun, Y., Mahmudi, M., Ricou, X., Durgesh, R., Chung, A. S., Hauswald, L., Pham, V. H., Mühlegg, M., Dorn, S., Fernandez, T., Jänicke, M., Mirashi, S., Savani, C., Sturm, M., Vorobiov, O., Oelker, M., Garreis, S., & Schuberth, P. (2020). A2D2: Audi autonomous driving dataset arXiv:2004.06320 [cs.CV]

Golestaneh, S.A., & Kitani, K. (2020). Importance of self-consistency in active learning for semantic segmentation. *In BMVC*

Hariharan, B., Arbelaez, P., Bourdev, L., Maji, S., & Malik, J. (2011). Semantic contours from inverse detectors. *In international conference on computer vision*

Houlsby, N., Huszár, F., Ghahramani, Z., & Lengyel, M. (2011). Bayesian active learning for classification and preference learning. ArXiv:1112.5745

Jia, R., Dao, D., Wang, B., Hubis, F.A., Hynes, N., Gürel, N.M., Li, B., Zhang, C., Song, D., & Spanos, C.J. (2019). Towards efficient data valuation based on the shapley value. *Proceedings of the twenty-second international conference on artificial intelligence and statistics*.

Johnson, E.H. (1966). Elementary applied statistics: For students in behavioral science. *Social forces*

Kendall, A., & Gal, Y. (2017). What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? *In NeurIPS*

Killamsetty, K., Sivasubramanian, D., Ramakrishnan, G., & Iyer, R. (2021). Glister: Generalization based data subset selection for efficient and robust learning. *Proceedings of the AAAI conference on artificial intelligence*.

Kim, K., Park, D., Kim, K. I., & Chun, S. Y. (2021). Task-aware variational adversarial active learning. *In proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*

Kirsch, A., Amersfoort, J., & Gal, Y. (2019). Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. In: Wallach, H., Larochelle, H., Beygelzimer, A., Alché-Buc, F., Fox, E., Garnett, R. (eds.) *Advances in neural information processing systems*. https://proceedings.neurips.cc/paper/2019/file/95323660ed2124450caaac2c46b5ed90-Paper.pdf

Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. *In NeurIPS*

Li, D., Wang, Z., Chen, Y., Jiang, R., Ding, W., & Okumura, M. (2024). A survey on deep active learning: Recent advances and new frontiers. *IEEE Transactions on Neural Networks and Learning Systems*. https://doi.org/10.1109/TNNLS.2024.3396463

Mackowiak, R., Lenz, P., Ghori, O., Diego, F., Lange, O., & Rother, C. (2018). Cereals: cost-effective region-based active learning for semantic segmentation. *In BMVC*.

Mittal, S., Tatarchenko, M., & Brox, T. (2019). Semi-supervised semantic segmentation with high-and low-level consistency. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 43*(4), 1369–1379.

Munjal, P., Hayat, N., Hayat, M., Sourati, J., & Khan, S. (2022). Towards robust and reproducible active learning using neural networks. *In proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*

Oliver, A., Odena, A., Raffel, C. A., Cubuk, E. D., & Goodfellow, I. (2018). Realistic evaluation of deep semi-supervised learning algorithms. *In NeurIPS*

Osband, I., Blundell, C., Pritzel, A., & Van Roy, B. (2016). Deep exploration via bootstrapped dqn. *In NeurIPS*

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., & Lerer, A. (2017). Automatic differentiation in pytorch

Paul, R., Feldman, D., Rus, D., & Newman, P. (2014). Visual precis generation using coresets. *In ICRA*.

Rangnekar, A., Kanan, C., & Hoffman, M. (2022). Semantic segmentation with active semi-supervised representation learning. arXiv. https://doi.org/10.48550/ARXIV.2210.08403 . https://arxiv.org/abs/2210.08403

Ravanbakhsh, M., Klein, T., Batmanghelich, K., & Nabi, M. (2019). Uncertainty-driven semantic segmentation through human-machine collaborative learning. *In international conference on medical imaging with deep learning: Extended abstract track*.

Ren, P., Xiao, Y., Chang, X., Huang, P.-Y., Li, Z., Gupta, B. B., Chen, X., & Wang, X. (2021). A survey of deep active learning. *ACM Computing Surveys (CSUR), 54*(9), 1–40.

Sener, O., & Savarese, S. (2017). Active learning for convolutional neural networks: A core-set approach. arXiv preprint arXiv:1708.00489

Seung, H. S., Opper, M., & Sompolinsky, H. (1992). Query by committee. *In annual workshop on computational learning theory*

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal, 27*, 379–423.

Shin, G., Xie, W., & Albanie, S. (2021). All you need are a few pixels: Semantic segmentation with pixelpick. *In proceedings of the IEEE/CVF international conference on computer vision (ICCV) workshops*

Shui, C., Zhou, F., Gagné, C., & Wang, B. (2020). Deep active learning: Unified and principled method for query and training. *In proceedings of the twenty third international conference on artificial intelligence and statistics*.

Simpson, A., Antonelli, M., Bakas, S., Bilello, M., Farahani, K., Ginneken, B., Kopp-Schneider, A., Landman, B., Litjens, G., Menze, B., Ronneberger, O., Summers, R., Bilic, P., Christ, P., Do, R., Gollub, M., Golia-Pernicka, J., Heckers, S., Jarnagin, W., McHugo, M., Napel, S., Vorontsov, E., Maier-Hein, L., & Cardoso, M. (February 2019). A large annotated medical image dataset for the development and evaluation of segmentation algorithms. Workingpaper

Sinha, S., Ebrahimi, S., & Darrell, T. (2019). Variational adversarial active learning. *In proceedings of the IEEE/CVF international conference on computer vision (ICCV)*

Wang, S., Li, Y., Ma, K., Ma, R., Guan, H., & Zheng, Y. (2020). Dual adversarial network for deep active learning. *In ECCV*

Wang, Y., Wang, H., Shen, Y., Fei, J., Li, W., Jin, G., Wu, L., Zhao, R., & Le, X. (2022). Semi-supervised semantic segmentation using unreliable pseudo labels. *In proceedings of the IEEE/CVF international conference on computer vision and pattern recognition (CVPR)*.

Wang, K., Zhang, D., Li, Y., Zhang, R., & Lin, L. (2017). Cost-effective active learning for deep image classification. *IEEE Transactions on Circuits and Systems for Video Technology, 27*, 2591.

Wang, K., Zhang, D., Li, Y., Zhang, R., & Lin, L. (2017). Cost-effective active learning for deep image classification. *IEEE Transactions on Circuits and Systems for Video Technology., 27*, 2591.

Wu, Z., Shen, C., & van den Hengel, A. (2019). Wider or deeper: Revisiting the resnet model for visual recognition. *Pattern Recognition, 90*, 119–133.

Xie, Q., Dai, Z., Hovy, E., Luong, M. -T., & Le, Q. V. (2019). Unsupervised data augmentation for consistency training. arXiv preprint arXiv:1904.12848

Yang, L., Zhang, Y., Chen, J., Zhang, S., & Chen, D.Z. (2017). Suggestive annotation: A deep active learning framework for biomedical image segmentation. *In medical image computing and computer assisted intervention - MICCAI*

Yoo, D., & Kweon, I.S. (2019). Learning loss for active learning. *In proceedings of the IEEE/CVF conference on computer vision and pattern recognition*

Yu, F., Xian, W., Chen, Y., Liu, F., Liao, M., Madhavan, V., & Darrell, T. (2018). BDD100K: A diverse driving video database with scalable annotation tooling. CoRR arXiv:1805.04687

Zagoruyko, S., & Komodakis, N. (2016). Wide residual networks. *In BMVC*

Zhan, X., Wang, Q., Huang, K.-h., Xiong, H., Dou, D., & Chan, A.B. (2022). A comparative survey of deep active learning. arXiv preprint arXiv:2203.13450

Zhang, Y., Ying, M. T. C., Yang, L., Ahuja, A. T., & Chen, D. Z. (2016). Coarse-to-fine stacked fully convolutional nets for lymph node segmentation in ultrasound images. *In IEEE international conference on bioinformatics and biomedicine (BIBM)*

Zhdanov, F. (2019). Diverse mini-batch active learning. CoRR arXiv:1901.05954