

# Parting with Illusions about Deep Active Learning

Sudhanshu Mittal

Maxim Tatarchenko

Özgün Çiçek

Thomas Brox

University of Freiburg

## 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 scheme used for deep active learning is below par. Current methods disregard some apparent parallel work in the closely related fields. Active learning methods are quite sensitive w.r.t. changes in the training procedure like data augmentation. They improve by a large-margin when integrated with semi-supervised learning, but barely perform better than the random baseline. We re-implement various latest active learning approaches for image classification and evaluate them under more realistic settings. We further validate our findings for semantic segmentation. Based on our observations, we realistically assess the current state of the field and propose a more suitable evaluation protocol.

## 1. Introduction

Supervised training of convolutional networks has shown remarkable success in various computer vision tasks. Its price is the collection of large datasets and their annotation. Especially the data annotation is a common bottleneck. Depending on the task, its cost may vary from a few seconds to a few hours per sample.

Active learning (AL) presumably mitigates this large annotation cost. It 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. The typical process in AL includes multiple cycles, where in each cycle a batch of samples is selected from the pool of still unlabeled data using a query function. The selected samples are manually annotated and are added to the labeled set. Then the model is re-trained. The process is repeated until the maximum annotation budget or the desired performance level is reached.

The appeal of the active learning idea has spawned a multitude of ConvNet-based AL methods. In this paper we aim to objectively assess the state of the field and challenge

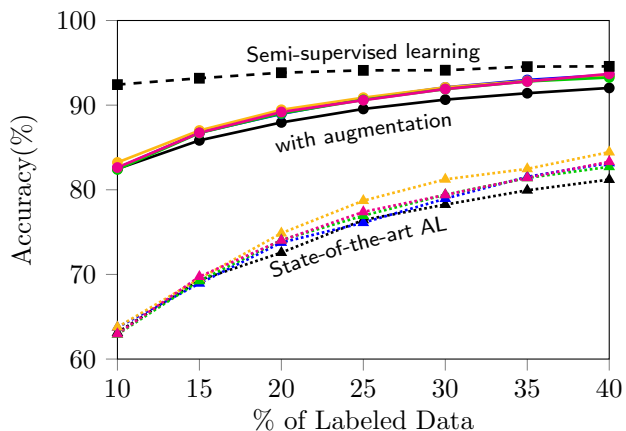


Figure 1. State-of-the-art active learning methods do not consistently use modern data augmentation techniques or advances in the closely related field of semi-supervised learning which leads to the wrong impression about the current state of the field. Results are shown for image classification on CIFAR 10.

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. Our study seeks answers to the following four scientific questions.

(1) Since a widely accepted evaluation protocol is missing, methods are often tested under incompatible circumstances: different architectures, different augmentation strategies, etc. 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 low-budget regime?

(4) Active learning is typically evaluated only on image classification tasks where manual labeling is relatively cheap compared to other tasks, *e.g.* semantic segmentation. Can active learning better exploit its potential in the more costly scope of semantic segmentation, where efficient annotation is practically more relevant?

In this work, keeping in mind the aforementioned questions, we perform an extensive comparison of existing approaches for image classification and semantic segmentation. Our experiments reveal that the progress recently made in the field of active learning is practically negligible when viewed under more realistic circumstances: in particular, using modern data augmentation and taking the advances of semi-supervised learning into account, see Figure 1. Based on our extensive study, we suggest a more suitable evaluation protocol.

## 2. Related Work

**Deep active learning (AL)** methods can be categorized into three types: uncertainty-based methods, representation-based methods and learning-based methods. Additionally, some methods have proposed a hybrid approach.

*Uncertainty-based methods* try to find the samples which are hard to learn. Several methods have been proposed to estimate uncertainty for neural networks using bayesian [8, 15, 16, 23] and non-bayesian approaches [25, 32]. Gal *et al.* [17] proposed to estimate posterior uncertainty using dropout for active learning. Wang *et al.* [41] used the entropy of the softmax output in a neural network as a proxy uncertainty measure to query samples. Beluch *et al.* [6] use ensemble method to estimate prediction uncertainty and selects new samples based on a statistical measure of committee disagreement called variation ratio [21]. They show this method outperforms all other uncertainty-based methods. *Representation-based methods* [36, 43], also referred as density-based methods, try to find a diverse set of samples that optimally represents the complete dataset distribution. Sener *et al.* [36] formulated the active learning problem as core-set selection and showed effectiveness for CNNs. *Learning-based approaches* [39, 44] use an auxiliary network module and loss function to learn a measure of information gain from new samples. Yoo *et al.* [44] proposed to learn a loss prediction module to predict target losses of unlabeled samples and selects samples with highest predicted loss. It can also be considered as a pseudo-uncertainty heuristic. Sinha *et al.* [39] 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 as a representation type method.

Many of the above mentioned approaches mainly focus on image classification. Lately, a few works have proposed

to solve tasks involving higher annotation cost like object detection [44], pedestrian detection [44], human pose estimation [27] and segmentation [20, 39]. We focus on semantic segmentation in this work, since creating segmentation masks is a highly expensive labeling task. This makes it one of the most relevant task for active learning. Suggestive Annotation [43], Cereals [28] and VAAL [39] are few works which have shown applicability of deep active learning for semantic segmentation. VAAL is a task-agnostic learning-based approach using adversarial training. Suggestive Annotation is a hybrid approach proposed for a binary segmentation problem. Cereals is a patch-based selection approach based on a hybrid heuristic of uncertainty, learned labeling cost model and spatial coherency of the image.

**Semi-supervised Learning (SSL)** methods make use of the unlabeled data for training the model. Effectively, this class of methods uses the same amount of information as the AL methods. SSL has recently seen a lot of progress due to consistency regularization [7, 42, 46]. Consistency regularization minimizes the discrepancy between class predictions of differently perturbed unlabeled image. Various additional schemes have been proposed to avoid overfitting and to improve training stability, such as temporal ensembling [24], student-teacher model [40], adversarial perturbation [30], self-supervision [46], data filtering [42], and snapshot ensembling [5]. In this work, we use a semi-supervised learning method UDA [42] for image classification.

A few recent works [19, 29] applied semi-supervised methods also to semantic segmentation. Mittal *et al.* [29] proposed s4GAN that uses a conditional GAN [18] to learn from unlabeled images, French *et al.* [14] used consistency regularization and Kalluri *et al.* [22] proposed a feature alignment objective to learning from unlabeled samples. We make use of s4GAN [29] in this work.

**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 [35, 36, 39, 41]. Sinha *et al.* [39] used unlabeled pool to learn its distribution against the distribution of labeled samples, but do not take its advantage to improve the feature representation of the target model itself. Sener *et al.* [36] have also previously shown the advantage of using the unlabeled pool for learning the target model. Wang *et al.* [41] also explored the usage of the most-certain samples from the unlabeled pool using pseudo-labeling, but the pseudo-labeling process can easily propagate erroneous labels if not tuned properly. Ravanbakhsh *et al.* [35] proposed a GAN-based approach to make use of the unlabelled pool and utilizes the discriminator score to query low-confident samples for active learning. Recently, two open-source concurrent works [2, 3] have also shown some similar findings as our work. However, they are restricted to only image classification.

Oliver *et al.* [31] raised concerns about the evaluation scheme for semi-supervised learning to help guide its applicability to real-world problems. In this work, we raise similar questions about unrealistic evaluation schemes for active learning by providing evidence through extensive experiments. In the following sections, we analyze the performance of active learning methods for image classification and semantic segmentation respectively.

### 3. Active Learning for Image Classification

In active learning, we usually start with a small set of labeled samples  $\mathcal{L}$  whose size is defined by the initial labeling budget  $\mathcal{B}_i$  and a large pool of unlabeled samples  $\mathcal{U}$ . In each cycle, a set of samples is selected from the unlabeled pool  $\mathcal{U}$  according to the sampling budget  $\mathcal{B}_s$  and added to the labeled set  $\mathcal{L}$  with the corresponding annotations provided by the oracle annotator. The selection of samples is performed using a query function, which can be learned using the full set of available samples ( $\mathcal{U} \cup \mathcal{L}$ ). This process is also called pool-based active learning. This acquisition step is iterated over several cycles until the objective is achieved.

In this section, we assess the performance of state-of-the-art AL methods for image classification and compare them with the state-of-the-art semi-supervised approach. We validate our experiments using at least one recent approach from each of three categories of AL methods as defined in the related work section.

#### 3.1. 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** [38] 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.

**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.* [6]. 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:

$$\text{varR} = 1 - \frac{f_m}{T}, \quad (1)$$

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.* [37]. 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.* [6], 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 [34]. The recent core-set approach proposed by Sener *et al.* [36] 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 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 [44] proposes a loss prediction module which 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 [42] is a semi-supervised learning method for image classification. It uses consistency regularization to learn from unlabeled samples along with AutoAugment [10] 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 [42].

### 3.2. Experiments and Results

#### 3.2.1 Evaluation protocol

**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 contain 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.

**Training Details.** For the network architecture, we consistently use the Wide-Resnet-Network [45] with depth=28

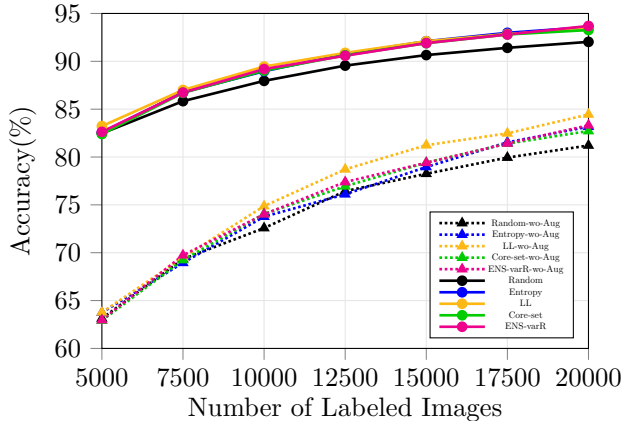


Figure 2. Using data augmentation on CIFAR-10 significantly improves the performance of active learning methods and makes the relative difference between them less pronounced. Results without augmentation are denoted as ‘X-w-o-Aug’.

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 SGD optimizer with a base learning rate of  $3e-2$ , momentum 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 [42]. Our implementation is based on the open source toolbox Pytorch [33].

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

### 3.2.2 Do AL methods work consistently well together with data augmentation?

Data augmentation is a widely accepted regularization technique, which increases the power of machine learning mod-

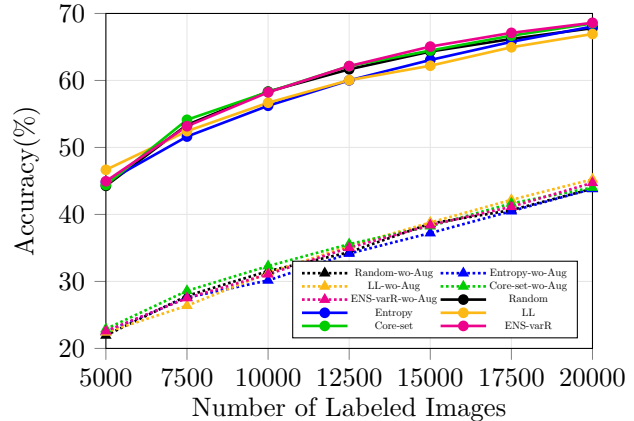


Figure 3. The performance of AL methods on CIFAR-100 improves significantly when using up-to-date image augmentation. Results without augmentation are denoted as ‘X-w-o-Aug’.

els particularly when there is little labeled data. Nevertheless, several latest AL works [6, 39] resort to either not using any augmentation during training, or only doing simplistic horizontal flipping. 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.* [10], cutout [11], 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 *et al.* [44]) 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 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 Figure 3.



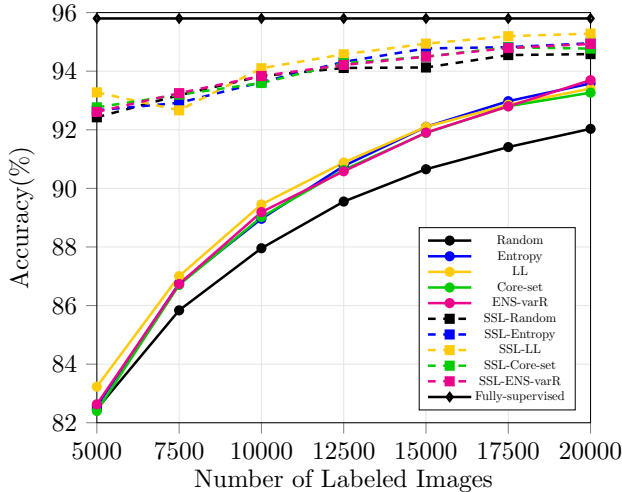


Figure 4. Combining AL methods with semi-supervised learning leads to significant performance improvement on CIFAR-10 compared to the raw AL case. Results shown in the large-budget setting with  $B_i = 5000$ ,  $B_s = 2500$ .

### 3.2.3 Does semi-supervised learning or active learning make better use of the pool of unlabeled data?

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. To this end, we employed the UDA semi-supervised learning method. 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. Data augmentation stays the same as in Sec. 3.2.2. We refer to the integrated methods as SSL-X, where X is the name of the AL method.

Figures 4 and 5 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;

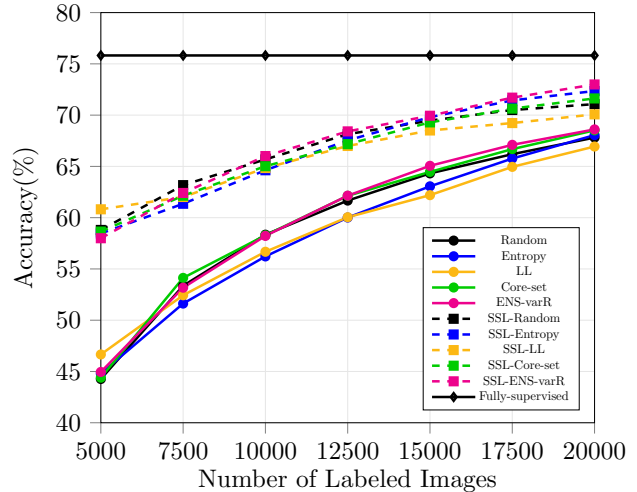


Figure 5. Integrating SSL and AL leads to overall performance improvement on CIFAR-100, however, not all combinations consistently outperform random sampling. Results shown in the large-budget setting with  $B_i = 5000$ ,  $B_s = 2500$ .

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 Figure 5.

### 3.2.4 Does active learning consistently outperform random sampling in low-budget regimes?

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 random selection of samples. Does active learning benefit 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 few labeled samples in the beginning is attractive. We explored such low-budget setting with  $B_i$  and  $B_s$  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 Figures 6 and 7. None of the active learning methods consistently outperforms the random baseline, neither on CIFAR-10 nor on CIFAR-100. This holds always 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

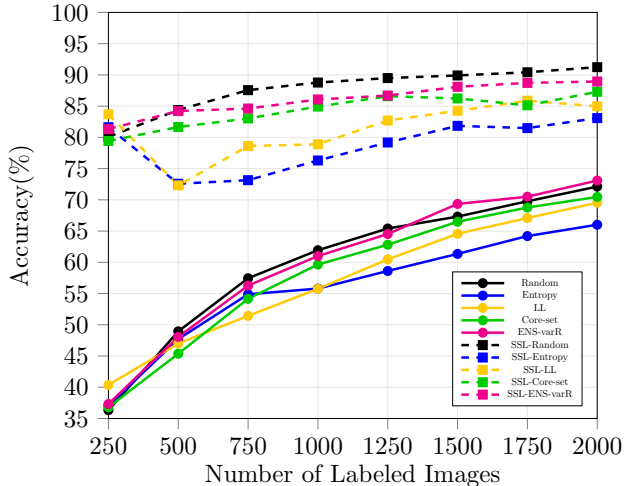


Figure 6. When evaluated in the low-budget regime ( $B_i = B_s = 250$ ) on CIFAR-10, integrated SSL-AL methods are still better than their raw counterparts, however, SSL with random sampling shows the best performance.

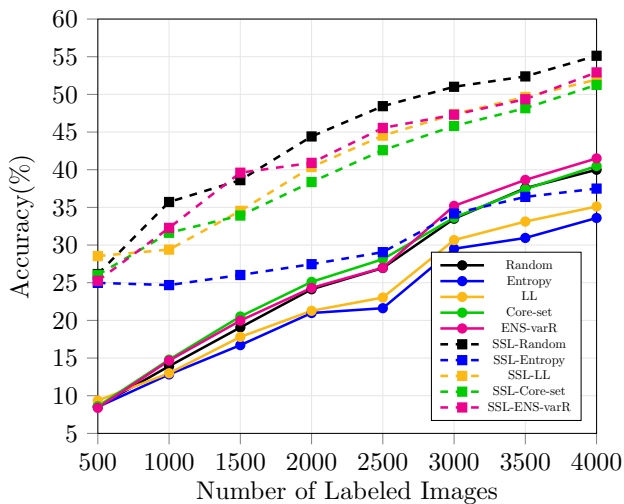


Figure 7. When evaluated in the low-budget regime ( $B_i = B_s = 500$ ) on CIFAR-100, most integrated SSL-AL methods are still better than their raw counterparts but nothing beats SSL with random sampling.

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.

### 3.2.5 Comparison to Transfer Learning

Oliver *et al.* [31] argued that transfer learning may 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 Im-

ageNet 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 Figure 8 and 9 respectively. This experiment shows that including an up-to-date semi-supervised learning algorithm into an active learning pipeline makes sense even when large pre-training data is available.

## 4. Active Learning for Semantic Segmentation

Image classification is a standard active learning task. However, with its relatively low annotation cost (1 click per image) it is not necessarily the most important one. In this section we aim to evaluate the applicability of active learning methods to a task with a significantly higher labeling cost - semantic segmentation. We adapt existing AL methods for classification to support semantic segmentation.

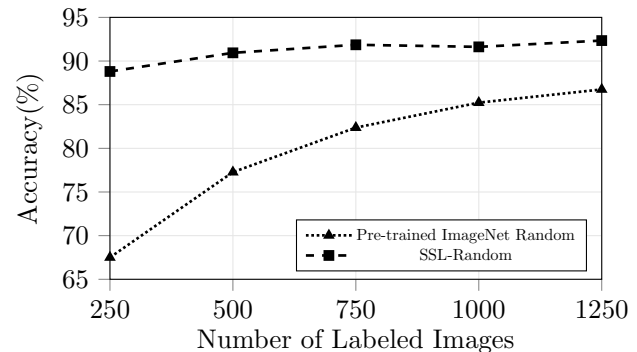


Figure 8. The SSL-Random baseline clearly outperforms a fine-tuned network pre-trained on ImageNet in the low-budget setting. Results shown on CIFAR-10.

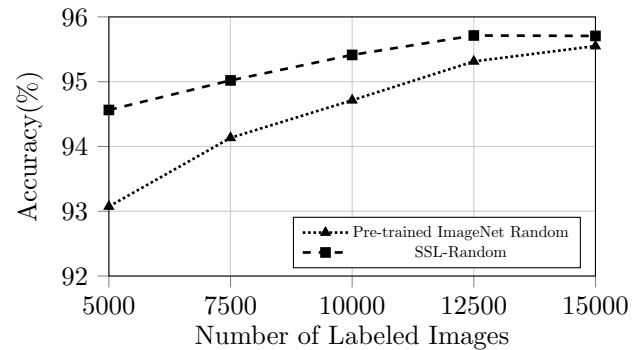


Figure 9. The SSL-Random baseline clearly outperforms a fine-tuned network pre-trained on ImageNet in the large-budget setting. Results shown on CIFAR-10.

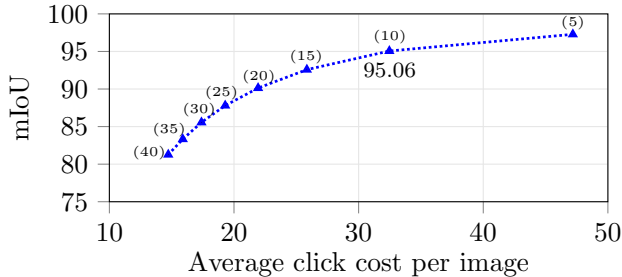


Figure 10. Trade-off between the polygon approximation quality and the annotation cost in clicks. Tolerance values used for each measurement are mentioned above the  $\blacktriangle$  markers.

#### 4.1. Annotation model

A conventional active learning setup includes a human in the loop who annotates the samples picked by the query function. Since training with an actual human annotator is prohibitively expensive, we simulated its actions during training. We used the number of clicks required to annotate the entire image as a proxy for the annotation cost. To do so, we approximate each connected component in the ground truth image with a polygon using the Ramer-Douglas-Peucker algorithm [12]. The approximation quality is controlled by a pre-defined pixel-level tolerance parameter. The total number of clicks per image is then calculated by adding up the number of vertices for all polygons in this image. We perform a grid search over different tolerance values ranging from 5 to 40 pixels to find a suitable value. Figure 10 shows the trade-off between the average click cost per image and the polygon approximation quality of annotations for different tolerance values. The trade-off between different tolerance values and labeling quality is shown in Figure 11. We select the pixel-level labeling tolerance of 10 pixels. The approximated labels retain 95.06% mIoU as compared to the original ground-truth labels. According to this approximation, an average image costs around 33 clicks to label.

#### 4.2. Baseline Methods

We evaluated two kinds of AL methods: uncertainty-based methods and learning-based methods.

**Random.** We consider the random sampling method as our first baseline.

**Entropy.** This uncertainty method is based on the softmax-entropy of the segmentation prediction. Different to the classification case, we need some integral uncertainty measure that aggregates per-pixel uncertainty values. There are a few heuristics for this [1, 43], but none of them is considered standard. We evaluated several simple heuristics including averaging and taking the maximum value over the image, and concluded that the best results are achieved by counting the number of pixels per-image with an uncer-

tainty value higher than a certain threshold. We use 0.6 as a threshold value, which was determined via grid search.

**Ensemble with Average Entropy (ENS-ent).** This second uncertainty-based method ENS-ent is based on the average entropy over the predictions from all members of the model ensemble. We used the same information accumulation heuristic as used for the Entropy method.

**Learn Loss (LL).** We adapted the LL [44] method from image classification to semantic segmentation. Since the original module is proposed for a resnet architecture and the segmentation network used in this work is also based on a resnet architecture, the exact method is directly adapted by reusing the original loss prediction module.

**Semi-supervised Learning (s4GAN).** To leverage unlabeled samples, we used the semi-supervised semantic segmentation method by Mittal *et al.* [29]. It has been shown to produce large performance gain with as few as 2% labeled samples on the PASCAL VOC dataset. We only used the s4GAN branch of the proposed SSL method, which can be trained in an end-to-end manner, and dropped the classification branch for simplicity. The s4GAN method is based on a conditional generative adversarial network, which uses the segmentation network as a generator. The discriminator of the s4GAN discriminates between the predicted and ground-truth segmentation masks. We used the same hyperparameters as provided by Mittal *et al.* [29] for our experiments. We also combine all the above mentioned AL methods with the s4GAN method and evaluate them in the active learning setting.

**SSL-D-score.** Inspired by Ravanbakhsh *et al.* [35], we propose to use the discriminator of the s4GAN network as a query function for sampling. The output of the discriminator varies between 0 and 1, where higher score is assigned to a higher quality of segmentation prediction. In other words, the discriminator of the s4GAN network acts as a critic which provides a higher rating for better segmentation quality. This heuristic selects the samples which are not

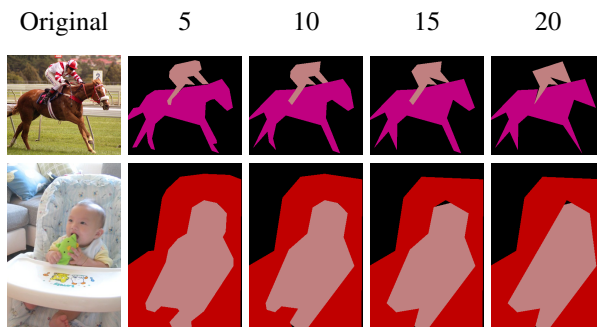


Figure 11. Labeling quality when using polygon approximation with different tolerance values (in pixels). We picked the tolerance value of 10 pixels for our experiments.

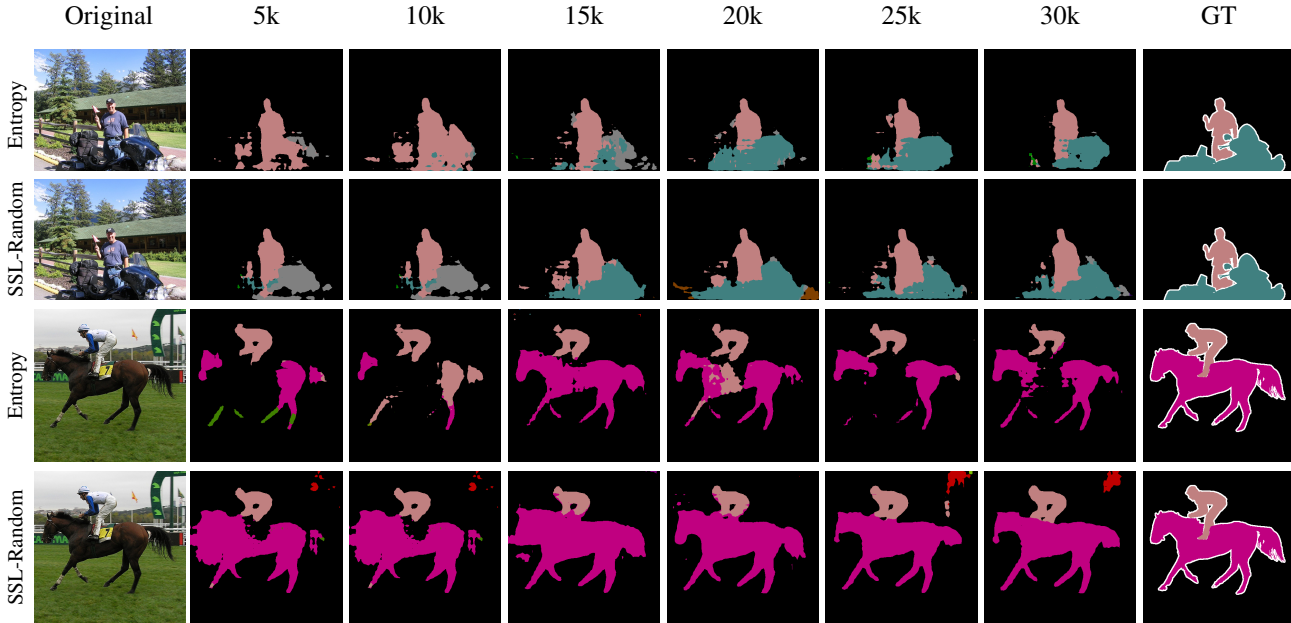


Figure 12. Qualitative semantic segmentation results on PASCAL-VOC at each cycle comparing the Entropy-Image method and the SSL-Random-Image baseline. The column headings indicate the click budget used to train the corresponding model.

well represented by the current learned model, which is indicated with lower rating. We refer to this semi-supervised approach for active learning as the SSL-D-score method.

### 4.3. Experimental design

We show the performance of the AL methods for semantic segmentation on PASCAL-VOC 2012 [13]. The dataset consists of 20 foreground classes and one background class. We use the augmented annotated dataset which contains 10582 training images and 1449 validation images.

**Data Setting.** In AL experiments for segmentation, we define the labeling cost in clicks. We use the initial labeling budget  $\mathcal{B}_i$  and subsequent sampling budget  $\mathcal{B}_s$  of 5000 clicks, which is approximately 1.5% of total labeling cost of the dataset. In the first cycle, randomly sampled images are completely labeled until  $\mathcal{B}_i$  is exhausted. In the subsequent cycles, an AL query method selects images based on a certain criterion and labels the picked image until  $\mathcal{B}_s$  is exhausted. We test all the segmentation AL methods for 5 sampling cycles. All the results are shown on the validation set.

**Training Details.** We used the DeepLabv2 [9] architecture for all the experiments. The DeepLabv2 model is pre-trained using Microsoft COCO [26] dataset. The network is optimized using a SGD optimizer with a base-learning rate of  $2.5e-4$ , momentum of 0.9 and a weight decay of  $5e-4$ . We use poly-learning policy similar to the original segmentation work [9]. We use a batch size of 10 and train each cycle for 150 epochs. The network gets an input image of size

$321 \times 321$ . The used DeepLabv2 version achieves a performance of 73.6 and 71.6 mIoU on original and approximated ground-truth labels respectively on the validation set. The combined semi-supervised active learning (SSL-AL) methods are trained for 10k iterations for each cycle with a batch size of 8. The training procedure and hyperparameters for semi-supervised learning are the same as in [29].

We evaluate AL methods for semantic segmentation in two different settings: (1) using standard augmentations and (2) using semi-supervised learning for training the target model. The mean performance is reported over 3 trials for all single model-based methods and over 2 trials for ensemble-based methods due to higher computation cost.

### 4.4. Results

#### 4.4.1 Is active learning beneficial in more labor-intense labeling tasks than image classification, e.g. semantic segmentation?

To find the relevance of active learning for semantic segmentation, we first train the model only using augmentations including random horizontal flipping and random resized cropping. Other geometry preserving augmentation like brightness, contrast, rotation, scaling do not show any measurable improvement for semantic segmentation [4]. In the results, the uncertainty method based on entropy performs best and shows an improvement of around 1.1% mIoU over the random sampling baseline after 5 AL sampling cycles. The LL method fails to outperform the random baseline approach. Corresponding performance curves are



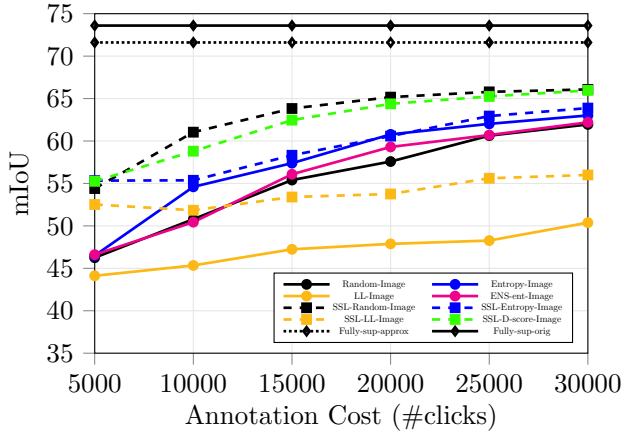


Figure 13. Integrated SSL-AL methods for semantic segmentation mostly perform better than their raw counterparts on PASCAL-VOC with  $\mathcal{B}_i = \mathcal{B}_s = 5000$  clicks ( $\approx 1.5\%$  of the dataset). None of the methods outperforms SSL with random sampling.

shown with solid lines in Figure 13.

Further, we use the unlabeled pool of images to train the target model for semantic segmentation and analyze the performance. We utilize the s4GAN model [29] to leverage the information from unlabeled samples. When used on top of the SSL method, all AL methods show a clear gain in performance. The performance of the random baseline (Random) when combined with s4GAN increases by the largest margin of 4.1% mIoU and reaches the overall best value. In addition, SSL-D-score heuristic also shows comparable performance to the random baseline after 5 sampling cycles, but does not bring any improvement over the SSL-Random baseline. The performance curves for all integrated methods are shown with dashed lines in Figure 13. Figure 12 shows the qualitative results at each sampling cycle, comparing the Entropy-Image method and the SSL-Random-Image baseline.

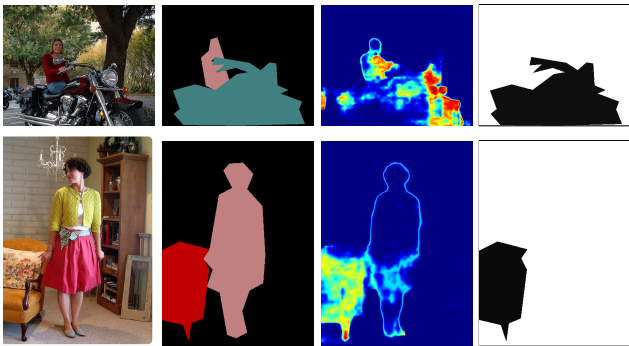


Figure 14. Image labeling in a polygon-level labeling regime. From left to right: Original image, approximated ground-truth, pixel-wise entropy and selected polygon for labeling based on the entropy heuristic.

#### 4.4.2 Polygon-level Labeling

Here, we explore whether labeling only a part of an image is more effective than labeling the complete image. We evaluate active learning methods for semantic segmentation, where only a region of an image is selected by the query function. This region is approximately labeled using a polygon by the annotation simulator. We evaluate methods where an image is selected randomly, but the polygon in the image is selected based on the active learning heuristic. We compare entropy-based and random polygon selection methods in both raw and SSL-integrated active learning settings.

**Experiment Details.** Entropy of a polygon is measured in a similar way as in the image-level labeling regime. We create a binary mask for the pixel-wise entropy based on a threshold and use the area of the high-entropy pixels as our selection heuristic. Only in the first cycle, images are completely labeled until the  $\mathcal{B}_i$  is covered. In the subsequent cycles, images are labeled polygon-wise until the sampling budget  $\mathcal{B}_s$  is exhausted. Figure 14 shows two examples of how polygon-level labeling regime works based on the entropy heuristic. The budget settings and the hyperparameters exactly match those from the image-level labeling regime.

**Results.** Entropy-based polygon selection approach is more effective than random polygon selection for the raw active learning (without SSL) setting. However, when combined with semi-supervised learning, both entropy (Random-Image-Entropy-Polygon) and random (Random-Image-Random-Polygon) polygon selection strategies perform very similarly. Results are shown in Figure 15. Moreover when all polygon-level labeling approaches are

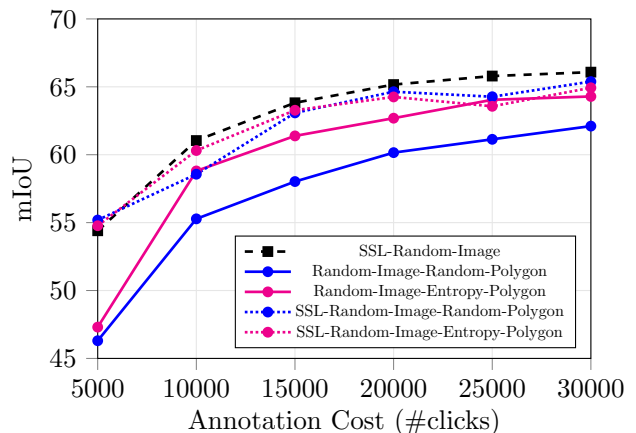


Figure 15. Active learning for semantic segmentation: comparison between SSL integrated with active learning (SSL-X) against the standard setting. Results shown on the PASCAL-VOC dataset with  $\mathcal{B}_i = \mathcal{B}_s = 5000$  clicks. The suffixes ‘Image’ and ‘Polygon’ refer to the image-level and polygon-level labeling regimes respectively.

compared with the image-level labeling approaches, we find SSL-Random-Image baseline even outperforms all the polygon-level active learning methods.

In this experiment, we also observed that the SSL-Random-Image baseline outperformed the SSL-Random-Image-Random-Polygon baseline, showing that image-level labeling is a more effective way of labeling an image.

## 5. Discussion

Our experiments provide strong evidence that the current evaluation protocol used in active learning is sub-optimal which in turn leads to wrong conclusions about the methods' performance and the state of the field in general.

Evaluating on 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 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 the 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 critically important for many real-world applications.

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

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

It would be interesting to know why AL often performs worse than random sampling and consistently does so in the low-budget regime. For now, we can only speculate. We believe that AL sampling introduces a bias into the distribution of annotated samples, i.e., the sampled distribution does not sufficiently match the true distribution anymore. The damage by this bias is larger than the positive effect of learning from "more interesting" samples. If this hypothesis is true, research in active learning should focus on ways that avoid any bias by the selection strategy.

Our results also indicate that adapting AL methods to tasks with higher labeling cost, e.g. semantic segmentation,

is a non-trivial problem. Although there is not enough empirical evidence for this, we speculate that such tasks can potentially benefit much more from employing informed sample selection strategies and thus define a promising direction for future research.

## Acknowledgements

This study was supported by the German Federal Ministry of Education and Research via the project Deep-PTL and by the Intel Network of Intelligent Systems.

## References

- [1] Hamed H. Aghdam, Abel Gonzalez-Garcia, Joost van de Weijer, and Antonio M. Lopez. Active learning for deep detection neural networks. In *ICCV*, 2019. 7
- [2] Anonymous. Consistency-based semi-supervised active learning: Towards minimizing labeling budget. In *Submitted to ICLR*, 2020. under review. 2
- [3] Anonymous. Rethinking deep active learning: Using unlabeled data at model training. In *Submitted to ICLR*, 2020. under review. 2
- [4] Anonymous. Semi-supervised semantic segmentation needs strong, high-dimensional perturbations. In *Submitted to ICLR*, 2020. under review. 8
- [5] Ben Athiwaratkun, Marc Finzi, Pavel Izmailov, and Andrew Gordon Wilson. There are many consistent explanations of unlabeled data: Why you should average. In *ICLR*, 2019. 2
- [6] William H. Beluch, Tim Genewein, Andreas Nrnberger, and Jan M. Khler. The power of ensembles for active learning in image classification. In *CVPR*, 2018. 2, 3, 4
- [7] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *NeurIPS*, 2019. 2
- [8] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural networks. In *ICML*, 2015. 2
- [9] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *TPAMI*, 2018. 8
- [10] Ekin D. Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V. Le. Autoaugment: Learning augmentation strategies from data. In *CVPR*, 2019. 3, 4
- [11] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with dropout. *arXiv preprint arXiv:1708.04552*, 2017. 4

- [12] David H. Douglas and Thomas K. Peucker. Algorithms for the Reduction of the Number of Points Required to Represent a Digitized Line or its Caricature. In *Classics in Cartography*. 2011. 7
- [13] Mark Everingham, Luc Gool, Christopher K. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 2010. 8
- [14] Geoffrey French, Timo Aila, Samuli Laine, Michal Mackiewicz, and Graham D. Finlayson. Consistency regularization and cutmix for semi-supervised semantic segmentation. *CoRR*, 2019. 2
- [15] Yarin Gal and Zoubin Ghahramani. Bayesian convolutional neural networks with Bernoulli approximate variational inference. In *ICLR-workshop track*, 2016. 2
- [16] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *ICML*, 2016. 2
- [17] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *ICML*, 2017. 2
- [18] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*. 2014. 2
- [19] W.-C. Hung, Y.-H. Tsai, Y.-T. Liou, Y.-Y. Lin, and M.-H. Yang. Adversarial learning for semi-supervised semantic segmentation. In *BMVC*, 2018. 2
- [20] S. D. Jain and K. Grauman. Active image segmentation propagation. In *CVPR*, 2016. 2
- [21] Elmer H. Johnson. Elementary applied statistics: For students in behavioral science. *Social Forces*, 1966. 2
- [22] Tarun Kalluri, Girish Varma, Manmohan Chandraker, and C.V. Jawahar. Universal semi-supervised semantic segmentation. In *ICCV*, 2019. 2
- [23] Alex Kendall and Yarin Gal. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? In *NeurIPS*, 2017. 2
- [24] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. In *ICLR (Poster)*, 2017. 2
- [25] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *NeurIPS*. 2017. 2
- [26] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollr, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014. 8
- [27] B. Liu and V. Ferrari. Active learning for human pose estimation. In *ICCV*, 2017. 2
- [28] Radek Mackowiak, Philip Lenz, Omair Ghori, Ferran Diego, Oliver Lange, and Carsten Rother. Cereals - cost-effective region-based active learning for semantic segmentation. In *BMVC*, 2018. 2
- [29] Sudhanshu Mittal, Maxim Tatarchenko, and Thomas Brox. Semi-supervised semantic segmentation with high- and low-level consistency. *arXiv preprint arXiv:1908.05724*, 2019. 2, 7, 8, 9
- [30] T. Miyato, S. Maeda, M. Koyama, and S. Ishii. Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *TPAMI*, 2019. 2
- [31] Avital Oliver, Augustus Odena, Colin A Raffel, Ekin Dogus Cubuk, and Ian Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. In *NeurIPS*. 2018. 3, 6
- [32] Ian Osband, Charles Blundell, Alexander Pritzel, and Benjamin Van Roy. Deep exploration via bootstrapped dqn. In *NeurIPS*. 2016. 2
- [33] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 4
- [34] R. Paul, D. Feldman, D. Rus, and P. Newman. Visual precis generation using coresets. In *ICRA*, 2014. 3
- [35] Mahdyar Ravanbakhsh, Tassilo Klein, Kayhan Batmanghelich, and Moin Nabi. Uncertainty-driven semantic segmentation through human-machine collaborative learning. In *International Conference on Medical Imaging with Deep Learning – Extended Abstract Track*, 2019. 2, 7
- [36] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *ICLR*, 2018. 2, 3
- [37] H. S. Seung, M. Opper, and H. Sompolinsky. Query by committee. In *Annual Workshop on Computational Learning Theory*, 1992. 3
- [38] Claude Elwood Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 1948. 3
- [39] Samarth Sinha, Sayna Ebrahimi, and Trevor Darrell. Variational adversarial active learning. In *ICCV*, 2019. 2, 4
- [40] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIPS*. 2017. 2

- [41] Keze Wang, Dongyu Zhang, Ya Li, Ruimao Zhang, and Liang Lin. Cost-effective active learning for deep image classification. *IEEE Trans. Cir. and Sys. for Video Technol.*, 2017. [2](#)
- [42] Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. Unsupervised data augmentation for consistency training. *arXiv preprint arXiv:1904.12848*, 2019. [2](#), [3](#), [4](#)
- [43] Lin Yang, Yizhe Zhang, Jianxu Chen, Siyuan Zhang, and Danny Z. Chen. Suggestive annotation: A deep active learning framework for biomedical image segmentation. In *Medical Image Computing and Computer Assisted Intervention - MICCAI*, 2017. [2](#), [7](#)
- [44] Donggeun Yoo and In So Kweon. Learning loss for active learning. In *CVPR*, 2019. [2](#), [3](#), [4](#), [7](#)
- [45] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *BMVC*, 2016. [3](#)
- [46] Xiaohua Zhai, Avital Oliver, Alexander Kolesnikov, and Lucas Beyer. S4l: Self-supervised semi-supervised learning. In *ICCV*, 2019. [2](#)