Unsupervised Selective Labeling for More Effective Semi-Supervised Learning

Xudong Wang*[0000−0002−4973−780X], Long Lian*[0000−0001−6098−189X], and Stella X. Yu[0000−0002−3507−5761]

UC Berkeley / ICSI

Abstract. Given an unlabeled dataset and an annotation budget, we study how to selectively label a fixed number of instances so that semi-supervised learning (SSL) on such a partially labeled dataset is most effective. We focus on *selecting* the right data to label, in addition to usual SSL's propagating labels from labeled data to the rest unlabeled data. This instance selection task is challenging, as without any labeled data we do not know what the objective of learning should be. Intuitively, no matter what the downstream task is, instances to be labeled must be *representative* and *diverse*: The former would facilitate label propagation to unlabeled data, whereas the latter would ensure coverage of the entire dataset. We capture this idea by selecting cluster prototypes, either in a pretrained feature space, or along with feature optimization, both without labels. Our unsupervised selective labeling consistently improves SSL methods over state-of-the-art active learning given labeled data, by 8∼25× in label efficiency. For example, it boosts FixMatch by 10% (14%) in accuracy on CIFAR-10 (ImageNet-1K) with 0.08% (0.2%) labeled data, demonstrating that small computation spent on selecting what data to label brings significant gain especially under a low annotation budget. Our work sets a new standard for practical and efficient SSL.

Keywords: semi-supervised learning · unsupervised selective labeling

1 Introduction

Deep learning's success on natural language understanding [21], visual object recognition [41], and object detection [31] follow a straightforward recipe: better model architectures, more data, and scalable computation [32, 36, 42, 73]. As training datasets get bigger, their full task annotation becomes infeasible [4, 63].

Semi-supervised learning (SSL) deals with learning from both a small amount of labeled data *and* a large amount of unlabeled data: Labeled data directly supervise model learning, whereas unlabeled data help learn a desirable model that makes consistent [4,5,43,58,63,65,69,72] and unambiguous [5,33,43] predictions.

Recent SSL methods approach fully supervised learning performance with a very small fraction of labeled data. For example, on ImageNet, SSL with 1%

* Equal contribution

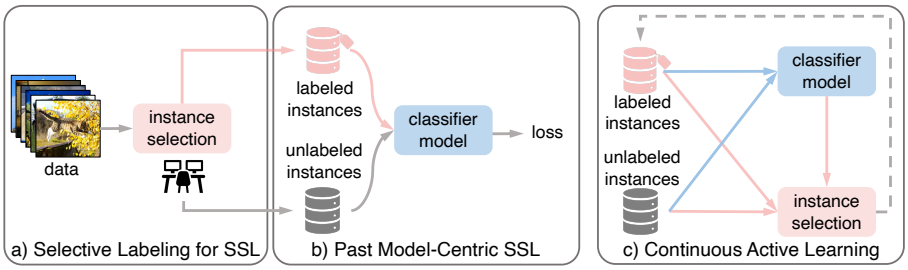


Fig. 1: Our unsupervised selective labeling is a novel aspect of semi-supervised learning (SSL) and different from active learning (AL). **a, b)** Existing SSL methods focus on optimizing the model *given* labeled and unlabeled data. Instead of such model-centric learning, we focus on optimizing the selection of training instances *prior to* their label acquisition. **c)** Existing AL methods alternate between classifier learning and instance selection, leveraging a classifier trained on initial labeled data and regularized on unlabeled data. In contrast, we select instances from unlabeled data without knowing the classification task.

labeled data, i.e., only 13 instead of around 1300 labeled images per class, captures 95% (76.6% out of 80.5% in terms of top-1 accuracy) of supervised learning performance with 100% fully labeled data [15].

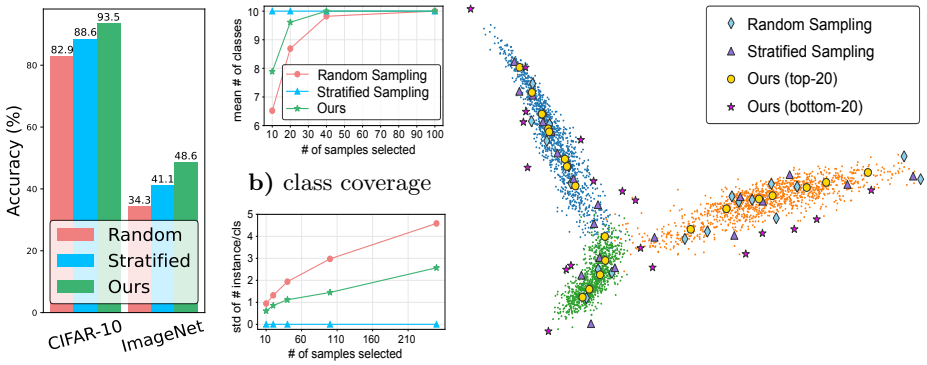
The lower the annotation level, the more important what the labeled instances are to SSL. While a typical image could represent many similar images, an odd-ball only represents itself, and labeled instances may even cover only part of the data variety, trapping a classifier in partial views with unstable learning and even model collapse.

A common assumption in SSL is that labeled instances are sampled randomly either over all the available data or over individual classes, the latter known as stratified sampling [4, 5, 63, 69]. Each method has its own caveats: Random sampling can fail to cover all semantic classes and lead to poor performance and instability, whereas stratified sampling is utterly unrealistic: If we can sample data by category, we would already have the label of every instance!

Selecting the right data to label for the sake of model optimization is not new. In fact, it is the focus of active learning (AL): Given an initial set of labeled data, the goal is to select an additional subset of data to label (Fig. 1) so that a model trained over such partially labeled data approaches that over the fully labeled data [26, 59, 75]. Unlabeled data can also be exploited for model training by combining AL and SSL, resulting in a series of methods called semi-supervised active learning (SSAL).

However, existing AL/SSAL methods have several shortcomings.

1. They often require randomly sampled labeled data to begin with, which is sample-inefficient in low labeling settings that SSL methods excel at [13].
2. AL/SSAL methods are designed with human annotators in a loop, working in multiple rounds of labeling and training. This could be cumbersome in low-shot scenario and leads to large labeling overhead.



a) accuracy **b)** class coverage **c)** data balance **d)** informativeness

Fig. 2: Our instance selection outperforms random and stratified sampling by selecting a diverse set of representative instances. **a)** The classification accuracy using SSL method FixMatch increases with our selectively labeled instances. **b)** Our method covers all the semantic classes with only a few instances. **c)** Our selection is far more balanced than random sampling. **d)** On a toy dataset of 3 classes in ImageNet, our top-ranked instances cover informative samples across the entire space, whereas our bottom-ranked instances tend to be outliers.

3. AL's own training pipeline with a human-in-the-loop design makes its integration into existing SSL code implementation hard [64].
4. The requested labels are tightly coupled with the model being trained so that labels need to be collected anew every time a model is trained with AL/SSAL.

We address *unsupervised selective labeling* for SSL (Fig. 1), in stark contrast with supervised data selection for AL, which is conditioned on an initial labeled set and for the benefit of a certain task. Given only an annotation budget and an unlabeled dataset, among many possible ways to select a fixed number of instances for labeling, which way would lead to the best SSL model performance when it is trained on such partially labeled data?

Our instance selection task is challenging, as without any labeled data we do not know what the objective of learning should be. Intuitively, no matter what the downstream task is, instances to be labeled must be *representative* and *diverse*: The former would facilitate label propagation to unlabeled data, whereas the latter would ensure coverage of the entire dataset. We capture this idea by selecting cluster prototypes, either in a pretrained feature space, or along with feature optimization, both without labels.

Our pipeline has three steps: 1) Unsupervised feature learning that maps data into a discriminative feature space. 2) Select instances for labeling for maximum representativeness and diversity, without or with additional optimization. 3) Apply SSL (e.g., [15, 63]) to the labeled data and the rest unlabeled data.

Fig. 2 shows that our method has many benefits over random or stratified sampling for labeled data selection, in terms of accuracy, coverage, balance over classes, and representativeness. As it selects informative instances without ini-

tial labels, it can not only integrate readily into existing SSL methods, but also achieve higher label efficiency than SSAL methods. While most AL/SSAL methods only work on small-scale datasets such as CIFAR [40], our method scales up easily to large-scale datasets such as ImageNet [57], taking less than an hour for our data selection on a commodity GPU server.

Our work sets a new standard for practical SSL with these contributions.

1. We systematically analyze the impact of different selective labeling methods on SSL under low-label settings, a previously ignored aspect of SSL.
2. We propose two unsupervised selective labeling methods that capture representativeness and diversity without or along with feature optimization.
3. We benchmark extensively on our data selection with various SSL methods, delivering much higher sample efficiency over sampling in SSL or AL/SSAL.
4. We release our toolbox with AL/SSL implementations and a unified data loader, including benchmarks, selected instance indices, and pretrained models that combine selective labeling with various methods for fair comparisons.

2 Selective Labeling for Semi-supervised Learning

Suppose we are given an unlabeled dataset of n instances and an annotation budget of m. Our task is to select m ($m \ll n$) instances for labeling, so that a SSL model trained on such a partially labeled dataset, with m instances labeled and $n-m$ unlabeled, produces the best classification performance.

Formally, let $\mathbb{D} = \{(x_i, y_i)\}_{i=1}^{n}$ denote n pairs of image x_i and its (*unknown*) class label y_i. Let $\mathbb{A}$ denote a size-m subset of $\mathbb{D}$ with *known* class labels. Our goal is to select $\mathbb{A} \subset \mathbb{D}$ for acquiring class labels, in order to maximize the performance of a given SSL model trained on labeled data $\mathbb{A}$ and unlabeled data $\mathbb{D} \setminus \mathbb{A}$.

Our unsupervised selective labeling is challenging, as we do not have any labels to begin with, i.e., we don't know what would make the SSL model perform the best. Our idea is to select m instances that are not only *representative* of most instances, but also *diverse* enough to broadly cover the entire dataset, so that we do not lose information prematurely before label acquisition.

Our SSL pipeline with selective labeling consists of three steps: **1)** unsupervised feature learning; **2)** unsupervised instance selection for annotation; **3)** SSL on selected labeled data $\mathbb{A}$ and remaining unlabeled data $\mathbb{D} \setminus \mathbb{A}$.

We propose two selective labeling methods in Step 2, training-free Unsupervised Selective Labeling (USL) and training-based Unsupervised Selective Labeling (USL-T), both aiming at selecting cluster prototypes in a discriminative feature space without label supervision.

2.1 Unsupervised Representation Learning

Our first step is to obtain lower-dimensional and semantically meaningful features with unsupervised contrastive learning [14,35,51,71], which maps x_i onto a d-dimensional hypersphere with L^2 normalization, denoted as $f(x_i)$. We use MoCov2 [17] (SimCLR [14] or CLD [68]) to learn representations on ImageNet (CIFAR [40]). See appendix for details.

2.2 Unsupervised Selective Labeling (USL)

We study the relationships between data instances using a weighted graph, where nodes $\{V_i\}$ denote data instances in the (normalized) feature space $\{f(x_i)\}$, and edges between nodes are attached with weights of pairwise feature similarity [7, 19, 25, 61], defined as $\frac{1}{D_{ij}}$, the inverse of feature distance D:

$$D_{ij} = \|f(x_i) - f(x_j)\|. \tag{1}$$

Intuitively, the smaller the feature distance, the better the class information can be transported from labeled nodes to unlabeled nodes. Given a labeling budget of m instances, we aim to select m instances that are not only similar to others, but also well dispersed to cover the entire dataset.

Representativeness: Select Density Peaks. A straightforward approach is to select well connected nodes to spread semantic information to nearby nodes. It corresponds to finding a density peak in the feature space. The K-nearest neighbor density (K-NN) estimation [28, 52] is formulated as:

$$p_{\text{KNN}}(V_i, k) = \frac{k}{n} \frac{1}{A_d \cdot D^d(V_i, V_{k(i)})} \tag{2}$$

where $A_d = \pi^{d/2}/\Gamma(\frac{d}{2}+1)$ is the volume of a unit d-dimensional ball, d the feature dimension, $\Gamma(x)$ the Gamma function, $k(i)$ instance i's kth nearest neighbor. p_{KNN} is very sensitive to noise, as it only takes the kth nearest neighbor into account. For robustness, we replace the kth neighbor distance $D(V_i, V_{k(i)})$ with the average distance $\bar{D}(V_i, k)$ to all k nearest neighbors instead:

$$\hat{p}_{\text{KNN}}(V_i, k) = \frac{k}{n} \frac{1}{A_d \cdot \bar{D}^d(V_i, k)}, \quad \text{where } \bar{D}(V_i, k) = \frac{1}{k} \sum_{j=1}^{k} D(V_i, V_{j(i)}). \tag{3}$$

We use $\hat{p}_{\text{KNN}}(V_i, k)$ to measure the *representativeness* of node V_i. Since only the relative ordering matters in our selection process, the density peak corresponds to the sample with maximum $\hat{p}_{\text{KNN}}(V_i, k)$ (i.e., maximum $1/\bar{D}(V_i, k)$).

Diversity: Pick One in Each Cluster. While instances of high feature density values are individually representative, a separate criterion is necessary to avoid repeatedly picking similar instances near the same density peaks (Fig. 3a). To select m diverse instances that cover the entire unlabeled dataset, we resort to K-Means clustering that partitions n instances into $m(\leq n)$ clusters, with each cluster represented by its centroid c [29, 47] and every instance assigned to the cluster of the nearest centroid. Formally, we seek m-way node partitioning $\mathbb{S} = \{S_1, S_2, ..., S_m\}$ that minimizes the within-cluster sum of squares [39]:

$$\min_{\mathbb{S}} \sum_{i=1}^{m} \sum_{V \in S_i} \|V - c_i\|^2 = \min_{\mathbb{S}} \sum_{i=1}^{m} |S_i| \text{Var}(S_i) \tag{4}$$

It is optimized iteratively with EM [48] from random initial centroids. We then pick the most representative instance of each cluster according to Eqn. 3.

Regularization: Inter-cluster Information Exchange. So far we use K-Means clustering to find m hard clusters, and then choose the representative of

a) local only **b)** local + global **c)** local + global + reg.

Fig. 3: **a)** Points at density peaks are individually representative of their local neighborhoods, but lack broad coverage of the entire set. **b)** Hard constraint by K-Means greatly depends on clustering quality and only partially alleviates the problem. **c)** Soft regularization leads to more uniform and diversified queries.

each cluster *independently*. This last step is sub-optimal, as instances of high density values could be located along cluster boundaries and close to instances in adjacent regions (Fig. 3b). We thus apply a regularizer to inform each cluster of other clusters' choices and iteratively diversify selected instances (Fig. 3c).

Specifically, let $\hat{\mathbb{V}}^t = \{\hat{V}_1^t, ..., \hat{V}_m^t\}$ denote the set of m instances selected at iteration t, $\hat{V}_i^t$ for clusters S_i, where $i \in \{1, \ldots, m\}$. For each candidate V_i in cluster S_i, the farther it is away from those in other clusters in $\hat{\mathbb{V}}^{t-1}$, the more diversity it creates. We thus minimize the total inverse distance to others in a regularization loss $\mathrm{Reg}(V_i, t)$, with a sensitivity hyperparameter α:

$$\mathrm{Reg}(V_i, t) = \sum_{\hat{V}_j^{t-1} \notin S_i} \frac{1}{\|V_i - \hat{V}_j^{t-1}\|^\alpha}. \tag{5}$$

This regularizer is updated with an exponential moving average:

$$\overline{\mathrm{Reg}}(V_i, t) = m_{\mathrm{reg}} \cdot \overline{\mathrm{Reg}}(V_i, t-1) + (1 - m_{\mathrm{reg}}) \cdot \mathrm{Reg}(V_i, t) \tag{6}$$

where m_{reg} is the momentum. At iteration t, we select instance i of the maximum *regularized utility* $U'(V_i, t)$ within each cluster:

$$U'(V_i, t) = U(V_i) - \lambda \cdot \overline{\mathrm{Reg}}(V_i, t) \tag{7}$$

where λ is a hyperparameter that balances diversity and individual representativeness, utility $U(V_i) = 1/\bar{D}(V_i, k)$. In practice, calculating distances between every candidate and every selected instance in $\hat{\mathbb{V}}^{t-1}$ is no longer feasible for a large dataset, so we only consider h nearest neighbors in $\hat{\mathbb{V}}^{t-1}$. $\hat{\mathbb{V}}^t$ at the last iteration is our final selection for labeling.

2.3 Training-Based Unsupervised Selective Labeling (USL-T)

Our USL is a simple yet effective *training-free* approach to selective labeling. Next we introduce an end-to-end *training-based* Unsupervised Selective Labeling (USL-T), an alternative that integrates instance selection into representation

learning and often leads to more balanced (Fig. 5) and more label-efficient (Table 2) instance selection. The optimized model implicitly captures semantics and provides a strong initialization for downstream tasks (Sec. 4.5).

Global Constraint via Learnable K-Means Clustering. Clustering in a given feature space is not trivial (Fig. 3c). We introduce a better alternative to K-Means clustering that jointly learns both the cluster assignment and the feature space for unsupervised instance selection.

Suppose that there are C centroids initialized randomly. For instance x with feature $f(x)$, we infer one-hot cluster assignment distribution $y(x)$ by finding the closest *learnable* centroid $c_i, i \in \{1,\ldots,C\}$ based on feature similarity s:

$$y_i(x) = \begin{cases} 1, & \text{if } i = \arg\min_{k\in\{1,\ldots,C\}} s(f(x), c_k) \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

We predict a soft cluster assignment $\hat{y}(x)$ by taking softmax over the similarity between instance x and each learnable centroid:

$$\hat{y}_i(x) = \frac{e^{s(f(x),c_i)}}{\sum_{j=1}^{C} e^{s(f(x),c_j)}}. \tag{9}$$

The hard assignment $y(x)$ can be regarded as pseudo-labels [43,63,67]. By minimizing $D_{\mathrm{KL}}(y(x)\|\hat{y}(x))$, the KL divergence between soft and hard assignments, we encourage not only each instance to become more similar to its centroid, but also the learnable centroid to become a better representative of instances in the cluster. With soft predictions, each instance has an effect on all the centroids.

Hardening soft assignments has a downside: Initial mistakes are hard to correct with later training, degrading performance. Our solution is to ignore ambiguous instances with maximal softmax scores below threshold τ:

$$L_{\mathrm{global}}(\{x_i\}_{i=1}^{n}) = \frac{1}{n} \sum_{\max(\hat{y}(x_i))\geq\tau} D_{\mathrm{KL}}(y(x_i)\|\hat{y}(x_i)) \tag{10}$$

where τ is the threshold hyper-parameter. This loss leads to curriculum learning: As instances are more confidently assigned to a cluster with more training, more instances get involved in shaping both feature $f(x)$ and clusters $\{c_i\}$.

Our global loss can be readily related to K-Means clustering.

Observation 1 *For $\tau = 0$ and fixed feature f, optimizing L_{global} is equivalent to optimizing K-Means clustering with a regularization term on inter-cluster distances that encourage additional diversity. See Appendix for derivations.*

Local Constraint with Neighbor Cluster Alignment. Our global constraint is the counterpart of K-Means clustering in USL. However, since soft assignments usually have low confidence scores for most instances at the beginning, convergence could be very slow and sometimes unattainable. We propose an additional local smoothness constraint by assigning an instance to the same cluster of its neighbors' in the unsupervisedly learned feature space to prepare confident predictions for the global constraint to take effect.

This simple idea as is could lead to two types of collapses: Predicting one big cluster for all the instances and predicting a soft assignment that is close to a uniform distribution for each instance. We tackle them separately.

1) For one-cluster collapse, we adopt a trick for long-tailed recognition [49] and adjust logits to prevent their values from concentrating on one cluster:

$$\hat{P}(z, \bar{z}) = z - \alpha \cdot \log \bar{z} \tag{11}$$

$$\bar{z} = \mu \cdot \sigma(z) + (1-\mu) \cdot \bar{z} \tag{12}$$

where α controls the intensity of adjustment, $\bar{z}$ is an exponential moving average of $\sigma(z)$, and $\sigma(\cdot)$ is the softmax function.

2) For even-distribution collapse, we use a sharpening function [2, 4, 5] to encourage the cluster assignment to approach a one-hot probability distribution, where a temperature parameter t determines the spikiness.

Both anti-collapse measures can be concisely captured in a single function $P(\cdot)$ that modifies and turns logits z into a reference distribution:

$$[P(z, \bar{z}, t)]_i = \frac{\exp(\hat{P}(z_i, \bar{z}_i)/t)}{\sum_j \exp(\hat{P}(z_j, \bar{z}_j/t))} \tag{13}$$

We now impose our local labeling smoothness constraints with such modified soft assignments between x_i and its randomly selected neighbor x'_i:

$$L_{\text{local}}(\{x_i\}_{i=1}^n) = \frac{1}{n} \sum_{i=1}^{n} D_{\text{KL}}(P(y(x'_i), \bar{y}(x'_i), t) || \hat{y}(x_i)). \tag{14}$$

We restrict x'_i to x's k nearest neighbors, selected according to the unsupervisedly learned feature prior to training and fixed for simplicity and efficiency.

We show that our local constraint prevents both collapses.

Observation 2 *Neither one-cluster nor even-distribution collapse is optimal to our local constraint, i.e.,* $P(y(x'), \bar{y}(x'), t) \neq \hat{y}(x)$. *See Appendix for more details.*

Our final loss adds up the global and local terms with loss weight λ:

$$L = L_{\text{global}} + \lambda L_{\text{local}} \tag{15}$$

Diverse and Representative Instance Selection in USL-T. Our USL-T is an end-to-end unsupervised feature learning method that directly outputs m clusters for selecting m *diverse* instances. For each cluster, we then select the most *representative* instance, characterized by its highest confidence score, i.e. $\max \hat{y}(x)$. Just as USL, USL-T improves model learning efficiency by selecting diverse representative instances for labeling, without any label supervision.

2.4 Distinctions and Connections With SSL/AL/SSAL

Table 1 compares our USL with related SSL, AL, and SSAL settings.

1. Our USL has the advantage of AL/SSAL that seeks optimal instances to label, yet does not require inefficient initial random samples or multiple rounds of human interventions. USL has high label efficiency for selected instances in low label settings and does not need to trade off annotation budget allocation between initial random sampling and several interim annotation stages.

Property	Semi-supervised Learning	Active Learning	Semi-supervised Active Learning	Ours
Uses no initial random labels	✗	✗	✗	✓
Actively queries for labels	✗	✓	✓	✓
Requires annotation only once	✓	✗	✗	✓
Leverages unlabeled data	✓	✗	✓	✓
Allows label reuse across runs	✓	✗	✗	✓

Table 1: Key properties of SSL, AL, SSAL, and our USL/USL-T pipelines. Among them, our approach is the only one that does not use any random labels.

2. Compared to AL, our USL also leverages unlabeled data. Compared to SSAL, USL is much easier to implement because we keep existing SSL implementation intact, while SSAL requires a human-in-the-loop pipeline. Consequently, unlike AL/SSAL where instance selection is coupled with the model to be trained, our selection is *decoupled* from the downstream SSL model. The same selection from USL works well even across different downstream SSL methods, enabling label reuse across different SSL experiments.
3. Most notably, our work is the first *unsupervised* selective labeling method on large-scale recognition datasets that requests annotation only *once*.

3 Related Work

Semi-supervised Learning (SSL) integrates information from small-scale labeled data and large-scale unlabeled data. *Consistency-based regularization* [58, 65, 72] applies a consistency loss by imposing invariance on unlabeled data under augmentations. *Pseudo-labeling* [4, 5, 43, 69] relies on the model's high confidence predictions to produce pseudo-labels of unlabeled data and trains them jointly with labeled data. FixMatch [63] integrates strong data augmentation [22] and pseudo-label filtering [46] and explores training on the most representative samples ranked by [10]. However, [10] is a supervised method that requires all labels. *Transfer learning* method SimCLRv2 [15] is a two-stage SSL method that applies contrastive learning followed by fine-tuning on labeled data. *Entropy-minimization* [5, 33] assumes that classification boundaries do not pass through the high-density area of marginal distributions and enforces confident predictions on unlabeled data. Instead of competing with existing SSL methods, our USL enables more effective SSL by choosing the right instances to label *for* SSL, without any prior semantic supervision.

Active Learning (AL) aims to select a small subset of labeled data to achieve competitive performance over supervised learning on fully labeled data [6, 20, 56]. *Traditional AL* has three major types [55, 60]: membership query synthesis [1], stream-based selective sampling [3, 23], and pool-based active learning [37, 50, 66, 70]. In *Deep AL*, Core-Set [59] approaches data selection as a set cover problem. [26] estimates distances from decision boundaries based on sensitivity to

adversarial attacks. LLAL [75] predicts target loss of unlabeled data parametrically and queries instances with the largest loss for labels. *Semi-supervised Active Learning* (SSAL) combines AL with SSL. [64] merges uncertainty-based metrics with MixMatch [5]. [30] merges consistency-based metrics with consistency-based SSL. AL/SSAL often rely on initial labeled data to learn both the model and the instance sampler, requiring multiple (e.g. 10) rounds of sequential annotation and significant modifications of existing annotation pipelines. Recent *few-label transfer* [45] leverages features from a large source dataset to select instances in a smaller target dataset for annotation. It also requires a seed instance per class to be pre-labeled in the target dataset, whereas we do not need supervision anywhere for our instance selection.

Deep Clustering. DeepCluster [11] also jointly learns features and cluster assignments with k-Means clustering. However, USL-T, with end-to-end backprop to jointly optimize classifiers and cluster assignments, is much more *scalable* and *easy to implement*. UIC/DINO [12,16] incorporate neural networks with categorical outputs through softmax, but both methods focus on learning feature or attention maps for downstream applications instead of acquiring a set of instances that are representative and diverse. Recently, SCAN/NNM/RUC [24, 53, 67] produce image clusters to be evaluated against semantic classes via Hungarian matching. However, such methods are often compared *against* SSL methods [67], whereas our work is *for* SSL methods. See appendix for more discussions about **self-supervised learning** and **deep clustering** methods.

4 Experiments

We evaluate our USL and USL-T by integrating them into both pseudo-label based SSL methods (FixMatch [63], MixMatch [5], or CoMatch [44]) and transfer-based SSL methods (SimCLRv2 and SimCLRv2-CLD [15,68]). We also compare against various AL/SSAL methods. Lastly, we show several intriguing properties of USL/USL-T such as generalizability.

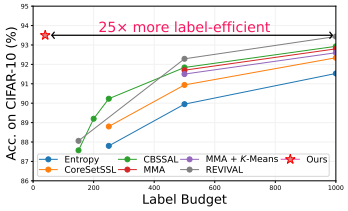
4.1 CIFAR-10

We compare against mainstream SSL methods such as FixMatch [63] and SimCLRv2-CLD [15,68] on extremely low-label settings to demonstrate our superior label efficiency. The labeling budget is 40 samples in total unless otherwise stated. Note that the self-supervised models used for instance selection are trained on CIFAR-10 from scratch entirely *without external data*. The SSL part, including backbone and hyperparameters, is untouched. See appendix for details.

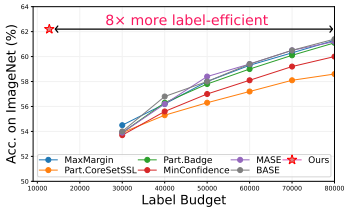
Comparison with AL and SSAL. Table 2 compares ours against various recent AL/SSAL methods in terms of sample efficiency and accuracy. AL methods operate at a much larger labeling budget than ours (187× more), because they rely only on labeled samples to learn both features and classification. SSAL methods make use of unlabeled samples and have higher label efficiency. However, we achieve much higher accuracy with fewer labels requested.

CIFAR-10	Budget	Acc (%)
Active Learning (AL)		
CoreSet [59]†	7500	85.4
VAAL [62]†	7500	86.8
UncertainGCN [9]†	7500	86.8
CoreGCN [9]†	7500	86.5
MCDAL [18]	7500	87.2
Semi-supervised Active Learning (SSAL)		
TOD-Semi [38]	7500	87.8
CoreSetSSL [59]‡	250	88.8
CBSSAL [30]	150	87.6
MMA [64]	500	91.7
MMA+k-means [64]	500	91.5
REVIVAL [34]	150	88.0
Selective Labeling		
FixMatch + USL (Ours)	**40**	90.4
FixMatch + USL (Ours)	**100**	**93.2**
FixMatch + USL-T (Ours)	**40**	**93.5**

Table 2: USL and USL-T greatly outperform AL/SSAL methods in accuracy and label efficiency on CIFAR-10. †, ‡: results from [38] and [30], respectively.



(a) CIFAR-10



(b) ImageNet

Fig. 4: Compared to SSAL, USL gets up to 25× higher label efficiency.

To tease apart whether our performance gains come from SSL or selective labeling, we tune recent AL/SSAL methods with their public implementations and run experiments with the same total budget, i.e. 40 samples in a 20 random + 20 selected setting. We then apply AL/SSAL selections to the same SSL for a fair comparison (Table 3).

While AL performs better than random selection in SimCLRv2-CLD, its advantage saturates on FixMatch. Since AL relies on labeled samples to learn the right features, with 20 random samples, it is very difficult to learn meaningful features for selection. Instead, AL could only learn a very coarse selection criterion and hence limited gains.

SSAL methods have greater gains on SimCLRv2-CLD. However, since SSAL still depends on initial random selections which seldom cover all 10 classes, these methods do not have an accurate knowledge of the full dataset in the low-label setting, where many rounds of queries are infeasible. That is, there is a serious trade-off in the low-label regime: Allowing more samples (e.g., 30) in the initial random selection for better coverage means less annotation budget for AL/SSAL selection (e.g., 10). Such a dilemma manifests itself in the imbalanced selection in Fig. 5 and the poor performance on FixMatch.

USL/USL-T as a Universal Method. In addition to mainstream SSL, we also use SimCLRv2, MixMatch [5], and SOTA CoMatch [44] for a comprehensive evaluation in Table 4. We observe significant accuracy gains on all of them.

CIFAR-10	S.v2-CLD	FixMatch
Random Selection	60.8	82.9
Stratified Selection†	66.5	88.6
UncertainGCN	63.0	77.3
CoreGCN	62.9	72.9
MMA^{+}‡	60.2	71.3
TOD-Semi	65.1	83.3
USL (Ours)	**76.6** ↑11.5	**90.4** ↑7.1
USL-T (Ours)	**76.1** ↑11.0	**93.5** ↑10.2

Table 3: The samples selected by USL and USL-T greatly outperform the ones from AL/SSAL on [15, 63, 68], with a budget of 40 labels on CIFAR-10. ‡: MMA^{+} is our improved MMA [64] based on FixMatch. †: not a fair baseline.

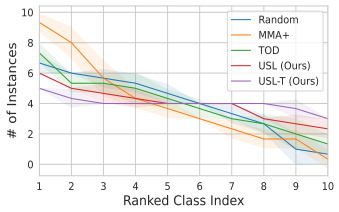


Fig. 5: Comparisons on the semantic class distributions of several methods over 3 runs. USL and USL-T get more balanced distribution.

CIFAR-10	MixMatch	SimCLRv2	SimCLRv2-CLD	FixMatch	CoMatch
Random	43.4	55.9	60.8	82.9	87.4
Stratified†	62.0	69.8	66.5	88.6	93.1
USL (Ours)	**61.6** ↑18.2	**69.1** ↑13.2	**76.6** ↑15.8	**90.4** ↑7.5	**93.4** ↑6.0
USL-T (Ours)	**66.0** ↑22.6	**71.5** ↑15.6	**76.1** ↑15.3	**93.5** ↑10.6	**93.0** ↑5.6

Table 4: USL/USL-T is a universal method that brings significant accuracy gains to various SSL methods. Experiments are conducted on CIFAR-10 with 40 labels. †: practically infeasible, as it assumes perfectly balanced labeled instances.

4.2 CIFAR-100

On CIFAR-100, we keep hyperparameters the same as the ones for CIFAR-10, except that we change the budget level to 400 to have 4 labels per class on average. Although we may benefit more from hyperparameter tuning, we already show *consistent gains* over other selection methods (Table 5).

4.3 ImageNet-100 and ImageNet-1k

To demonstrate our effectiveness on large-scale datasets, we benchmark on 100 random classes of ImageNet [67] and the full ImageNet [57].

ImageNet-100. On SimCLRv2 with a budget of 400 labels in total, we outperform baselines by 6.1% in this extremely low-label setting (Table 6).

ImageNet-1k: Setup. We experiment on SimCLRv2 and FixMatch with 1% (12,820 labels) and 0.2% (2,911 labels) labeled data. We also design a variant of our method that utilizes features provided by CLIP [54]. CLIP is trained on uncurated internet-crawled data in a wide range of domains. Following [8], we initialize FixMatch parameters with MoCov2. See appendix for more details.

ImageNet-1k: Comparing With AL/SSAL Methods. As most AL/SSAL methods in Table 2 do not scale to ImageNet, we compare our USL with SSAL methods specifically designed for ImageNet-scale settings [27]. Fig. 4b shows our

CIFAR-100	S.v2-CLD Acc	FixMatch Acc
Random Selection	26.5	48.7
Stratified Selection†	30.6	51.2
USL (Ours)	**33.0** ↑**6.5**	**55.1** ↑**6.4**
USL-T (Ours)	**36.9** ↑**10.4**	**55.7** ↑**7.0**

Table 5: By selecting informative samples to label, USL and USL-T greatly improve performance of SSL methods on CIFAR-100 with 400 labels. †: practically infeasible, as it assumes perfectly balanced labeled instances.

ImageNet-100	SimCLRv2 Acc
Random	62.2
Stratified†	65.1
USL (Ours)	**67.5** ↑**5.3**
USL-T (Ours)	**68.3** ↑**6.1**

Table 6: USL and USL-T scale well to high dimensional image inputs with many classes on ImageNet-100 [67]. †: practically infeasible.

	SimCLRv2		FixMatch	
ImageNet-1k	1%	0.20%	1%	0.20%
Random	49.7	33.2	58.8	34.3
Stratified†	52.0	36.4	60.9*	41.1
USL-MoCo (Ours)	51.5 ↑1.8	39.8 ↑6.6	61.6 ↑2.8	**48.6** ↑**14.3**
USL-CLIP (Ours)	**52.6** ↑**2.9**	**40.4** ↑**7.2**	**62.2** ↑**3.4**	47.5 ↑13.2

Table 7: Our proposed methods scale well on large-scale dataset ImageNet [57]. *: reported in [8]. USL-MoCo and USL-CLIP use MoCov2 features and CLIP features, respectively, to perform selective labeling. †: not a fair comparison.

8× improvement in terms of label efficiency. Table 7 shows that our approach provides up to 14.3% (3.4%) gains in the 0.2% (1%) SSL setting.

ImageNet-1k: USL-CLIP. Table 7 shows samples selected according to both MoCov2 and CLIP features boost SSL performance. USL-MoCo performs 1.1% better than USL-CLIP in the FixMatch setting. We hypothesize that it is, in part, due to a mismatch between parameter initialization (MoCov2) and the feature space used for the sampling process (CLIP). However, for 1% case, USL-CLIP performs 0.6% better than USL-MoCo, showing a slight advantage of a model trained with sufficient general knowledge and explicit semantics.

4.4 Strong Generalizability

Cross-dataset Generalizability with CLIP. Since CLIP does not use ImageNet samples in training and the downstream SSL task is not exposed to the CLIP model either, USL-CLIP's result shows strong cross-dataset generalizability in Table 7. It means that: **1)** When a new dataset is collected, we could use a general multi-modal model to skip self-supervised pretraining; **2)** Unlike AL where sample selection is strictly coupled with model training, our annotated instances work *universally* rather than with only the model used to select them.

Cross-domain Generalizability. Such generalizability also holds *across domains*. We use a CLD model trained on CIFAR-10 to select 40 labeled instances in medical imaging dataset BloodMNIST [74]. Although our model has *not* been

Weights	Selection Method	Accuracy
SimCLR [14]	Random	55.9
SimCLR [14]	USL-T (Ours)	71.5
CLD [68]	USL-T (Ours)	77.2
USL-T (Ours)	USL-T (Ours)	**85.4** ↑**8.2**

Table 8: The backbone weights learned as a by-product in USL-T capture more semantic information, thereby working as a good initialization.

Hyperparam	CIFAR-10/100	ImageNet-100/1k
Adjustment Factor α	5	2.5
Temperature t	0.25	0.5
Loss Term Weight λ	5	0.5
Neighborhood Size k	20	
Momentum μ	0.5	

Table 9: Hyperparams for USL-T. Hyperparams for USL are in appendix.

trained on any medical images, our model with FixMatch performs 10.9% (7.6%) better than random (stratified) sampling. See appendix for more details.

4.5 USL-T for Representation Learning

Our USL-T updates feature backbone weights during selective labeling. The trained weights are not used as a model initializer in the downstream SSL experiments for fair comparisons. However, we discover surprising generalizability that greatly exceeds self-supervised learning models under the SimCLRv2 setting. Specifically, we compare the performance of classifiers that are initialized with various model weights and are optimized on samples selected by different methods. Table 8 shows that, even with these strong baselines, initializing the model with our USL-T weights surpasses baselines by 8.2%.

4.6 Hyperparameters and Run Time

Table 9 shows that our hyperparameters generalize within small-scale and large-scale datasets. Our computational overhead is negligible. On ImageNet, we only introduce about 1 GPU hour for selective labeling, as opposed to 2300 GPU hours for the subsequent FixMatch pipeline. See appendix for more analysis, including formulations and visualizations.

5 Summary

Unlike existing SSL methods that focus on algorithms that better integrate labeled and unlabeled data, our selective-labeling is the first to focus on unsupervised data selection for labeling and enable more effective subsequent SSL. By choosing a diverse representative set of instances for annotation, we show significant gains in annotation efficiency and downstream accuracy, with remarkable selection generalizability within and across domains.

Acknowledgements. The authors thank Alexei Efros and Trevor Darrell for helpful discussions and feedback on this work in their classes.

References

1. Angluin, D.: Queries and concept learning. Machine learning **2**(4), 319–342 (1988) 9
2. Assran, M., Caron, M., Misra, I., Bojanowski, P., Joulin, A., Ballas, N., Rabbat, M.: Semi-supervised learning of visual features by non-parametrically predicting view assignments with support samples. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8443–8452 (2021) 8
3. Atlas, L.E., Cohn, D.A., Ladner, R.E.: Training connectionist networks with queries and selective sampling. In: Advances in neural information processing systems. pp. 566–573. Citeseer (1990) 9
4. Berthelot, D., Carlini, N., Cubuk, E.D., Kurakin, A., Sohn, K., Zhang, H., Raffel, C.: Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. arXiv preprint arXiv:1911.09785 (2019) 1, 2, 8, 9
5. Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., Raffel, C.: Mixmatch: A holistic approach to semi-supervised learning. arXiv preprint arXiv:1905.02249 (2019) 1, 2, 8, 9, 10, 11
6. Bilgic, M., Getoor, L.: Link-based active learning. In: NIPS Workshop on Analyzing Networks and Learning with Graphs. vol. 4 (2009) 9
7. Bondy, J.A., Murty, U.S.R., et al.: Graph theory with applications, vol. 290. Macmillan London (1976) 5
8. Cai, Z., Ravichandran, A., Maji, S., Fowlkes, C., Tu, Z., Soatto, S.: Exponential moving average normalization for self-supervised and semi-supervised learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 194–203 (2021) 12, 13
9. Caramalau, R., Bhattarai, B., Kim, T.K.: Sequential graph convolutional network for active learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9583–9592 (2021) 11
10. Carlini, N., Erlingsson, U., Papernot, N.: Distribution density, tails, and outliers in machine learning: Metrics and applications. arXiv preprint arXiv:1910.13427 (2019) 9
11. Caron, M., Bojanowski, P., Joulin, A., Douze, M.: Deep clustering for unsupervised learning of visual features. In: ECCV (2018) 10
12. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9650–9660 (2021) 10
13. Chan, Y.C., Li, M., Oymak, S.: On the marginal benefit of active learning: Does self-supervision eat its cake? In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 3455–3459. IEEE (2021) 2
14. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020) 4, 14
15. Chen, T., Kornblith, S., Swersky, K., Norouzi, M., Hinton, G.: Big self-supervised models are strong semi-supervised learners. arXiv preprint arXiv:2006.10029 (2020) 2, 3, 9, 10, 12
16. Chen, W., Pu, S., Xie, D., Yang, S., Guo, Y., Lin, L.: Unsupervised image classification for deep representation learning. In: European Conference on Computer Vision. pp. 430–446. Springer (2020) 10

17. Chen, X., Fan, H., Girshick, R., He, K.: Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297 (2020) 4
18. Cho, J.W., Kim, D.J., Jung, Y., Kweon, I.S.: Mcdal: Maximum classifier discrepancy for active learning. arXiv preprint arXiv:2107.11049 (2021) 11
19. Chung, F.R., Graham, F.C.: Spectral graph theory. No. 92, American Mathematical Soc. (1997) 5
20. Cohn, D., Atlas, L., Ladner, R.: Improving generalization with active learning. Machine learning **15**(2), 201–221 (1994) 9
21. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P.: Natural language processing (almost) from scratch. Journal of machine learning research **12**(ARTICLE), 2493–2537 (2011) 1
22. Cubuk, E.D., Zoph, B., Shlens, J., Le, Q.V.: Randaugment: Practical automated data augmentation with a reduced search space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 702–703 (2020) 9
23. Dagan, I., Engelson, S.P.: Committee-based sampling for training probabilistic classifiers. In: Machine Learning Proceedings 1995, pp. 150–157. Elsevier (1995) 9
24. Dang, Z., Deng, C., Yang, X., Wei, K., Huang, H.: Nearest neighbor matching for deep clustering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13693–13702 (2021) 10
25. Deo, N.: Graph theory with applications to engineering and computer science. Networks **5**(3), 299–300 (1975) 5
26. Ducoffe, M., Precioso, F.: Adversarial active learning for deep networks: a margin based approach. arXiv preprint arXiv:1802.09841 (2018) 2, 9
27. Emam, Z.A.S., Chu, H.M., Chiang, P.Y., Czaja, W., Leapman, R., Goldblum, M., Goldstein, T.: Active learning at the imagenet scale. arXiv preprint arXiv:2111.12880 (2021) 12
28. Fix, E., Hodges, J.L.: Discriminatory analysis. nonparametric discrimination: Consistency properties. International Statistical Review/Revue Internationale de Statistique **57**(3), 238–247 (1989) 5
29. Forgy, E.W.: Cluster analysis of multivariate data: efficiency versus interpretability of classifications. biometrics **21**, 768–769 (1965) 5
30. Gao, M., Zhang, Z., Yu, G., Arık, S.Ö., Davis, L.S., Pfister, T.: Consistency-based semi-supervised active learning: Towards minimizing labeling cost. In: European Conference on Computer Vision. pp. 510–526. Springer (2020) 10, 11
31. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 580–587 (2014) 1
32. Goodfellow, I., Bengio, Y., Courville, A.: Deep learning. MIT press (2016) 1
33. Grandvalet, Y., Bengio, Y., et al.: Semi-supervised learning by entropy minimization. CAP **367**, 281–296 (2005) 1, 9
34. Guo, J., Shi, H., Kang, Y., Kuang, K., Tang, S., Jiang, Z., Sun, C., Wu, F., Zhuang, Y.: Semi-supervised active learning for semi-supervised models: Exploit adversarial examples with graph-based virtual labels. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2896–2905 (2021) 11
35. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9729–9738 (2020) 4
36. Hestness, J., Narang, S., Ardalani, N., Diamos, G., Jun, H., Kianinejad, H., Patwary, M., Ali, M., Yang, Y., Zhou, Y.: Deep learning scaling is predictable, empirically. arXiv preprint arXiv:1712.00409 (2017) 1

37. Huang, S.J., Jin, R., Zhou, Z.H.: Active learning by querying informative and representative examples. Advances in neural information processing systems **23**, 892–900 (2010) 9
38. Huang, S., Wang, T., Xiong, H., Huan, J., Dou, D.: Semi-supervised active learning with temporal output discrepancy. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3447–3456 (2021) 11
39. Kriegel, H.P., Schubert, E., Zimek, A.: The (black) art of runtime evaluation: Are we comparing algorithms or implementations? Knowledge and Information Systems **52**(2), 341–378 (2017) 5
40. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009) 4
41. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems **25**, 1097–1105 (2012) 1
42. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. nature **521**(7553), 436–444 (2015) 1
43. Lee, D.H., et al.: Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In: Workshop on challenges in representation learning, ICML. vol. 3 (2013) 1, 7, 9
44. Li, J., Xiong, C., Hoi, S.C.: Comatch: Semi-supervised learning with contrastive graph regularization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9475–9484 (2021) 10, 11
45. Li, S., Chen, D., Chen, Y., Yuan, L., Zhang, L., Chu, Q., Liu, B., Yu, N.: Improve unsupervised pretraining for few-label transfer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10201–10210 (2021) 10
46. Liu, B., Wu, Z., Hu, H., Lin, S.: Deep metric transfer for label propagation with limited annotated data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. pp. 0–0 (2019) 9
47. Lloyd, S.: Least squares quantization in pcm. IEEE transactions on information theory **28**(2), 129–137 (1982) 5
48. McLachlan, G.J., Krishnan, T.: The EM algorithm and extensions, vol. 382. John Wiley & Sons (2007) 5
49. Menon, A.K., Jayasumana, S., Rawat, A.S., Jain, H., Veit, A., Kumar, S.: Long-tail learning via logit adjustment. arXiv preprint arXiv:2007.07314 (2020) 8
50. Miao, Z., Liu, Z., Gaynor, K.M., Palmer, M.S., Yu, S.X., Getz, W.M.: Iterative human and automated identification of wildlife images. Nature Machine Intelligence **3**(10), 885–895 (2021) 9
51. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018) 4
52. Orava, J.: K-nearest neighbour kernel density estimation, the choice of optimal k. Tatra Mountains Mathematical Publications **50**(1), 39–50 (2011) 5
53. Park, S., Han, S., Kim, S., Kim, D., Park, S., Hong, S., Cha, M.: Improving unsupervised image clustering with robust learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12278–12287 (2021) 10
54. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. arXiv preprint arXiv:2103.00020 (2021) 12
55. Ren, P., Xiao, Y., Chang, X., Huang, P.Y., Li, Z., Chen, X., Wang, X.: A survey of deep active learning. arXiv preprint arXiv:2009.00236 (2020) 9

56. Roy, N., Mccallum, A.: Toward optimal active learning through sampling estimation of error reduction. Proceedings of the 18th International Conference on Machine Learning (08 2001) 9
57. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV) **115**(3), 211–252 (2015). https://doi.org/10.1007/s11263-015-0816-y 4, 12, 13
58. Sajjadi, M., Javanmardi, M., Tasdizen, T.: Regularization with stochastic transformations and perturbations for deep semi-supervised learning. arXiv preprint arXiv:1606.04586 (2016) 1, 9
59. Sener, O., Savarese, S.: Active learning for convolutional neural networks: A core-set approach. arXiv preprint arXiv:1708.00489 (2017) 2, 9, 11
60. Settles, B.: Active learning literature survey (2009) 9
61. Shi, J., Malik, J.: Normalized cuts and image segmentation. IEEE Transactions on pattern analysis and machine intelligence **22**(8), 888–905 (2000) 5
62. Sinha, S., Ebrahimi, S., Darrell, T.: Variational adversarial active learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5972–5981 (2019) 11
63. Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C.A., Cubuk, E.D., Kurakin, A., Li, C.L.: Fixmatch: Simplifying semi-supervised learning with consistency and confidence. Advances in Neural Information Processing Systems **33** (2020) 1, 2, 3, 7, 9, 10, 12
64. Song, S., Berthelot, D., Rostamizadeh, A.: Combining mixmatch and active learning for better accuracy with fewer labels. arXiv preprint arXiv:1912.00594 (2019) 3, 10, 11, 12
65. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. pp. 1195–1204 (2017) 1, 9
66. Tong, S., Koller, D.: Support vector machine active learning with applications to text classification. Journal of machine learning research **2**(Nov), 45–66 (2001) 9
67. Van Gansbeke, W., Vandenhende, S., Georgoulis, S., Proesmans, M., Van Gool, L.: Scan: Learning to classify images without labels. In: European Conference on Computer Vision. pp. 268–285. Springer (2020) 7, 10, 12, 13
68. Wang, X., Liu, Z., Yu, S.X.: Unsupervised feature learning by cross-level instance-group discrimination. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12586–12595 (2021) 4, 10, 12, 14
69. Wang, X., Wu, Z., Lian, L., Yu, S.X.: Debiased learning from naturally imbalanced pseudo-labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14647–14657 (2022) 1, 2, 9
70. Wei, K., Iyer, R., Bilmes, J.: Submodularity in data subset selection and active learning. In: International Conference on Machine Learning. pp. 1954–1963. PMLR (2015) 9
71. Wu, Z., Xiong, Y., Yu, S.X., Lin, D.: Unsupervised feature learning via non-parametric instance discrimination. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3733–3742 (2018) 4
72. Xie, Q., Dai, Z., Hovy, E., Luong, T., Le, Q.: Unsupervised data augmentation for consistency training. Advances in Neural Information Processing Systems **33** (2020) 1, 9

73. Xie, Q., Luong, M.T., Hovy, E., Le, Q.V.: Self-training with noisy student improves imagenet classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10687–10698 (2020) 1
74. Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., Pfister, H., Ni, B.: Medmnist v2: A large-scale lightweight benchmark for 2d and 3d biomedical image classification. arXiv preprint arXiv:2110.14795 (2021) 13
75. Yoo, D., Kweon, I.S.: Learning loss for active learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 93–102 (2019) 2, 10