Unsupervised Selective Labeling for More Effective Semi-Supervised Learning

Anonymous ECCV CVinW submission

Paper ID 6

Abstract. Given an unlabeled dataset and an annotation budget, we study how to selectively label a budgeted 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 for real-world applications.

1 Introduction

Deep learning's success on natural language understanding [15], visual object recognition [26], and object detection [20] follow a straightforward recipe: better model architectures, more data, and scalable computation [21, 23, 27, 42]. As training datasets get bigger, their full task annotation becomes infeasible [1, 37].

Semi-supervised learning (SSL) deals with learning from both a small amount of labeled data *and* a large amount of unlabeled data. In SSL, the lower the annotation level, the more important what the labeled instances are to good generalization. While a typical image could represent many similar images that we will counter in downstream, an odd-ball only represents itself. Labeled instances may even only cover part of the data variety, trapping a classifier in partial views with unstable learning.

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 [1, 2, 37, 41]. Each method has its own caveats: Random

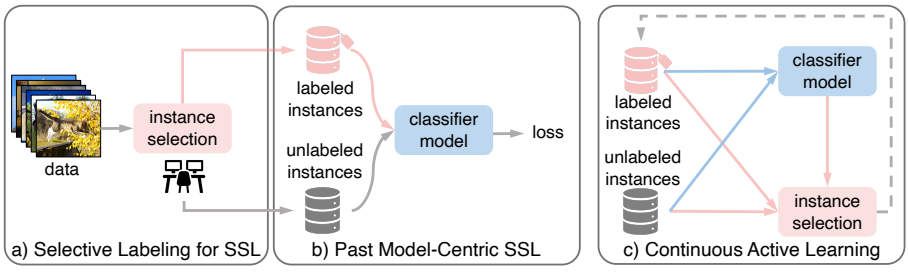


Fig. 1: Our unsupervised selective labeling is a novel aspect of semi-supervised learning (SSL) and different from active learning (AL). **a, b)** Existing SSL methods focus on optimizing the model *given* labeled and unlabeled data, while we focus on optimizing the selection of training instances *prior to* label acquisition. **c)** Existing AL methods require initial randomly-selected labeled data, while we select instances from unlabeled data without knowing the classification task.

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	✓	✗	✗	✓

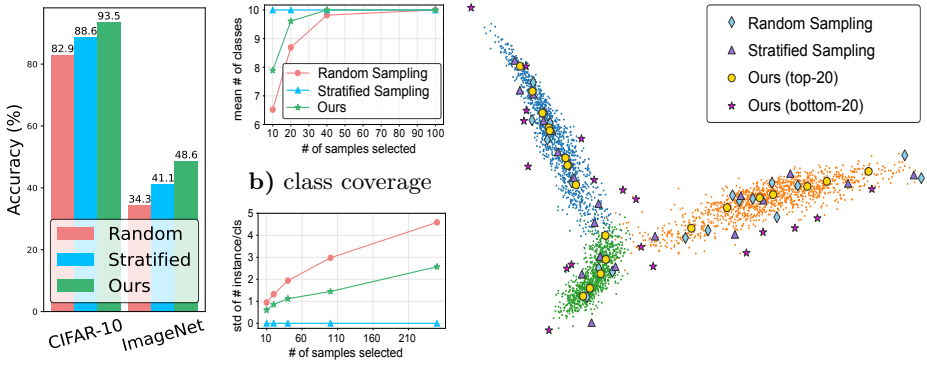
Table 1: Key properties of SSL, AL, SSAL, and our USL/USL-T pipelines. Among them, our approach is the only one that does not use any random labels.

sampling can fail to cover all semantic classes and lead to poor performance and instability, whereas stratified sampling is utterly *unrealistic in the wild*: If we can sample data by category, we would already have the label of every instance!

We address *unsupervised selective labeling* for SSL (Fig. 1) close to the wild: 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 trained on such partially labeled data?

Our work differs from active learning (AL) methods that they often require randomly sampled labeled data to begin with, which is sample-inefficient in low-label settings that SSL methods excel at [8]. Most notably, our work is the first *unsupervised* selective labeling method on large-scale recognition datasets that requests annotation only *once* (see Table 1 for comparisons on key properties).

Fig. 2 shows that our method has many benefits over random or stratified sampling for labeled data selection in accuracy, coverage, balance over classes, and representativeness. As it selects informative instances without initial 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 [25], our method scales up easily to *large-scale datasets* that are often encountered in the wild, as the selection runs within an hour on a commodity GPU server on ImageNet [34].



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)** SSL classification accuracy increases with our selectively labeled instances. **b)** Our method covers all the semantic classes with 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.

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 demonstrate our method's domain transfer ability to select samples in medical imaging with a model that never saw medical images.

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 produces the best classification performance.

Since we do not have any labels to begin with, our idea is to select m instances that are not only *representative* of most instances, but also *diverse* to 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 and remaining unlabeled data.

We propose two selective labeling methods in Step 2, training-free Unsupervised Selective Labeling (USL) and the training-based variant (USL-T). With the former leveraging self-supervised pretrained features [9,12,40] and the latter jointly optimizes the feature space and clusters, both *without* label supervision.

2.1 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 $\{f(x_i)\}$ obtained by self-supervised learning methods (e.g. MoCov2 [12], SimCLR [9] or CLD [40]), and edges between nodes are attached with weights of pairwise feature similarity [4, 13, 17, 36], 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 (representative), but also well dispersed to cover the entire dataset (diverse).

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. We use a robust variant of K-nearest neighbor density estimator [19, 30] to measure the *representativeness* and select the nodes with max density, formulated as:

$$\hat{p}_{\text{KNN}}(V_i, k) = \frac{k}{n} \frac{1}{A_d \cdot \bar{D}^d(V_i, k)}, \quad \text{where } \bar{D}(V_i, k) = \frac{1}{k} \sum_{j=1}^{k} D(V_i, V_{j(i)}). \tag{2}$$

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.

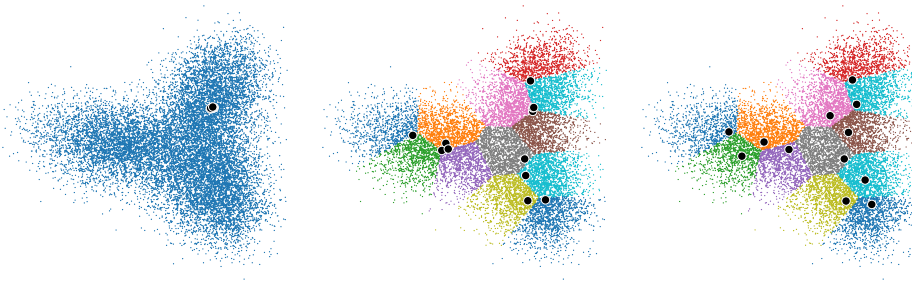
Diversity: Pick One in Each Cluster. To select m diverse instances that cover the entire unlabeled dataset, we resort to K-Means clustering that partitions n instances into $m(\leq n)$ clusters. Formally, we seek m-way node partitioning $\$ = \{S_1, S_2, ..., S_m\}$ that minimizes the within-cluster sum of squares [24]:

$$\min_{\$} \sum_{i=1}^{m} \sum_{V \in S_i} \|V - c_i\|^2 = \min_{\$} \sum_{i=1}^{m} |S_i| \text{Var}(S_i) \tag{3}$$

We then pick the most representative instance of each cluster according to Eqn. 2.

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



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.

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:

$$\mathrm{Reg}(V_i, t) = \sum_{\hat{V}_j^{t-1} \notin S_i} \frac{1}{\|V_i - \hat{V}_j^{t-1}\|^\alpha}. \tag{4}$$

This regularizer is updated with an exponential moving average:

$$\overline{\mathrm{Reg}}(V_i, t) = m_{\mathrm{reg}} \cdot \overline{\mathrm{Reg}}(V_i, t-1) + (1 - m_{\mathrm{reg}}) \cdot \mathrm{Reg}(V_i, t) \tag{5}$$

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:

$$U'(V_i, t) = U(V_i) - \lambda \cdot \overline{\mathrm{Reg}}(V_i, t) \tag{6}$$

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.

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

We also 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 label-efficient instance selection.

Global Constraint via Learnable K-Means Clustering. Clustering in a given feature space is not trivial (Fig. 3c). We introduce a learnable 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 and predict a soft cluster assignment $\hat{y}(x)$ by taking softmax over the similarity between instance x and each learnable centroid:

$$y_i(x) = [i = \arg\min_{k\in\{1,\ldots,C\}} s(f(x), c_k)], \quad \hat{y}_i(x) = \text{softmax}(e^{s(f(x),c)})_i$$

The hard assignment $y(x)$ can be regarded as pseudo-labels [28, 37, 39]. By minimizing the KL divergence between soft and hard assignments, $D_{\text{KL}}(y(x)\|\hat{y}(x))$, we encourage each instance to become more similar to its centroid and the centroid to become a better representative of instances in the cluster. With soft predictions, each sample affects on every centroid. For robust training, we only take pseudo-labels from confident predictions with confidence above τ:

$$L_{\text{global}}(\{x_i\}_{i=1}^n) = \frac{1}{n} \sum_{\max(\hat{y}(x_i))\geq\tau} D_{\text{KL}}(y(x_i)\|\hat{y}(x_i)) \tag{7}$$

where τ is the threshold hyperparameter. This loss leads to curriculum learning that gradually allows more instances to take part in training.

Local Constraint with Neighbor Cluster Alignment. Since soft assignments usually have low confidence scores for most instances at the beginning, convergence with global constraint 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.

However, 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.

Therefore, we applied logit adjustment [29] to the output logits to prevent one-cluster collapse and a sharpening function to prevent even-distribution collapse. Both the logit adjustment and sharpening function can be concisely captured in a single function $P(\cdot)$ that turns logits z into a reference distribution, with $\hat{P}(\cdot,\cdot)$ the logit adjustment operator and $\bar{z}$ the moving average of z:

$$[P(z,\bar{z},t)]_i = \frac{\exp(\hat{P}(z_i,\bar{z}_i)/t)}{\sum_j \exp(\hat{P}(z_j,\bar{z}_j/t))} \tag{8}$$

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

$$L_{\text{local}}(\{x_i\}_{i=1}^n) = \frac{1}{n}\sum_{i=1}^{n} D_{\text{KL}}(P(y(x_i'),\bar{y}(x_i'),t)||\hat{y}(x_i)). \tag{9}$$

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.

Our final loss adds up both terms with loss weight λ: $L = L_{\text{global}} + \lambda L_{\text{local}}$.

Diverse and Representative 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. $\max \hat{y}(x)$.

3 Related Work

Semi-supervised Learning (SSL) integrates information from small-scale labeled data and large-scale unlabeled data. *Pseudo-labeling* [1, 2, 28, 41] obtains pseudo-labels on unlabeled data from a model's confident predictions. *Transfer learning* method SimCLRv2 [10] is a two-stage method that fine-tunes self-supervised learning models on labeled data. *Entropy-minimization* [2, 22] assumes that classification boundaries do not pass through high-density area. Instead of competing with SSL methods, our USL enables more effective SSL by choosing the right instances to label *for* SSL without supervision.

Active Learning (AL) aims to select a subset of data to query labels to achieve competitive performance over full supervised learning [3, 14, 33]. In *Deep AL*, Core-Set [35] approaches data selection as a set cover problem. [18] estimates distances from decision boundaries based on sensitivity to adversarial attacks. LLAL [44] predicts target loss of unlabeled data and queries instances with the largest loss for labels. *Semi-supervised AL* (SSAL) combines AL with SSL.

Deep Clustering. Methods such as [6,7,11,16,31,39] also jointly learns features and cluster assignments. However, such methods are often compared *against* SSL methods [39], while our work is designed *for* SSL methods.

4 Experiments

We evaluate our USL and USL-T by integrating them into both pseudo-label based SSL method FixMatch [37] and transfer-based SSL methods (SimCLRv2/ SimCLRv2-CLD [10, 40]). We also compare against various AL/SSAL methods and show intriguing generalizability of USL(-T).

CIFAR-10. We compare with mainstream SSL methods FixMatch [37] and SimCLRv2-CLD [10, 40] on low-label settings (40 labeled samples) to demonstrate our superior label efficiency. The self-supervised weights used for instance selection are trained on CIFAR-10 from scratch *without external data*. The SSL part is untouched. Our selection leads to 10.2% (11.5%) SSL accuracy improvement combined with FixMatch [37] (SimCLRv2-CLD [10, 40]), compared to selections from AL/SSAL methods (Tab. 2). It is also more balanced (Fig. 4).

ImageNet-1k. On our benchmark on ImageNet [34], we use either MoCov2 [12] or CLIP [32] features as the first step of our selection. We evaluate on SSL methods SimCLRv2 and FixMatch with 1% (12,820 labels) and 0.2% (2,911 labels) labeled data. Tab. 3 shows that our approach provides up to 14.3% (3.4%) gains in the 0.2% (1%) SSL setting.

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 2: The samples selected by USL and USL-T greatly outperform the ones from AL/SSAL on [10, 37, 40], with a budget of 40 labels on CIFAR-10. ‡: MMA+ is our improved MMA [38] based on FixMatch. †: not a fair baseline.

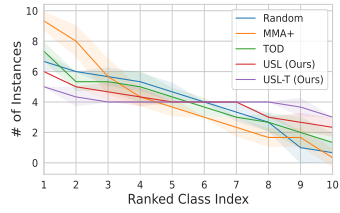


Fig. 4: Comparisons on the semantic class distributions of several methods over 3 runs. USL and USL-T get more balanced distribution.

ImageNet-1k	SimCLRv2 1%	SimCLRv2 0.20%	FixMatch 1%	FixMatch 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 3: Our proposed methods scale well on large-scale dataset ImageNet [34]. *: reported in [5]. USL-MoCo and USL-CLIP use MoCov2 features and CLIP features, respectively, to perform selective labeling. †: not a fair comparison.

4.1 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 Tab. 3. 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 holds *across domains*. We use a CLD model trained on CIFAR-10 to select 40 labeled instances in medical imaging dataset BloodMNIST [43]. Random selections, stratified selections, and USL selections obtain 77.17%, 80.46%, and 88.06% accuracy, respectively. Although our model has *not* been trained on any medical images, our model with FixMatch performs 10.9% (7.6%) better than random (stratified) sampling.

5 Summary

Unlike existing SSL methods that improve models and training algorithms, USL 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.

References

1. 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, 7
2. 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, 7
3. Bilgic, M., Getoor, L.: Link-based active learning. In: NIPS Workshop on Analyzing Networks and Learning with Graphs. vol. 4 (2009) 7
4. Bondy, J.A., Murty, U.S.R., et al.: Graph theory with applications, vol. 290. Macmillan London (1976) 4
5. 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) 8
6. Caron, M., Bojanowski, P., Joulin, A., Douze, M.: Deep clustering for unsupervised learning of visual features. In: ECCV (2018) 7
7. 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) 7
8. 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
9. 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) 3, 4
10. 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) 7, 8
11. 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) 7
12. Chen, X., Fan, H., Girshick, R., He, K.: Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297 (2020) 3, 4, 7
13. Chung, F.R., Graham, F.C.: Spectral graph theory. No. 92, American Mathematical Soc. (1997) 4
14. Cohn, D., Atlas, L., Ladner, R.: Improving generalization with active learning. Machine learning **15**(2), 201–221 (1994) 7
15. 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
16. 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) 7
17. Deo, N.: Graph theory with applications to engineering and computer science. Networks **5**(3), 299–300 (1975) 4

18. Ducoffe, M., Precioso, F.: Adversarial active learning for deep networks: a margin based approach. arXiv preprint arXiv:1802.09841 (2018) 7
19. Fix, E., Hodges, J.L.: Discriminatory analysis. nonparametric discrimination: Consistency properties. International Statistical Review/Revue Internationale de Statistique **57**(3), 238–247 (1989) 4
20. 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
21. Goodfellow, I., Bengio, Y., Courville, A.: Deep learning. MIT press (2016) 1
22. Grandvalet, Y., Bengio, Y., et al.: Semi-supervised learning by entropy minimization. CAP **367**, 281–296 (2005) 7
23. 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
24. 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) 4
25. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009) 2
26. 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
27. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. nature **521**(7553), 436–444 (2015) 1
28. 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) 6, 7
29. 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) 6
30. Orava, J.: K-nearest neighbour kernel density estimation, the choice of optimal k. Tatra Mountains Mathematical Publications **50**(1), 39–50 (2011) 4
31. 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) 7
32. 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) 7
33. 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) 7
34. 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 2, 7, 8
35. Sener, O., Savarese, S.: Active learning for convolutional neural networks: A core-set approach. arXiv preprint arXiv:1708.00489 (2017) 7
36. Shi, J., Malik, J.: Normalized cuts and image segmentation. IEEE Transactions on pattern analysis and machine intelligence **22**(8), 888–905 (2000) 4

37. 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, 6, 7, 8
38. Song, S., Berthelot, D., Rostamizadeh, A.: Combining mixmatch and active learning for better accuracy with fewer labels. arXiv preprint arXiv:1912.00594 (2019) 8
39. 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) 6, 7
40. 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) 3, 4, 7, 8
41. 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, 7
42. 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
43. 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) 8
44. 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) 7