DASO: Distribution-Aware Semantics-Oriented Pseudo-label for Imbalanced Semi-Supervised Learning

Youngtaek Oh[1], Dong-Jin Kim[2], and In So Kweon[1]

[1] Korea Advanced Institute of Science and Technology, Daejeon, Republic of Korea
{youngtaek.oh,iskweon77}@kaist.ac.kr
[2] Hanyang University, Seoul, Republic of Korea
djnjusa@gmail.com

Abstract. Semi-supervised learning (SSL) is far from real-world application due to severely biased pseudo-labels caused by (1) class imbalance and (2) class distribution mismatch between labeled and unlabeled data. This paper addresses such a relatively under-explored problem. First, we propose a general pseudo-labeling framework that class-adaptively blends the semantic pseudo-label from a similarity-based classifier to the linear one from the linear classifier, after making the observation that both types of pseudo-labels have complementary properties in terms of bias. We further introduce a novel semantic alignment loss to establish balanced feature representation to reduce the biased predictions. We term the whole framework as **D**istribution-**A**ware **S**emantics-**O**riented (DASO) Pseudo-label. We conduct extensive experiments in a wide range of imbalanced benchmarks and demonstrate that DASO reliably improves SSL learners especially when both (1) class imbalance and (2) distribution mismatch dominate.

Keywords: Semi-supervised Learning, Long-tailed Learning, Distribution Mismatch

1 Introduction

Semi-supervised learning (SSL) [4] has shown to be promising for leveraging unlabeled data to reduce the data cost [3,2,19]. The common approach is to produce *pseudo-labels* for unlabeled data based on model's predictions and utilize them for regularizing model training [13,17,19]. Although adopted in a variety of tasks, these algorithms often assume class-balanced data, while many real-world datasets exhibit *long-tailed* distributions [9]. With class-imbalanced data, pseudo-labels become severely biased to the majority classes due to confirmation bias [1]. Such pseudo-labels can further bias the model during training.

In this work, we present a new imbalanced SSL method for alleviating the bias in pseudo-labels, while discarding the common assumption that the class distribution of unlabeled data is the same with the label distribution. To this end, as shown in Fig. 1, we observe that semantic pseudo-labels [11] obtained from a

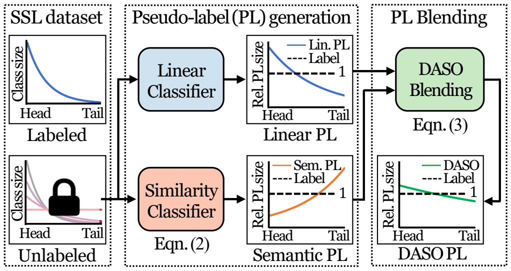


Fig. 1: DASO reduces the overall bias in pseudo-labels (PL) from unlabeled data by blending two complementary PLs from different classifiers.

similarity-based classifier [18] are biased towards minority classes as opposed to linear classifier-based pseudo-labels [17,19] being biased towards head classes. We draw the key inspiration from those complementary properties of two different types of pseudo-labels to develop a new pseudo-labeling scheme.

In this regard, we introduce a generic imbalanced SSL framework termed Distribution-Aware Semantics-Oriented (DASO) Pseudo-label. We propose to blend the linear and semantic pseudo-labels in different proportions for each class to reduce the overall bias. As such, without resorting to any class priors for the unlabeled data, DASO can reliably bring performance gain.

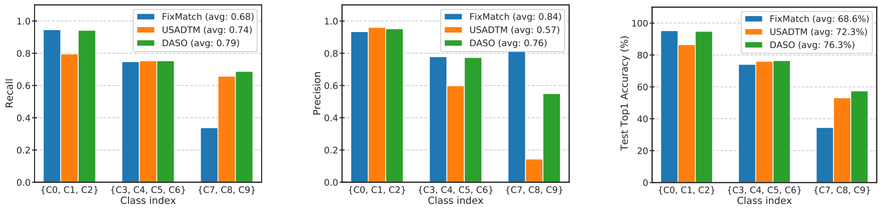
We further propose a simple yet effective semantic alignment loss to establish balanced feature representation. We consistently assign two different views of an unlabeled sample in *feature space* to the same prototype. These enhanced feature representations not only help linear classifier produce less biased predictions, but can also be reused for semantic pseudo-labels from similarity-based classifier.

The efficacy of DASO is extensively justified with the imbalanced versions of benchmarks: CIFAR-10/100 [15] and STL-10 [5]. We even test DASO with Semi-Aves [20], closely related to real-world scenarios. As such, DASO consistently benefits under various distributions of unlabeled data and degrees of imbalance, demonstrating to be a truly generic framework.

2 Proposed Method

2.1 Preliminaries

Problem setup. We consider K-class semi-supervised learning that leverages both labeled data $\mathcal{X} = \{(x_n, y_n)\}_{n=1}^{N}$ and unlabeled data $\mathcal{U} = \{u_m\}_{m=1}^{M}$ to train a model f. Note that the model $f = f_{\phi}^{\text{cls}} \circ f_{\theta}^{\text{enc}}$ consists of a feature encoder f_{θ}^{enc} followed by a linear classifier f_{ϕ}^{cls}, where θ and ϕ are the set of parameters of f_{θ}^{enc} and f_{ϕ}^{cls}. The input image x is paired with the label y to learn $\mathcal{L}_{\text{cls}}$ (*e.g.*, cross-entropy) from the prediction $f(x)$. For the unlabeled data, a pseudo-label $\hat{p} \in \mathbb{R}^K$ is assigned to learn the unsupervised loss $\mathcal{L}_u = \Phi_u(\hat{p}, f(u))$, where Φ_u can be implemented via entropy [10] or consistency regularization [16,21], depending on the SSL learner. For FixMatch [19] as an example, the pseudo-label $\hat{p} = \text{OneHot}\left(\text{argmax}_k\, p_k^{(w)}\right)$ with $p^{(w)} = f\left(\mathcal{A}_w(u)\right)$ provides the target for



(a) Recall of PL.

(b) Precision of PL.

(c) Test accuracy.

Fig. 2: Analysis on recall and precision of pseudo-labels (PL) and the corresponding test accuracy. Although USADTM [11] improves the recall of minority classes, the precision of those classes is significantly reduced. In contrast, DASO improves the recall of minority classes while sustaining the precision.

the prediction $p^{(s)} = f\left(\mathcal{A}_s(u)\right)$ with some confident ones to the cross-entropy loss $\mathcal{H}$, where $\mathcal{A}_w$ and $\mathcal{A}_s$ correspond to weak augmentation (*e.g.*, random flip and crop) and advanced augmentation (*e.g.*, RandAugment [6]), respectively.

Imbalanced semi-supervised learning. Let us denote N_k and M_k as the number of labeled and unlabeled examples respectively in class k. The degree of imbalance for each data is characterized by the imbalance ratio, γ_l or γ_u, where we assume $\gamma_l = \frac{\max_k N_k}{\min_k N_k} \gg 1$. γ_u is specified in the same way using the actual labels without access during training. As note, class distribution of $\mathcal{U}$ (*e.g.*, γ_u) can significantly diverge from $\mathcal{X}$ in practice, and such varying distributions greatly affect the performances. In this regard, our goal is to debias the pseudo-labels with class-imbalanced data, while maintaining the performances of SSL algorithms with various, but still *unknown* class distribution of unlabeled data.

Trade-offs between linear and semantic pseudo-label. As shown in Fig. 2, we compare FixMatch [19] and USADTM [11] using linear and semantic pseudo-label respectively. From Figs. 2a and 2b, FixMatch achieves high recall in majority classes while low recall but high precision in the minorities, suggesting that actual minority class examples are biased towards head classes. In contrast, for USADTM, the actual majorities are biased towards minority classes. This is because the precision of tail classes has decreased significantly in Fig. 2b, while the recall has increased in sacrifice of the recall from head classes in Fig. 2a.

2.2 DASO Pseudo-label Framework

Framework overview. Without loss of generality, we consider DASO built on top of FixMatch [19] for convenience. First, the linear and semantic pseudo-label, $\hat{p}$ and $q^{(w)}$ are produced with a feature $z^{(w)} = f_\theta^{\text{enc}}(\mathcal{A}_w(u))$ from the linear and similarity-based classifier, respectively. Then the final pseudo-label $\hat{p}'$ is obtained from the distribution-aware blending process using $\hat{p}$ and $q^{(w)}$, and it provides the target to $\mathcal{L}_u = \Phi_u(\hat{p}', p)$ instead of linear pseudo-label in the existing SSL learner. In case of FixMatch, the prediction of u corresponds to $p = p^{(s)} = f(\mathcal{A}_s(u))$. For the semantic alignment loss, the semantic pseudo-label $q^{(w)}$ provides the target for $q^{(s)}$ to the cross-entropy, where $q^{(s)}$ is the result of the similarity-based classifier with $z^{(s)} = f_\theta^{\text{enc}}(\mathcal{A}_s(u))$. Note that we denote $q^{(w)}$ as $\hat{q}$ for simplicity, unless confusion arises.

Balanced prototype generation. To execute a similarity-based classifier for obtaining the semantic pseudo-label, we first build a set of class prototypes $\mathbf{C} = \{c_k\}_{k=1}^K$ from $\mathcal{X}$, similar to [11]. In detail, we build a dictionary of memory queue $\mathbf{Q} = \{Q_k\}_{k=1}^K$ where each key corresponds to the class and Q_k denotes a memory queue for class k with the fixed size $|Q_k|$. The class prototype c_k for every class k is efficiently calculated by averaging the feature points in the queue Q_k, where we update Q_k for all k at every step by pushing new features from labeled data in the batch and discarding the most old ones when Q_k is full.

Linear and semantic pseudo-label generation. We obtain linear pseudo-label $\hat{p}$ as: $\hat{p} = \sigma(f_\phi^{\text{cls}}(z^{(w)}))$. Semantic pseudo-label $\hat{q}$ is obtained from the similarity classifier that measures the per-class similarity of a feature point z of either $z^{(w)}$ or $z^{(s)}$ to the prototypes $\mathbf{C}$: $q = \sigma\left(\text{sim}(z, \mathbf{C})\ /\ T_{\text{proto}}\right)$, where $\text{sim}(\cdot, \cdot)$ corresponds to cosine similarity, and T_{proto} is a temperature hyper-parameter. Note that $\hat{p}$ is biased towards head classes while $\hat{q}$ is the vice versa.

Distribution-aware blending. To obtain unbiased pseudo-label $\hat{p}'$, the semantic pseudo-label $\hat{q}$ should be exploited *differently* across the class. To this end, we increase the exposure of the component of $\hat{q}$ when $\hat{p}$ is more biased to the head classes. Formally, we blend them with a set of distribution-aware weights $\upsilon = \{\upsilon_k\}_{k=1}^K$ to reduce the bias that might occur when using either $\hat{p}$ or $\hat{q}$:

$$\hat{p}' = (1 - \upsilon_{k'})\,\hat{p} + \upsilon_{k'}\hat{q}, \tag{1}$$

where $\upsilon_k = \frac{1}{\max_k \hat{m}_k^{1/T_{\text{dist}}}} \left(\hat{m}_k^{1/T_{\text{dist}}}\right)$ and k' is the class prediction from $\hat{p}$. Note that $\hat{m}$ is the normalized class distribution of the current pseudo-labels and T_{dist} is a hyper-parameter that intercedes the optimal trade-offs between $\hat{p}$ and $\hat{q}$. Overall, in terms of the linear pseudo-label, the minority pseudo-labels will remain as minority, while pseudo-labels predicted as majority will be likely to recover the original classes thanks to large $\upsilon_{k'}$. This makes DASO flexible to various distributions of $\mathcal{U}$ without resorting to any distribution.

Semantic alignment loss. To establish balanced feature representations, we propose new semantic alignment loss. In high-level, we align each unlabeled sample u to the most similar prototype used in the similarity classifier, by imposing *consistent assignment* for two augmented views $\mathcal{A}_w(u)$ and $\mathcal{A}_s(u)$ to the same c_k in feature space. Note $\hat{q}$ provides the target for $q^{(s)}$ with cross-entropy $\mathcal{H}$:

$$\mathcal{L}_{\text{align}} = \mathcal{H}\left(\hat{q},\, q^{(s)}\right). \tag{2}$$

Finally, the enhanced representation can implicitly guide the classifier f_ϕ^{cls} to produce less biased predictions in general.

Total objective. DASO can easily couple with other SSL algorithms with the modified pseudo-label, where the final DASO objective is as below:

$$\mathcal{L}_{\text{DASO}} = \mathcal{L}_{\text{cls}} + \lambda_u \mathcal{L}_u + \lambda_{\text{align}} \mathcal{L}_{\text{align}}, \tag{3}$$

where $\mathcal{L}_{\text{cls}}$ and $\mathcal{L}_u$ come from the base SSL learner, and $\mathcal{L}_{\text{align}}$ is newly introduced from DASO. Note that $\mathcal{L}_u$ takes the blended pseudo-label in Eq. (1).

	CIFAR10-LT				CIFAR100-LT			
	$\gamma=\gamma_l=\gamma_u=100$		$\gamma=\gamma_l=\gamma_u=150$		$\gamma=\gamma_l=\gamma_u=10$		$\gamma=\gamma_l=\gamma_u=20$	
Algorithm	$N_1=500$ $M_1=4000$	$N_1=1500$ $M_1=3000$	$N_1=500$ $M_1=4000$	$N_1=1500$ $M_1=3000$	$N_1=50$ $M_1=400$	$N_1=150$ $M_1=300$	$N_1=50$ $M_1=400$	$N_1=150$ $M_1=300$
FixMatch [19]	67.8 ±1.13	77.5 ±1.32	62.9 ±0.36	72.4 ±1.03	45.2 ±0.55	56.5 ±0.06	40.0 ±0.96	50.7 ±0.25
w/ DARP [14]	74.5 ±0.78	77.8 ±0.63	67.2 ±0.32	73.6 ±0.73	49.4 ±0.20	58.1 ±0.44	43.4 ±0.87	52.2 ±0.66
w/ CReST+ [22]	**76.3** ±0.86	78.1 ±0.42	67.5 ±0.45	73.7 ±0.34	44.5 ±0.94	57.4 ±0.18	40.1 ±1.28	52.1 ±0.21
w/ DASO (Ours)	76.0 ±0.37	**79.1** ±0.75	**70.1** ±1.81	**75.1** ±0.77	**49.8** ±0.24	**59.2** ±0.35	**43.6** ±0.09	**52.9** ±0.42

Table 1: Comparison of accuracy (%) on CIFAR10/100-LT under $\gamma_l = \gamma_u$ setup.

	CIFAR10-LT ($\gamma_l \neq \gamma_u$)				STL10-LT ($\gamma_u = N/A$)			
	$\gamma_u=1$ (uniform)		$\gamma_u=1/100$ (reversed)		$\gamma_l=10$		$\gamma_l=20$	
Algorithm	$N_1=500$ $M_1=4000$	$N_1=1500$ $M_1=3000$	$N_1=500$ $M_1=4000$	$N_1=1500$ $M_1=3000$	$N_1=150$ $M=100k$	$N_1=450$ $M=100k$	$N_1=150$ $M=100k$	$N_1=450$ $M=100k$
FixMatch [19]	73.0 ±3.81	81.5 ±1.15	62.5 ±0.94	71.8 ±1.70	56.1 ±2.32	72.4 ±0.71	47.6 ±4.87	64.0 ±2.27
w/ DARP [14]	82.5 ±0.75	84.6 ±0.34	70.1 ±0.22	80.0 ±0.93	66.9 ±1.66	75.6 ±0.45	59.9 ±2.17	72.3 ±0.60
w/ CReST [22]	83.2 ±1.67	87.1 ±0.28	70.7 ±2.02	**80.8** ±0.39	61.7 ±2.51	71.6 ±1.17	57.1 ±3.67	68.6 ±0.88
w/ CReST+ [22]	82.2 ±1.53	86.4 ±0.42	62.9 ±1.39	72.9 ±2.00	61.2 ±1.27	71.5 ±0.96	56.0 ±3.19	68.5 ±1.88
w/ DASO (Ours)	**86.6** ±0.84	**88.8** ±0.59	**71.0** ±0.95	80.3 ±0.65	**70.0** ±1.19	**78.4** ±0.80	**65.7** ±1.78	**75.3** ±0.44

Table 2: Comparison of accuracy (%) for imbalanced SSL methods on CIFAR10-LT and STL10-LT under $\gamma_l \neq \gamma_u$ setup.

3 Experiments

3.1 Experimental Setup

Datasets. We conduct SSL experiments with various scenarios where the class distribution of unlabeled data can deviate from that of labeled data. We adopt CIFAR-10/100 [15] and STL-10 [5] typically adopted in SSL literature [19]. We make the imbalanced versions by exponentially decreasing the amount of samples per class following [7,14]. We also consider Semi-Aves [20], which is the large-scale collection of bird species with natural long-tailed distribution.

Baseline methods. We mainly adopt *FixMatch* [19] as baseline and consider *DARP* [14] and *CReST* [22] for comparison.

Training and evaluation. We train Wide ResNet-28-2 [23] on CIFAR10/100-LT and STL10-LT . For Semi-Aves, we fine-tune ResNet-34 [12] pre-trained on ImageNet [8]. For evaluation, we measure the top-1 accuracy.

3.2 Results on CIFAR10/100-LT and STL10-LT.

In case of $\gamma_l = \gamma_u$. We compare imbalanced SSL methods: DARP [14] and CReST+ [22] with the proposed DASO on FixMatch. Remarkably, DASO shows comparable or even better results in most setups with significant gains compared to baseline FixMatch, although DARP and CReST+ even push the predictions of unlabeled data to the label distribution using the assumption $\gamma_l = \gamma_u$ (*i.e.*, distribution alignment [2]). This verifies the efficacy of DASO for debiasing, even without resorting to the label distribution.

In case of $\gamma_l \neq \gamma_u$. For CIFAR10-LT, we consider two extreme cases for the class distribution of unlabeled data: uniform ($\gamma_u = 1$) and flipped long-tail ($\gamma_u = 1/100$) with respect to the labeled data. For STL10-LT, since we cannot control

Benchmark	Semi-Aves			
	$\mathcal{U}=\mathcal{U}_{\text{in}}$		$\mathcal{U}=\mathcal{U}_{\text{in}}+\mathcal{U}_{\text{out}}$	
Method	Last Top1	Med20 Top1	Last Top1	Med20 Top1
FixMatch [19]	53.8 ±0.17	53.8 ±0.13	45.7 ±0.89	46.1 ±0.50
w/ DARP [14]	52.3 ±0.48	52.1 ±0.48	46.3 ±0.70	46.4 ±0.61
w/ CReST [22]	52.1 ±0.36	52.2 ±0.27	43.6 ±0.69	43.6 ±0.68
w/ CReST+ [22]	53.9 ±0.38	53.8 ±0.38	45.1 ±1.09	45.2 ±1.00
w/ DASO (Ours)	**54.5** ±0.08	**54.6** ±0.12	**47.9** ±0.41	**47.9** ±0.38

Table 3: Accuracy on Semi-Aves [20]. DASO shows the best among imbalanced SSL methods. DASO also performs well in presence of large $\mathcal{U}_{\text{out}}$.

	$\mathcal{L}_{\text{align}}$	C10	STL10
FixMatch	✗	68.25	55.53
DASO	✗	70.98	61.64
FixMatch	✓	73.15	58.51
DASO	✓	**75.97**	**70.21**

Table 4: Ablation study on blending and the semantic alignment loss.

	C10	STL10
$\upsilon_k = 0$	73.15	58.51
$\upsilon_k = 1$	72.35	62.60
$\upsilon_k = 0.5$	72.96	64.21
DASO	**75.97**	**70.21**

Table 5: Ablation study on pseudo-label blending strategy.

the size and imbalance of unlabeled data due to unknown labels, we instead set $\gamma_l \in \{10, 20\}$ with the whole fixed unlabeled data. Table 2 summarizes the results of imbalanced SSL methods under the setups.

Surprisingly, DASO outperforms other baselines by significant margins in most cases. Though DARP [14] estimates the distribution of unlabeled data in advance as prior, the estimation accuracy decreases as using less labels for training. Under $\gamma_l \neq \gamma_u$, we evaluate both CReST and CReST+ [22]. Clearly, resorting to the label distributions as the prior for unlabeled data in CReST+ rather harms the accuracy since the assumption of $\gamma_l = \gamma_u$ is violated. The accuracy loss becomes more severe under $\gamma_u = 1/100$.

By virtue of debiased pseudo-labels from DASO, the abundant minority-class unlabeled samples are correctly used. Consequently, the results confirm that conditioning on a certain distribution for unlabeled data (*e.g.*, $\gamma_u = \gamma_l$) is undesirable in imbalanced SSL, and DASO greatly reduces the bias in presence of distribution mismatch, even without access to the distribution.

3.3 Results on Large-Scale Semi-Aves

We test DASO on a realistic Semi-Aves benchmark [20]. Both labeled data ($\mathcal{X}$) and unlabeled data ($\mathcal{U}$) show long-tailed distributions, while $\mathcal{U}$ contains large *open-set* examples ($\mathcal{U}_{\text{out}}$) that do not belong to any of the classes in $\mathcal{X}$. The results are shown in Table 3. We report both cases: $\mathcal{U} = \mathcal{U}_{\text{in}}$ and $\mathcal{U} = \mathcal{U}_{\text{in}} + \mathcal{U}_{\text{out}}$, where $\mathcal{U}_{\text{in}}$ contains examples that share the class of $\mathcal{X}$.

In case of $\mathcal{U} = \mathcal{U}_{\text{in}}$. As it has the distribution gap between $\mathcal{X}$ and $\mathcal{U}$, DARP [14] and CReST [22] show only a slight gain or even unsatisfactory performances compared to FixMatch [19]. In contrary, DASO shows the best performance among the baselines with favorable improvements upon FixMatch.

In case of $\mathcal{U} = \mathcal{U}_{\text{in}} + \mathcal{U}_{\text{out}}$. Since $\mathcal{U}$ contains large amount of *open-set* class examples, performance drop is observed consistently across all baselines. Among them, DASO shows the best performance with favorable gain. DARP [14] is slightly helpful for optimization. Concerning CReST and CReST+ [22], they rather performs poorly than FixMatch due to noisy predictions from $\mathcal{U}_{\text{out}}$. As such, DASO has superiority in the challenging but practical scenario of long-tailed distributions, even in presence of large amount of open-set examples.

3.4 Ablation Study

Component analysis. Table 4 studies the two major components of DASO: distribution-aware pseudo-label blending and the semantic alignment loss. Both blending mechanism and $\mathcal{L}_{\text{align}}$ provides significant gain over FixMatch. For example, the blending and $\mathcal{L}_{\text{align}}$ achieve about 6% and 3% absolute gain, respectively, and combining both shows 15.7% gain in total on STL10.

Effect of pseudo-label blending. Table 5 studies the different way of pseudo-label blending on DASO with *constant* weights. Due to the bias in the pseudo-labels, using either linear ($\upsilon_k = 0$) or semantic ($\upsilon_k = 1$) pseudo-label leads to a marginal gain. In addition, blending them with the same ratio ($\upsilon_k = 0.5$) shows the lower performance compared to our final DASO, which demonstrates that distribution-aware class-adaptive blending is crucial for imbalanced SSL.

4 Conclusion

We proposed a novel distribution-aware semantics-oriented (DASO) pseudo-label for imbalanced semi-supervised learning. DASO adaptively blends the linear and semantic pseudo-labels within each class to mitigate the overall bias across the class. Moreover, we introduced semantic alignment loss. From extensive experiments, we showed the efficacy of DASO on challenging and realistic setups, especially when class imbalance and class distribution mismatch dominate.

Remarks This paper is a re-publishing (summary presentation) of the paper which has been published in *2022 IEEE CVF Computer Vision and Pattern Recognition Conference (CVPR)* by request of the IW-FCV2023 program committee to share the research results.

References

1. Eric Arazo, Diego Ortego, Paul Albert, Noel E O'Connor, and Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *International Joint Conference on Neural Networks (IJCNN)*, 2020. 1
2. David Berthelot, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. Remixmatch: Semi-supervised learning with distribution matching and augmentation anchoring. In *International Conference on Learning Representations (ICLR)*, 2020. 1, 5
3. David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2019. 1
4. Olivier Chapelle, Bernhard Scholkopf, and Alexander Zien. Semi-supervised learning. *IEEE Transactions on Neural Networks*, 2009. 1
5. Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2011. 2, 5
6. Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Advances in Neural Information Processing Systems (NIPS)*, 2020. 3

7. Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 5
8. Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 5
9. Qi Dong, Shaogang Gong, and Xiatian Zhu. Imbalanced deep learning by minority class incremental rectification. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2018. 1
10. Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In *Advances in Neural Information Processing Systems (NIPS)*, 2005. 2
11. Tao Han, Junyu Gao, Yuan Yuan, and Qi Wang. Unsupervised semantic aggregation and deformable template matching for semi-supervised learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2020. 1, 3, 4
12. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 5
13. Dong-Jin Kim, Jinsoo Choi, Tae-Hyun Oh, and In So Kweon. Image captioning with very scarce supervised data: Adversarial semi-supervised learning approach. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019. 1
14. Jaehyung Kim, Youngbum Hur, Sejun Park, Eunho Yang, Sung Ju Hwang, and Jinwoo Shin. Distribution aligning refinery of pseudo-label for imbalanced semi-supervised learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2020. 5, 6
15. Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. *Technical report*, 2009. 2, 5
16. Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. In *International Conference on Learning Representations (ICLR)*, 2016. 2
17. Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, 2013. 1, 2
18. Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2017. 2
19. Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *Advances in Neural Information Processing Systems (NIPS)*, 2020. 1, 2, 3, 5, 6
20. Jong-Chyi Su and Subhransu Maji. The semi-supervised inaturalist-aves challenge at fgvc7 workshop, 2021. 2, 5, 6
21. Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems (NIPS)*, 2017. 2
22. Chen Wei, Kihyuk Sohn, Clayton Mellina, Alan Yuille, and Fan Yang. Crest: A class-rebalancing self-training framework for imbalanced semi-supervised learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 5, 6
23. Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *British Machine Vision Conference (BMVC)*, 2016. 5