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

Youngtaek Oh[1] Dong-Jin Kim[2] In So Kweon[1]
[1]KAIST, South Korea. [2]UC Berkeley / ICSI, CA.
[1]{youngtaek.oh, iskweon}@kaist.ac.kr [2]djkim93@berkeley.edu

Abstract

*The capability of the traditional semi-supervised learning (SSL) methods 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 from the classifier. 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: CIFAR10/100-LT, STL10-LT, and large-scale long-tailed Semi-Aves with open-set class, and demonstrate that, the proposed DASO framework reliably improves SSL learners with unlabeled data especially when both (1) class imbalance and (2) distribution mismatch dominate.*

1. Introduction

Semi-supervised learning (SSL) [7] has shown to be promising for leveraging unlabeled data to reduce the cost of constructing labeled data [4, 5, 36, 40, 57] and even boost the performance at scale [29, 49, 67, 68]. The common approach of these algorithms is to produce *pseudo-labels* for unlabeled data based on model's predictions and utilize them for regularizing model training [29, 38, 57]. Although adopted in a variety of tasks, these algorithms often assume class-balanced data, while many real-world datasets exhibit *long-tailed* distributions [3, 18, 31, 32]. With class-imbalanced data, the class distribution of pseudo-labels from unlabeled data becomes severely biased to the majority classes due to confirmation bias [2]. Such biased pseudo-labels can further bias the model during training.

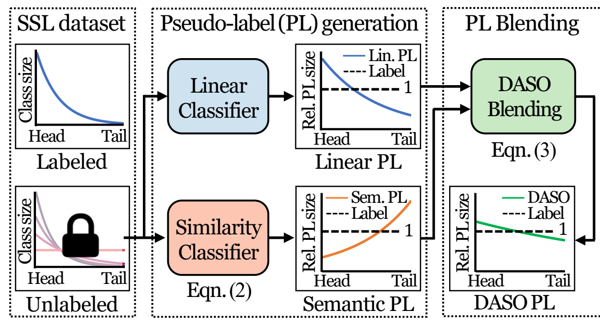


Figure 1. Glimpse of the DASO framework. DASO reduces the overall bias in pseudo-labels (PL) from unlabeled data by blending two complementary PLs from different classifiers. Note that bias is conceptually illustrated as relative PL size (Rel. PL size), meaning that pseudo-label size is normalized by actual label size.

Many methods of handling class-imbalanced labels have been proposed in the supervised learning community, but little interest has been made in re-balancing pseudo-labels in SSL. Recent studies have explored this *imbalanced SSL* setting, where as a reference to the class distribution of unlabeled data, it is often assumed that it is the same as the class distribution of labels [33, 64], or a separate distribution estimate is required [33]. However, the actual class distribution of unlabeled data is unknown without the labels. For example, unlabeled data may have large class distribution gap from labeled data, including many samples in novel classes not defined in the label set [58]. As we elaborate in Sec. 4, the bias of pseudo-labels also depends on such class distribution mismatch between labeled and unlabeled data, and using inaccurate estimates or wrong assumptions about the unlabeled data cannot be helpful under imbalanced SSL.

In this work, we present a new imbalanced SSL method specifically tailored for alleviating the bias in pseudo-labels under class-imbalanced data, 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 [22] obtained from a similarity-based classifier [56] are biased towards minority classes as opposed to linear classifier-based pseudo-labels [38, 57] being biased towards head classes. As illustrated in Sec. 3.2, 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 in Sec. 3.3. Building upon the existing SSL learner, we propose to blend the linear and semantic pseudo-labels in different proportions for each class to reduce the overall bias. This blending strategy can provide a more balanced supervision than simply using either of the pseudo-label. The primary novelty comes from the scheduling of the weights for mixing the pseudo-labels. Specifically, we dynamically adjust the relative weights of semantic pseudo-labels to be blended so that linear pseudo-labels are less biased according to the current class distribution of pseudo-labels. By virtue of such mechanism, without resorting to any class priors for the unlabeled data, DASO reliably brings performance gain even with substantial class distribution mismatch between labeled and unlabeled data.

We further propose a simple yet effective semantic alignment loss to establish balanced feature representation via *balanced* class prototypes, which is the extension of the consistency regularization framework in [57, 66] onto feature space. We align the unlabeled data onto each of the similar prototypes, by consistently assigning 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. We validate the semantic alignment loss is useful under imbalanced SSL, especially helpful for DASO.

The efficacy of DASO is extensively justified with the imbalanced versions of benchmarks: CIFAR-10/100 [35] and STL-10 [12] in Sec. 4. We even test DASO with large-scale long-tailed Semi-Aves [58] with open-set classes in unlabeled data, 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 that works well on top of diverse frameworks such as existing SSL learners and even other re-balancing frameworks for labels and SSL.

The key contributions in our work can be summarized as follows: (1) We propose a novel pseudo-labeling framework, DASO, for debiasing pseudo-labels by class-adaptively blending two complementary types of pseudo-labels observing current class distribution of pseudo-labels. (2) DASO introduces semantic alignment loss to further alleviate the bias from high-quality feature representation, by aligning each unlabeled example to the similar prototype. (3) DASO readily integrates with other frameworks to show significant performance improvements under diverse imbalanced SSL setup, including the most practical scenario.

2. Related Work

Class-imbalanced learning. Datasets that well capture the dynamic nature of *real-world* exhibit *class-imbalanced*, or *long-tailed* distributions [21,61]. Learning on such datasets has been a great challenge to deep neural networks, since they cannot generalize well to the rare classes [3]. Conventional approaches to combat the imbalance include data re-sampling [1,8,34], cost-sensitive re-weighting [6,14,47], and decoupling the representation and the classifier [27,71]. Recently, learning expert models across classes [62, 65] and re-balancing with the data distribution in loss computation phase [25, 43, 51] are also shown to be effective. On the other hand, [42, 69] leveraged unlabeled data for class-imbalanced learning. Unlike all the aforementioned methods, we focus on alleviating the bias of pseudo-labels in semi-supervised learning due to class imbalanced labels and distribution mismatch between labeled and unlabeled data.

Semi-supervised learning (SSL). SSL aims to learn from both labeled and unlabeled data. For unlabeled data, SSL generates targets (*e.g.*, pseudo-labels) from model predictions via *pseudo-labeling* [29, 38], *consistency regularization* [44, 59], and combinations of them [4, 5, 30, 36] under *cluster assumption* [7]. However, pseudo-labels can be biased with class-imbalanced data [33], which harm the model when utilized. Some works deal with such issue via loss re-weighting [26, 29, 39], optimization [33], data re-sampling [64], and meta-learning sample importance [52,53]. However, class distribution of unlabeled data either unknown or different from the labeled one can also exacerbate the bias, limiting the applicability of such methods. In this aspect, we devise a new pseudo-labeling method that handles such challenging but practical scenarios.

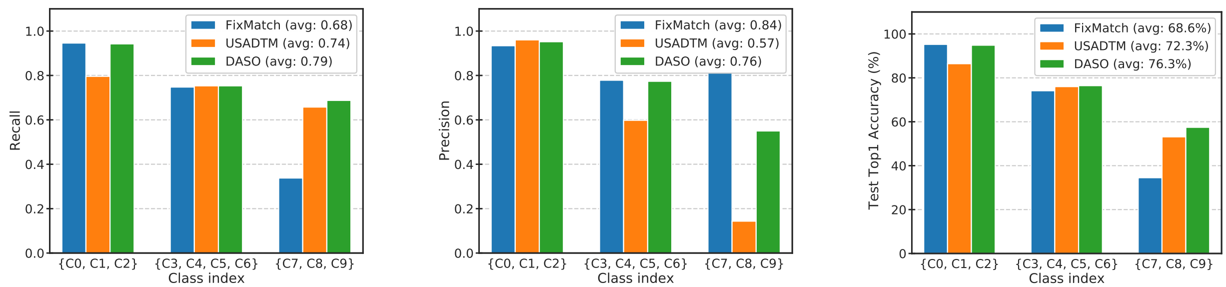
3. Proposed Method

3.1. Preliminaries

Problem setup. We consider K-class semi-supervised image classification 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[1] $\hat{p} \in \mathbb{R}^K$ is assigned to learn the unsupervised loss $\mathcal{L}_u = \Phi_u(\hat{p}, f(u))$, where Φ_u can be implemented via entropy [19] or consistency regularization [37, 59], depending on the SSL learner.

For FixMatch [57] as an example, the pseudo-label $\hat{p} =$ OneHot $\left(\text{argmax}_k\, p_k^{(w)}\right)$ with $p^{(w)} = f(\mathcal{A}_w(u))$ provides

[1]In this work, we assume it includes both one-hot form and soft form cases: $\Sigma_k \hat{p}_k = 1$ where $\hat{p}_k \in [0, 1]$.



(a) Recall of pseudo-labels.

(b) Precision of pseudo-labels.

(c) Class-wise test accuracy.

Figure 2. Analysis on recall and precision of pseudo-labels and the corresponding test accuracy. Note that the class index from x-axis is sorted by the class size; `C0` and `C9` are the head and tail classes, respectively. Although USADTM [22] 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, which leads to higher test accuracy of those classes. More analyses with various SSL methods are provided in Appendix E.1.

the target for the prediction $p^{(s)} = f\left(\mathcal{A}_s(u)\right)$ with some confident ones to the cross-entropy loss $\mathcal{H}$ as follows:

$$\Phi_u(\hat{p},\, p^{(s)}) = \mathbb{1}\left(\max_k p_k^{(w)} \geq \tau\right) \mathcal{H}\left(\hat{p},\, p^{(s)}\right), \qquad (1)$$

where $\mathcal{A}_w$ and $\mathcal{A}_s$ correspond to weak augmentation (*e.g.*, random flip and crop) and advanced augmentation (*e.g.*, RandAugment [13] followed by Cutout [17]), 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 $\gamma_l = \frac{\max_k N_k}{\min_k N_k} \gg 1$ under imbalanced SSL. γ_u is specified in the same way using the actual labels without access during training. It is worth noting that the class distribution of $\mathcal{U}$ (*e.g.*, γ_u) may be either similar to $\mathcal{X}$, or significantly divergent in practice, and such varying distributions greatly affect the SSL performances with the same $\mathcal{X}$ as shown in Table 3. In this regard, our goal is to produce debiased pseudo-labels with class-imbalanced data, while maintaining the performances of SSL algorithms with various, but still *unknown* class distribution of unlabeled data.

3.2. Motivation

Linear and semantic pseudo-label. Pseudo-labeling based on linear classifier (*i.e.*, fc layer), which has been widely adopted by pseudo-label-based algorithms [10,30–32] especially for SSL [38,57], can produce biased pseudo-labels towards majority classes with class-imbalanced data. We abbreviate this type of pseudo-labels as linear pseudo-labels. Instead, pseudo-labels can be obtained from similarity-based classifier [15, 54] by measuring the similarity of a given representation (*e.g.*, prototypes [56]) to an unlabeled sample in feature space, which we call simply *semantic pseudo-labels*. As note, similarity-based classifier has been widely adopted for reducing biased predictions [27,41,50]. In SSL, USADTM [22] utilizes semantic pseudo-labeling method. As following, we conduct a simple experiment to explore each aspect of linear and semantic pseudo-labels.

Trade-offs between linear and semantic pseudo-label. As shown in Fig. 2, we compare FixMatch [57] and USADTM [22] using linear and semantic pseudo-label respectively, under *imbalanced* SSL setup. From Figs. 2a and 2b, the linear pseudo-labels from FixMatch achieve 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 semantic pseudo-labels from 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. Comparing the test accuracy from Fig. 2c, USADTM shows relatively increased overall test accuracy compared to FixMatch by virtue of more abundant minority pseudo-labels, while losing the accuracy on the head. In other words, the overall increase in accuracy is limited when only using semantic pseudo-labels. We provide two lessons from the simple experiment in Fig. 2, as summarized by:

1. Semantic pseudo-labels are *reversely biased towards the tail side*, which lead to the limited accuracy gain.

2. The linear and semantic pseudo-labels have the *complementary properties* useful for reducing the overall bias.

These empirical findings motivate us to exploit the linear and semantic pseudo-labels *differently* in different classes for debiasing. For example, as the linear pseudo-label for a sample u points to the majorities, more semantic pseudo-label component should contribute to the final pseudo-label to prevent the false positives towards the head, and the vice versa when the linear pseudo-label predicts u as minority.

We also present the result of our solution, DASO, in Fig. 2, where the recall of the final pseudo-label has increased but the overall pseudo-labels are still not biased towards the minority classes, unlike USADTM. Thanks to such unbiased pseudo-labels between the head and tail classes obtained by properly blending two pseudo-labels, the overall test accuracy also increased a lot from Fig. 2c.

3.3. DASO Pseudo-label Framework

We propose DASO, a generic framework for imbalanced SSL with two novel contributions as (1) distribution-aware blending for the linear and semantic pseudo-labels and (2) semantic alignment loss, which are described as follows.

Framework overview. Without loss of generality, we consider DASO built on top of FixMatch [57] for convenience in notations, while DASO can easily integrate with other SSL learners as shown in Tables 1 and 3. 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 [22]. 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.

The prototype representation can also be imbalanced using class-imbalanced labeled data. To prevent such biased prototypes, we additionally propose *balancing* the prototypes compared to [22] in two ways. First, instead of the size of Q_k in proportional to the class frequency, we fix the size of Q_k for all k to the same amount as L. By averaging the same number of features from each class, we can compensate for the prototypes especially for the minority classes, with earlier samples remaining in Q_k. Secondly, we adopt momentum encoder $f_{\theta'}^{\text{enc}}$ when extracting the features for prototype generation inspired by [23]. Note that $f_{\theta'}^{\text{enc}}$ has the same architecture with f_θ^{enc}, but θ' is the exponential moving average (EMA) of θ with momentum ratio ρ, *i.e.*, $\theta' \leftarrow \rho\theta' + (1-\rho)\theta$. This stabilizes the movement of each prototype in feature space across iteration by slowing the pace of network parameter updates. We will verify the effectiveness of balanced prototypes in Table 7.

Linear and semantic pseudo-label generation. We obtain the linear pseudo-label $\hat{p}$ using the linear classifier followed by softmax activation: $\hat{p} = \sigma(f_\phi^{\text{cls}}(z^{(w)}))$. The semantic pseudo-label $\hat{q}$ is obtained from the similarity-based classifier that measures the per-class similarity of a query feature point z of either $z^{(w)}$ or $z^{(s)}$ to the *balanced prototypes* $\mathbf{C}$:

$$q = \sigma\left(\text{sim}(z, \mathbf{C}) \,/\, T_{\text{proto}}\right), \tag{2}$$

where $\text{sim}(\cdot, \cdot)$ corresponds to cosine similarity, and T_{proto} is a temperature hyper-parameter for the classifier. Note that $\hat{p}$ is biased towards head classes while $\hat{q}$ is the vice versa.

Distribution-aware blending. To obtain class-specific unbiased pseudo-label $\hat{p}'$, the semantic pseudo-label $\hat{q}$ should be exploited *differently* across the class. To this end, we propose a novel blending method for pseudo-labels, where 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 $\hat{q}$:

$$\hat{p}' = (1 - \upsilon_{k'})\,\hat{p} + \upsilon_{k'}\hat{q}, \tag{3}$$

where k' is the class prediction from $\hat{p}$, and each υ_k is derived as $\upsilon_k = \frac{1}{\max_k \hat{m}_k^{1/T_{\text{dist}}}} \left(\hat{m}_k^{1/T_{\text{dist}}}\right)$. Note that $\hat{m}$ is the normalized class distribution of the current pseudo-labels, which is the accumulation of $\hat{p}'$ over a few previous iterations 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 $\upsilon_{k'}$.

Note that we dynamically adjust the set of weights υ that determines relative intensity of $\hat{q}$ in Eq. (3), based on the current bias of pseudo-labels $\hat{m}$. This makes DASO flexible to various distributions of $\mathcal{U}$ without resorting to any pre-defined distribution. For example, even under the same prediction of $\hat{p}$ for a head class, more $\hat{q}$ is blended when the current model is more biased. Similarly, a concurrent work [63] accumulates predictions for adaptive debiasing.

Semantic alignment loss. To establish more balanced feature representations, we propose new semantic alignment loss for regularizing the feature encoder f_θ^{enc}. It extends the consistency training framework with two asymmetric augmentations $\mathcal{A}_w$ and $\mathcal{A}_s$ like [57, 66] onto feature space. In high-level, we align each unlabeled sample u to the most similar prototype used in the similarity-based 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 $\hat{q}$ is reused to provide the target for $q^{(s)}$ with the cross-entropy loss $\mathcal{H}$:

$$\mathcal{L}_{\text{align}} = \mathcal{H}\left(\hat{q},\, q^{(s)}\right), \tag{4}$$

where $q^{(s)}$ is from the similarity-based classifier by passing through $z^{(s)} = f_\theta^{\text{enc}}(\mathcal{A}_s(u))$ to Eq. (2). Since $\mathcal{L}_{\text{align}}$ relates unlabeled data to the label space through consistently assigning to $\mathbf{C}$ constructed from labeled features, such enhanced representation can implicitly guide the classifier f_ϕ^{cls}

to produce less biased predictions in general, where we validate the efficacy of $\mathcal{L}_{\text{align}}$ in Secs. 4.4 and 4.5 respectively.

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

$$\mathcal{L}_{\text{DASO}} = \mathcal{L}_{\text{cls}} + \lambda_u \mathcal{L}_u + \lambda_{\text{align}} \mathcal{L}_{\text{align}}, \tag{5}$$

where both $\mathcal{L}_{\text{cls}}$ and $\mathcal{L}_u$ with λ_u come from the base SSL learner, and $\mathcal{L}_{\text{align}}$ is newly introduced from DASO. Note that $\mathcal{L}_u$ takes the proposed blended pseudo-label in Eq. (3) instead of the original linear pseudo-label of the learner. We emphasize that DASO is also applicable to traditional SSL algorithms for performance gain without $\mathcal{L}_{\text{align}}$ due to the absence of $\mathcal{A}_s$ in the algorithm, as validated in Table 3.

4. Experiments

4.1. Experimental Setup

To ensure reproducibility[2], all the settings of DASO and other baseline methods are clarified in Appendix C.3.

Datasets. We conduct SSL experiments with various scenarios where the class distribution of unlabeled data is not just limited to the class distribution of labeled data. To accommodate such conditions, we adopt CIFAR-10/100 [35] and STL-10 [12] typically adopted in SSL literature [57]. We make the imbalanced versions by exponentially decreasing the amount of samples per class [14]. Following [33], we denote the head class size as N_1 (M_1), and the imbalance ratio as γ_l (γ_u) for the labeled (unlabeled) data respectively. Note that γ_l and γ_u can vary independently, and we specify 'LT' for those imbalanced variants. We also consider Semi-Aves benchmark [58] for practical setup, which is the large-scale collection of bird species with natural long-tailed distribution. Its unlabeled data also show long-tailed distribution, and include large portion of examples in broader categories compared to samples in labeled data (*e.g.*, *open-set*). For more details, see Appendix C.1.

Baseline methods. We consider *Supervised* baseline, learning cross-entropy with only labeled data. For using unlabeled data, we mainly adopt *FixMatch* [57] for its simplicity and powerful performances. To extensively validate our proposed method in terms of re-balancing, we mainly compare it with the following re-balancing algorithms on top of FixMatch. Note that the results with other baseline SSL algorithms are provided in Table 3 and the Appendix D.3. We consider *logit adjustment* (LA) [43] for balancing labels. Note that LA can also be applied to SSL methods for re-balancing using labels. For re-balancing in unlabeled data similar to our framework, *DARP* [33] and *CReST* [64] are compared. We also experiment with the recently proposed *ABC* [39] that performs single unified re-balancing using both labeled and unlabeled data simultaneously.

[2] Code is available at: https://github.com/ytaek-oh/daso.

Training and evaluation. We have re-implemented all the baseline methods using PyTorch [48] and conducted experiments under the same codebase for fair comparison, as suggested by [45]. We train Wide ResNet-28-2 [70] on CIFAR10/100-LT and STL10-LT as a backbone. For training Semi-Aves, we fine-tune the ResNet-34 [24] pre-trained on ImageNet [16]. To evaluate, we use the EMA network with the parameters updating every steps, following [5, 33]. As note, the class score is measured via learned linear classifier at inference time. We measure the top-1 accuracy on the test data every epoch and finally obtain the median of the accuracy values during the last 20 evaluations [5]. When reporting the results, we compute the mean and standard deviation of three independent runs.

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

As the main results, we first consider the case when the distribution of labeled data and unlabeled data is the same (*e.g.*, $\gamma = \gamma_l = \gamma_u$) in Table 1, which is the ideal case for SSL. In Table 2, we relax such assumption and test imbalanced SSL methods under practical yet challenging scenarios with diverse unlabeled data distributions (*e.g.*, $\gamma_l \neq \gamma_u$).

In case of $\gamma_l = \gamma_u$. We compare the proposed DASO with several baseline methods, with or without class re-balancing in Table 1. For *Supervised* case, even if Logit Adjustment (LA) [43] is applied, the performances are rather limited compared to even naïve SSL method (*i.e.*, FixMatch [57]).

We then compare imbalanced SSL methods: DARP [33] and CReST+ [64] 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 [4]). This verifies the efficacy of DASO for debiasing pseudo-labels, even without resorting to the label distribution.

To validate DASO can reliably benefit from re-balancing labels for debiasing pseudo-labels, we further compare imbalanced SSL methods on label re-balancing FixMatch via LA [43] (noted as FixMatch + LA). The results show DASO performs the best in most of the setups. It is noticeable that LA with DASO always improves performances compared to both FixMatch w/ DASO and FixMatch + LA cases.

Finally, we consider ABC [39] in the bottom of Table 1. It jointly trains the SSL learner and the auxiliary balanced classifier (ABC) using both labeled and unlabeled data with *linear pseudo-labels*, while the ABC is opted for evaluation. We find that training ABC can readily be extended by just replacing the *linear pseudo-label* for ABC with DASO pseudo-label (3). Finally, DASO can be significantly pushed by combining with ABC [39] (*i.e.*, 13% gain upon FixMatch for CIFAR-10). It verifies the flexibility of DASO on any baselines regardless of re-balancing methods.

Algorithm	CIFAR10-LT $\gamma=\gamma_l=\gamma_u=100$ $N_1=500$ $M_1=4000$	$N_1=1500$ $M_1=3000$	$\gamma=\gamma_l=\gamma_u=150$ $N_1=500$ $M_1=4000$	$N_1=1500$ $M_1=3000$	CIFAR100-LT $\gamma=\gamma_l=\gamma_u=10$ $N_1=50$ $M_1=400$	$N_1=150$ $M_1=300$	$\gamma=\gamma_l=\gamma_u=20$ $N_1=50$ $M_1=400$	$N_1=150$ $M_1=300$
Supervised	47.3 ±0.95	61.9 ±0.41	44.2 ±0.33	58.2 ±0.29	29.6 ±0.57	46.9 ±0.22	25.1 ±1.14	41.2 ±0.15
w/ LA [43]	53.3 ±0.44	70.6 ±0.21	49.5 ±0.40	67.1 ±0.78	30.2 ±0.44	48.7 ±0.89	26.5 ±1.31	44.1 ±0.42
FixMatch [57]	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 [33]	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+ [64]	**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
FixMatch + LA [43]	75.3 ±2.45	82.0 ±0.36	67.0 ±2.49	78.0 ±0.91	47.3 ±0.42	58.6 ±0.36	41.4 ±0.93	53.4 ±0.32
w/ DARP [33]	76.6 ±0.92	80.8 ±0.62	68.2 ±0.94	76.7 ±1.13	50.5 ±0.78	59.9 ±0.32	**44.4** ±0.65	53.8 ±0.43
w/ CReST+ [64]	76.7 ±1.13	81.1 ±0.57	**70.9** ±1.18	77.9 ±0.71	44.0 ±0.21	57.1 ±0.55	40.6 ±0.55	52.3 ±0.20
w/ DASO (Ours)	**77.9** ±0.88	**82.5** ±0.08	70.1 ±1.68	**79.0** ±2.23	**50.7** ±0.51	**60.6** ±0.71	44.1 ±0.61	**55.1** ±0.72
FixMatch + ABC [39]	78.9 ±0.82	**83.8** ±0.36	66.5 ±0.78	80.1 ±0.45	47.5 ±0.18	59.1 ±0 .21	41.6 ±0.83	53.7 ±0.55
w/ DASO (Ours)	**80.1** ±1.16	83.4 ±0.31	**70.6** ±0.80	**80.4** ±0.56	**50.2** ±0.62	**60.0** ±0.32	**44.5** ±0.25	**55.3** ±0.53

Table 1. Comparison of accuracy (%) with combinations of re-balancing methods on CIFAR10/100-LT under $\gamma_l = \gamma_u$ setup. Our DASO consistently improves the performance over all the baselines without or with re-balancing, even with ABC [39] designed for imbalanced SSL. We indicate the best results for each division as bold. More results including new baseline methods are provided in Appendix D.1.

Algorithm	CIFAR10-LT ($\gamma_l \neq \gamma_u$) $\gamma_u=1$ (uniform) $N_1=500$ $M_1=4000$	$N_1=1500$ $M_1=3000$	$\gamma_u=1/100$ (reversed) $N_1=500$ $M_1=4000$	$N_1=1500$ $M_1=3000$	STL10-LT ($\gamma_u=$ *N/A*) $\gamma_l=10$ $N_1=150$ $M=100k$	$N_1=450$ $M=100k$	$\gamma_l=20$ $N_1=150$ $M=100k$	$N_1=450$ $M=100k$
FixMatch [57]	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 [33]	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 [64]	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+ [64]	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. For CIFAR10-LT, γ_l is fixed to 100, and γ_u is unknown for STL10-LT. Our DASO consistently shows significant gains on FixMatch [57] without resorting to any class prior under diverse class distribution mismatches between labeled and unlabeled data. We indicate the best results as bold.

In case of $\gamma_l \neq \gamma_u$. The class distribution of unlabeled data could be either unknown or arguably different from that of the labeled data in real-world (*e.g.*, $\gamma_l \neq \gamma_u$). To simulate such scenarios, 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 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. Note that more comparisons of SSL methods with different re-balancing techniques (*i.e.*, LA [43] and ABC [39]) are presented in Appendix D.2.

Surprisingly, DASO outperforms other baselines by significant margins in most cases. For example, DASO shows 13.6% and 18.1% of absolute gain from FixMatch upon CIFAR-10 ($\gamma_u = 1$) and STL-10 ($\gamma_l = 20$), respectively. Though DARP [33] estimates the distribution of unlabeled data in advance as the prior, the estimation accuracy decreases as using less labels for training. Under $\gamma_l \neq \gamma_u$, we evaluate both CReST with self-training only and CReST+ with progressive distribution alignment [64]. Clearly, resorting to the label distributions as the prior for unlabeled data in CReST+ rather harms the accuracy compared to CReST, since the assumption of $\gamma_l = \gamma_u$ is violated. In particular, when the class distribution of unlabeled data is completely inverted ($\gamma_u = 1/100$), the accuracy loss becomes more severe, resulting in little gain over FixMatch.

By virtue of debiased pseudo-labels from DASO, the abundant minority-class unlabeled samples are correctly used despite class-imbalanced labels. 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.

DASO on other SSL learner. To verify DASO is a generic pseudo-labeling framework, we evaluate DASO based on other SSL algorithms including MeanTeacher [59], MixMatch [5], and ReMixMatch [4] in Table 3. As note, MeanTeacher and MixMatch only perform pseudo-label blending (3) without semantic alignment loss (4) due to the absence of $\mathcal{A}_s$. For CIFAR10-LT, we set $\gamma_l = 100$ and for CIFAR100-LT and STL10-LT, we set $\gamma_l = 10$. We observe that DASO greatly improves the performances for all the se-

Algorithm	C10-LT $N_1 = 1500$ $M_1 = 3000$ $\gamma_u = 100$	C10-LT $\gamma_u = 1$	C100-LT $N_1 = 150$ $M_1 = 300$ $\gamma_u = 10$	STL10-LT $N_1 = 450$ $M = 100k$ γ_u : *N/A*
Mean Teacher [59]	68.6 ±0.88	46.4 ±0.98	52.1 ±0.09	54.6 ±1.17
w/ DASO (Ours)	**70.7** ±0.59	**87.6** ±0.27	**52.5** ±0.37	**78.4** ±0.80
MixMatch [5]	65.7 ±0.23	35.7 ±0.69	54.2 ±0.47	52.7 ±1.42
w/ DASO (Ours)	**70.9** ±1.91	**73.4** ±2.05	**55.6** ±0.49	**68.4** ±0.71
ReMixMatch [4]	77.0 ±0.55	60.4 ±0.70	61.5 ±0.57	71.9 ±0.86
w/ DASO (Ours)	**80.2** ±0.68	**90.5** ±0.35	**62.1** ±0.69	**80.9** ±0.55

Table 3. Comparison of accuracy (%) from DASO upon other SSL methods: MeanTeacher [59], MixMatch [5], and ReMixMatch [4]. DASO improves the performances in all the setups.

Benchmark / Method	Semi-Aves $\mathcal{U} = \mathcal{U}_{\text{in}}$ Last Top1	$\mathcal{U} = \mathcal{U}_{\text{in}}$ Med20 Top1	$\mathcal{U} = \mathcal{U}_{\text{in}} + \mathcal{U}_{\text{out}}$ Last Top1	$\mathcal{U} = \mathcal{U}_{\text{in}} + \mathcal{U}_{\text{out}}$ Med20 Top1
Supervised	41.7 ±0.32	41.7 ±0.32	41.7 ±0.32	41.7 ±0.32
FixMatch [57]	53.8 ±0.17	53.8 ±0.13	45.7 ±0.89	46.1 ±0.50
w/ DARP [33]	52.3 ±0.48	52.1 ±0.48	46.3 ±0.70	46.4 ±0.61
w/ CReST [64]	52.1 ±0.36	52.2 ±0.27	43.6 ±0.69	43.6 ±0.68
w/ CReST+ [64]	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 4. Comparison of accuracy (%) on Semi-Aves benchmark [58]. DASO shows the best performance among state-of-the-art imbalanced SSL methods. Moreover, DASO still performs well in presence of massive open-set class examples $\mathcal{U}_{\text{out}}$.

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

Table 5. Ablation study on pseudo-label blending and semantic alignment loss $\mathcal{L}_{\text{align}}$.

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

Table 6. Ablation study on the pseudo-label blending strategy with $\mathcal{L}_{\text{align}}$ applied.

bal.	EMA	C10	STL10
✗	✗	74.98	68.54
✓	✗	74.54	70.01
✗	✓	75.01	69.49
✓	✓	**75.97**	**70.21**

Table 7. Ablation study on balancing prototypes and using EMA encoder on DASO.

	C10	STL10
$T_{\text{dist}} = 0.3$	73.97	**70.21**
$T_{\text{dist}} = 0.5$	74.47	68.35
$T_{\text{dist}} = 1.0$	74.82	65.96
$T_{\text{dist}} = 1.5$	**75.97**	64.54

Table 8. Ablation study on T_{dist} for DASO. We select T_{dist} by 1.5 and 0.3 each.

tups, and notably, it achieves 2.05$\times$ accuracy compared to MixMatch and brings 29.1% absolute gain in ReMixMatch on CIFAR10-LT under $\gamma_u = 1$. This implies that DASO noticeably helps SSL algorithms in general to benefit from unlabeled data under imbalanced SSL setup. As note, we show the comparison of imbalanced SSL methods built on other SSL learner (*e.g.*, ReMixMatch [4]) in Appendix D.3.

4.3. Results on Large-Scale Semi-Aves

We test DASO on a realistic Semi-Aves benchmark [58]. 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 4. 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 $\mathcal{X}$. We measure the performances by top-1 accuracy, reporting the one in the final (Last Top1) and the median values in last 20 epochs (Med20 Top1), following [45]. More details on this dataset can be found in Appendix C.1.

In case of $\mathcal{U} = \mathcal{U}_{\text{in}}$. As it has the distribution gap between $\mathcal{X}$ and $\mathcal{U}$, baseline DARP [33] and CReST [64] with inadequate class prior from $\mathcal{X}$ show only a slight gain or even unsatisfactory performances compared to FixMatch [57]. 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, as similar observations are made in [9, 20, 46]. Among the baselines, DASO shows the best performance with favorable gain. The results suggest that DARP [33] is slightly helpful when both $\mathcal{U}_{\text{in}}$ and $\mathcal{U}_{\text{out}}$ are considered altogether for optimization. Concerning CReST and CReST+ [64] with self-training, due to noisy predictions from $\mathcal{U}_{\text{out}}$ for constructing datasets for the next generation, they rather performs poorly than FixMatch. 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. To understand this, we further provide the analyses on the confidence plots with or without DASO using each of $\mathcal{U}_{\text{in}}$ and $\mathcal{U}_{\text{out}}$ in Appendix E.5.

4.4. Ablation Study

We conduct ablation studies to understand why DASO reliably provides improvements to baseline methods. To accommodate both $\gamma_l = \gamma_u$ and $\gamma_l \neq \gamma_u$ cases, we consider FixMatch on CIFAR10-LT with $N_1 = 500$, $\gamma = 100$ (noted as C10) and STL10-LT with $N_1 = 150$, $\gamma_l = 10$ (noted as STL10) respectively to evaluate each aspect of DASO.

Component analysis. Table 5 studies the two major components of DASO: distribution-aware pseudo-label blending and the semantic alignment loss. From the table, 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. The results confirm that both class-adaptively blending linear and semantic pseudo-labels and the semantic alignment loss are important for reducing bias under imbalanced SSL.

Effect of pseudo-label blending. Table 6 studies the different way of pseudo-label blending on DASO with *constant* weights. Due to the bias in the pseudo-labels, using either linear ($v_k = 0$) or semantic ($v_k = 1$) pseudo-label leads to

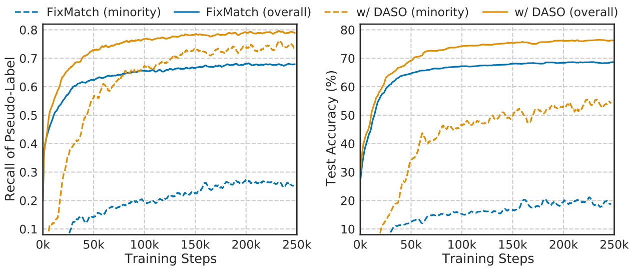


Figure 3. Train curves for the recall of pseudo-labels (left) and the test accuracy (right) on CIFAR10-LT. DASO significantly remedies the bias of pseudo-labels on minority classes, and such unbiased pseudo-labels lead to large gains on the test accuracy.

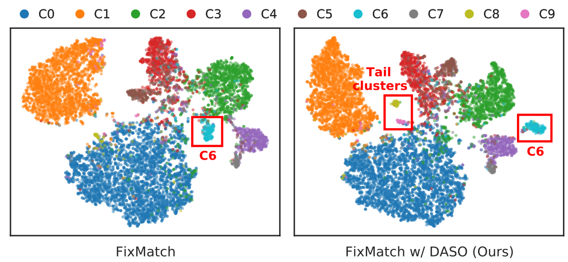


Figure 4. Comparison of t-SNE visualization of unlabeled data from FixMatch (left) and FixMatch w/ DASO (right). Learning with DASO helps the model to establish tail-class clusters in feature space, which can further reduce the biases from the classifier.

a marginal gain. In addition, blending them with the same ratio ($v_k = 0.5$) shows the lower performance compared to our final DASO, which demonstrates that distribution-aware class-adaptive blending is crucial for imbalanced SSL.

Effect of balanced prototype. Table 7 studies the different design choices of DASO in prototype generation: balanced prototypes (noted as bal.) with EMA encoder (noted as EMA). When generating class prototypes, using class-imbalanced queue without EMA encoder leads to worse performance. In contrary, DASO with both balanced queue using EMA encoder shows the best performance, showing that both correspond to the valid components for the balanced prototypes from imbalanced labeled data.

Ablation study on T_{dist}. In Table 8, we study the effect of the temperature hyper-parameter T_{dist} to compute the weights for pseudo-label blending described in Eq. (3). We empirically find that, for CIFAR-10 and STL-10, $T_{\text{dist}} = 1.5$ and $T_{\text{dist}} = 0.3$ show the best performance respectively.

4.5. Detailed Analysis

In this section, we qualitatively analyze how DASO improves the performance under imbalanced SSL setup. We consider FixMatch [57] without and with DASO trained on CIFAR10-LT with $\gamma = 100$ and $N_1 = 500$. Note that Appendix E includes analyses in more various setups.

Unbiased pseudo-label improves test accuracy. We visualize the train curves for the recall of pseudo-labels and the test accuracy values in Fig. 3. We denote those for the minorities (*e.g.*, last 20% classes) as dashed lines. From the left of Fig. 3, DASO significantly raises the final recall for the tail classes, which is 3$\times$ compared to that of FixMatch. From the right, both minority and overall test accuracy values in final greatly improved by virtue of the less biased pseudo-labels towards the head classes, which are nearly 3$\times$ and 9% compared to those of FixMatch, respectively.

Tail-class clusters are better identified. To verify the efficacy of reducing the bias, we present t-SNE [60] visualizations of the encoders' outputs on $\mathcal{U}$ from FixMatch and w/ DASO respectively. As shown in Fig. 4, tail class examples (*e.g.*, C8 and C9) from FixMatch are scattered to the majority classes. From the right, however, the clusters of tail are clearly recognized as indicated. In addition, the separability of C6 is improved. Thanks to such well identified tail-class clusters from DASO, the actual minority unlabeled examples are correctly leveraged to learn the unbiased model.

5. Discussion

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 balanced prototypes and semantic alignment loss. From extensive experiments, we showed the efficacy of DASO on various challenging and realistic setups, especially when class imbalance and class distribution mismatch dominate.

Potential societal impact. The proposed solution can contribute to solving various social problems attributed to imbalance in real-world, such as gender, racial or religious bias, by improving the fairness of classifiers using unlabeled data. Also, our method can contribute to the active learning research [11, 28, 55], which can also suffer from the bias. However, the proposed algorithm should be carefully considered as it can be used to raise other fairness issues such as over-balance or discrimination against minorities.

Limitations. DASO modulates the debiased pseudo-labels by introducing a hyper-parameter T_{dist}, which is effective and efficient than estimating the class distribution of unlabeled data. However, T_{dist} can be highly dependent on each data and distribution. As mentioned in [45], tuning such hyper-parameter is not straightforward under label-scarce setting, which is the common concern in SSL literature.

Acknowledgements

This research was supported by the National Research Foundation of Korea (NRF)'s program of developing and demonstrating innovative products based on public demand funded by the Korean government (Ministry of Science and ICT (MSIT)) (No. NRF-2021M3E8A2100445).

References

[1] Shin Ando and Chun Yuan Huang. Deep over-sampling framework for classifying imbalanced data. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 770–785, 2017. 2

[2] 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)*, pages 1–8, 2020. 1

[3] Samy Bengio. Sharing representations for long tail computer vision problems. In *ACM International Conference on Multimodal Interaction*, pages 1–1, 2015. 1, 2

[4] 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, 2, 5, 6, 7

[5] 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)*, volume 32, pages 5049–5059, 2019. 1, 2, 5, 6, 7

[6] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems (NIPS)*, volume 32, pages 1567–1578, 2019. 2

[7] Olivier Chapelle, Bernhard Scholkopf, and Alexander Zien. Semi-supervised learning. *IEEE Transactions on Neural Networks*, 20(3):542–542, 2009. 1, 2

[8] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002. 2

[9] Yanbei Chen, Xiatian Zhu, Wei Li, and Shaogang Gong. Semi-supervised learning under class distribution mismatch. In *AAAI Conference on Artificial Intelligence (AAAI)*, volume 34, pages 3569–3576, 2020. 7

[10] Jae Won Cho, Dong-Jin Kim, Jinsoo Choi, Yunjae Jung, and In So Kweon. Dealing with missing modalities in the visual question answer-difference prediction task through knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 3

[11] Jae Won Cho, Dong-Jin Kim, Yunjae Jung, and In So Kweon. Mcdal: Maximum classifier discrepancy for active learning. *IEEE transactions on neural networks and learning systems*, 2022. 8

[12] 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)*, volume 15, pages 215–223, 2011. 2, 5

[13] 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

[14] 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)*, pages 9268–9277, 2019. 2, 5

[15] Piew Datta and Dennis Kibler. Symbolic nearest mean classifiers. In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 82–87, 1997. 3

[16] 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)*, pages 248–255, 2009. 5

[17] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. 3

[18] Qi Dong, Shaogang Gong, and Xiatian Zhu. Imbalanced deep learning by minority class incremental rectification. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 41(6):1367–1381, 2018. 1

[19] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In *Advances in Neural Information Processing Systems (NIPS)*, volume 17, pages 281–296, 2005. 2

[20] Lan-Zhe Guo, Zhen-Yu Zhang, Yuan Jiang, Yu-Feng Li, and Zhi-Hua Zhou. Safe deep semi-supervised learning for unseen-class unlabeled data. In *International Conference on Machine Learning (ICML)*, pages 3897–3906, 2020. 7

[21] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5356–5364, 2019. 2

[22] 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)*, volume 33, pages 9972–9982, 2020. 1, 3, 4

[23] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9729–9738, 2020. 4

[24] 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)*, pages 770–778, 2016. 5

[25] Youngkyu Hong, Seungju Han, Kwanghee Choi, Seokjun Seo, Beomsu Kim, and Buru Chang. Disentangling label distribution for long-tailed visual recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6626–6636, June 2021. 2

[26] Minsung Hyun, Jisoo Jeong, and Nojun Kwak. Class-imbalanced semi-supervised learning. *arXiv preprint arXiv:2002.06815*, 2020. 2

[27] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *International Conference on Learning Representations (ICLR)*, 2020. 2, 3

[28] Dong-Jin Kim, Jae Won Cho, Jinsoo Choi, Yunjae Jung, and In So Kweon. Single-modal entropy based active learning for

visual question answering. In *British Machine Vision Conference (BMVC)*, 2021. 8

[29] 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, 2

[30] Dong-Jin Kim, Jinsoo Choi, Tae-Hyun Oh, Youngjin Yoon, and In So Kweon. Disjoint multi-task learning between heterogeneous human-centric tasks. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2018. 2, 3

[31] Dong-Jin Kim, Xiao Sun, Jinsoo Choi, Stephen Lin, and In So Kweon. Detecting human-object interactions with action co-occurrence priors. In *European Conference on Computer Vision (ECCV)*, 2020. 1, 3

[32] Dong-Jin Kim, Xiao Sun, Jinsoo Choi, Stephen Lin, and In So Kweon. Acp++: Action co-occurrence priors for human-object interaction detection. *IEEE Transactions on Image Processing (TIP)*, 30:9150–9163, 2021. 1, 3

[33] 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. 1, 2, 5, 6, 7

[34] Jaehyung Kim, Jongheon Jeong, and Jinwoo Shin. M2m: Imbalanced classification via major-to-minor translation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13896–13905, 2020. 2

[35] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. *Technical report*, 2009. 2, 5

[36] Chia-Wen Kuo, Chih-Yao Ma, Jia-Bin Huang, and Zsolt Kira. Featmatch: Feature-based augmentation for semi-supervised learning. In *European Conference on Computer Vision (ECCV)*, volume 18, pages 479–495, 2020. 1, 2

[37] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. In *International Conference on Learning Representations (ICLR)*, 2016. 2

[38] 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, 3

[39] Hyuck Lee, Seungjae Shin, and Heeyoung Kim. Abc: Auxiliary balanced classifier for class-imbalanced semi-supervised learning. *arXiv preprint arXiv:2110.10368*, 2021. 2, 5, 6

[40] Junnan Li, Caiming Xiong, and Steven Hoi. Comatch: Semi-supervised learning with contrastive graph regularization. In *IEEE International Conference on Computer Vision (ICCV)*, 2021. 1

[41] Junnan Li, Caiming Xiong, and Steven Hoi. Mopro: Webly supervised learning with momentum prototypes. In *International Conference on Learning Representations (ICLR)*, 2021. 3

[42] Yunru Liu, Tingran Gao, and Haizhao Yang. Selectnet: Learning to sample from the wild for imbalanced data training. In *Mathematical and Scientific Machine Learning*, volume 107, pages 193–206, 2020. 2

[43] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *International Conference on Learning Representations (ICLR)*, 2021. 2, 5, 6

[44] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 41(8):1979–1993, 2018. 2

[45] Avital Oliver, Augustus Odena, Colin A Raffel, Ekin Dogus Cubuk, and Ian Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. In *Advances in Neural Information Processing Systems (NIPS)*, volume 31, pages 3235–3246, 2018. 5, 7, 8

[46] Jongjin Park, Sukmin Yun, Jongheon Jeong, and Jinwoo Shin. Opencos: Contrastive semi-supervised learning for handling open-set unlabeled data. *arXiv preprint arXiv:2107.08943*, 2021. 7

[47] Seulki Park, Jongin Lim, Younghan Jeon, and Jin Young Choi. Influence-balanced loss for imbalanced visual classification. In *IEEE International Conference on Computer Vision (ICCV)*, pages 735–744, 2021. 2

[48] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems (NIPS)*, volume 32, page 8026–8037, 2019. 5

[49] Hieu Pham, Qizhe Xie, Zihang Dai, and Quoc V Le. Meta pseudo labels. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11557–11568, 2021. 1

[50] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2001–2010, 2017. 3

[51] Jiawei Ren, Cunjun Yu, shunan sheng, Xiao Ma, Haiyu Zhao, Shuai Yi, and hongsheng Li. Balanced meta-softmax for long-tailed visual recognition. In *Advances in Neural Information Processing Systems (NIPS)*, volume 33, pages 4175–4186, 2020. 2

[52] Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *International Conference on Machine Learning (ICML)*, pages 4334–4343. PMLR, 2018. 2

[53] Zhongzheng Ren, Raymond Yeh, and Alexander Schwing. Not all unlabeled data are equal: Learning to weight data in semi-supervised learning. In *Advances in Neural Information Processing Systems (NIPS)*, volume 33, pages 21786–21797, 2020. 2

[54] Ruslan Salakhutdinov and Geoff Hinton. Learning a nonlinear embedding by preserving class neighbourhood struc-

ture. In *Artificial Intelligence and Statistics*, pages 412–419. PMLR, 2007. 3

[55] Inkyu Shin, Dong-Jin Kim, Jae Won Cho, Sanghyun Woo, KwanYong Park, and In So Kweon. Labor: Labeling only if required for domain adaptive semantic segmentation. In *IEEE International Conference on Computer Vision (ICCV)*, 2021. 8

[56] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems (NIPS)*, volume 30, pages 4077–4087, 2017. 1, 3

[57] 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, 4, 5, 6, 7, 8

[58] Jong-Chyi Su and Subhransu Maji. The semi-supervised inaturalist-aves challenge at fgvc7 workshop, 2021. 1, 2, 5, 7

[59] 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)*, volume 30, pages 1195–1204, 2017. 2, 6, 7

[60] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 8

[61] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8769–8778, 2018. 2

[62] Xudong Wang, Long Lian, Zhongqi Miao, Ziwei Liu, and Stella Yu. Long-tailed recognition by routing diverse distribution-aware experts. In *International Conference on Learning Representations (ICLR)*, 2021. 2

[63] Xudong Wang, Zhirong Wu, Long Lian, and Stella X Yu. Debiased learning from naturally imbalanced pseudo-labels for zero-shot and semi-supervised learning. *arXiv preprint arXiv:2201.01490*, 2022. 4

[64] 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. 1, 2, 5, 6, 7

[65] Liuyu Xiang, Guiguang Ding, and Jungong Han. Learning from multiple experts: Self-paced knowledge distillation for long-tailed classification. In *European Conference on Computer Vision (ECCV)*, pages 247–263, 2020. 2

[66] Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. Unsupervised data augmentation for consistency training. In *Advances in Neural Information Processing Systems (NIPS)*, 2020. 2, 4

[67] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10687–10698, 2020. 1

[68] I Zeki Yalniz, Hervé Jégou, Kan Chen, Manohar Paluri, and Dhruv Mahajan. Billion-scale semi-supervised learning for image classification. *arXiv preprint arXiv:1905.00546*, 2019. 1

[69] Yuzhe Yang and Zhi Xu. Rethinking the value of labels for improving class-imbalanced learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2020. 2

[70] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *British Machine Vision Conference (BMVC)*, pages 87.1–87.12, 2016. 5

[71] Boyan Zhou, Quan Cui, Xiu-Shen Wei, and Zhao-Min Chen. Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9719–9728, 2020. 2