CReST: A Class-Rebalancing Self-Training Framework for Imbalanced Semi-Supervised Learning

Chen Wei[1]* Kihyuk Sohn[2] Clayton Mellina[2] Alan Yuille[1] Fan Yang[2]
[1]Johns Hopkins University [2]Google Cloud AI

Abstract

Semi-supervised learning on class-imbalanced data, although a realistic problem, has been under studied. While existing semi-supervised learning (SSL) methods are known to perform poorly on minority classes, we find that they still generate high precision pseudo-labels on minority classes. By exploiting this property, in this work, we propose Class-Rebalancing Self-Training (CReST), a simple yet effective framework to improve existing SSL methods on class-imbalanced data. CReST iteratively retrains a baseline SSL model with a labeled set expanded by adding pseudo-labeled samples from an unlabeled set, where pseudo-labeled samples from minority classes are selected more frequently according to an estimated class distribution. We also propose a progressive distribution alignment to adaptively adjust the rebalancing strength dubbed CReST+. We show that CReST and CReST+ improve state-of-the-art SSL algorithms on various class-imbalanced datasets and consistently outperform other popular rebalancing methods. Code has been made available at https://github.com/google-research/crest.

1. Introduction

Semi-supervised learning (SSL) utilizes unlabeled data to improve model performance and has achieved promising results on standard SSL image classification benchmarks [34, 25, 43, 2, 39, 47]. A common assumption, which is often made implicitly during the construction of SSL benchmark datasets, is that the class distribution of labeled and/or unlabeled data are balanced. However, in many realistic scenarios, this assumption holds untrue and becomes the primary cause of poor SSL performance [5, 22].

Supervised learning on imbalanced data has been widely explored. It is commonly observed that models trained on imbalanced data are biased towards *majority classes* which have numerous examples, and away from *minority classes* which have few examples. Various solutions have been proposed to help alleviate bias, such as re-sampling [3, 4], re-weighting [9, 6], and two-stage training [20, 52]. All these methods rely on labels to re-balance the biased model.

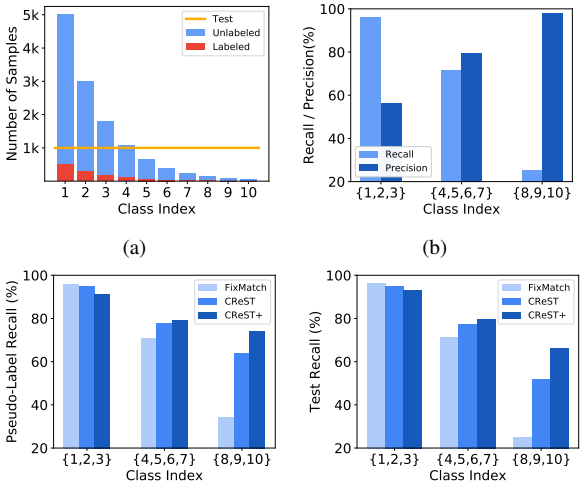


Figure 1. Experimental results on CIFAR10-LT. (a) Both labeled and unlabeled sets are class-imbalanced, where the most majority class has 100× more samples than the most minority class. The test set remains balanced. (b) Precision and recall of a FixMatch [39] model. Although minority classes have low recall, they obtain high precision. (c) & (d) The proposed CReST and CReST+ improve the quality of pseudo-labels (c) and thus the recall on the balanced test set (d), especially on minority classes.

In contrast, SSL on imbalanced data has been understudied. In fact, data imbalance poses further challenges in SSL where missing label information precludes rebalancing the unlabeled set. Pseudo-labels for unlabeled data generated by a model trained on labeled data are commonly leveraged in SSL algorithms. However, pseudo-labels can be problematic if they are generated by an initial model trained on imbalanced data and biased toward majority classes: subsequent training with such biased pseudo-labels intensifies the bias and deteriorates the model quality. Apart from a few recent works [22, 48], the majority of existing SSL algorithms [2, 1, 46, 39] have not been thoroughly evaluated on imbalanced data distributions.

In this work, we investigate SSL in the context of class-

*Work done while an intern at Google.

imbalanced data in which both labeled and unlabeled sets have roughly the same imbalanced class distributions, as illustrated in Fig. 1(a). We observe that the undesired performance of existing SSL algorithms on imbalanced data is mainly due to low recall on minority classes. Our method is motivated by the further observation that, despite this, precision on minority classes is surprisingly high. In Fig. 1(b), we show predictions on a CIFAR10-LT dataset produced by FixMatch [39], a representative SSL algorithm with state-of-the-art performance on balanced benchmarks. The model obtains high recall on majority classes but suffers from low recall on minority classes, which results in low accuracy overall on the balanced test set. However, the model has almost perfect precision on minority classes, suggesting that the model is conservative in classifying samples into minority classes, but once it makes such a prediction we can be confident it is correct. Similar observations are made on other SSL methods, and on supervised learning [19].

With this in mind, we introduce a class-rebalancing self-training scheme (CReST) which re-trains a baseline SSL model after adaptively sampling pseudo-labeled data from the unlabeled set to supplement the original labeled set. We refer to each fully-trained baseline model as a *generation*. After each generation, pseudo-labeled samples from the unlabeled set are added into the labeled set to retrain an SSL model. Rather than updating the labeled set with all pseudo-labeled samples, we instead use a *stochastic* update strategy in which samples are selected with higher probability if they are predicted as minority classes, as those are more likely to be correct predictions. The updating probability is a function of the data distribution estimated from the labeled set. In addition, we extend CReST to CReST+ by incorporating distribution alignment [1] with a temperature scaling factor to control its alignment strength over generations, so that predicted data distributions are more aggressively adjusted to alleviate model bias. As shown in Fig. 1(c) and 1(d), the proposed strategy reduces the bias of pseudo-labeling and improves the class-balanced test set accuracy as a result.

We show in experiments that CReST and CReST+ improve over baseline SSL methods by a large margin. On CIFAR-LT [9, 6], our method outperforms FixMatch [39] under different imbalance ratios and label fractions by as much as 11.8% in accuracy. Our method also outperforms DARP [22], a state-of-the-art SSL algorithm designed for learning from imbalanced data, on both MixMatch [2] and FixMatch [39] by up to 4.0% in accuracy. To further test the efficacy of the proposed method on large-scale data, we apply our method on ImageNet127 [17], a naturally imbalanced dataset created from ImageNet [11] by merging classes based on the semantic hierarchy, and get 7.9% gain on recall. Extensive ablation study further demonstrates that our method particularly helps improve recall on minority classes, making it a viable solution for imbalanced SSL.

2. Related work

2.1. Semi-supervised learning

Recent years have observed a significant advancement of SSL research [26, 25, 32, 2, 1, 47, 46, 39]. Many of these methods share similar basic techniques, such as entropy minimization [13], pseudo-labeling, or consistency regularization, with deep learning. Pseudo-labeling [26, 39] trains a classifier with unlabeled data using pseudo-labeled targets derived from the model's own predictions. Relatedly, [25, 2, 46, 1, 47] use a model's predictive probability with temperature scaling as a soft pseudo-label. Consistency regularization [36, 25, 32] learns a classifier by promoting consistency in predictions between different views of unlabeled data, either via soft [25, 32, 2, 46] or hard [39] pseudo-labels. Effective methods of generating multiple views include input data augmentations of varying strength [12, 8, 1], standard dropout within network layers [40], and stochastic depth [16]. The performance of most recent SSL methods relies on the quality of pseudo-labels. However, none of aforementioned works have studied SSL in the class-imbalanced setting, in which the quality of pseudo-labels is significantly threatened by model bias.

2.2. Class-imbalanced supervised learning

Research on class-imbalanced supervised learning has attracted increasing attention. Prominent works include re-sampling [7, 3, 4, 14] and re-weighting [21, 9, 6, 41] which re-balance the contribution of each class, while others focus on re-weighting each instance [27, 38, 35, 19]. Some works [50, 44, 23, 28, 29] aim to transfer knowledge from majority classes to minority classes. A recent trend of work proposes to decouple the learning of representation and classifier [52, 20, 42]. These methods assume all labels are available during training and their performance is largely unknown under SSL scenarios.

2.3. Class-imbalanced semi-supervised learning

While SSL has been extensively studied, it is under-explored regarding class-imbalanced data. Recently, Yang and Xu [48] argued that leveraging unlabeled data by SSL and self-supervised learning can benefit class-imbalanced learning. Hyun *et al*. [18] proposed a suppressed consistency loss to suppress the loss on minority classes. Kim *et al*. [22] proposed Distribution Aligning Refinery (DARP) to refine raw pseudo-labels via a convex optimization. In contrast, we boost the quality of the model's raw pseudo-labels directly via an class-rebalancing sampling strategy and a progressive distribution alignment strategy. DARP also discussed another interesting setting where labeled and unlabeled data do not share the same class distribution, while in this work we focus on the scenario when labeled and unlabeled data have roughly the same distribution.

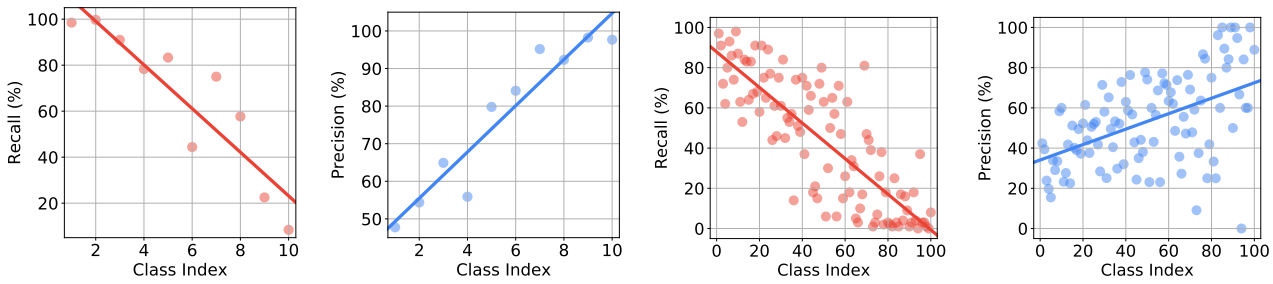


Figure 2. Bias of a FixMatch [39] model on class-imbalanced data. **Left**: Per-class recall and precision on CIFAR10-LT. **Right**: Per-class recall and precision on CIFAR100-LT. The class index is sorted by the number of examples in descending order. While the conventional assumption might be that the performance of the majority classes is better than that of the minority classes, we find it only partially true. The model obtains high recall but low precision on majority classes, while obtaining low recall but high precision on minority classes. See more details in Sec. 3.2.

3. Class-Imbalanced SSL

In this section, we first set up the problem and introduce baseline SSL algorithms. Next, we investigate the biased behavior of existing SSL algorithms on class-imbalanced data. Based on these observations, we propose a class-rebalancing self-training framework (CReST) that takes advantage of, rather than suffers from, the model's bias to alleviate the performance degeneration on minority classes. In addition, we extend distribution alignment [1] and integrate it as CReST+ to further improve the quality of online pseudo-labeling.

3.1. Problem setup and baselines

We first set up the problem of class-imbalanced semi-supervised learning. For an L-class classification task, there is a labeled set $\mathcal{X} = \{(x_n, y_n) : n \in (1, \dots, N)\}$, where $x_n \in \mathbb{R}^d$ are training examples and $y_n \in \{1, \dots, L\}$ are corresponding class labels. The number of training examples in $\mathcal{X}$ of class l is denoted as N_l, *i.e.*, $\sum_{l=1}^{L} N_l = N$. Without loss of generality, we assume that the classes are sorted by cardinality in descending order, *i.e.*, $N_1 \geq N_2 \geq \cdots \geq N_L$. The marginal class distribution of $\mathcal{X}$ is skewed, *i.e.*, $N_1 \gg N_L$. We measure the degree of class imbalance by imbalance ratio, $\gamma = \frac{N_1}{N_L}$. Besides the labeled set $\mathcal{X}$, an unlabeled set $\mathcal{U} = \{u_m \in \mathbb{R}^d : m \in (1, \dots, M)\}$ that shares the same class distribution as $\mathcal{X}$ is also provided. The label fraction $\beta = \frac{N}{N+M}$ measures the percentage of labeled data. Given class-imbalanced sets $\mathcal{X}$ and $\mathcal{U}$, our goal is to learn a classifier $f : \mathbb{R}^d \to \{1, \dots, L\}$ that generalizes well under a class-*balanced* test criterion.

Many state-of-the-art SSL methods [39, 47] utilize unlabeled data by assigning a pseudo-label with the classifier's prediction $\hat{y}_m = f(u_m)$. The classifier is then optimized on both labeled and unlabeled samples with their corresponding pseudo-labels. Therefore, the quality of pseudo-labels is crucial to the final performance. These algorithms work successfully on standard class-balanced datasets since the quality of the classifier — and thus its online pseudo-labels — improves for all classes over the course of training. However, when the classifier is biased at the beginning due to a skewed class distribution, the online pseudo-labels of unlabeled data can be even more biased, further aggravating the class-imbalance issue and resulting in severe performance degradation on minority classes.

3.2. A closer look at the model bias

Previous works [9, 6] introduce long-tailed versions of CIFAR [24] datasets with various class-imbalanced ratios to evaluate class-imbalanced fully-supervised learning algorithms. We extend this protocol by retaining a fraction of training samples as labeled and the remaining as unlabeled. We test FixMatch [39], one of the state-of-the-art SSL algorithms designed for class-balanced data. Fig. 2 shows test recall and precision of each class on CIFAR10-LT with imbalance ratio $\gamma = 100$, label fraction $\beta = 10\%$, and CIFAR100-LT with imbalance ratio $\gamma = 50$, label fraction $\beta = 30\%$.

First, as shown in the first and third plots of Fig. 2, FixMatch achieves very high recall on majority classes and poor recall on minority classes, which is consistent with the conventional wisdom. For example, the recall of the most and second most majority classes of CIFAR10-LT is 98.5% and 99.7%, respectively, while the model recognizes only 8.4% of samples correctly from the most minority class. In other words, the model is highly biased towards majority classes, resulting in poor recall averaged over all classes which is also known as accuracy as the test set is balanced.

Despite the low recall, the minority classes maintain surprisingly high precision as in the second and fourth plots of Fig. 2. For example, the model achieves 97.7% and 98.3% precision, respectively, on the most and the second most minority classes of CIFAR10-LT, while only achieving relatively low precision on majority classes. This indicates that many minority class samples are predicted as one of the majority classes.

While the conventional wisdom may suggest that the performance of the majority classes is better than that of the

minority classes, we find that it is only partly true: the biased model learned on class-imbalanced data indeed performs favorably on majority classes in terms of recall, but favors minority classes in terms of precision. Similar observations are made on other SSL algorithms, and also on fully-supervised class-imbalanced learning [19]. This empirical finding motivates us to exploit the high precision of minority classes to alleviate their recall degradation. To achieve this goal, we introduce CReST, a class-rebalancing self-training framework illustrated in Fig. 3.

3.3. Class-rebalancing self-training

Self-training [37, 49] is an iterative method widely used in SSL. It trains the model for multiple generations, where each generation involves two steps. First, the model is trained on the labeled set to obtain a teacher model. Second, the teacher model's predictions are used to generate pseudo-labels $\hat{y}_m$ for unlabeled data u_m. The pseudo-labeled set $\hat{\mathcal{U}} = \{(u_m, \hat{y}_m)\}_{m=1}^{M}$ is included into the labeled set, *i.e.*, $\mathcal{X}' = \mathcal{X} \cup \hat{\mathcal{U}}$, for the next generation.

To accommodate the class-imbalance, we propose two modifications to the self-training strategy. First, instead of solely training on the labeled data, we use SSL algorithms to exploit both labeled and unlabeled data to get a better teacher model in the first step. More importantly, in the second step, rather than including every sample in $\hat{\mathcal{U}}$ in the labeled set, we instead expand the labeled set with a selected subset $\hat{\mathcal{S}} \subset \hat{\mathcal{U}}$, *i.e.*, $\mathcal{X}' = \mathcal{X} \cup \hat{\mathcal{S}}$. We choose $\hat{\mathcal{S}}$ following a class-rebalancing rule: the less frequent a class l is, the more unlabeled samples that are predicted as class l are included into the pseudo-labeled set $\hat{\mathcal{S}}$.

We estimate the class distribution from the labeled set. Specifically, unlabeled samples that are predicted as class l are included into $\hat{\mathcal{S}}$ at the rate of

$$\mu_l = \left(\frac{N_{L+1-l}}{N_1}\right)^{\alpha}, \tag{1}$$

where $\alpha \geq 0$ tunes the sampling rate and thus the size of $\hat{\mathcal{S}}$. For instance, for a 10-class imbalanced dataset with imbalance ratio of $\gamma = \frac{N_1}{N_{10}} = 100$, we keep all samples predicted as the most minority class since $\mu_{10} = (\frac{N_{10+1-10}}{N_1})^{\alpha} = 1$. While for the most majority class, $\mu_1 = (\frac{N_{10+1-1}}{N_1})^{\alpha} = 0.01^{\alpha}$ of samples are selected. When $\alpha = 0$, $\mu_l = 1$ for all l, then all unlabeled samples are kept and the algorithm falls back to the conventional self-training. When selecting pseudo-labeled samples in each class, we take the most confident ones.

The motivation of our CReST strategy is two-fold. First, as observed in Sec. 3.2, the precision of minority classes is much higher than that of majority classes, hence minority class pseudo-labels are less risky to include in the labeled set. Second, adding samples to minority classes is more critical due to data scarcity. With more samples from minority classes added, the labeled set is more class-balanced, which leads to a less biased classifier for online pseudo-labeling in the subsequent generation. Note that there are other ways of sampling the pseudo-labels in a class-balancing fashion and we provide a practical and effective example.

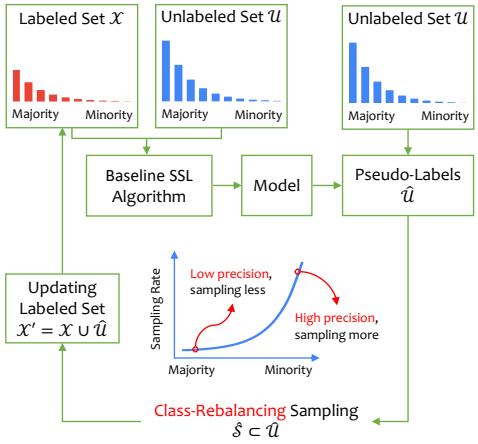


Figure 3. CReST (Class-Rebalancing Self-Training) alternatingly trains a baseline SSL algorithm on both labeled and unlabeled data and expands the labeled set by sampling pseudo-labeled unlabeled data. Sampling rates for majority and minority classes are adaptively determined based on the quality of pseudo-labels. See text for details.

3.4. Progressive distribution alignment

We further improve the quality of online pseudo-labels by additionally introducing progressive distribution alignment into CReST and distinguish it as CReST+.

While first introduced for class-balanced SSL, Distribution Alignment (DA) [1] fits with class-imbalanced scenarios particularly well. It aligns the model's predictive distribution on unlabeled samples with the labeled training set's class distribution $p(y)$. Let $\tilde{p}(y)$ be the moving average of the model's predictions on unlabeled examples. DA first scales the model's prediction $q = p(y|u_m; f)$ for an unlabeled example u_m by the ratio $\frac{p(y)}{\tilde{p}(y)}$, aligning q with the target distribution $p(y)$. It then re-normalizes the scaled result to form a valid probability distribution: $\tilde{q} = \text{Normalize}(q\frac{p(y)}{\tilde{p}(y)})$, where $\text{Normalize}(x)_i = x_i / \sum_j x_j$. $\tilde{q}$ is used as the label guess for u_m instead of q.

To further enhance DA's ability to handle class-imbalanced data, we extend it with temperature scaling. Specifically, we add a tuning knob $t \in [0, 1]$ that controls the class-rebalancing strength of DA. Instead of directly taking $p(y)$ as target, we use a temperature-scaled distribution $\text{Normalize}(p(y)^t)$. When $t = 1$, we recover DA. When $t < 1$, the temperature-scaled distribution becomes smoother and balances the model's predictive distribution more aggressively. When $t = 0$, the target distribution be-

	CIFAR10-LT						CIFAR100-LT			
	$\beta=10\%$			$\beta=30\%$			$\beta=10\%$		$\beta=30\%$	
Method	$\gamma=50$	$\gamma=100$	$\gamma=200$	$\gamma=50$	$\gamma=100$	$\gamma=200$	$\gamma=50$	$\gamma=100$	$\gamma=50$	$\gamma=100$
FixMatch [39]	$79.4_{\pm 0.65}$	$66.3_{\pm 1.74}$	$59.7_{\pm 0.74}$	$81.9_{\pm 0.30}$	$73.1_{\pm 0.58}$	$64.7_{\pm 0.69}$	$33.7_{\pm 0.94}$	$28.3_{\pm 0.66}$	$43.1_{\pm 0.24}$	$38.6_{\pm 0.45}$
w/ CReST	$83.8_{\pm 0.45}$	$75.9_{\pm 0.62}$	$64.1_{\pm 0.23}$	$84.2_{\pm 0.13}$	$77.6_{\pm 0.86}$	$67.7_{\pm 0.82}$	$37.4_{\pm 0.29}$	$32.1_{\pm 1.52}$	$45.6_{\pm 0.19}$	$40.2_{\pm 0.53}$
w/ CReST+	$\mathbf{84.2}_{\pm 0.39}$	$\mathbf{78.1}_{\pm 0.84}$	$\mathbf{67.7}_{\pm 1.39}$	$\mathbf{84.9}_{\pm 0.27}$	$\mathbf{79.2}_{\pm 0.20}$	$\mathbf{70.5}_{\pm 0.56}$	$\mathbf{38.8}_{\pm 1.03}$	$\mathbf{34.6}_{\pm 0.74}$	$\mathbf{46.7}_{\pm 0.34}$	$\mathbf{42.0}_{\pm 0.44}$

Table 1. Classification accuracy (%) on CIFAR10-LT and CIFAR100-LT under various label fraction β and imbalance ratio γ. The numbers are averaged over 5 different folds. Models with CReST are trained for 15 generations. Models with CReST+ are trained for 6 generations.

comes uniform.

While using a smaller t can benefit a single generation under a class-balanced test criterion, it is less desirable for multiple generations of self-training since it affects the quality of pseudo-labels. Specifically, applying a $t<1$ enforces the model's predictive distribution to be more balanced than the class distribution of the training set, leading the model to predict minority classes more frequently. However, on an imbalanced training set with few samples of minority classes, such pseudo-labeling tends to be over-balanced, *i.e.*, more samples are wrongly predicted as minority classes. This decreases the high precision of minority classes, interfering with our ability to exploit it to produce better pseudo-labels.

To handle this, we propose to progressively increase the strength of class-rebalancing by decreasing t over generations. Specifically, we set t by a linear function of the current generation g which indexes from 0:

$$t_g = (1 - \frac{g}{G}) \cdot 1.0 + \frac{g}{G} \cdot t_{\text{min}}\,, \tag{2}$$

where $G+1$ is the total number of generations and t_{min} is the temperature used for the last generation. This progressive schedule for t enjoys both high precision of pseudo-labels in early generations, and stronger class-rebalancing in late generations. It also speeds up the iterative training, obtaining better results with fewer generations of training. See Sec. 4.3 for empirical analysis.

4. Experiments

4.1. CIFAR-LT

Datasets. We first evaluate the efficacy of the proposed method on long-tailed CIFAR10 (CIFAR10-LT) and long-tailed CIFAR100 (CIFAR100-LT) introduced in [9, 6]. On these datasets, training images are randomly discarded per class to maintain a pre-defined imbalance ratio γ. Specifically, $N_l = \gamma^{-\frac{l-1}{L-1}} \cdot N_1$ while $N_1 = 5000$, $L = 10$ for CIFAR10-LT and $N_1 = 500$, $L = 100$ for CIFAR100-LT. We randomly select $\beta = 10\%$ and 30% of samples from training data to create the labeled set, and test imbalance ratio $\gamma = 50$, 100 and 200 for CIFAR10-LT and $\gamma = 50$ and 100 for CIFAR100-LT. The test set remains untouched and balanced, so that the evaluated criterion, accuracy on the test set, is class-balanced.

Setup. We use Wide ResNet-28-2 [51] following [33, 39] as the backbone. We evaluate our method on FixMatch and MixMatch. For each generation, the model is trained for 2^{16} steps when using FixMatch as the baseline SSL algorithm and 2^{17} steps for MixMatch. We use a cosine learning rate decay [30, 39] whose formulation is provided in the supplementary material. Other hyper-parameters for each training generation are untouched. For CReST and CReST+ related hyper-parameters, we set $\alpha = 1/3$, $t_{\text{min}} = 0.5$ for FixMatch and $\alpha = 1/2$, $t_{\text{min}} = 0.8$ for MixMatch. CReST takes 15 generations, while CReST+ only takes 6 generations accelerated by progressive distribution alignment. The hyper-parameters are selected based on a single fold of CIFAR10-LT with $\gamma = 100$ and $\beta = 10\%$. We evaluate the model on the test dataset every 2^{10} steps and report the average test accuracy of the last 5 evaluations. Each algorithm is tested under 5 different folds of labeled data and we report the mean and the standard deviation of accuracy on the test set. Following [2] and [39], we report final performance using an exponential moving average of model parameters.

Main results. First, we compare our model with baseline FixMatch, and present the results in Table 1. Although FixMatch performs reasonably well on imbalance ratio $\gamma = 50$, its accuracy decreases significantly with increasing imbalance ratio. In contrast, CReST improves the accuracy of FixMatch on all evaluated settings and achieves as much as 9.6% absolute performance gain. When incorporating progressive distribution alignment, our CReST+ model is able to further boost the performance on all settings by another few points, resulting in 3.0% to 11.8% absolute accuracy improvement compared to baseline FixMatch.

The accuracy of all compared methods improves with increasing number of labeled samples, but CReST consistently outperforms the baseline. This indicates that CReST can better utilize labeled data to reduce model bias under imbalanced class-distribution.

We also observe that our model works particularly well and achieves 11.8% and 6.1% accuracy gain for imbalance ratio $\gamma = 100$ with 10% and 30% labeled data, respectively. We hypothesize the reason is that our model finds more cor-

rectly pseudo-labeled samples to augment the labeled set when the imbalance ratio is moderate. However, when imbalance ratio is very high, *e.g.*, $\gamma = 200$, our model's capability is constrained by insufficient number of training samples from minority classes.

Comparison with baselines. We further report the performance of other SSL baselines in Table 2. For fair comparison, all algorithms are trained for 6×2^{16} steps. This leads to 6 generations for CReST and CReST+ on a FixMatch base with 2^{16} steps each generation, and 3 generations for CReST and CReST+ on a MixMatch base with 2^{17} steps each generation. Other models that do not use self-training are trained for a single generation with 6×2^{16} steps.

Method	$\gamma = 50$	$\gamma = 100$	$\gamma = 200$
Pseudo-Labeling [26]	52.5$_{\pm 0.74}$	46.5$_{\pm 1.29}$	42.0$_{\pm 1.39}$
Mean Teacher [43]	57.1$_{\pm 3.00}$	48.1$_{\pm 0.71}$	45.1$_{\pm 1.28}$
MixMatch [2]	69.1$_{\pm 1.18}$	60.4$_{\pm 2.24}$	54.5$_{\pm 1.87}$
w/ CReST	69.8$_{\pm 1.06}$	60.5$_{\pm 1.56}$	55.2$_{\pm 2.25}$
w/ CReST+	76.7$_{\pm 0.35}$	66.1$_{\pm 0.79}$	57.6$_{\pm 1.30}$
FixMatch [39]	80.1$_{\pm 0.44}$	67.3$_{\pm 1.19}$	59.7$_{\pm 0.63}$
w/ CB [9]	80.2$_{\pm 0.45}$	67.6$_{\pm 1.88}$	60.8$_{\pm 0.26}$
w/ RS [3, 4]	80.2$_{\pm 0.78}$	69.6$_{\pm 1.30}$	60.9$_{\pm 1.25}$
w/ DA [1] ($t = 1.0$)	80.2$_{\pm 0.45}$	69.7$_{\pm 1.27}$	62.0$_{\pm 0.84}$
w/ DA [1] ($t = 0.5$)	82.4$_{\pm 0.33}$	73.6$_{\pm 0.63}$	63.7$_{\pm 1.17}$
w/ LA [31]	83.2$_{\pm 0.87}$	70.4$_{\pm 2.90}$	62.4$_{\pm 1.24}$
w/ CReST	83.2$_{\pm 0.37}$	74.8$_{\pm 1.09}$	63.4$_{\pm 0.32}$
w/ CReST+	84.2$_{\pm 0.39}$	78.1$_{\pm 0.84}$	67.7$_{\pm 1.39}$
w/ CReST+ & LA	**85.6**$_{\pm 0.36}$	**81.2**$_{\pm 0.70}$	**71.9**$_{\pm 2.24}$

Table 2. We compare CReST and CReST+ with baseline methods including different SSL algorithms and typical class-rebalancing techniques designed for fully-supervised learning. For fair comparison, all models are measured at the same number of training steps. See text for details. Three imbalance ratios γ with $\beta = 10\%$ labels are evaluated. Numbers are averaged over 5 different folds.

Method	$\gamma = 50$	$\gamma = 100$	$\gamma = 150$
Supervised	65.2$_{\pm 0.05}$	58.8$_{\pm 0.13}$	55.6$_{\pm 0.43}$
MixMatch [2]	73.2$_{\pm 0.56}$	64.8$_{\pm 0.28}$	62.5$_{\pm 0.31}$
w/ DARP [22]	75.2$_{\pm 0.47}$	67.9$_{\pm 0.14}$	65.8$_{\pm 0.52}$
w/ CReST	78.4$_{\pm 0.36}$	70.0$_{\pm 0.49}$	64.7$_{\pm 0.96}$
w/ CReST+	**79.0**$_{\pm 0.26}$	**71.9**$_{\pm 0.33}$	**68.3**$_{\pm 0.57}$
FixMatch [39]	79.2$_{\pm 0.33}$	71.5$_{\pm 0.72}$	68.4$_{\pm 0.15}$
w/ DARP [22]	81.8$_{\pm 0.24}$	75.5$_{\pm 0.05}$	70.4$_{\pm 0.25}$
w/ CReST	83.0$_{\pm 0.39}$	75.7$_{\pm 0.38}$	70.8$_{\pm 0.25}$
w/ CReST+	**83.9**$_{\pm 0.14}$	**77.4**$_{\pm 0.36}$	**72.8**$_{\pm 0.58}$

Table 3. Accuracy (%) under DARP's protocol [22] on CIFAR10. See the supplementary material for dataset details. Three imbalance ratios γ are evaluated. Numbers are averaged over 5 runs.

Method	Gen_1	Gen_2	Gen_3
Supervised (100% labels)	75.8	-	-
Supervised (10% labels)	46.0	-	-
FixMatch (10% labels)	65.8	-	-
w/ DA ($t = 0.5$)	69.1	-	-
w/ CReST	65.8	67.6	67.7
w/ CReST+	68.3	70.7	**73.7**

Table 4. Evaluating the proposed method on ImageNet127 with $\beta = 10\%$ samples are labeled. We retrain FixMatch models for 3 generations with our CReST and CReST+.

We first directly evaluate several classic SSL methods on class-imbalanced datasets, including Pseudo-Labeling [26], Mean Teacher [43], MixMatch [2] and FixMatch [39]. All the SSL baselines suffer from low accuracy due to imbalanced data, and the accuracy drop becomes more pronounced with increasing imbalance ratio. On MixMatch, the improvement provided by CReST is modest mainly due to the schedule constraint. Providing more generation budget, the results of MixMatch with CReST can be further improved. Among these algorithms, FixMatch achieves the best performance, so we take it as the baseline for various rebalancing methods.

We consider typical class-rebalancing methods designed for fully-supervised learning that can be directly applied in SSL algorithms including 1) Class-Balanced loss (CB) [9], a representative of re-weighting strategies in which labeled examples are re-weighted according to the inverse of the effective number of samples in each class; 2) Re-Sampling (RS) [3, 4], a representative of re-sampling strategies in which each labeled example is sampled with probability proportional to the inverse sample size of its class. We also consider Distribution Alignment (DA) [1] as described in Sec. 3.4 and Logit Adjustment (LA) [31], an ad-hoc post-processing technique to enhance models' discriminative ability on minority classes by adjusting the logits of model predictions. While CB, RS, DA and LA all improve accuracy over SSL baselines, the gain is relatively small. With CReST and CReST+, we successfully improve the accuracy for all imbalance ratios by at most 10.8% over FixMatch, outperforming all compared SSL baselines and class-rebalancing methods. Finally, applying LA as the post-processing correction of our CReST+ models further gives consistent accuracy gains, producing the best results.

Comparison with DARP. We directly compare with DARP [22], the most recent state-of-the-art SSL algorithm specifically designed for imbalanced data. Both DARP and our method are built upon MixMatch and FixMatch as drop-in additions to standard SSL algorithms. We apply our method on exactly the same datasets used in DARP and present the results in Table 3. Details of the dataset construction are provided in the supplementary material. For

all three imbalance ratios, our model consistently achieves up to 4.0% accuracy gain over DARP on MixMatch, and up to 2.4% accuracy gain on FixMatch.

4.2. ImageNet127

Datasets. We also evaluate CReST on ImageNet127 [17, 45] to verify its performance on large-scale datasets. ImageNet127 is originally introduced in [17], where the 1000 classes of ImageNet [11] are grouped into 127 classes based on their top-down hierarchy in WordNet. It is a naturally imbalanced dataset with imbalance ratio $\gamma = 286$. Its most majority class "mammal" consists of 218 original classes and 277,601 training images. While its most minority class "butterfly" is formed by a single original class with 969 training examples. We randomly select $\beta = 10\%$ training samples as the labeled set and keep the test set unchanged. Due to class grouping, the test set is not balanced. Therefore, we compute averaged class recall instead of accuracy to achieve a balanced metric.

We note that there are other large-scale datasets like iNaturalist [10] and ImageNet-LT [29] which often serve as testbeds for fully-supervised long-tailed recognition algorithms. However, these datasets contain too few examples of minority classes to form a statistically meaningful dataset and draw reliable conclusions for semi-supervised learning. For example, there are only 5 examples in the most minority class of the ImageNet-LT dataset.

Setup. We use ResNet50 [15] as the backbone. The hyper-parameters for each training generation are adopted from the original FixMatch paper. The model is self-trained for 3 generations with $\alpha = 0.7$ and $t_{\text{min}} = 0.5$.

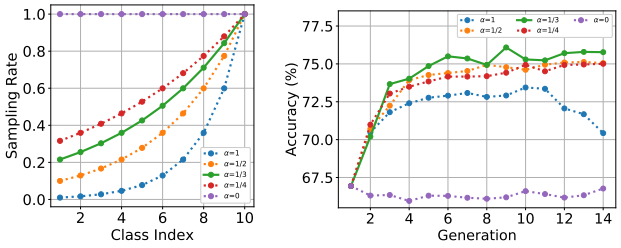
Results. We report results in Table 4. Supervised learning with 100% and 10% labeled training examples and DA with temperature scaling are also presented as reference. Comparing with the baseline FixMatch, both CReST and CReST+ progressively improve over 3 generations of self-training, while CReST+ provides 7.9% absolute gain in the end, which verifies the efficacy of our proposed method.

4.3. Ablation study

We perform an extensive ablation study to evaluate and understand the contribution of each critical component in CReST and CReST+. The experiments in this section are all performed with FixMatch on CIFAR10-LT with imbalance ratio $\gamma = 100$, label fraction $\beta = 10\%$ and a single fold of labeled data.

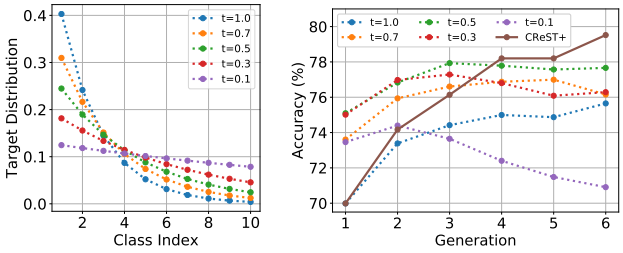
Effect of sampling rate. CReST introduces the sampling rate hyper-parameter α that controls the per-class sampling rate and the number of selected pseudo-labeled samples to be included in the labeled set. In Fig. 4 we show how α influences performance over generations.

When $\alpha = 0$, our method falls back to conventional self-training, which expands the labeled set with all unlabeled examples and their corresponding predicted labels. However, conventional self-training does not produce a performance gain over multiple generations, showing that simply applying self-training can not provide performance improvement. In contrast, with our class-rebalancing sampling strategy ($\alpha > 0$), the accuracy can be improved by iterative model retraining.



(a) (b)

Figure 4. Effect of α across multiple generations on CIFAR10-LT ($\gamma = 100$, $\beta = 10\%$) in CReST. (a) Illustration of how α influences sampling rate. (b) Test accuracy over generations with different α. When $\alpha = 0$, the method falls back to conventional self-training with all the unlabeled examples and corresponding pseudo-labels added into the labeled set, showing no improvement after generations of retraining, whereas our class-rebalancing sampling ($\alpha > 0$) helps.



(a) (b)

Figure 5. Effect of temperature t across multiple generations on CIFAR10-LT ($\gamma = 100$, $\beta = 10\%$). (a) Illustration of how t controls the target distribution of distribution alignment. (b) Test accuracy over generations with different constant t and our CReST+ using progressive t. Compared to using a constant t, CReST+ achieves the best final accuracy by progressing from $t = 0$ to $t_{\text{min}} = 0.5$ over 6 generations.

As illustrated in Fig. 4(a), smaller α means more pseudo-labeled samples are added into the labeled set, which enlarges the labeled set but adversely introduces more low-quality pseudo-labels. On the other hand, larger α biases pseudo-labeled samples towards minority classes. As a result, the class-rebalancing sampling can be too strong with large α, leading to imbalance in the reversed direction, towards the original minority classes. This is the case for $\alpha = 1$ where, after the 10-*th* generation, the model becomes increasingly biased towards minority classes and suffers from performance degradation on majority classes, result-

Method / Class	Split	1	2	3	4	5	6	7	8	9	10	Avg.
FixMatch [39]	test	**98.7**	**99.5**	**90.0**	**83.5**	85.0	47.6	69.9	59.0	8.9	7.2	64.9
w/ CReST	test	97.7	98.3	88.8	81.9	**88.2**	59.7	79.5	61.2	47.0	47.9	75.0
		-1.0	-1.2	-1.2	-1.6	+3.2	+12.1	+9.6	+2.2	+38.1	+40.7	+10.1
w/ CReST+	test	93.8	97.7	87.3	76.9	87.5	**69.2**	**84.9**	**67.9**	**60.3**	**70.8**	**79.6**
		-4.9	-1.8	-2.7	-6.6	+2.5	+21.6	+15.0	+8.9	+51.4	+63.6	+14.7
FixMatch [39]	unlabeled	**98.5**	**99.1**	**90.0**	**84.0**	84.7	49.7	64.9	65.6	14.9	22.2	67.4
w/ CReST	unlabeled	97.8	96.8	**90.0**	82.9	87.4	62.4	79.3	64.8	60.8	66.7	78.9
		-0.7	-2.3	0	-1.1	+2.7	+12.7	+14.4	-0.8	+45.9	+44.5	+11.5
w/ CReST+	unlabeled	92.2	95.7	86.1	76.7	**87.6**	**68.1**	**85.1**	**71.2**	**75.7**	**75.6**	**81.4**
		-6.3	-3.4	-3.9	-7.3	+2.9	+18.4	+20.2	+5.6	+60.8	+53.4	+14.0

Table 5. Per-class recall (%) on the balanced test set and the imbalanced unlabeled set of CIFAR10-LT ($\gamma = 100$, $\beta = 10\%$). Our strategies compromise small loss on majority classes for significant gain on minority classes, leading to improved averaged recall over all classes.

ing in decreased accuracy. For example, from the 10-*th* generation to the last generation, the recall of the most minority classes increases by a large margin from 55.0% to 71.1%, while 7 of the other 9 classes suffer from severe recall degradation, resulting in 3.0% drop of the class-balanced test set accuracy. Empirically, we find $\alpha = 1/3$ achieves a balance between the quality of pseudo-labels and the class-rebalancing strength across different imbalance ratios and label fractions on long-tailed CIFAR datasets.

Effect of progressive temperature scaling. The proposed adaptive distribution alignment used in CReST+ introduces another hyper-parameter, temperature t, that scales the target distribution. We first illustrate in Fig. 5(a) how temperature t smooths the target distribution in distribution alignment so that smaller t provides stronger re-balancing strength. In Fig. 5(b), we study the effect of using a constant temperature and our proposed progressive temperature scaling in which t gradually decreases from 1.0 to $t_{\min} = 0.5$ across generations of self-training.

First, we notice that $t = 0.5$ provides the best *single* generation accuracy of 75.1% among all tested temperature values. This suggests that the model can benefit from class re-balancing with a properly "smoothed" target distribution compared with 70.0% accuracy of the original distribution alignment whose temperate t is fixed to 1.0. Further decreasing t to 0.1 results in lower accuracy, as the target distribution is overly smoothed, which introduces more pseudo-labeling errors.

Over multiple generations of self-training, using a constant t is not optimal. Although a relatively small t (*e.g.*, 0.5) can give better performance in early generations, it can not provide further gains through continuing self-training due to the decreased pseudo-label quality. When t is lower than 0.5, performance can even degrade after certain later generations. In contrast, the proposed CReST+, which progressively enhances the distribution alignment strength, provides the best accuracy at the last generation.

Per-class performance. To show the source of accuracy improvements, in Table 5 we present per-class recall on the balanced test set of CIFAR10-LT with imbalance ratio 100 and label fraction 10%. Both CReST and CReST+ sacrifice a few points of accuracy on four majority classes but provide significant gains on the other six minority classes, obtaining better performance over all classes. We also include the results on the imbalanced unlabeled set. The results are particularly similar to those of the test set with mild drop on majority classes and remarkable improvement on minority classes. This suggests that our method indeed improves the quality of pseudo-labels, which can be transferred to better generalization on a balanced test criterion.

5. Conclusion

In this work, we present a class-rebalancing self-training framework, named CReST, for imbalanced semi-supervised learning. CReST is motivated by the observation that existing SSL algorithms produce high precision pseudo-labels on minority classes. CReST iteratively refines a baseline SSL model by supplementing the labeled set with high quality pseudo-labels, where minority classes are updated more aggressively than majority classes. Over generations of self-training, the model becomes less biased towards majority classes and focuses more on minority classes. We also extend distribution alignment to progressively increase its class-rebalancing strength over generations and denote the combined method CReST+. Extensive experiments on long-tailed CIFAR datasets and ImageNet127 dataset demonstrate that the proposed CReST and CReST+ improve baseline SSL algorithms by a large margin, and consistently outperform state-of-the-art rebalancing methods.

Acknowledgments. We thank Zizhao Zhang, Yin Cui and Prannay Khosla for valuable advice. We thank Alex Kurakin for helping experiments on ImageNet, Jinsung Yoon for the proofread of our manuscript. This work is partially supported by ONR N00014-20-1-2206 to CW and AY.

References

[1] 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 *ICLR*, 2020. 1, 2, 3, 4, 6

[2] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. MixMatch: A holistic approach to semi-supervised learning. In *NeurIPS*, 2019. 1, 2, 5, 6

[3] Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 2018. 1, 2, 6

[4] Jonathon Byrd and Zachary Lipton. What is the effect of importance weighting in deep learning? In *ICML*, 2019. 1, 2, 6

[5] Saul Calderon-Ramirez, Armaghan Moemeni, David Elizondo, Simon Colreavy-Donnelly, Luis Fernando Chavarria-Estrada, and Miguel A Molina-Cabello. Correcting data imbalance for semi-supervised covid-19 detection using x-ray chest images. *arXiv preprint arXiv:2008.08496*, 2020. 1

[6] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *NeurIPS*, 2019. 1, 2, 3, 5

[7] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. SMOTE: synthetic minority oversampling technique. *Journal of Artificial Intelligence Research*, 2002. 2

[8] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *CVPR Workshop*, 2020. 2

[9] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *CVPR*, 2019. 1, 2, 3, 5, 6

[10] Yin Cui, Yang Song, Chen Sun, Andrew Howard, and Serge Belongie. Large scale fine-grained categorization and domain-specific transfer learning. In *CVPR*, 2018. 7

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009. 2, 7

[12] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. 2

[13] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In *NeurIPS*, 2005. 2

[14] Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, 2009. 2

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 7

[16] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. Deep networks with stochastic depth. In *ECCV*, 2016. 2

[17] Minyoung Huh, Pulkit Agrawal, and Alexei A Efros. What makes imagenet good for transfer learning? *arXiv preprint arXiv:1608.08614*, 2016. 2, 7

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

[19] Muhammad Abdullah Jamal, Matthew Brown, Ming-Hsuan Yang, Liqiang Wang, and Boqing Gong. Rethinking class-balanced methods for long-tailed visual recognition from a domain adaptation perspective. In *CVPR*, 2020. 2, 4

[20] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *ICLR*, 2020. 1, 2

[21] Salman H Khan, Munawar Hayat, Mohammed Bennamoun, Ferdous A Sohel, and Roberto Togneri. Cost-sensitive learning of deep feature representations from imbalanced data. *IEEE transactions on neural networks and learning systems*, 2017. 2

[22] 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 *NeurIPS*, 2020. 1, 2, 6

[23] Jaehyung Kim, Jongheon Jeong, and Jinwoo Shin. M2m: Imbalanced classification via major-to-minor translation. In *CVPR*, 2020. 2

[24] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. 2009. 3

[25] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016. 1, 2

[26] Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *ICML Workshop*, 2013. 2, 6

[27] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017. 2

[28] Jialun Liu, Yifan Sun, Chuchu Han, Zhaopeng Dou, and Wenhui Li. Deep representation learning on long-tailed data: A learnable embedding augmentation perspective. In *CVPR*, 2020. 2

[29] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X. Yu. Large-scale long-tailed recognition in an open world. In *CVPR*, 2019. 2, 7

[30] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *ICLR*, 2017. 5

[31] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *ICLR*, 2021. 6

[32] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *TPAMI*, 2018. 2

[33] Avital Oliver, Augustus Odena, Colin A Raffel, Ekin Dogus Cubuk, and Ian Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. In *NeurIPS*, 2018. 5

[34] Antti Rasmus, Mathias Berglund, Mikko Honkala, Harri Valpola, and Tapani Raiko. Semi-supervised learning with ladder networks. In *NeurIPS*, pages 3546–3554, 2015. 1

[35] Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *ICML*, 2018. 2

[36] Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. In *NeurIPS*, 2016. 2

[37] H Scudder. Probability of error of some adaptive pattern-recognition machines. *IEEE Transactions on Information Theory*, 1965. 4

[38] Jun Shu, Qi Xie, Lixuan Yi, Qian Zhao, Sanping Zhou, Zongben Xu, and Deyu Meng. Meta-weight-net: Learning an explicit mapping for sample weighting. In *NeurIPS*, 2019. 2

[39] 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 *NeurIPS*, 2020. 1, 2, 3, 5, 6, 8

[40] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *JMLR*, 2014. 2

[41] Jingru Tan, Changbao Wang, Buyu Li, Quanquan Li, Wanli Ouyang, Changqing Yin, and Junjie Yan. Equalization loss for long-tailed object recognition. In *CVPR*, 2020. 2

[42] Kaihua Tang, Jianqiang Huang, and Hanwang Zhang. Long-tailed classification by keeping the good and removing the bad momentum causal effect. In *NeurIPS*, 2020. 2

[43] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIPS*, 2017. 1, 6

[44] Yiru Wang, Weihao Gan, Jie Yang, Wei Wu, and Junjie Yan. Dynamic curriculum learning for imbalanced data classification. In *ICCV*, 2019. 2

[45] Zhirong Wu, Alexei A Efros, and Stella X Yu. Improving generalization via scalable neighborhood component analysis. In *ECCV*, 2018. 7

[46] Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V. Le. Unsupervised data augmentation for consistency training. In *NeurIPS*, 2020. 1, 2

[47] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *CVPR*, 2020. 1, 2, 3

[48] Yuzhe Yang and Zhi Xu. Rethinking the value of labels for improving class-imbalanced learning. In *NeurIPS*, 2017. 1, 2

[49] David Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *33rd annual meeting of the association for computational linguistics*, 1995. 4

[50] Xi Yin, Xiang Yu, Kihyuk Sohn, Xiaoming Liu, and Manmohan Chandraker. Feature transfer learning for deep face recognition with under-represented data. In *CVPR*, 2018. 2

[51] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *BMVC*, 2016. 5

[52] Boyan Zhou, Quan Cui, Xiu-Shen Wei, and Zhao-Min Chen. Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In *CVPR*, 2020. 1, 2