This WACV 2023 paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Dynamic Re-weighting for Long-tailed Semi-supervised Learning

Hanyu Peng, **Weiguo Pian**, **Mingming Sun**, **Ping Li**
Cognitive Computing Lab
Baidu Research
No.10 Xibeiwang East Road, Beijing 100193, China
10900 NE 8th St. Bellevue, Washington 98004, USA
{hanyu.peng0510,wibergpian,sunmingming01,pingli98}@gmail.com

Abstract

Semi-supervised Learning (SSL) reduces significant human annotations by simply demanding a small number of labelled samples and a large number of unlabelled samples. The research community has often developed SSL regarding the nature of a balanced data set; in contrast, real data is often imbalanced or even long-tailed. The need to study SSL under imbalance is therefore critical. In this paper, we essentially extend FixMatch (a SSL method) to the imbalanced case. We find that the unlabeled data is as well highly imbalanced during the training process; in this respect we propose a re-weighting solution based on the effective number. Furthermore, since prediction uncertainty leads to temporal variations in the number of pseudo-labels, we are innovative in proposing a dynamic re-weighting scheme on the unlabeled data. The simplicity and validity of our method are backed up by experimental evidence. Especially on CIFAR-10, CIFAR-100, ImageNet127 data sets, our approach provides the strongest results against previous methods across various scales of imbalance.

1. Introduction

It is known that deep learning is data-starving [13, 9, 27, 23, 5, 11]. Its overwhelming success in various fields was made possible by the massive amount of labeled data [8, 23, 13, 9, 5], which is typically manually costly. Such a boom brings a tremendous expense that might be unaffordable in some areas, such as for video annotation [40, 24, 51, 10] and natural language processing [44, 32, 35] which requires high semantic richness. In view of emerging from the dilemma, semi-supervised learning (SSL) is an extremely promising solution. Its strength resides in the possibility of delivering impressive performance with only a limited amount of labeled data, together with a large amount of unlabeled data, sometimes even matching that of supervised algorithms [2, 1, 37]. Cutting-edge SSL routines are consistency-based algorithms that rely on data augmentation, with the goal of minimizing the distance between augmented samples and the raw samples measured in the output space [2, 1, 12]. Among the representative algorithms are FixMatch [37], MixMatch [2], ReMixMatch [1], etc.

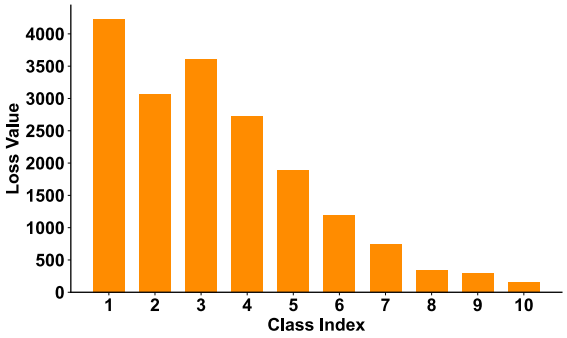


Figure 1: The sum of training losses value for each class of unlabeled data on CIFAR-10 data set with extremely imbalanced training data.

For the most part, SSL algorithms are structured around the assumption that labeled and unlabeled data are well-balanced [42]. Unfortunately, however, real-world data usually follows a long-tailed distribution [3, 19, 15, 18] (The same can be claimed for being imbalanced). It is mainly the case if the model is developed on an imbalanced training set, but validated on a balanced testing set [19, 18, 15, 25]. The performance of the minority class will be abysmal, whereas the performance of the majority class is somehow more

The work of Weiguo Pian was conducted in 2021 at Baidu. He is currently a PhD student at the University of Luxembourg.

favorable. Worse still, the issue will be amplified in SSL, as the number of labeled samples in SSL is relatively small compared to unlabeled samples. There is still little associated work in this area, and this little, or hardly any, consideration [20, 42] leaves us actively pondering how to resolve the issue.

This paper focuses on a trendy SSL model, FixMatch, and we dive into the qualities of unlabeled data that FixMatch presents during its training process: imbalance and the temporal variation. While the first point is pretty understandable, given that if the labeled data is imbalanced, it is ideally also imbalanced for the labeled data insofar with the unlabeled and labeled data being distributionally aligned. Such imbalance appears not just in the number of each class, but is also evident in the sum of the loss values for each class in the unlabelled data. In Figure 1, we visualize the loss values on unlabeled data of CIFAR-10, and we can spot that the loss values also show an extremely imbalanced pattern. It is envisaged that the optimization of the model will be strongly orientated towards the head classes whilst the tail classes tend to be more rarely optimized. There is another unexpected quality that we find concerning temporal variation. In this context, we suggest borrowing the concept of effective number to reweight the loss functions of both labeled and unlabeled data. Nevertheless, owing to the prediction uncertainty in the unlabeled data, pseudo-label will also differ across training epochs. According to this characteristic, we propose a pseudo number and reformulate the effective number within the meaning of unlabeled data. Along with the experimental results testifying our method's validity, the best results so far have been achieved.

In all, **our contributions** include the following:

- We find prediction uncertainty leads to temporal variations for unlabeled data. So we propose a dynamic re-weighting scheme on the unlabeled data.
- On the ground of effective number, we offer to reweight both labeled and unlabeled data. Additionally, pseudo numbers in unlabeled data evolve dynamically in time.
- Empirical results highlight more than just the simplicity and power of our approach, also superior performances on the real-world data sets compared to state-of-the-art methods.

2. Literature Review

2.1. Imbalanced Classification

Recently, numerous works have focused on learning a robust classification model from the data with the imbalanced distribution. Some of these works try to solve this problem by over-sampling few-shot classes or under-sampling many-shot classes [4, 14], which aims to balance the sampling number over classes. In another way, several works have been proposed to solve this challenge with a robust model directly. Kang *et al.* [19] proposed a two-stage training strategy for representation learning to tackle the imbalanced problem. Jamal *et al.* [15] considered this problem within the perspective of domain adaptation. In parallel, Kang *et al.* [18] further explored the impact of balanced feature spaces when learning from imbalanced data, and proposed a self-supervised learning-based approach to learn more robust representation for imbalanced classification. Besides, some works also try to improve the model's performance when training on imbalanced data equipped with well-designed loss function [50, 3, 6, 33] or meta learning [36, 41] algorithm.

2.2. Semi-supervised Learning

SSL has attracted an amount of attention with the advances in deep learning areas. Unlike conventional supervised learning, it can train a well-generalized model with little manual annotations versus the amount of unlabeled data. Due to its advantage of saving labeling expense, SSL has been applied in several learning tasks, such as image classification [46, 7, 45], object detection [38, 17], segmentation [43, 31], domain adaptation [34, 48]. More recently, SSL methods usually use pseudo-labels which are generated by the model for unlabeled data [2, 1, 37]. Specifically, MixMatch [2] applies mixup [49] for data augmentation for both labeled and unlabeled data. Based on MixMatch, a new approach called ReMixMatch [1] was proposed with an augmentation anchoring and a distribution alignment to improve the performance further. After that, Sohn *et al.* [37] proposed FixMatch, which uses two separate data augmentation methods to ensure consistency regularization. Besides, it also uses pseudo-labels generated from the model for training on unlabeled data. In this paper, we marry SSL to imbalanced cases.

2.3. Semi-supervised Long-tailed Learning

More recently, with the development of SSL and the deeper delving of long-tailed problems, a new problem appeared in the field of SSL, that is, the long-tailed distribution problem in SSL. One of the first methods tried to solve this problem is called DARP [20], which tries to reduce the biases dominated by many shot classes when generating pseudo-labels for the classes with few samples by a distribution aligning refinery approach. After that, Wei *et al.* [42] observed that the

high precision of minority classes, and based on this, they proposed an SSL-based method for long-tailed classification by assigning more confidence for minority classes' pseudo-labels when selecting unlabeled data from the unlabeled set.

3. Preliminary

We provide the notational definition, then the problem description for long-tailed SSL, and then we outline the popular SSL method, FixMatch [37], as a prelude, since we are primarily extending FixMatch to the imbalanced scenario.

3.1. Notations

In this paper, we follow the following notation conventions. Lowercase typeface letters (x) represent scalar, lowercase bold typeface letters ($\boldsymbol{x}$) represent vectors, uppercase bold typeface letters ($\boldsymbol{X}$) represent matrix.

3.2. Problem Description

SSL, as the term implies, its training data set $\mathcal{D}^s = \mathcal{D}^{s,l} \cup \mathcal{D}^{s,u}$, $\mathcal{D}^{s,l} = \{(\boldsymbol{x}_i^{s,l}, y_i^{s,l})\}_{i=1}^{N^{s,l}}$, $\mathcal{D}^{s,u} = \{(\boldsymbol{x}_j^{s,u}, y_j^{s,u})\}_{j=1}^{N^{s,u}}$ usually incorporates both labeled and unlabeled data for C-class classification with input dimension d and their associated labels, where $\mathcal{D}^{s,l}$ is the labeled training set, and $\mathcal{D}^{s,u}$ is the unlabled training set. The high cost of labeling data results in the number of unlabeled data, generally being much greater than those labeled, that is to say, $N^{s,u} \gg N^{s,l}$, we shall indicate the ratio of two by $\gamma = \frac{N^{s,u}}{N^{s,l}}$. To quote more, we denote the number of samples per category for the labeled data in the training data as $N_c^{s,l}$, *i.e.*, $\sum_{c=1}^{C} N_c^{s,l} = N^{s,l}$. Once we get to imbalance and even the long-tail as a nature of the data distribution, the imbalance can be expressed as an metric $R^l = \frac{N_{max}^{s,l}}{N_{min}^{s,l}}$, typically, $R^l \gg 1$, here $N_{max}^{s,l}$ and $N_{min}^{s,l}$ represent the number of samples in the dominating/minority of categories respectively. By contrast, the validation and testing sets $\mathcal{D}^v, \mathcal{D}^t$ are usually evenly distributed. Alternatively we suppose from here, the proportion of imbalance R^u in the unlabeled data is aligned with what is in the labeled data.

3.3. A Quick Recap of FixMatch

FixMatch [37] is a pseudo-labeling-based method for SSL tasks; it applies two different data augmentation methods: weak and strong augmentation, to generate augmented samples for unlabeled data, with the goal to ensure the consistency regularization between them. Briefly speaking, given a mini-batch consisting of labeled data $\mathcal{X} = \{(\boldsymbol{x}_i^l, y_i^l); i \in (1, ..., B^l)\}$ and unlabeled data $\mathcal{U} = \{\boldsymbol{x}_j^u; j \in (1, ..., B^u)\}$, the loss function for labeled data $\mathcal{X}$ can be expressed as:

$$\mathcal{L}_s = \frac{1}{B^l} \sum_{i=1}^{B^l} \mathcal{H}(y_i^l, f(\alpha(\boldsymbol{x}_i^l); \boldsymbol{\theta})), \tag{1}$$

where $\mathcal{H}(\cdot, \cdot)$ represents the cross-entropy loss function, $f(\cdot; \boldsymbol{\theta}) : \mathbb{R}^d \to \mathbb{R}^C$ denotes a neural network parameterized by $\boldsymbol{\theta}$, and $\alpha(\cdot)$ is the weak augmentation method. For the unlabeled data $\mathcal{U}$, an analogous loss function would follow; however, owing to the lack of labels for $\mathcal{U}$, the training cannot be conducted under normal training procedures. To tackle the issue, getting the pseudo-label is the well-known general strategy, the first step in the pseudo-labeling process is to obtain the corresponding prediction via weak argumentation.

$$\tilde{\boldsymbol{y}}_j^u = f(\alpha(\boldsymbol{x}_j^u); \boldsymbol{\theta}), \tag{2}$$

where $\tilde{\boldsymbol{y}}_j^u \in \mathbb{R}^C$ stands for the soft pseudo-label. Then we can simply access the hard pseudo-label $\hat{y}_j^u$ by taking the maximum value of $\boldsymbol{y}_j^u$ as follows:

$$\hat{y}_j^u = \max(\tilde{\boldsymbol{y}}_j^u). \tag{3}$$

Based on this, the unlabeled loss used to calculate the distance between hard pseudo-labels and strongly-augmented soft pseudo-labels can be expressed as:

$$\mathcal{L}_u = \frac{1}{B^u} \sum_{j=1}^{B^u} \mathbb{1}(\hat{y}_j^u \geq \tau) \mathcal{H}(\hat{y}_j^u, f(\mathcal{A}(\boldsymbol{x}_j^u); \boldsymbol{\theta})), \tag{4}$$

where τ is a scalar hyperparameter that is employed as the threshold above which the pseudo-label is retained, $\mathcal{A}(\cdot)$ denotes the strong-augmentation method, and $\mathbb{1}(\cdot)$ indicates whether a condition holds. The final loss function of FixMatch, through the introduction of auxiliary parameter λ, can be provided as:

$$\begin{aligned} \mathcal{L} &= \mathcal{L}_s + \lambda \mathcal{L}_u \\ &= \frac{1}{B^l} \sum_{i=1}^{B^l} \mathcal{H}(y_i^l, f(\alpha(\boldsymbol{x}_i^l); \boldsymbol{\theta})) \qquad (5) \\ &\quad + \lambda \frac{1}{B^u} \sum_{j=1}^{B^u} \mathbb{1}(\hat{y}_j^u \geq \tau) \mathcal{H}(\hat{y}_j^u, f(\mathcal{A}(\boldsymbol{x}_j^u); \boldsymbol{\theta})). \qquad (6) \end{aligned}$$

This concludes the profile of FixMatch. In the expansion of the approach to long-tailed domains, it differs mainly in that the labeled and unlabeled data suffer from imbalanced and even long-tailed characteristics. Note that FixMatch can yield pseudo-label $\hat{y}^u$ for associated sample $\boldsymbol{x}^u$. This makes it easy to count the

number of categories $\hat{N}^{s,u}$ in the unlabeled data, which we call the **pseudo number** in this paper. We define $\hat{R}^u = \frac{\hat{N}^{s,u}_{max}}{\hat{N}^{s,u}_{min}}$, $\hat{N}^{s,u}_{max}$ and $\hat{N}^{s,u}_{min}$ denote the maximum and minimum number of $\hat{N}^{s,u}_c, \forall c := 1, 2, ..., C$.

4. Motivation

In SSL, there are considerably more unlabeled data than labeled data [37, 2, 1, 20, 42]; therefore, unlabeled data is a huge wealth that ought not to be disregarded and is definitely valuable to be explored to optimize model performance. *Due to the imbalanced nature of the unlabeled data as well, it is precisely what motivates us.* However, this problem is pretty much left unresolved in terms of performance. Notwithstanding, we are naturally tempted to wonder: in the presence of imbalance, just what are the exceptional characteristics of the unlabeled data throughout training? As well, the implications of these anomalies for the model. In turn, what should we do to manage the negative effects on performance?

You cannot conceive the response immediately, and on the other hand, the intrinsic qualities can only be understood by trial and error. Towards this goal, we train the Wide ResNet-28-2 [30] for 300 epochs on CIFAR-10 [22] data set with $R^l = 150$. We also log the values of numbers of each class in $\mathcal{D}^{s,u}$ and $\hat{R}^u$ every 5 epochs.

Figure 2 exhibits the numbers of each class in $\mathcal{D}^{s,u}$ and $\hat{R}^u$ evolving with the training step as the training is progressed. Clearly, it can be seen on the left subfigure that the unlabeled data also suffers from *imbalanced nature*. Aside from that, on the right subfigure, it is also interesting to point out that $\hat{R}^u$ *is a dynamic value, at different training steps.* Understandably, there is an imbalanced behavior in the unlabeled data, with the exception that one might be curious as to why $\hat{R}^u$ changes over time. Practically speaking, this is also sensible, seeing as the model parameters are in constant change throughout the training process, so, accordingly, the pseudo-labels of each sample will change as well.

5. Method

There is now time for us to formalize the issue of imbalance in unlabeled data. As a matter of fact, the literature on imbalance is full of potential solutions to this issue. Work of interest in this regard appears in the subsection 2.1. Below, we approach the problem with a simple yet powerful approach. In fact, it is also possible to combine this with other reweighting methods, if the dynamic reweighting feature can be incorporated. Be aware that solutions are numerous, but all are designed to *cope with the imbalance properties in unlabeled data.* To be more specific, we have been inspired by the concept of effective number, as developed in [6]. This concept is defined as follows:

Definition 1 *Effective Number for* $\mathcal{D}^{s,l}$. $e^l_c = (1 - \beta^{N^{s,l}_c})/(1 - \beta)$, *where* $\beta = (N^{s,l}_c - 1)/N^{s,l}_c$.

where e^l_c represents the effective number (expected volume) of class c in labeled set. The above definition states that e^l_c grows exponentially with $N^{s,l}_c$. Moreover, $\beta \in [0, 1)$ shapes the degree of this increase. Figure 3 presents the curve of the change in effective number with β. $\beta = 0$ and $\beta \to 1$ correspond to two extreme cases. To be specific, $\beta = 0$ means that all samples contribute equally, $\beta \to 1$ is equivalent to reweighing by inverse frequency of class numbers. It is possible to strike a balance between these two scenarios by adapting the applicable $\beta \in [0, 1)$ to various tasks and the data sets.

Cui *et al.*[6] explained for this: With an increased sample size, probably the new samples created will be nearly similar to the already existing ones. Besides, the neural networks are trained with a large amount of data augmentation, such as random cropping, rescaling and simple transformations, which are applied to the input data. Under these circumstances, all augmented examples are also considered to be identical to the original examples. Moreover, on this basis, they propose class-balanced loss for a couple of prototypical loss functions, such as cross-entropy loss function, which can be written as:

$$\mathcal{H}(y^{s,l}, \tilde{\boldsymbol{y}}^{s,l}) = -\frac{1}{e^l_c} \log \left(\frac{\exp(\tilde{\boldsymbol{y}}^{s,l}_c)}{\sum_{i=1}^{C} \exp(\tilde{\boldsymbol{y}}^{s,l}_i)} \right). \quad (7)$$

where $\tilde{\boldsymbol{y}}^{s,l} = f(\boldsymbol{x}^{s,l}; \boldsymbol{\theta})$. These conclusions, by the way, only remain valid if the data are labeled. For unlabeled data, we need to recalibrate the strategy to define the effective number of samples. Also we extend the definition in Definition 1 to be given as follows:

Definition 2 *Effective Number for* $\mathcal{D}^{s,u}$. $e^u_c = (1 - \beta^{\hat{N}^{s,u}_c})/(1 - \beta)$, *where* $\beta = (\hat{N}^{s,u}_c - 1)/\hat{N}^{s,u}_c$.

Building on this definition, likewise, we merge the effective number into the loss function with regard to the unlabeled data, namely Eq. (4). A little further, one can readily write the following loss function

$$\mathcal{L}_u = \frac{1}{B^u} \sum_{j=1}^{B^u} \frac{1}{e^u_c} \mathbb{1}(\hat{y}^u_j \geq \tau) \mathcal{H}(\hat{y}^u_j, f(\mathcal{A}(\boldsymbol{x}^u_j); \boldsymbol{\theta})). \quad (8)$$

We can temper the imbalance in unlabeled data by tuning the value of β to cover imbalances of varying

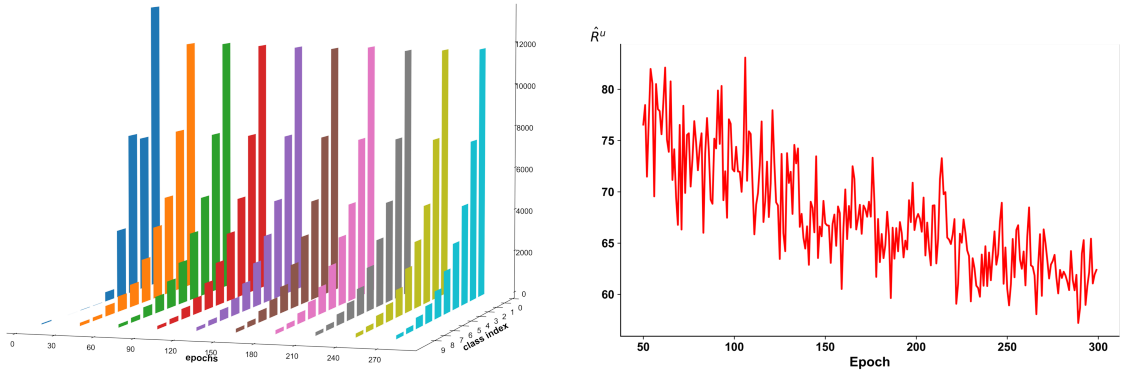


Figure 2: Exhibition on the dynamics of the numbers of each class in $\mathcal{D}^{s,u}$ and $\hat{R}^u$ over training iterations. The left figure shows the pseudo number in each category as the epoch changes throughout the training process. The figure on the right displays the change in the value of $\hat{R}^u$ from epoch 50 to the end of the training period.

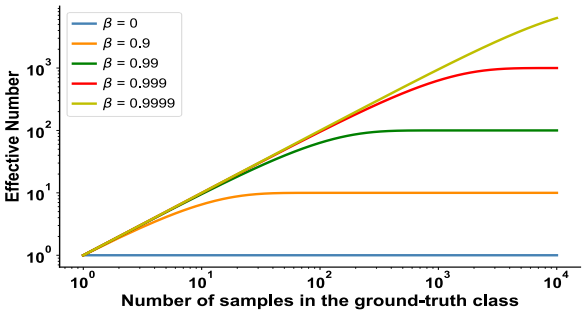


Figure 3: Visualization of class-balanced term at different values of β with changes in the number of samples of the ground-truth class.

degrees. Plus, we mentioned that we earlier observed that pseudo number $\hat{N}_c^{s,u}$ of each class is time-dynamic. Hence accordingly, we would also dynamically compute the value of $\frac{1-\beta}{1-\beta^{\hat{N}_c^{s,u}}}$. With labeled data, we similarly employ the effective number in Definition 1 to tackle the imbalance. *The difference, however, comes from the observation that we have fixed $N_c^{s,l}$ here, without any changes over time.* To conclude, our overall loss function is

$$\mathcal{L} = \frac{1}{B^l}\sum_{i=1}^{B^l}\frac{1}{e_c^l}\mathbb{1}(y_i^l = c)\mathcal{H}(y_i^l, f(\alpha(\boldsymbol{x}_i^l);\boldsymbol{\theta})) \quad (9)$$

$$+\lambda\frac{1}{B^u}\sum_{j=1}^{B^u}\frac{1}{e_c^u}\mathbb{1}(\hat{y}_j^u \geq \tau)\mathcal{H}(\hat{y}_j^u, f(\mathcal{A}(\boldsymbol{x}_j^u);\boldsymbol{\theta})). \quad (10)$$

For the sake of greater visibility of our algorithm, *the entire algorithmic workflow is outlined in Appendix.* So formal that we give the name of our method in this paper to **DRw** (**D**ynamic **R**e-**w**eighting for long-tailed SSL)

5.1. Limitations

We found that the batch size needs to be large, preferably with more labeled data in a batch that can guide the learning of unlabeled data. A batch with smaller size containing less labeled data than necessary can not achieve this goal. Now we have a batch size of 64, and we found that a larger batch size yielded better results, but required more iterations and consumed more memory.

6. Experiments

In this section, we first introduce the settings of the data sets and baselines. Besides, *we also give the implementation details of the baseline methods and our proposed methods in the appendix.* After that, we report the main results of our proposed method on the data set compared with the baselines. Moreover, we have done some parameter studies and ablation studies of our approach. Furthermore, we also conduct the experiment to compare with other typical re-weighting methods, *i.e.* Focal loss. Finally, we give the visualization of the loss values and the testing accuracy.

6.1. Data Sets and Baselines

6.1.1 Data Sets

In our experiments, we generated the long-tailed version of CIFAR-10 [22] and CIFAR-100 [22] as semi-supervised data sets. Specifically, just recall to ourselves, before we denoted the number of the labeled data for class c in the training set as $N_c^{s,l}$, where $c \in \{1, ..., C\}$ and C is the total number of classes. Without loss of generality, we assume that $N_1^{s,l} \geq ... \geq N_C^{s,l}$. Besides, following [6, 20], we determine the number of classes in the labeled set varying imbalance ratio R^l, typically where $R^l \geq 1$. That is, we set the data num-

ber of class c to $N_c^{s,l} = N_1^{s,l} \cdot R^{l-\frac{c-1}{C-1}}$. Naturally the imbalance can be regulated by R^l, thereby accessing the extent to which the model generalizes. For unlabeled data, likewise, we denote the number of class c as $N_c^{s,u}$ where $c \in \{1, ..., C\}$ and we assume that $N_1^{s,u} \geq ... \geq N_C^{s,u}$. As the profile in [6, 20], we also use a parameter R^u, where $R^u \geq 1$, to control the imbalance of unlabeled data in training set, which can be denoted as $N_c^{s,u} = N_1^{s,l} \cdot R^{u-\frac{c-1}{C-1}}$. Here, we set $N_1^{s,l}$ and $N_1^{s,u}$ to 1500 and 3000 respectively for CIFAR-10, and to 150 and 300 for CIFAR-100. Besides, we assume the labeled data and unlabeled data have the same imbalance ratio, so that $R^l = R^u$, and the imbalance ratio is set to 50, 100, 150 for CIFAR-10, and 10, 20 for CIFAR-100. In the testing stage, we use the *balanced accuracy* (bACC) as the evaluation metric. For the definition of bACC, we take binary classification as an example. bACC $= 0.5(t_p/N_p + t_n/N_n)$, Acc $= ((t_p + t_n)/(N_p + N_n))$, N_p and N_n are the numbers of positive and negative samples, while t_p and t_n are the numbers of true positive and true negative. *In the context of imbalanced classification, standard accuracy can be biased to the majority class, therefore bACC is a better indicator.* In addition, we report the mean and standard deviation for our proposed methods, as well as the baselines.

6.1.2 Baseline Methods

We compared our method with 1) **Vanilla:** The basic backbone model without any other techniques. 2) **Re-sampling** [16]**:** A supervised learning method with a re-balancing sampling strategy to make each class equally sampled for training. 3) **LDAM-DRW** [3]**:** A supervised learning re-balancing method with label-distribution-aware margin loss to encourage larger margins for minority classes. 4) **cRT** [19]**:** A two stage training approach to separately learn the representation and the classifier for long-tailed classification. 5) **VAT** [29]**:** A SSL method with adversarial learning. 6) **Mean-Teacher** [39]**:** A SSL approach that uses the ensemble of previous models' weight to construct the teacher model to generate targets for unlabeled data. 7) **MixMatch** [2]**:** A SSL method that applies mixup [49] for data augmentation. 8) **FixMatch** [37]**:** A SSL approach with consistency regularization. 9) **DARP** [20]**:** A SSL method for long-tailed learning using by distribution aligning refinery approach. 10) **CReST** [42]**:** A SSL method for long-tailed learning by giving more confidence to pseudo-labels generated from minority classes.

6.2. Main Results

Table 1 shows the main evaluation results of our approach and the baselines on CIFAR-10 and CIFAR-100 data sets. The results illustrate the superiority of our proposed method over baselines. Specifically, for the results on CIFAR-10 data set, our proposed method achieves the best results under the balance ratios of 50, 100 and 150, with the accuracy of **84.7%**, **79.3%** and **74.1%** respectively, which outperforms the second-best method FixMatch + CReST with the accuracy improvements of **0.8%**, **1.9%** and **1.3%** under the three balance ratios respectively. For the CIFAR-100 data set, our proposed approach also has the best performance with the accuracy of **61.8%** and **57.1%**, under the balance ratios of 10, 20 respectively. Compared with FixMatch + DARP, the second-best method on CIFAR-100, our method has the accuracy improvements of **0.7%** and **2.2%** under the two balance ratios, respectively. *To further compare with CReST and show the scalability of our method, we test the effectiveness of our method by adding LA (Logit Adjustment) [28] as in CReST. We show better performance again with our approach.*

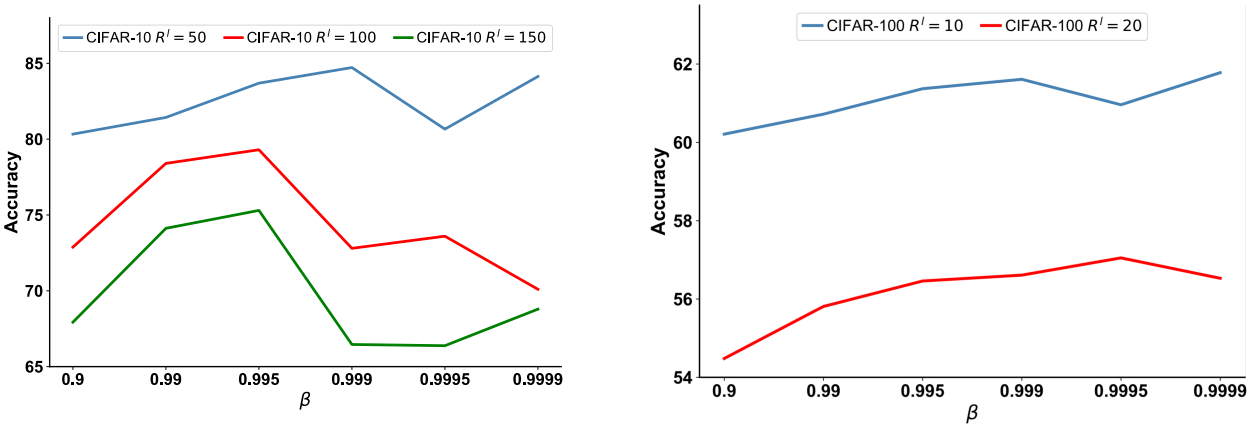
We likewise desire to display the capabilities of our approach on large-scale data sets. Akin to CReST, we carry out experiments on ImageNet127 data set. ImageNet127 consolidates the 1000 classes in ImageNet into 127 classes. We stick to the same network structure, we shall employ instance-wise accuracy on the imbalanced validation set as in CReST. *Please note iNaturalist and ImageNet-LT contain too few examples of minority classes to draw reliable conclusions.* ResNet50 is used as backbone in our experiment, we also apply FixMatch as a fundamental method. 10% of the training samples are marked as labeled data. The results are listed in Table 2, it can be seen that our approach also delivers superior performance.

6.3. Parameter Study

Figure 4a shows the results of the parameter study of our proposed method. As it illustrates, each line denotes the evaluation results of our method regarding β under a fixed balance ratio R^l. Furthermore, we study the results of different balance ratios 50, 100, and 150. From the results, we can see that under the imbalance ratio of 50, our method achieves the best performance when β is set to 0.999, and when the imbalance ratio is set to 100 and 150, our method achieves the best performance with $\beta = 0.995$. We recommend that practitioners choose a larger beta if the β is extremely imbalanced, and a smaller β if the data is less imbalanced. After determining a suitable range for β, we can use the grid search to find the best β on the

Table 1: Comparison with advanced methods on two datasets, CIFAR-10 and CIFAR-100, under different imbalanced ratio R^l's. SSL means whether the algorithm is a semi-supervised learning algorithm, and RB means whether rebalancing techniques are used

Algorithm	SSL	RB	CIFAR-10 ($R^l = R^u$)			CIFAR-100 ($R^l = R^u$)	
			$R^l = 50$	$R^l = 100$	$R^l = 150$	$R^l = 10$	$R^l = 20$
Vanilla	-	-	$65.2_{\pm 0.05}$	$58.8_{\pm 0.13}$	$55.6_{\pm 0.43}$	$55.9_{\pm 0.12}$	$49.5_{\pm 0.03}$
Re-sampling [16]	-	✓	$64.3_{\pm 0.48}$	$55.8_{\pm 0.47}$	$52.2_{\pm 0.05}$	$54.6_{\pm 0.05}$	$48.1_{\pm 0.17}$
LDAM-DRW [3]	-	✓	$68.9_{\pm 0.07}$	$62.8_{\pm 0.17}$	$57.9_{\pm 0.20}$	$55.7_{\pm 0.75}$	$50.4_{\pm 0.32}$
cRT [19]	-	✓	$67.8_{\pm 0.13}$	$63.2_{\pm 0.45}$	$59.3_{\pm 0.10}$	$56.2_{\pm 0.36}$	$50.7_{\pm 0.11}$
VAT [29]	✓	-	$70.6_{\pm 0.29}$	$62.6_{\pm 0.40}$	$57.9_{\pm 0.42}$	$54.6_{\pm 0.06}$	$48.5_{\pm 0.16}$
Mean-Teacher [39]	✓	-	$68.8_{\pm 1.05}$	$60.9_{\pm 0.33}$	$54.5_{\pm 0.22}$	$54.1_{\pm 0.13}$	$48.2_{\pm 0.13}$
MixMatch [2]	✓	-	$73.2_{\pm 0.56}$	$64.8_{\pm 0.28}$	$62.5_{\pm 0.31}$	$60.1_{\pm 0.39}$	$53.4_{\pm 0.04}$
MixMatch + DARP [20]	✓	-	$75.2_{\pm 0.47}$	$67.9_{\pm 0.14}$	$65.8_{\pm 0.52}$	$60.9_{\pm 0.24}$	$54.8_{\pm 0.27}$
MixMatch + CReST [42]	✓	-	$78.4_{\pm 0.36}$	$70.0_{\pm 0.49}$	$64.7_{\pm 0.96}$	-	-
MixMatch + CReST+ [42]	✓	-	$79.0_{\pm 0.26}$	$71.9_{\pm 0.33}$	$68.3_{\pm 0.57}$	-	-
ReMixMatch [1]	✓	-	$81.5_{\pm 0.26}$	$73.8_{\pm 0.38}$	$69.9_{\pm 0.47}$	$59.2_{\pm 0.03}$	$53.5_{\pm 0.03}$
ReMixMatch + DARP [20]	✓	-	$82.1_{\pm 0.14}$	$75.8_{\pm 0.09}$	$71.0_{\pm 0.27}$	$59.8_{\pm 0.20}$	$54.4_{\pm 0.07}$
FixMatch [37]	✓	-	$79.2_{\pm 0.33}$	$71.5_{\pm 0.72}$	$68.4_{\pm 0.15}$	$60.1_{\pm 0.05}$	$54.0_{\pm 0.04}$
FixMatch + DARP [20]	✓	-	$81.8_{\pm 0.24}$	$75.5_{\pm 0.05}$	$70.4_{\pm 0.25}$	$61.1_{\pm 0.23}$	$54.9_{\pm 0.05}$
FixMatch + CReST [42]	✓	-	$83.0_{\pm 0.39}$	$75.7_{\pm 0.38}$	$70.8_{\pm 0.25}$	-	-
FixMatch + CReST+ [42]	✓	-	$83.9_{\pm 0.14}$	$77.4_{\pm 0.36}$	$72.8_{\pm 0.58}$	-	-
FixMatch + CReST+ LA [42]	✓	-	$85.6_{\pm 0.36}$	$81.2_{\pm 0.70}$	$76.5_{\pm 0.40}$	-	-
DRw	✓	-	**84.7**$_{\pm 0.41}$	**79.3** $_{\pm 0.27}$	**74.1** $_{\pm 0.54}$	**61.8** $_{\pm 0.29}$	**57.1** $_{\pm 0.33}$
DRw + LA	✓	-	**86.5**$_{\pm 0.56}$	**82.4** $_{\pm 0.16}$	**77.8** $_{\pm 0.48}$	**62.7** $_{\pm 0.37}$	**58.4** $_{\pm 0.19}$



(a) (b)

Figure 4: Visualization of the change in accuracy with β for both CIFAR-10 and CIFAR-100 data sets.

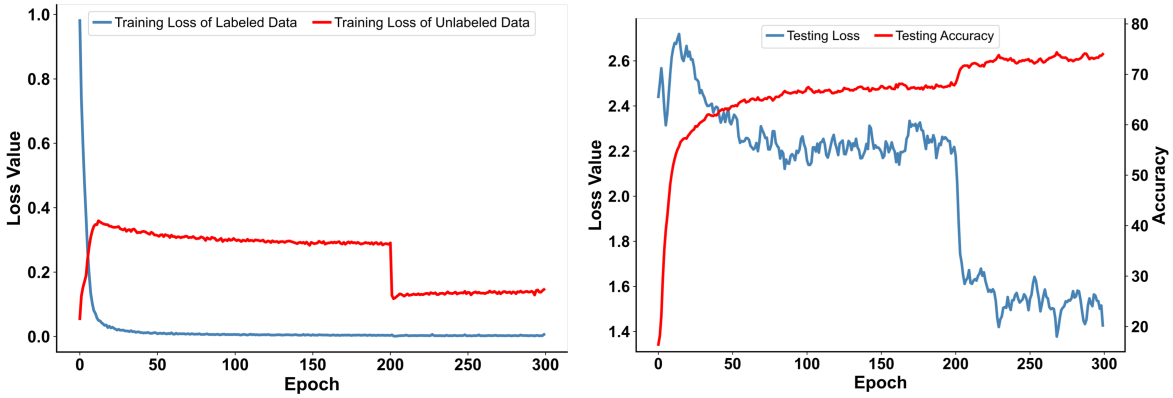
validation data set within this range.

6.4. Compare with Focal Loss

To compare our method with other re-weighting methods, we also conduct experiments with a typical re-weighting method Focal loss [26] on CIFAR-10 data set. The Focal loss can be denoted as $\mathcal{L}_{focal} = -\alpha(1-\hat{y})^{\gamma}\log\hat{y}$, where $\hat{y}$ denotes outputs of the model produced by a *softmax* activation function. α and γ are the hyperparameters to re-weight the positive and negative samples. In our experiments of Focal loss, we fix α to 1.0 and adjust γ with 1 and 2. Table 3 shows the results of Focal loss compared with our method. We can see that our method outperforms Focal loss significantly under different settings with varying imbalance ratios.

6.5. Ablation Study

To evaluate the impacts of different re-weighting strategies, we construct the following variants of our



(a) (b)

Figure 5: (a) Visualization of training loss of labeled and unlabeled data sets on CIFAR-10 data set. (b) Visualization of testing loss and accuracy of unlabeled data set on CIFAR-10 data set. We start performing dynamic re-weighting for unlabeled data from epoch 200

Table 2: Experimental result on ImageNet127 data set, 10% of the training samples are marked as labeled data. ResNet-50 is used as the backbone model.

Data set	Accuracy
FixMatch	65.8
CReST+	73.3
DRw	75.2

Table 3: Comparison with the classical method Focal Loss on the CIFAR-10 data set, with different imbalance coefficients R^l.

Algorithm	CIFAR-10					
	$R^l = 50$		$R^l = 100$		$R^l = 150$	
Focal loss	$\gamma = 1$	$\gamma = 2$	$\gamma = 1$	$\gamma = 2$	$\gamma = 1$	$\gamma = 2$
	80.1	78.0	73.6	71.2	68.8	66.0
DRw	**84.7**		**79.3**		**74.1**	

method: 1) **None Re-weighting (NRw):** a variant without any re-weighting strategy. 2) **Fixed Re-weighting for Labeled Data (FRwL):** a variant with fixed re-weighting strategy for labeled data. 3) **Fixed Re-weighting for Unlabeled Data (FRwU):** a variant with fixed re-weighting strategy for unlabeled data. 4) **Fixed Re-weighting for Both Labeled and Unlabeled Data (FRwLU):** a variant with fixed re-weighting strategy for both labeled and unlabeled data. 5) **Dynamic Re-weighting (DRw):** our proposed dynamic re-weighting method that uses fixed re-weighting for labeled data and dynamic re-weighting for unlabeled data.

Table 4 shows the evaluation results of the variants methods on CIFAR-10 and CIFAR-100 data set. The results illustrate that our proposed DRw achieves the best performance compared with fixed re-weighting and none re-weighting variants.

Table 4: The strength of variants of our method.

Variants	CIFAR-10 (R^l=R^u)			CIFAR-100 (R^l=R^u)	
	R^l=50	R^l=100	R^l=150	R^l=10	R^l=20
NRw	81.8	75.5	70.4	61.1	54.9
FRwL	82.8	76.2	71.5	60.6	55.2
FRwU	81.5	75.8	71.3	60.8	54.9
FRwLU	82.1	77.0	70.7	61.3	54.8
DRw	**84.7**	**79.3**	**74.1**	**61.8**	**57.1**

6.6. Visualization on Loss Values and Accuracy

Figure 5a shows the loss values of both labeled and unlabeled data regarding training epochs on CIFAR-10. From the figure, we can see that the training loss of unlabeled data decreases significantly at epoch 200. Figure 5b illustrates the testing loss and testing accuracy regarding training epochs on CIFAR-10. The testing loss has a significant decrease while the testing accuracy shows an obvious increase at epoch 200.

7. Conclusion

We propose a **D**ynamic **R**e-**w**eighting method called **DRw** for long-tailed semi-supervised method. More specifically, we apply a fixed re-weighting method for labeled data. While for unlabeled data, instead, we propose to utilize the dynamic re-weighting scheme to tackle the temporal variation problem of pseudo-labels, which brings a new solution to this task. Further, we conduct extensive experiments on CIFAR-10, CIFAR-100 and ImageNet127 data sets, the results show the superiority of DRw over existing methods.

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 *International Conference on Learning Representations*, 2020.

[2] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in Neural Information Processing Systems*, 32, 2019.

[3] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in Neural Information Processing Systems*, 32:1567–1578, 2019.

[4] 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.

[5] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[6] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9268–9277, 2019.

[7] Dengxin Dai and Luc Van Gool. Ensemble projection for semi-supervised image classification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2072–2079, 2013.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. IEEE, 2009.

[9] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[10] Wolfgang Fuhl, Nora Castner, Lin Zhuang, Markus Holzer, Wolfgang Rosenstiel, and Enkelejda Kasneci. Mam: Transfer learning for fully automatic video annotation and specialized detector creation. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018.

[11] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012.

[12] Chengyue Gong, Dilin Wang, and Qiang Liu. Alphamatch: Improving consistency for semi-supervised learning with alpha-divergence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13683–13692, 2021.

[13] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[14] Haibo He, Yang Bai, Edwardo A Garcia, and Shutao Li. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*, pages 1322–1328. IEEE, 2008.

[15] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7610–7619, 2020.

[16] Nathalie Japkowicz. The class imbalance problem: Significance and strategies. In *Proc. of the Int'l Conf. on Artificial Intelligence*, volume 56. Citeseer, 2000.

[17] Jisoo Jeong, Seungeui Lee, Jeesoo Kim, and Nojun Kwak. Consistency-based semi-supervised learning for object detection. *Advances in neural information processing systems*, 32:10759–10768, 2019.

[18] Bingyi Kang, Yu Li, Sa Xie, Zehuan Yuan, and Jiashi Feng. Exploring balanced feature spaces for representation learning. In *International Conference on Learning Representations*, 2021.

[19] 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*, 2020.

[20] 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. *Advances in Neural Information Processing Systems*, 33, 2020.

[21] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

[22] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[23] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

[24] Alina Kuznetsova, Aakrati Talati, Yiwen Luo, Keith Simmons, and Vittorio Ferrari. Efficient video annotation with visual interpolation and frame selection guidance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3070–3079, 2021.

[25] Shuang Li, Kaixiong Gong, Chi Harold Liu, Yulin Wang, Feng Qiao, and Xinjing Cheng. Metasaug:

Meta semantic augmentation for long-tailed visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5212–5221, 2021.

[26] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[27] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[28] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. *arXiv preprint arXiv:2007.07314*, 2020.

[29] 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*, 41(8):1979–1993, 2018.

[30] Avital Oliver, Augustus Odena, Colin A Raffel, Ekin Dogus Cubuk, and Ian Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. *Advances in Neural Information Processing Systems*, 31:3235–3246, 2018.

[31] Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12674–12684, 2020.

[32] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[33] Jiawei Ren, Cunjun Yu, Shunan Sheng, Xiao Ma, Haiyu Zhao, Shuai Yi, and Hongsheng Li. Balanced meta-softmax for long-tailed visual recognition. In *Proceedings of Neural Information Processing Systems(NeurIPS)*, 2020.

[34] Kuniaki Saito, Donghyun Kim, Stan Sclaroff, Trevor Darrell, and Kate Saenko. Semi-supervised domain adaptation via minimax entropy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8050–8058, 2019.

[35] Or Sharir, Barak Peleg, and Yoav Shoham. The cost of training nlp models: A concise overview. *arXiv preprint arXiv:2004.08900*, 2020.

[36] 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. *Advances in Neural Information Processing Systems*, 32:1919–1930, 2019.

[37] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in Neural Information Processing Systems*, 33, 2020.

[38] Yuxing Tang, Josiah Wang, Boyang Gao, Emmanuel Dellandréa, Robert Gaizauskas, and Liming Chen. Large scale semi-supervised object detection using visual and semantic knowledge transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2119–2128, 2016.

[39] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in Neural Information Processing Systems*, 30, 2017.

[40] Carl Vondrick, Donald Patterson, and Deva Ramanan. Efficiently scaling up crowdsourced video annotation. *International journal of computer vision*, 101(1):184–204, 2013.

[41] Renzhen Wang, Kaiqin Hu, Yanwen Zhu, Jun Shu, Qian Zhao, and Deyu Meng. Meta feature modulator for long-tailed recognition. *arXiv preprint arXiv:2008.03428*, 2020.

[42] Chen Wei, Kihyuk Sohn, Clayton Mellina, Alan Yuille, and Fan Yang. Crest: A class-rebalancing self-training framework for imbalanced semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10857–10866, 2021.

[43] Yunchao Wei, Huaxin Xiao, Honghui Shi, Zequn Jie, Jiashi Feng, and Thomas S Huang. Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7268–7277, 2018.

[44] Thomas Wolf, Julien Chaumond, Lysandre Debut, Victor Sanh, Clement Delangue, Anthony Moi, Pierric Cistac, Morgan Funtowicz, Joe Davison, Sam Shleifer, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, 2020.

[45] Hao Wu and Saurabh Prasad. Semi-supervised deep learning using pseudo labels for hyperspectral image classification. *IEEE Transactions on Image Processing*, 27(3):1259–1270, 2017.

[46] Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. *Advances in Neural Information Processing Systems*, 33, 2020.

[47] Ziyu Xu, Chen Dan, Justin Khim, and Pradeep Ravikumar. Class-weighted classification: Trade-offs and robust approaches. In *International Conference on Machine Learning*, pages 10544–10554. PMLR, 2020.

[48] Luyu Yang, Yan Wang, Mingfei Gao, Abhinav Shrivastava, Kilian Q Weinberger, Wei-Lun Chao, and Ser-Nam Lim. Deep co-training with task decomposition

for semi-supervised domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8906–8916, 2021.

[49] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018.

[50] Xiao Zhang, Zhiyuan Fang, Yandong Wen, Zhifeng Li, and Yu Qiao. Range loss for deep face recognition with long-tailed training data. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5409–5418, 2017.

[51] Yizhou Zhou, Xiaoyan Sun, Dong Liu, Zhengjun Zha, and Wenjun Zeng. Adaptive pooling in multi-instance learning for web video annotation. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 318–327, 2017.