BaCon: Boosting Imbalanced Semi-supervised Learning via Balanced Feature-Level Contrastive Learning

Qianhan Feng[1,2], **Lujing Xie**[3], **Shijie Fang**[2,4*], **Tong Lin**[1,2†]

[1]National Key Laboratory of General Artificial Intelligence, China
[2]School of Intelligence Science and Technology, Peking University
[3]Yuanpei College, Peking University
[4]Google, Shanghai, China
{fengqianhan, lujing_xie}@stu.pku.edu.cn, shijiefang@google.com, lintong@pku.edu.cn

Abstract

Semi-supervised Learning (SSL) reduces the need for extensive annotations in deep learning, but the more realistic challenge of imbalanced data distribution in SSL remains largely unexplored. In Class Imbalanced Semi-supervised Learning (CISSL), the bias introduced by unreliable pseudo-labels can be exacerbated by imbalanced data distributions. Most existing methods address this issue at instance-level through reweighting or resampling, but the performance is heavily limited by their reliance on biased backbone representation. Some other methods do perform feature-level adjustments like feature blending but might introduce unfavorable noise. In this paper, we discuss the bonus of a more balanced feature distribution for the CISSL problem, and further propose a **Ba**lanced Feature-Level **Con**trastive Learning method (**BaCon**). Our method directly regularizes the distribution of instances' representations in a well-designed contrastive manner. Specifically, class-wise feature centers are computed as the positive anchors, while negative anchors are selected by a straightforward yet effective mechanism. A distribution-related temperature adjustment is leveraged to control the class-wise contrastive degrees dynamically. Our method demonstrates its effectiveness through comprehensive experiments on the CIFAR10-LT, CIFAR100-LT, STL10-LT, and SVHN-LT datasets across various settings. For example, BaCon surpasses instance-level method FixMatch-based ABC on CIFAR10-LT with a 1.21% accuracy improvement, and outperforms state-of-the-art feature-level method CoSSL on CIFAR100-LT with a 0.63% accuracy improvement. When encountering more extreme imbalance degree, BaCon also shows better robustness than other methods.

Introduction

Recently, several Semi-supervised Learning (SSL) methods have been proposed to alleviate the burden of time-consuming data labeling. Most of these methods incorporate unlabeled samples into model training using consistency constraints and pseudo-labeling (Sohn et al. 2020; Zhang et al. 2021). However, these methods are often studied under the assumption of equally distributed unlabeled data, which may not hold true in realistic scenarios. Although some Class Imbalanced Semi-supervised Learning (CISSL) methods have been proposed, most of them resample and redistribute at the instance level, ignoring the upstream representation that has a significant impact. Even if some methods (Fan et al. 2022) make up for this limitation with feature blending from the perspective of representation, it may not only bring more feature noise but also be subject to the spurious prior knowledge of the same distribution of labeled data and unlabeled data. The issue of biased feature representation in the CISSL setting has yet to be thoroughly explored.

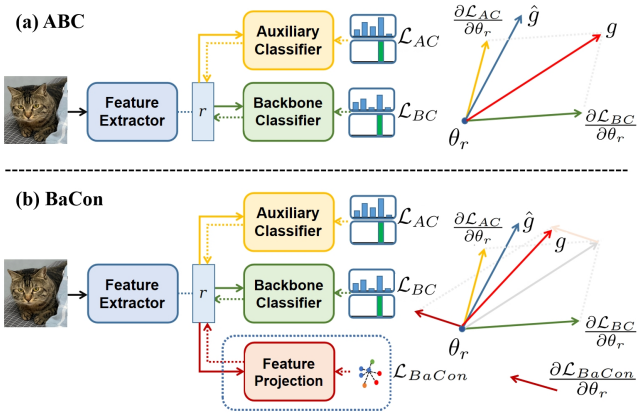


Figure 1: Gradient of ABC on representation layer is still biased. In contrast, BaCon provides an extra gradient that narrows the gap between the estimated gradient g and the ideal optimal gradient $\hat{g}$. r is the feature representation.

In this paper, we first examine the limitations of the state-of-the-art instance-level CISSL method called ABC (Auxiliary Balanced Classifier). ABC adds an auxiliary classification head onto the backbone classifier, and uses inversely proportional Bernoulli mask to achieve balanced learning. There are two sources of gradients that propagate back into the upstream feature extractor: one from the original backbone classifier and the other from the auxiliary classifier. As shown in Figure 1(a), the effect of the combination of these two gradients on the representation layer is still very different from the ideal optimal gradient direction. This

*Work done as a master student at Peking University.
†Corresponding Aurthor.

would result in the uneven distribution of representations, and directly leads to the degraded performance of the overall CISSL method.

Then, we propose to solve this problem from the aspect of feature distribution, and add an additional contrastive loss to directly perform distribution regularization of the feature representations provided by the feature extractor (Figure 1(b)). To be more specific, a projection head is firstly used to map the representation r into another contrastive space, and then the corresponding features of reliable instances are recorded. To implement global positive contrastive learning, the feature centers of each category are calculated to act as the positive anchors in the proposed contrastive loss. Furthermore, a Reliable Negative Selection method (RNS) is designed to easily find adequate and reliable negative samples within a mini-batch for contrastive learning.

However, simply applying equal degree of contrast to each category of imbalanced data sizes cannot fully achieve the desired balanced feature distribution. To tackle this issue, a Balanced Temperature Adjusting mechanism (BTA) is proposed to dynamically adjust the temperature coefficient in the contrastive loss according to the fluctuant distribution, achieving self-adaptive learning. To sum up, our work has the following contributions:

- We discuss the limitations of previous instance-level CISSL methods from the perspective of representation distribution. Based on this, we propose a contrastive learning method to directly regularize feature-level distribution.
- We design an easy and reasonable way to construct positives and negatives for contrastive learning. And a self-adaptive mechanism is proposed to adjust the class-wise learning degree based on the imbalanced distribution.
- Extensive experiments across various datasets and settings demonstrate the effectiveness of our method.

Related Work

Semi-supervised Learning

Semi-supervised Learning has a long history of research (Zhu 2005; Ouali, Hudelot, and Tami 2020). In deep learning, many SSL algorithms have been proposed under the paradigms of mean teacher(Tarvainen and Valpola 2017), pseudo labeling (Lee et al. 2013) and consistency regularization (Berthelot et al. 2019; Xie et al. 2020; Wang et al. 2023). ReMixMatch (Berthelot et al. 2020) introduces Distribution Alignment, which enforces that the aggregate of predictions on unlabeled data matches the distribution of the provided labeled data. In FixMatch (Sohn et al. 2020), the weakly-augmented unlabeled example is first fed to the model to obtain the reliable pseudo-label. Then consistency regularization is performed between the prediction of the strongly-augmented version and the pseudo label. FlexMatch (Zhang et al. 2021) further proposes an adaptive threshold mechanism according to the different learning stages and categories. SimMatch (Zheng et al. 2022) considers semantic similarity and instance similarity simultaneously to encourage consistent prediction. Besides, explicit consistency regularization is also widely explored (Laine and Aila 2017; Miyato et al. 2019; Ganev and Aitchison 2021).

However, these methods are designed under the balanced data distribution, and would fail miserably when encountering with imbalanced training data.

Class Imbalanced Semi-supervised Learning

Recent works have made great progress in addressing the issue of CISSL. DARP (Kim et al. 2020) softly refines the pseudo-labels generated from the biased model by solving a convex optimization problem. Wei et al. (Wei et al. 2021) observe that models typically exhibit high precision but low recall on minority classes, consequently proposing a reverse sampling methodology. The state-of-the-art instance-level method ABC (Lee, Shin, and Kim 2021) utilizes an auxiliary classifier to rebalance the class distribution. Meanwhile, DASO (Oh, Kim, and Kweon 2022) establishes a similarity-based classifier and a linear classifier, blending the pseudo-label in accordance with the pseudo-label distribution. CoSSL (Fan et al. 2022) decouples representation and classifier learning, increases the data diversity of minority classes by feature blending, but it brings a lot of noise and relies on the prior of similar distribution between labeled and unlabeled data.

While most of these methods concentrate on instance-level design, it has been observed that incorporating a self-supervision framework can consistently enhance the final performance (Yang and Xu 2020). We argue that due to the fact that the classifier is biased towards the majority classes, the feature extractor may not learn high-quality balanced representations. Thus, we are inspired to further optimize the representations during training.

Contrastive Learning

Contrastive Learning, successful in self-supervised vision tasks using various strategies (van den Oord, Li, and Vinyals 2018; Chen et al. 2020; Tian, Krishnan, and Isola 2020; He et al. 2020; Caron et al. 2020; Grill et al. 2020; Chen and He 2021). Particularly, the instance discrimination task (Wu et al. 2018) with NCE loss (Gutmann and Hyvärinen 2010) and InfoNCE loss (van den Oord, Li, and Vinyals 2018) have led to milestone works (He et al. 2020; Chen et al. 2020). These works also propose techniques like momentum encoder, memory bank and projection head.

Despite leveraging novel and successful ideas in Self-supervised Learning, Semi-supervised Learning methods have limited access to labeled data. CoMatch (Li, Xiong, and Hoi 2021) regularizes the structure of embeddings to smooth the class probabilities in a contrastive manner. U^2PL (Wang et al. 2022b) makes use of the most reliable predictions to generate positive samples, while using the least reliable predictions to create a negative sample memory bank to address ambiguous boundaries in Semi-supervised Semantic Segmentation. With the help of labeled data, contrastive methods in semi-supervised field can be more natural and general, in contrast to considering a single example as a class in self-supervised methods.

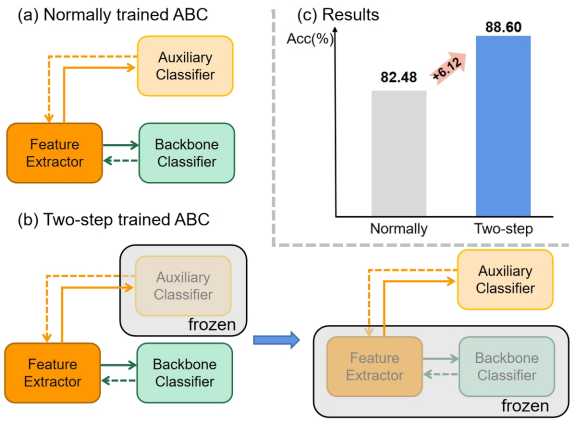


Figure 2: FixMatch-based ABC is trained in two steps as shown in (b) as the comparison of the normally trained one in (a). Remarkable improvement can be observed in (c).

Proposed Method

Problem Setup

For a K-class classification task, suppose we have a labeled dataset $\mathcal{X} = \{(x_n, y_n) : n \in (1, ..., N)\}$, where $x_n \in \mathbb{R}^d$ is the n-th labeled data sample, and $y_n \in \{1, ..., K\}$ is the corresponding label. Additionally, an unlabeled dataset $\mathcal{U} = \{(u_m) : m \in (1, ..., M)\}$ is also provided, where $u_m \in \mathbb{R}^d$ is the m-th unlabeled instance. In SSL setting, only a few data points have accessible labels, while the rest are unlabeled. We denote the ratio of the amount of labeled data as $\beta = \frac{N}{M+N}$, which is generally small since data labeling is expensive. Then, the number of labeled data in the k-th class is denoted as N_k, and $\sum_k^K N_k = N$. Apart from the above setups, the imbalance degree across classes is also a key consideration. We assume that the K classes are sorted in descending order, i.e. $N_1 > N_2 > \cdots > N_K$, and the imbalance degree of the dataset is represented by the imbalance ratio $\gamma = \frac{N_1}{N_K}$. We do not follow prior works in assuming that $\mathcal{X}$ and $\mathcal{U}$ share a similar distribution, this means that any extreme distributions can be possible. For example, the imbalance ratio of labeled and unlabeled data could be even inverse, i.e. $\gamma_L = 1/\gamma_U$. Regarding the test time, we evaluate the effectiveness of our method on a class-balanced dataset.

Motivation

Before introducing our method, we begin by discussing the limitations of the state-of-the-art instance-level approach ABC. ABC introduces an auxiliary classifier besides the one in backbone model, and generates an inversely proportional Bernoulli mask according to the predicted distribution. This mask is applied to the auxiliary classifier to provide a more balanced learning. However, the upstream feature extractor receives an imbalanced gradient from the backbone classifier head simultaneously, which may cause an annoying problem. This conflicting gradient pushes the representation learning away from the optimal gradient direction, thereby constraining the overall performance of ABC.

To investigate the potential benefits of a more balanced representation, we conduct a set of comparison experiments on the standard CIFAR10-LT dataset, which will be introduced in the experiment section. As a baseline, we simultaneously train the backbone model and the auxiliary classifier for 300,000 iterations. In the comparison setting, the backbone model is first trained on a balanced dataset with the same total number of samples as CIFAR10-LT for 300,000 iterations, aiming to learn a more balanced representation parameters. The auxiliary head is then trained on imbalanced CIFAR10-LT for another 300,000 iterations, using the frozen backbone parameters.

The results of this exploration experiment are presented in Figure 2. It is obvious that when the auxiliary classifier learns on an imbalanced representation, only an accuracy of 82.48% is achieved. As a comparison, the accuracy surges to 88.60% when it learns on the frozen balanced representation layer. This intriguing outcome indicates that methods like ABC are considerably limited by the imbalanced learned representation. However, it is impossible to produce balanced instance distribution in practice as we did in former discussion. To overcome this challenge, we propose to directly regularize class-wise feature of the samples to have a more balanced distribution, thus facilitating downstream classification.

Base SSL Algorithm

Our feature-level contrastive learning method can be viewed as a plug-in component, and we design to combine it with the state-of-the-art instance-level method ABC for better performance. ABC attaches an auxiliary classifier on backbone SSL algorithms like FixMatch and ReMixMatch, and we take FixMatch as an example here. Apart from the basic classification loss $\mathcal{L}_S$ calculated from the prediction of labeled data $\alpha(x_b)$, FixMatch also uses the reliable prediction beyond threshold of weakly augmented version of unlabeled input $\alpha(u_b)$ to produce pseudo label $\hat{q}_b$, in order to enable consistency regularization unsupervised loss $\mathcal{L}_U$ on the strongly augmented version of the same input $\mathcal{A}(u_b)$. The $\mathcal{L}_S$ and $\mathcal{L}_U$ can be formulated as:

$$\mathcal{L}_S = \frac{1}{B_l} \sum_{b=1}^{B_l} \mathbf{H}(p_s(y|\alpha(x_b)), y_b), \tag{1}$$

$$\mathcal{L}_U = \frac{1}{B_u} \sum_{b=1}^{B_u} \mathbf{H}(p_s(y|\mathcal{A}(x_b)), \hat{q}_b), \tag{2}$$

where B_l and B_u are the number of labeled and unlabeled data within a mini-batch, and $\mathbf{H}$ is the cross-entropy loss, $p_s(y|\alpha(x_b))$ represents predicted confidence distribution of augmented labeled data $\alpha(x_b)$.

The auxiliary classifier in ABC performs similar supervised and consistency-based unsupervised learning, the only difference is that a Bernoulli mask $\mathcal{M}$ inversely proportional to the predicted category size is applied. The classification loss of auxiliary head on labeled data x_b is:

$$\mathcal{L}_{cls} = \frac{1}{B_l} \sum_{b=1}^{B_l} \mathcal{M}(x_b) \mathbf{H}(p_a(y|\alpha(x_b)), y_b), \tag{3}$$

$$\mathcal{M}(x_b) = \mathcal{B}\left(\frac{N_L}{N_{y_b}}\right), \tag{4}$$

The auxiliary head also makes prediction for weakly augmented version of unlabeled data $\alpha(u_b)$ and strongly augmented version $\mathcal{A}(u_b)$. $\mathcal{B}$ is the Bernoulli distribution generator.

$$\mathcal{L}_{consis} = \frac{1}{B_u}\sum_{b=1}^{B_u} \mathcal{M}(u_b)\mathbf{I}(\max(q_b) > \tau)\mathbf{H}(p_a(y|\mathcal{A}(x_b)), \hat{q}_b), \tag{5}$$

$$\mathcal{M}(u_b) = \mathcal{B}\left(\frac{N_L}{N_{\hat{q}_b}}\right), \tag{6}$$

where $\mathbf{I}$ is the indicator function, $max(q_b)$ is the highest predicted assignment probability for any class, and τ is the confidence threshold set to 0.95 by default.

Finally, the total loss for backbone SSL algorithm with the auxiliary classifier is formulated as:

$$\mathcal{L}_{back} = \mathcal{L}_S + \mathcal{L}_U + \mathcal{L}_{cls} + \mathcal{L}_{consis}. \tag{7}$$

Contrastive Learning

In the Motivation section, we have shown that a more balanced feature distribution can facilitate downstream classification task. Therefore, we propose to directly regularize the learned representation distribution using contrastive learning method.

The feature-level distribution regularization is achieved by a novel contrastive loss. Similar to InfoNCE Loss(van den Oord, Li, and Vinyals 2018), we introduce the balanced feature-level contrastive loss, which is expressed as follows:

$$\mathcal{L}_{BaCon} = -\frac{1}{B}\sum_{k=1}^{K}\sum_{b=1}^{B_k} \log \frac{e^{\langle f_b, Anc_k\rangle/\hat{\tau}}}{e^{\langle f_b, Anc_k\rangle/\hat{\tau}} + \lambda\sum_{q=1}^{B_{\bar{k}}} e^{\langle f_b, f_q\rangle/\tau}}, \tag{8}$$

where B is the size of a mini-batch during training, B_k is the number of samples predicted to belong to class k. Besides, f_b and Anc_k stand for the representation of current instance and the its positive anchor target. $\langle\cdot,\cdot\rangle$ represents the cosine similarity function. $\hat{\tau}$ and τ are temperatures of positive and negative pairs. $\bar{k}$ stands for the subset in mini-batch that confidently does not belong to class k, and $\mathcal{B}_{\bar{k}}$ is its size while f_m is the negative representation. λ is the weight that controls the learning of negative samples. After presenting the loss, we begin to introduce details.

To get the representation of each input instance, we attach a linear projection head directly to the representation layer of the backbone. Feature of each instance is projected into another high-dimension space $\mathbb{D}$ where we perform the contrastive learning:

$$f_b = \mathcal{P}(\mathcal{F}(x_b)), f \in \mathbb{D}, \tag{9}$$

in which $\mathcal{P}$ is the linear projection.

After obtaining new representations, we maintain a memory bank $\mathcal{S}_b$ to store the features, which is updated every time when the data is sampled. In order to make the memory bank stable and reliable, only features of samples whose highest confidence score surpass the predefined threshold τ_{th} can be recorded:

$$f_b \mapsto \mathcal{S}_b, \max(\sigma(\mathcal{H}_A(\mathcal{F}(x_b))) > \tau_{th}, \tag{10}$$

where $\mathcal{H}_A$ is the auxiliary classifier and σ is the $SoftMax$ function. τ_{th} is set to 0.98 by default.

Moreover, another memory bank $\mathcal{S}_K$ is also built to record samples' corresponding category, where the ground-truth label of labeled data and predicted result of unlabeled data from the auxiliary head is saved. Then, we calculate the feature center of each class by averaging representations according to the latest memory banks, taking the feature centers to be the anchor points Anc_k which are the positive learning targets in $\mathcal{L}_{BaCon}$,

$$Anc_k = \frac{1}{N_k}\sum_{n=1}^{N_k} f_n, f_n \in \mathcal{S}_b\{k\}, \tag{11}$$

$\mathcal{S}_b\{k\}$ represents the subset in $\mathcal{S}_b$ that are predicted to belong to class k, and N_k is the size of it.

Reliable Negative Selection

The selection of negative samples is challenging. Under the paradigm of pseudo labeling, if labeled instances are solely used for promoting the accuracy of negative samples, the size of negative batch would be reduced greatly. This could result in huge bias we intend to eliminate originally. On the other hand, if all samples that are not identified as current class are treated as negative samples for the sake of larger size of negatives, a lot of noise would be introduced, which will also make contrastive learning intractable.

To tackle this dilemma, we propose the Reliable Negative Selection (RNS) scheme. In RNS, the representation of labeled data whose predicted confidence is above τ_{th} and belongs to other category rather than class k is selected into reliable negatives.

As for the unlabeled data, RNS first sorts the output confidence scores by descending order and assigns an index $Idx(q)$ to the corresponding category q. Then, RNS searches for the representations of samples whose confidence score for class k is outside of the top n in the queue to be viewed as reliable unlabeled negatives, i.e.

$$f_b \mapsto \bar{k}, Idx(k) > n, \tag{12}$$

where n is set to 3 by default.

In most SSL methods, many data points with low confidence scores are excluded from training, which in turn aggravates the lack of accessible data. Although samples with low confidence of class k cannot directly provide information about what features of class k should be like, they can indeed reliably provide key information about what the feature should not be like. With RNS, we reach a balance point where more data points are included and at the same time reliable contrastive information is also acquired.

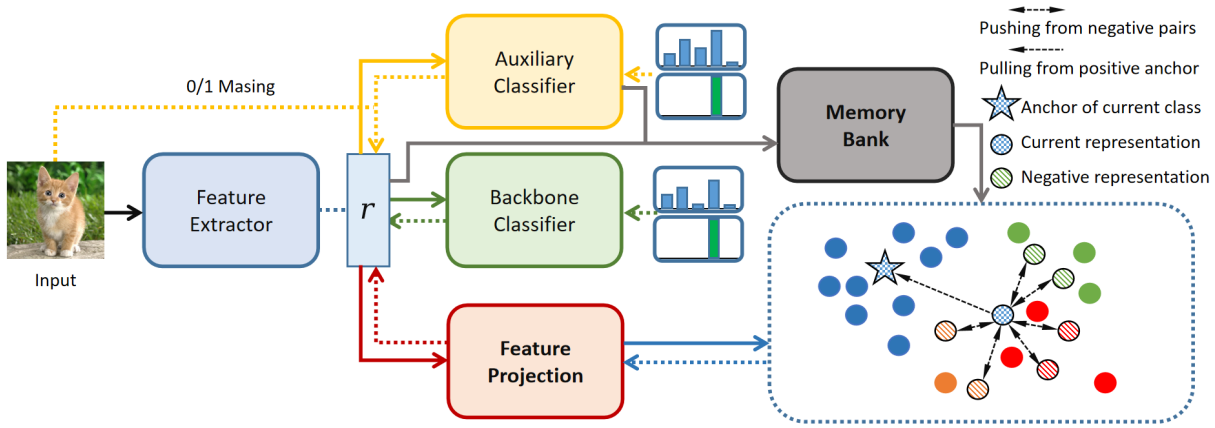


Figure 3: Overall training procedure of BaCon. The circle and star filled with squares represent the current representation and the corresponding positive anchor point, respectively. Circles filled with slashes in different colors represent negative instance features belonging to different classes in the current mini-batch.

However, the size of $\bar{k}$ can fluctuate significantly. To handle this problem, we make the weight λ batchsize-related to stabilize the loss, thus the second term of denominator in Equation (8) can be formulated as:

$$\lambda \sum_{m=1}^{B_{\bar{k}}} e^{\langle f_b, f_m \rangle / \tau} = \frac{B}{B_{\bar{k}}} \sum_{m=1}^{B_{\bar{k}}} e^{\langle f_b, f_m \rangle / \tau}. \tag{13}$$

Balanced Temperature Adjusting

Although the attraction and the repelling in contrastive learning already provide a good regularization for feature distribution, we argue that the degree of class-wise clustering should be different. The attraction from the positive anchors of tail classes should be weaker for the following reason: the feature center might be biased from the ground-truth center because it is calculated by averaging on a small number of instances. Learning towards a less reliable target should be more careful.

With this reasonable analysis, we propose a Balanced distribution-related Temperature Adjusting method (BTA) to perform dynamic class-wise temperature modulation of positive pairs in $\mathcal{L}_{BaCon}$. The ratio of number of each category N_c to the global maximum value $\max\{N_C\}$ is first calculated. In contrastive learning, temperature τ is used to scale the degree of learning. The smaller τ is, the greater the mutual attraction or repulsion would be. So we multiply it by a coefficient that is negatively correlated with the distribution ratio to dynamically control the contrastive learning.

In the later stage of training, the feature centers of different categories should be regionally stable. Therefore, the difference in temperature coefficients should be gradually reduced, so the final modulation can be formulated as:

$$\hat{\tau}_c = \tau \cdot \left[1 - (\frac{t}{T})^2 \cdot \sqrt{\frac{N_c}{\max\{N_C\}}} \cdot \eta \right], N_c \in \{N_C\}, \tag{14}$$

where τ is the default temperature and η controls the sensitivity of the temperature. t is the current iteration number while T is the total training iterations.

Training and Inference

During training, the backbone SSL algorithm and auxiliary classification head are trained first to warmup the memory banks, and the $\mathcal{L}_{BaCon}$ is added to the total loss after warmup. The total loss can be formulated as

$$\begin{aligned} \mathcal{L} &= \mathcal{L}_{back} + \mathbf{1}(t) \cdot \mathcal{L}_{BaCon} \\ &= \mathcal{L}_S + \mathcal{L}_U + \mathcal{L}_{cls} + \mathcal{L}_{consis} + \mathbf{1}(t) \cdot \mathcal{L}_{BaCon}, \end{aligned} \tag{15}$$

where $\mathbf{1}(t)$ is 1 if the warmup is end and otherwise 0.

During inference time, only the prediction of the auxiliary classifier is used to select most likely category.

Experiments

Implement Details

We first construct imbalance datasets based on several benchmark datasets including CIFAR10, CIFAR100 (Krizhevsky, Hinton et al. 2009), STL10 (Coates, Ng, and Lee 2011) and SVHN (Netzer et al. 2011). We consider long-tail (LT) distribution as the imbalance type, where the data size for each category decreases exponentially in order. Combined with the imbalance ratio γ mentioned in problem setup before, the data number of each class can be expressed as $N_k = N_1 \times \gamma^{-\frac{k-1}{K-1}}$, where $\gamma = \frac{N_1}{N_K}$ as mentioned before. To construct standard CIFAR10-LT and SVHN-LT dataset for our main experiment, we set $N_1 = 1000$, $\gamma = 100$ and $\beta = 20\%$. We also try to evaluate our algorithm on a larger imbalance dataset, by following Lee et al. (Lee, Shin, and Kim 2021) to set $N_1 = 200$, $\gamma = 20$ and $\beta = 40\%$ for standard CIFAR100-LT. In building standard STL10-LT dataset, N_1 is set to 150 and $\gamma = 10$ for labeled data, but β is not used here since we do not have access to the ground-truth labels of unlabeled data in STL10's training set.

For base network structure, we follow Lee et al.(Lee, Shin, and Kim 2021) to use Wide ResNet-28-2 (Zagoruyko

Method	CIFAR10-LT $\gamma = 100, \beta = 20\%$	CIFAR100-LT $\gamma = 20, \beta = 40\%$	STL10-LT $\gamma_L = 10$	SVHN-LT $\gamma = 100, \beta = 20\%$
Supervised Only	57.16 ± 0.35	45.72 ± 0.12	45.23 ± 0.23	85.92 ± 0.11
FixMatch(Sohn et al. 2020)	75.30 ± 0.37	53.94 ± 0.09	67.16 ± 0.36	92.63 ± 0.15
w/ DASO(Oh, Kim, and Kweon 2022)	74.78 ± 0.21	54.83 ± 0.19	68.69 ± 0.15	90.24 ± 0.27
w/ DebiasPL(Wang et al. 2022a)	75.25 ± 0.21	54.90 ± 0.11	65.96 ± 0.23	92.42 ± 0.24
w/ CReST(Wei et al. 2021)	76.62 ± 0.12	54.83 ± 0.10	66.45 ± 0.09	93.25 ± 0.07
w/ CReST+PDA(Wei et al. 2021)	78.64 ± 0.40	55.01 ± 0.12	67.17 ± 0.04	93.23 ± 0.12
w/ DARP(Kim et al. 2020)	78.00 ± 0.33	55.63 ± 0.07	62.43 ± 0.10	92.49 ± 0.16
w/ Adsh(Guo and Li 2022)	78.72 ± 0.36	53.97 ± 0.12	70.44 ± 0.09	92.71 ± 0.13
w/ SAW(Lai et al. 2022)	80.12 ± 0.59	55.87 ± 0.05	70.51 ± 0.21	92.92 ± 0.09
w/ ABC(Lee, Shin, and Kim 2021)	83.25 ± 0.77	56.91 ± 0.02	71.23 ± 0.04	94.15 ± 0.04
w/ CoSSL(Fan et al. 2022)	84.09 ± 0.16	57.33 ± 0.05	70.95 ± 0.17	93.39 ± 0.05
w/ BaCon(Ours)	**84.46** ± 0.15	**57.96** ± 0.26	**71.55** ± 0.09	**94.54** ± 0.06
ReMixMatch(Berthelot et al. 2020)	77.96 ± 0.24	56.12 ± 0.12	66.97 ± 0.04	92.35 ± 0.05
w/ DASO(Oh, Kim, and Kweon 2022)	78.86 ± 0.15	57.67 ± 0.20	65.38 ± 0.18	92.49 ± 0.17
w/ DebiasPL(Wang et al. 2022a)	78.14 ± 0.08	56.85 ± 0.19	64.90 ± 0.35	92.41 ± 0.09
w/ CReST(Wei et al. 2021)	79.38 ± 0.17	59.14 ± 0.11	65.56 ± 0.10	93.63 ± 0.06
w/ CReST+PDA(Wei et al. 2021)	79.91 ± 0.20	59.78 ± 0.23	67.57 ± 0.11	93.74 ± 0.15
w/ DARP(Kim et al. 2020)	77.80 ± 0.18	57.21 ± 0.21	65.93 ± 0.16	92.47 ± 0.04
w/ SAW(Lai et al. 2022)	81.71 ± 0.38	32.53 ± 0.67	66.07 ± 0.26	93.42 ± 0.51
w/ ABC(Lee, Shin, and Kim 2021)	84.49 ± 0.24	59.92 ± 0.01	67.24 ± 1.02	94.03 ± 0.18
w/ CoSSL(Fan et al. 2022)	84.93 ± 0.02	**60.46** ± 0.15	68.73 ± 0.77	92.26 ± 0.03
w/ BaCon(Ours)	**85.05** ± 0.09	60.15 ± 0.05	**69.26** ± 0.83	**94.35** ± 0.11

Table 1: Overall results under different imbalance datasets with various semi-supervised learning algorithms. The results are reported according to balance accuracy(%). Labeled data and unlabeled data share the same imbalance degree γ in CIFAR10-LT, CIFAR100-LT and SVHN-LT datasets. But in STL10, only the imbalance ratio γ_L of labeled data is available.

and Komodakis 2016). The projection head is implemented with a single linear layer of 32-dimension. We implement all the algorithms based on USB (Wang, Chen, and Fan 2022) framework and use a single RTX 3090 GPU to train models. SGD is used to optimize parameters. Each mini-batch includes 64 labeled samples and $64\times uratio$ unlabeled samples, and $uratio$ varies for different base SSL algorithms. The learning rate is initially set as $\eta_0 = 0.03$ with a cosine learning rate decay schedule as $\eta = \eta_0 \cos(\frac{7\pi t}{16T})$. The total number od training for each algorithm is 300,000, and the first 100,000 is warmup stage by default. We report the mean balanced accuracy as well as standard deviation of three trials.

Main Results

We conduct main experiments on standard CIFAR10-LT, CIFAR100-LT, STL10-LT and SVHN-LT datasets. We select FixMatch and ReMixMatch as backbone SSL algorithms and apply several edge-cutting CISSL algorithms including ours over them. The results are reported in Table 1. In each dataset, fully supervised method using only labeled data performs poorly. FixMatch and ReMixMatch take a step forward with the help of using unlabeled data, but there is still a lot of room for improvement. Methods like DASO and DebiasPL make progress on some datasets, but do not perform stably well on others. ABC and CoSSL are typical representatives of existing instance-level method and feature-level method respectively, and they steadily outperform the above mentioned methods. However, it is clear that BaCon achieves new state-of-the-art results across multiple settings. For example, BaCon outperforms CoSSL by 0.37% on CIFAR10-LT based on FixMatch. Also, on more challenging STL10-LT, FixMatch-based BaCon reaches an accuracy of 71.55% which exceeds ABC by 0.32%. On CIFAR100-LT with larger number of categories, our FixMatch version method stably obtains the best result of 57.96%, which is 0.63% higher than CoSSL. Based on ReMixMatch, Bacon maintains a gap of 0.53% on STL10-LT compared with CoSSL and 2.02% compared with ABC.

Different Imbalance Degree

The imbalance degree of the dataset is a key factor that affects the performance of CISSL algorithms. Methods that have good results under one imbalance degree may fail to maintain them under a steeper distribution.

Here we evaluate the robustness of our method across different imbalance distribution on CIFAR10-LT and compare with other algorithms. Except for the setting shown in Table 1, a more extreme imbalance ratio $\gamma_L = \gamma_U = 150$ is also implemented. What's more, we also introduce an unusual setting where the imbalance ratios of labeled data and unlabeled data are inversely proportional, to be more specific, $\gamma_L = 1/\gamma_U = 100$ and $\gamma_L = 1/\gamma_U = 150$. Such an unusual setting may appear counter-intuitive at first glance, but it can test the robustness of the algorithms and whether it strongly relies on prior knowledge of the data distribution.

Method	$\gamma_L = 100$ $\gamma_U = 100$	$\gamma_L = 100$ $\gamma_U = 1/100$	$\gamma_L = 150$ $\gamma_U = 150$	$\gamma_L = 150$ $\gamma_U = 1/150$
FixMatch	75.66	56.35	73.45	62.30
w/ CReST+	79.14	66.47	74.51	62.75
w/ ABC	82.48	81.14	79.41	78.84
w/ CoSSL	83.94	71.99	81.83	74.14
w/ BaCon	**84.61**	**83.80**	**81.99**	**82.35**

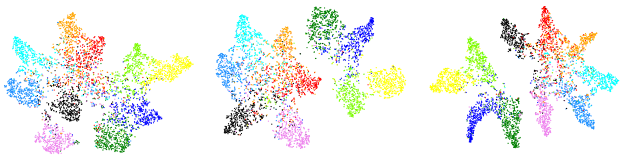
Table 2: Ablation studies on different imbalance degree. γ_L and γ_U represents the imbalance ratio of labeled and unlabeled data respectively. CReST+: CReST+PDA.

Contra	RNS	Naive BTA	Decay BTA	Acc(%)
✓	✓			84.30
✓		✓		83.87
✓	✓	✓		84.15
✓	✓		✓	**84.61**

Table 3: Ablation for proposed components. Contra: Contrastive loss. Naive BTA: BTA without iteration decay.

	Identity	Nonlinear	32-D	128-D	512-D
Acc(%)	82.86	83.95	**84.61**	83.48	82.64

Table 4: Ablation studies of projection mode for contrastive learning space. Identity: identity mapping of representation.



(a) FixMatch (b) FixMatch+ABC (c) Ours

Figure 4: t-SNE visualization results of balanced test set learned by algorithms trained on CIFAR10-LT.

We fix N_1 to be 4000 as same as in Table 1, and only change γ_L and γ_U, while other settings remain unchanged. The results are reported in Table 2. A huge drop in performance can be witnessed on CoSSL when it is trained on a inversely proportional dataset. For example, it achieves an accuracy of 83.94% when $\gamma_L = \gamma_U = 100$, but suffers from a huge decrease of 11.95% to only 71.99% when trained on $\gamma_L = 1/\gamma_U = 100$. ABC shows a better stability, but leaves with some room for improvement. BaCon not only obtains the best results cross these four settings, but also shows a great robustness against the fluctuation of imbalance degree. For example, BaCon produces an accuracy of 83.80% on $\gamma_L = 1/\gamma_U = 100$ and 82.35% on $\gamma_L = 1/\gamma_U = 150$, which are 2.66% and 3.51% higher than the second best.

Ablation Studies

To evaluate the effectiveness of each component we propose in BaCon, we conduct a series of ablation studies on standard CIFAR10-LT based on FixMatch.

We first test the performance of backbone SSL with only an auxiliary classifier attached to it, and then add each component one by one. The results are shown in Table 3. Baseline with the auxiliary classifier achieves an accuracy of 83.95%, and with a contrastive loss that uses RNS the accuracy raises to 84.30%. However, a naive BTA with no iteration decay brings no gain, and the accuracy drops to 83.87% after RNS is abandoned. Finally, when the iteration decay is used in BTA, we witness the best performance of 84.61%. These experiments can verify the effectiveness of our proposed methods.

In addition, we further explore different projection method towards target contrastive space. As comparisons, we choose identity projection, linear projections of different mapping dimensions, and a nonlinear layer constructed by adding $ReLU$ function to the 32-D linear projection. The results are shown in Table 4. It is obvious that the identity projection largely brings down the performance and 32-D linear projection obtains the highest result. Nonlinear projection does not perform well here, and we argue that this is because nonlinear layers can filter out task-related information, and it is more likely to be useful in pretraining task as in (Chen et al. 2020).

Qualitative Analysis

BaCon builds more balanced representations for downstream classification. To verify this, we present the visualization of t-distributed stochastic neighbor embedding (t-SNE) (Van der Maaten and Hinton 2008) for the representations of the balanced CIFAR-10 test set learned by FixMatch, FixMatch-based ABC and FixMatch-based Bacon respectively, and the results are shown in Figure 4. Normal SSL algorithm FixMatch fails to produce separable boundaries. With distribution-related 0/1 mask, ABC presents clearer boundaries than FixMatch, but there is still a great deal of confusion among most categories. BaCon directly implements feature-level clustering, and produces more separable representations. This further proves the gains of our feature-level approach.

Conclusion

In this paper, we discuss the limitation of instance-level CISSL method with biased representation. Next, we propose our feature-level contrastive learning method BaCon to deal with this problem. BaCon projects backbone's representation to another feature space and computes class-wise feature centers as positive anchors. In addition, the proposed RNS method is used to efficiently find sufficient reliable negative samples. Considering the imbalanced distribution, a balanced temperature regulation mechanism is also designed. Finally we show that extensive experiments have verified the effectiveness of our method.

Acknowledgments

This work was supported by National Key R&D Program of China (No.2018AAA0100300) and the funding of China Tower.

References

Berthelot, D.; Carlini, N.; Cubuk, E. D.; Kurakin, A.; Sohn, K.; Zhang, H.; and Raffel, C. 2020. ReMixMatch: Semi-Supervised Learning with Distribution Matching and Augmentation Anchoring. In *International Conference on Learning Representations*.

Berthelot, D.; Carlini, N.; Goodfellow, I.; Papernot, N.; Oliver, A.; and Raffel, C. A. 2019. Mixmatch: A holistic approach to semi-supervised learning. In *Advances in Neural Information Processing Systems 32*, 5050–5060.

Caron, M.; Misra, I.; Mairal, J.; Goyal, P.; Bojanowski, P.; and Joulin, A. 2020. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems 33, virtual*.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607.

Chen, X.; and He, K. 2021. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 15750–15758.

Coates, A.; Ng, A. Y.; and Lee, H. 2011. An Analysis of Single-Layer Networks in Unsupervised Feature Learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *JMLR Proceedings*, 215–223.

Fan, Y.; Dai, D.; Kukleva, A.; and Schiele, B. 2022. Cossl: Co-learning of representation and classifier for imbalanced semi-supervised learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14574–14584.

Ganev, S.; and Aitchison, L. 2021. Semi-supervised learning objectives as log-likelihoods in a generative model of data curation. arXiv:2008.05913.

Grill, J.-B.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P.; Buchatskaya, E.; Doersch, C.; Avila Pires, B.; Guo, Z.; Gheshlaghi Azar, M.; et al. 2020. Bootstrap your own latent-a new approach to self-supervised learning. In *Advances in Neural Information Processing Systems 33, virtual*.

Guo, L.; and Li, Y. 2022. Class-Imbalanced Semi-Supervised Learning with Adaptive Thresholding. In *International Conference on Machine Learning*, 8082–8094.

Gutmann, M.; and Hyvärinen, A. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 297–304. JMLR Workshop and Conference Proceedings.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9729–9738.

Kim, J.; Hur, Y.; Park, S.; Yang, E.; Hwang, S. J.; and Shin, J. 2020. Distribution Aligning Refinery of Pseudo-label for Imbalanced Semi-supervised Learning. In *Advances in Neural Information Processing Systems 33, virtual*.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning Multiple Layers of Features from Tiny Images. *Technical Report*.

Lai, Z.; Wang, C.; Gunawan, H.; Cheung, S. S.; and Chuah, C. 2022. Smoothed Adaptive Weighting for Imbalanced Semi-Supervised Learning: Improve Reliability Against Unknown Distribution Data. In *International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, 11828–11843.

Laine, S.; and Aila, T. 2017. Temporal Ensembling for Semi-Supervised Learning. In *5th International Conference on Learning Representations*.

Lee, D.-H.; et al. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, 896. Atlanta.

Lee, H.; Shin, S.; and Kim, H. 2021. ABC: Auxiliary Balanced Classifier for Class-imbalanced Semi-supervised Learning. In *Advances in Neural Information Processing Systems 34, virtual*, 7082–7094.

Li, J.; Xiong, C.; and Hoi, S. C. 2021. Comatch: Semi-supervised learning with contrastive graph regularization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9475–9484.

Miyato, T.; Maeda, S.; Koyama, M.; and Ishii, S. 2019. Virtual Adversarial Training: A Regularization Method for Supervised and Semi-Supervised Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8): 1979–1993.

Netzer, Y.; Wang, T.; Coates, A.; Bissacco, A.; Wu, B.; and Ng, A. Y. 2011. Reading Digits in Natural Images with Unsupervised Feature Learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*.

Oh, Y.; Kim, D.; and Kweon, I. S. 2022. DASO: Distribution-Aware Semantics-Oriented Pseudo-label for Imbalanced Semi-Supervised Learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA*, 9776–9786.

Ouali, Y.; Hudelot, C.; and Tami, M. 2020. An Overview of Deep Semi-Supervised Learning. arXiv:2006.05278.

Sohn, K.; Berthelot, D.; Carlini, N.; Zhang, Z.; Zhang, H.; Raffel, C.; Cubuk, E. D.; Kurakin, A.; and Li, C. 2020. FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence. In *Advances in Neural Information Processing Systems, Virtual*.

Tarvainen, A.; and Valpola, H. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems 30*, 1195–1204.

Tian, Y.; Krishnan, D.; and Isola, P. 2020. Contrastive multiview coding. In *European Conference on Computer Vision*, 776–794.

van den Oord, A.; Li, Y.; and Vinyals, O. 2018. Representation Learning with Contrastive Predictive Coding. arXiv:1807.03748.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Wang, X.; Wu, Z.; Lian, L.; and Yu, S. X. 2022a. Debiased Learning from Naturally Imbalanced Pseudo-Labels. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14627–14637.

Wang, Y.; Chen, H.; and Fan, Y. e. a. 2022. USB: A Unified Semi-supervised Learning Benchmark for Classification. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Wang, Y.; Chen, H.; Heng, Q.; Hou, W.; Fan, Y.; Wu, Z.; Wang, J.; Savvides, M.; Shinozaki, T.; Raj, B.; Schiele, B.; and Xie, X. 2023. FreeMatch: Self-adaptive Thresholding for Semi-supervised Learning. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Wang, Y.; Wang, H.; Shen, Y.; Fei, J.; Li, W.; Jin, G.; Wu, L.; Zhao, R.; and Le, X. 2022b. Semi-Supervised Semantic Segmentation Using Unreliable Pseudo-Labels. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4238–4247.

Wei, C.; Sohn, K.; Mellina, C.; Yuille, A. L.; and Yang, F. 2021. CReST: A Class-Rebalancing Self-Training Framework for Imbalanced Semi-Supervised Learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, 10857–10866.

Wu, Z.; Xiong, Y.; Yu, S. X.; and Lin, D. 2018. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3733–3742.

Xie, Q.; Dai, Z.; Hovy, E. H.; Luong, T.; and Le, Q. 2020. Unsupervised Data Augmentation for Consistency Training. In *Advances in Neural Information Processing Systems, Virtual*.

Yang, Y.; and Xu, Z. 2020. Rethinking the value of labels for improving class-imbalanced learning. In *Advances in Neural Information Processing Systems 33, virtual*.

Zagoruyko, S.; and Komodakis, N. 2016. Wide Residual Networks. In *Proceedings of the British Machine Vision Conference 2016*.

Zhang, B.; Wang, Y.; Hou, W.; Wu, H.; Wang, J.; Okumura, M.; and Shinozaki, T. 2021. FlexMatch: Boosting Semi-Supervised Learning with Curriculum Pseudo Labeling. In *Advances in Neural Information Processing, Virtual*, 18408–18419.

Zheng, M.; You, S.; Huang, L.; Wang, F.; Qian, C.; and Xu, C. 2022. Simmatch: Semi-supervised learning with similarity matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14471–14481.

Zhu, X. J. 2005. Semi-supervised learning literature survey. *Technical Report*.