DC-SSL: Addressing Mismatched Class Distribution in Semi-supervised Learning

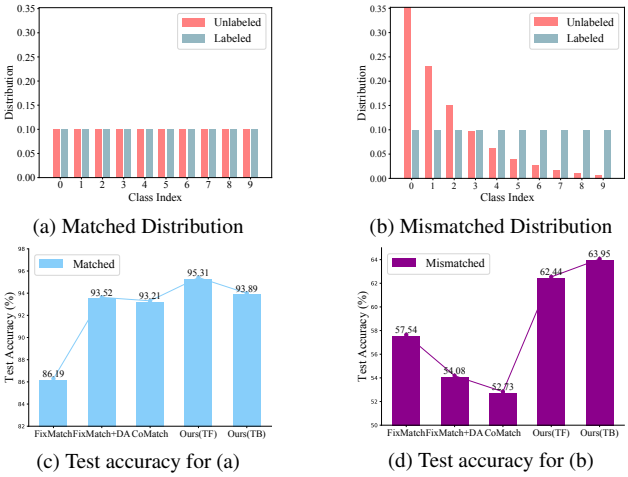
Zhen Zhao[1] Luping Zhou[1]* Yue Duan[2] Lei Wang[3] Lei Qi[4] Yinghuan Shi[2]*
[1]University of Sydney [2]Nanjing University [3]University Of Wollongong [4]Southeast University

Abstract

Consistency-based Semi-supervised learning (SSL) has achieved promising performance recently. However, the success largely depends on the assumption that the labeled and unlabeled data share an identical class distribution, which is hard to meet in real practice. The distribution mismatch between the labeled and unlabeled sets can cause severe bias in the pseudo-labels of SSL, resulting in significant performance degradation. To bridge this gap, we put forward a new SSL learning framework, named Distribution Consistency SSL (DC-SSL), which rectifies the pseudo-labels from a distribution perspective. The basic idea is to directly estimate a reference class distribution (RCD), which is regarded as a surrogate of the ground truth class distribution about the unlabeled data, and then improve the pseudo-labels by encouraging the predicted class distribution (PCD) of the unlabeled data to approach RCD gradually. To this end, this paper revisits the Exponentially Moving Average (EMA) model and utilizes it to estimate RCD in an iteratively improved manner, which is achieved with a momentum-update scheme throughout the training procedure. On top of this, two strategies are proposed for RCD to rectify the pseudo-label prediction, respectively. They correspond to an efficient training-free scheme and a training-based alternative that generates more accurate and reliable predictions. DC-SSL is evaluated on multiple SSL benchmarks and demonstrates remarkable performance improvement over competitive methods under matched- and mismatched-distribution scenarios.

1. Introduction

Recent consistency-based semi-supervised learning (SSL) methods have seen fast progress and shown competitive performance to supervised learning [21, 22]. These methods commonly utilize the model trained on labeled samples to generate pseudo-labels on unlabeled samples, and then enforce prediction consistency against their corresponding perturbed variants. An implicit assumption in such methods is that the labeled and unlabeled data share the same class distribution. However, such a strong assumption cannot hold in real practice. The scarcity of labeled samples or the sampling errors can inevitably lead to a distribution mismatch between the labeled and unlabeled data. This could, unfortunately, invalidate most of the advanced SSL methods.



(a) Matched Distribution (b) Mismatched Distribution
(c) Test accuracy for (a) (d) Test accuracy for (b)

Figure 1. (a) and (b) show the class distributions on CIFAR10 in the matched and mismatched distributions settings, respectively. (c) and (d) show the corresponding test performance on the recent SOTA SSL methods and our proposed DC-SSL with training-free (TF) and training-based (TB) strategies.

To illustrate this problem, this paper conducted a performance comparison under matched and mismatched distribution scenarios. As shown in Figure 1c, two state-of-the-art (SOTA) SSL methods, FixMatch [26] and CoMatch [18], can achieve promising results on CIFAR-10 with only 40 labeled samples when the labeled and unlabeled class distributions are matched, *e.g.*, a high test accuracy of 93.21% of CoMatch. However, when there exists a distribution mismatch as shown in Figure 1b, the test accuracy can drop sharply by around 30% on FixMatch and severely more than 40% on CoMatch. It is because the pseudo-labels on the unlabeled set are severely biased and

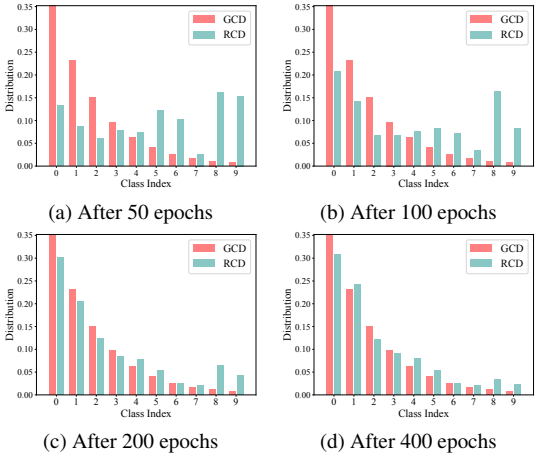
*Corresponding authors (luping.zhou@sydney.edu.au). The authors thank Australian Research Council (ARC DP200103223), National Key Research and Development Program of China (2019YFC0118300), and NSFC Major Program (62192783) for the support of this work.

unreliable in a mismatched distribution setting, resulting in a significant performance degradation.

Inspired by distribution alignment (DA) [4], we aim to improve the biased pseudo-labels from a distribution perspective. The basic logic is to modify the pseudo-labels by encouraging the predicted class distribution (PCD) of the unlabeled data to be close to the underlying ground-truth class distribution (GCD) across the training. However, the existing works using DA [4, 11, 18, 28] widely assumed that the labeled and unlabeled data fall in the same class distribution, and therefore took the provided labeled class distribution (LCD) as the GCD on the unlabeled set to rectify pseudo-labels. As shown in Figure 1c, built into FixMatch, although DA significantly improves the performance in the matched distribution setting (*i.e.* LCD=GCD), it causes severe negative impact under the mismatched distribution scenario (*i.e.* LCD≠GCD) with a sharp accuracy drop as shown in Figure 1d. A key rescue and challenge is to employ an accurate distribution to guide PCD on the unlabeled data, whereas the unlabeled GCD is commonly unknown and the known LCD is biased and unreliable.

To address the above limitations, we propose a simple but effective method, named Distribution Consistency SSL (DC-SSL), which can effectively rectify the pseudo-labels from a distribution perspective. The design of DC-SSL is based on two main components. **First**, instead of using LCD, DC-SSL directly estimates a reference class distribution (RCD) from the unlabeled data, which is regarded as a surrogate of the unknown GCD. To this end, we revisit the exponentially moving averaged (EMA) model in SSL and carefully study i) why the EMA model is employed merely for the testing instead of the training process in recent SOTA SSL methods [1, 13, 14, 18, 26], and ii) how the EMA model can benefit the distribution estimation on unlabeled samples. Based on this investigation, we design our framework to involve EMA to estimate a robust RCD by a momentum-updated scheme over historical label predictions. As shown in Figure 2, the estimated RCD gradually approaches GCD with the progression of the training procedure. **Second**, on top of the estimated distributions, two direct and indirect updating strategies are proposed, respectively, to modify the pseudo-labels, corresponding to the training-free and the training-based strategies shown in Figure 3. The training-free (TF) strategy directly modifies the pseudo-labels by scaling them with a ratio of RCD to PCD, while the training-based (TB) strategy minimizes a distribution consistency loss between PCD and RCD to indirectly enhance the SSL performance. Both strategies are orthogonal to existing consistency-based SSL methods and can be easily applied with minimal change of implementation.

Despite of it simplicity, our method can consistently improve the SOTA SSL methods, especially when the labeled and unlabeled data follow different distributions. For example, in conventional matched distribution settings, DC-SSL (TF) can achieve a higher average accuracy of 95.31% on CIFAR10 (40 labels) compared to the previous SOTA of 93.21% and the baseline FixMatch of 86.19%. In the mismatched settings, our methods consistently outperform other SSL methods, *e.g.*, DC-SSL (TB) can obtain an average accuracy of 63.95% on CIFAR10 in a mismatched setting as in Figure 1b, compared to Fixmatch of 57.54% and CoMatch of 52.73%. Our main contributions are summarized as follows:



(a) After 50 epochs (b) After 100 epochs

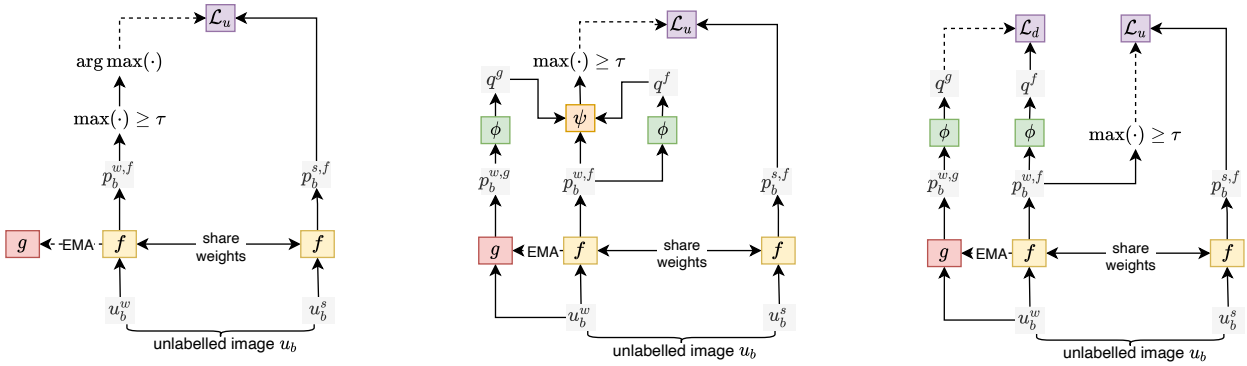
(c) After 200 epochs (d) After 400 epochs

Figure 2. (a)-(d) compares the RCD in DC-SSL (TB) and GCD at different training stages with the mismatched setting in Fig. 1b.

- We revisit the EMA model in SSL and observe that it can be helpful in estimating unlabeled class distributions, although it may not produce more accurate high-confidence pseudo-labels directly.
- We propose a new method, DC-SSL, to enhance SSL performance from a distribution perspective. Two effective strategies are designed to improve the pseudo-labels by encouraging PCD of unlabeled data to approach an iteratively-improved RCD gradually.
- Our method can obtain new SOTA performance across different amounts of labeled data on standard SSL image classification benchmarks under both matched and mismatched distribution scenarios.

2. Related work

Semi-supervised Learning. The key of SSL is to leverage unlabeled data and cooperate with few labeled data to train models. Recent studies have mainly focused on using pseudo-labels for unlabeled data and achieved great success. In specific, self-training-based approaches [6, 17, 19, 23, 30] use the model's predictions on unlabeled data and add high-confidence ones to the labeled data to retrain the model in a two-stage manner. Differently, recent consistency-based approaches [16, 18, 20, 26, 27, 29] can si-



(a) Consistency-based SSL

(b) Training-free (TF) strategy

(c) Training-based (TB) strategy

Figure 3. (a) shows the diagram of FixMatch, a widely adopted consistency-based SSL method. (b) and (c) are our proposed two strategies to enforce distribution consistency on top of FixMatch. Sepcifically, u_b^w and u_b^s are the weakly and strongly augmented variants of an unlabeled image u_b, respectively. f denotes the network model and g is the EMA of f. p is the network's probability prediction and τ is a high-confidence threshold. q represents the class distribution derived from historical predictions by the scheme ϕ. Without introducing new network components, our models estimate class distributions on unlabeled data, and enforce distribution consistency by either the training-free update denoted as ψ in (b) or the training-based consistency loss denoted by $\mathcal{L}_d$ in (c). Dash lines indicate "stop gradient".

multaneously train models on labeled and unlabeled data and achieve competitive performance to supervised learning. The work in [3] initially proposed the idea of consistency regularization, which enforces the prediction consistency on two augmented views derived from the same instance. Early extensions like PI-Model [16] and Mean-Teacher [27] intended to improve the quality of pseudo-labels by saving several checkpoints or maintaining an EMA teacher model. The work in [5] proposed a hybrid framework MixMatch and involved generic regularization techniques like mixup [31]. After that, [29] proved a critical conclusion that using strong data augmentations can significantly promote SSL performance. Later SSL studies like ReMixMatch [4] and noisy-student [30] exploited this finding and integrated techniques like sharpening, and entropy minimization [12] into an unified framework, resulting in better performance. Furthermore, FixMatch [26] inherited previous findings and significantly simplified the hybrid framework, but achieved the state-of-the-art performance. Most recent studies tend to integrate other advanced deep learning techniques into SSL. The work [25] adopted the uncertainty evaluation to further select more accurate pseudo-labels. The work [1] used transfer learning to enhance the SSL performance. Most complicated, the work [18] unified the ideas of consistency regularization, entropy minimization, contrastive learning, distribution alignment, and graph-based SSL, and proposed CoMatch to jointly train two contrastive representations on unlabeled data and smooth the pseudo-labels under the help of a large memory bank. Differently, our methods only require minimal changes to the fundamental consistency-based SSL methods but achieve the new SOTA performance under the same setting (*i.e.*, with matched distribution). The very recent work [28] investigated the imbalanced SSL where both labeled and unlabeled data are long-tailed distributed in a same manner. However, none of these works have studied the situation that the labeled and unlabeled data follow two different class distribution, where such mismatched setting will significantly degrade the SSL performance.

Distribution alignment. Distribution alignment [7] (DA) and has become an important component in the recent state-of-the-art semi-supervised learning (SSL) methods. ReMixMatch [4] was the first one that introduced the idea of distribution alignment in SSL by encouraging the distribution of predictions on unlabeled data to be close to the distribution of provided labeled data or some pre-known distribution. This technique has then been widely utilized in the latest studies for either balanced SSL [11, 18, 26] or imbalanced SSL settings [28]. However, the success of DA relies heavily on a strong assumption that the potential class distribution of the unlabeled data is identical to the marginal class distribution of the provided labeled data. Unfortunately, such an assumption cannot always hold in practice, especially when the amount of labeled data is severely scarce. Therefore, we get rid of this assumption in our methods and propose the distribution estimation directly from the unlabeled data.

3. Method

In an N-class classification task, the labeled data D_x and unlabeled data D_u are given to train a model with the embedding function $f(\cdot)$. In a mini-batch, suppose we have B labeled samples, $\mathcal{X} = \{(x_b, y_b)|(x_b, y_b) \in D_x\}_{b=1}^{B}$, and μB unlabeled samples, $\mathcal{U} = \{u_b|u_b \in D_u\}_{b=1}^{\mu B}$, where μ

represents the size ratio of $\mathcal{U}$ to $\mathcal{X}$. In most SSL studies, the total loss can be formulated as:

$$\mathcal{L} = \mathcal{L}_x(\mathcal{X}) + \lambda_u \mathcal{L}_u(\mathcal{U}), \tag{1}$$

where $\mathcal{L}_x$ is a supervised loss and $\mathcal{L}_u$ is an unsupervised loss within a mini-batch, measured on $\mathcal{X}$ and $\mathcal{U}$ respectively. λ_u is a weighting parameter to balance the relative importance between the labeled and the unlabeled data. Commonly, $\mathcal{L}_x$ can be obtained by

$$\mathcal{L}_x = \frac{1}{B}\sum_{b=1}^{B} H(y_b, f(x_b)), \tag{2}$$

where H denotes the cross entropy loss. Whereas, the form of $\mathcal{L}_u$ depends on specific SSL methods. In this section, we first review how $\mathcal{L}_u$ is formulated in the backbone consistency-based SSL learner, FixMatch. After that, we introduce the crucial components in our method on top of the backbone: RCD estimation and two updating strategies.

3.1. Backbone SSL learner

Recent consistency-based SSL methods typically use weakly-augmented unlabeled images to generate pseudo-labels and enforce consistency against their corresponding strongly-augmented variants. As shown in Figure 3a, u_b^w and u_b^s are obtained through weakly and strongly augmented operations on an unlabeled instance u_b. The weakly augmented operations consists of standard flip-and-shift augmentation strategies, while the strongly augmented operations usually refer to RandAugment [10] or CTAugment [4]. Subsequently, the model f outputs probability predictions $p_b^{w,f}$ and $p_b^{s,f}$ for u_b^w and u_b^s, respectively. As the most simplified but effective consistency-based SSL method, FixMatch [26] adopted a fixed high-confidence threshold to alleviate the confirmation bias [2] of pseudo-labels. Given a predefined high-confidence threshold τ, the unsupervised loss in FixMatch can be calculated as,

$$\mathcal{L}_u = \frac{1}{\mu B}\sum_{b=1}^{\mu B} \mathbf{1}(\max(p_b^{w,f}) \geq \tau) H(\hat{p}_b^{w,f}, p_b^{s,f}), \tag{3}$$

where $\hat{p}_b^{w,f} = \arg\max(p_b^{w,f})$ denotes the hard pseudo-labels (*i.e.*, in a one-hot form) for unlabeled samples, and the operation $\mathbf{1}(\cdot)$ retains the pseudo-labels whose maximum probability is higher than the threshold τ. Besides, an exponential-moving-averaging model g is maintained along with the model f. However, in FixMatch, g is only used for the testing process and independent from the training process, as in many recent SSL methods.

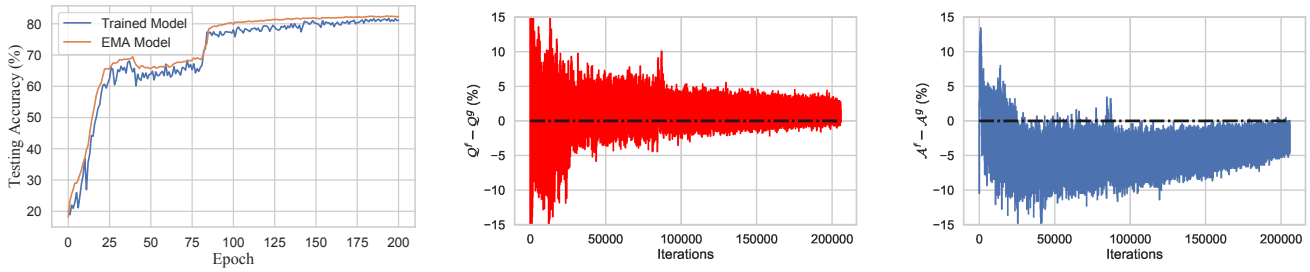
3.2. Distribution Estimations

Properly estimating the class distribution (*i.e.*, frequency of each class on unlabeled data) is the most important problem in our design. Inspired by distribution alignment [4], our primary idea is to encourage the predicted class distribution (PCD) on unlabeled data to be close to the ground-truth class distribution (GCD). However, the lack of label information makes this GCD unknown and challenging to obtain. Almost all existing works, either in balanced SSL [18] or imbalanced SSL tasks [25], adopt the marginal distribution of the provided labeled data as the GCD of the unlabeled data, which will inevitably produce severely biased pseudo-labels, and largely degrade the SSL performance in mismatched distribution settings. Differently, in our work, instead of relying on labeled data, we purely work on unlabeled data to propose a referenced class distribution (RCD) as a surrogate of GCD. Specifically, we carefully involve the EMA model during the training period to estimate the RCD on unlabeled data. As shown in Figure 2, the iteratively-improved RCD can be gradually approaching the GCD across the training process. In this section, we first revisit the EMA model in SSL and then describe the momentum-updated scheme to estimate the distribution from the model's predictions.

3.2.1 Revisiting the EMA model

In the literature, an EMA model with a typical decay of 0.999 is widely adopted in SSL methods for performance enhancement. To investigate its effectiveness, based on FixMatch and using CIFAR-10 with 40 labeled samples, we compare the test accuracies of the trained model f and the EMA model g across different training epochs. As shown in Figure 4a, unsurprisingly, the EMA model g can consistently outperform the trained model f. Based on this, we revisit the EMA model in details by answering two questions the the following.

Question 1: Since the EMA model can achieve a higher test accuracy, will it be beneficial to directly exploit the predictions of the EMA model as pseudo-labels for training? Surprisingly, the answer is NO. In recent SSL studies [4, 13, 18, 26], the EMA model is only used for testing rather than proposing pseudo-labels. However, the potential reasons are not clearly explained in the literature. Thus we perform another experiment to directly use the EMA model's predictions as pseudo-labels. However, this method significantly degrades the SSL performance, achieving a testing accuracy of 45.31% compared to 82.50% of the original FixMatch. We then explore the reasons in term of the accuracy of high-confidence pseudo-labels throughout the training, denoted by $\mathcal{Q}$. As shown in Figure 4b, we measure the accuracy difference of the high-confidence pseudo-labels from f and g throughout a same training process, i.e. $\mathcal{Q}^f - \mathcal{Q}^g$. As seen, $\mathcal{Q}^f$ is higher than $\mathcal{Q}^g$ for above 70% of the training period. Therefore, directly using the EMA model's predictions leads to poor quality of the high-confidence pseudo-labels, which explains why recent SSL



(a) Test accuracy

(b) Accuracy of high-confidence pseudo-labels

(c) Accuracy of all generated pseudo-labels

Figure 4. (a) Comparison of testing accuracy between the trained model f and its corresponding EMA model g. (b) Accuracy difference ($\mathcal{Q}^f - \mathcal{Q}^g$) of the high-confidence pseudo-labels in a mini-batch between f and g at each iteration. Statistically, g obtains a lower accuracy than f at about 70% iterations. (c) Accurate difference ($\mathcal{A}^f - \mathcal{A}^g$) of all pseudo-labels between f and g at each iteration. The model g can generate more accurate pseudo-labels in 96% iterations.

methods exclude the EMA model in the training process.

Question 2: How can our method use the EMA model to estimate a better class distribution on unlabeled data? By further analyzing the above experimental results, we find that, compared with f, although EMA model g obtains a lower accuracy on high-confidence predictions, it can produce a higher accuracy on all unlabeled data (with both high-confidence and low-confidence predictions), *i.e.*, obtaining larger amounts of accurate predictions. Let $\mathcal{A}$ be the pseudo-label accuracy on all unlabeled data in a mini-batch instead of just high-confidence ones. We investigate $\mathcal{A}^f - \mathcal{A}^g$ across the training process in Figure 4c. It is observed that in most iterations, g can achieve a higher value of $\mathcal{A}$ (see the negative values of $\mathcal{A}^f - \mathcal{A}^g$), *i.e.* more accurate predictions. That is indeed what we need for better distribution estimation, since the class distribution ought to be estimated on the whole unlabeled data rather than just the high-confidence ones. Therefore, we can rely on the EMA model's predictions to make a better class distribution estimation on unlabeled data. In the supplement, we also show the same observations with different settings and datasets.

Then can we directly use all predictions from the EMA model as pseudo-labels of the unlabeled data to train models? No, it will also largely decrease the test accuracy due to the well-known issue in SSL, i.e., the confirmation bias [2]. Combining the entropy minimization [12], it is claimed in [26] and [2] that retaining only the pseudo-labels with high-confidence predictions can effectively alleviate the bias. In the following section, we provide our solution to estimate the class distribution by the predictions of EMA model.

In summary, we observe that the EMA model can achieve a higher accuracy of pseudo-labels on all unlabeled data but a lower accuracy on high-confidence ones.

3.2.2 Estimating distribution from predictions

The next problem is how we derive the class distribution from EMA's predictions on unlabeled data. Since the class distribution between different mini-batches can vary considerably, a natural way to improve the estimation is to involve multiple mini-batches. As proposed in ReMixMatch [4], a direct way to estimate the class distribution is to average over historical predictions. However, such a method requires maintaining a memory bank to store the model's predictions from the most recent K mini-batches. More importantly, it ignores temporal differences among historical predictions, i.e., the more recent predictions are more accurate throughout the training. Therefore, we adopt a momentum-updated strategy, denoted by ϕ in Figure 3b and Figure 3c, to estimate the class distribution, requiring calculations only on the current mini-batch. ϕ is essentially a weighted averaging scheme and will assign higher weights on more recent predictions. Given the prediction results $\{p_b^{w,f}\}_{b=1}^{\mu B}$ on the trained model f within a mini-batch, its corresponding class distribution q^f can be estimated as

$$q^f := \alpha\, q^f + \frac{(1-\alpha)}{\mu B} \sum_{b=1}^{\mu B} p_b^{w,f}, \tag{4}$$

where α is a momentum coefficient. In such ways, we cannot only decrease the memory cost but also prioritize the most recent predictions. Likewise, given the EMA model's prediction $\{p_b^{w,f}\}_{b=1}^{\mu B}$, we can obtain another distribution estimation, q^g,

$$q^g := \alpha\, q^g + \frac{(1-\alpha)}{\mu B} \sum_{b=1}^{\mu B} p_b^{w,g}. \tag{5}$$

3.3. Updating strategies

At each mini-batch, we produce two distribution estimations from the unlabeled samples: 1) the predicted class distribution (PCD), q^f, estimated by the trained model via Eq. (4), and 2) the reference class distribution (RCD), q^g, derived by the EMA model via Eq. (5). Based on q^f and q^g, we design two alternative training strategies to improve pseudo-labels either directly or indirectly.

3.3.1 Training-free Strategy

Inspired by ReMixMatch [4], we design a training-free strategy to enhance the quality of pseudo-labels from a distribution perspective. We measure the distribution dissimilarity between RCD and PCD by a ratio q^g/q^f. Then, the training-free strategy, denoted by ψ in Figure 3b, can be performed via two steps: 1) revise the pseudo-label by the distribution dissimilarity ratio, and 2) normalize the revised pseudo-label in a valid probability form. Consequently, the ultimate pseudo-label $\bar{p}_b^{w,f}$ can be calculated as

$$\bar{p}_b^{w,f} = \text{Normalize}(\frac{q^g}{q^f} p_b^{w,f}), \tag{6}$$

where $\text{Normalize}(x_i) = x_i / \sum x_i$. Then the unsupervised loss $\mathcal{L}_u^{tf}$ in this strategy is,

$$\mathcal{L}_u^{tf} = \frac{1}{\mu B} \sum_{b=1}^{\mu B} \mathbf{1}(\max(p_b^{w,f}) \geq \tau) H(\bar{p}_b^{w,f}, p_b^{s,f}). \tag{7}$$

To the end, the total loss for this strategy is $\mathcal{L}_x + \lambda_u \mathcal{L}_u^{tf}$. No additional training efforts are introduced by this strategy.

3.3.2 Training-based Strategy

As shown in Figure 3c, we also propose a training-based strategy to encourage PCD to gradually approach RCD. Specifically, given RCD and PCD, we can minimize a distribution consistency loss $\mathcal{L}_d$:

$$\mathcal{L}_d = H(p^g, p^f), \tag{8}$$

where we use the cross entropy loss $H(\cdot, \cdot)$ to measure the discrepancy between the two distributions. Besides, we also reserve the consistency loss at the instance level,

$$\mathcal{L}_u^{tb} = \frac{1}{\mu B} \sum_{b=1}^{\mu B} \mathbf{1}(\max(p_b^{w,f}) \geq \tau) H(p_b^{w,f}, p_b^{s,f}), \tag{9}$$

where we use the soft pseudo-labels $p_b^{w,f}$ for calculations compared to the hard labels $\hat{p}_b^{w,f}$ used in Eq. (3). In summary, the total loss is,

$$\mathcal{L} = \mathcal{L}_x + \lambda_u \mathcal{L}_u^{tb} + \lambda_d \mathcal{L}_d, \tag{10}$$

where λ_u and λ_d are two weights of the consistency loss at the instance level and at the distribution level, respectively.

Remarks: Our proposed DC-SSL is conceptually analogous to an Expectation-Maximization (EM) procedure. In the E-step, DC-SSL produces distribution estimations p^g and p^f by taking f and g as available models with fixed parameters. In the M-step, DC-SSL updates the models f and g by minimizing the total loss in Eq. (1) or Eq. (10) on top of the two distributions estimated in the E-step. The algorithm can alternately improve the distribution estimations and the trained models.

4. Experiments

This section presents our experimental setup and implementation details, followed by extensive evaluations of our methods with mismatched and matched class distributions.

4.1. Experimental setup

Dataset and Backbone. We evaluate our methods on four SSL image classification benchmarks, CIFAR-10 [15], CIFAR-100 [15], Mini-Imagenet [24], and STL-10 [9]. Of these, CIFAR-10 and CIFAR-100 contain 50,000 32x32 training images and 10000 32x32 testing images, with 10 and 100 classes, respectively. STL-10 is composed of 5,000 labeled images of size 96x96 from 10 classes, along with 10,000 unlabeled images. Mini-Imagenet consists of 50000 training images and 10000 testing images, evenly distributed across 100 classes. For fair comparison [18, 26], we use Wide ResNet-28-2 for CIFAR-10, Wide ResNet-28-8 for CIFAR-100, ResNet-18 for Mini-Imagenet and STL-10, respectively. We use Fixmatch as our backbone (the fundamental consistency-based SSL method) and compare our methods with multiple SSL baselines.

Mismatched Settings. Since the original datasets are all class-balanced, we sample the training images to investigate two mismatched cases: 1) balanced labeled samples with imbalanced unlabeled samples, and 2) balanced unlabeled samples with imbalanced labeled samples. Inspired by CIFAR-LT [8], we utilize an exponential function to mimic the imbalanced distribution. For imbalanced labeled samples, we use $\Gamma_i = \Gamma_0 \gamma_x^{-\frac{i}{N-1}}, i \in [0, N-1]$ to generate the labeled number for the i_{th} class. We use different Γ_0 to investigate different scale of imbalance, while the γ_x is calculated by the constraint $\sum_i \Gamma_i = |D_x|$. On the other hand, we refer to CIFAR-LT [8] to generate imbalanced unlabeled samples, with $M_i = M_{max} \gamma_u^{-\frac{i}{N-1}}$, where M_{max} is set as the image number of the i_{th} class in the original datasets. By adjusting the value of γ_u for difference scales of imbalance, we control the degree of distribution mismatch between the labeled and unlabeled samples, i.e., the larger the γ_u , the higher the severity of distribution mismatch.

Parameters. Our proposed methods introduce two new hyper-parameters: the momentum coefficient α for both strategies and the loss weight λ_d for the training-based (TS) strategy. By default, we simply set $\alpha = 0.999$, and $\lambda_d = 1.0$. Ablation studies on these parameters are provided in the next section. The default values of other training hyper-parameters are $B = 64, \mu = 7, \lambda_u = 1, \tau = 0.9$. We train our methods for 512 epochs and utilize a SGD optimizer with a momentum of 0.9 and a weight decay of 5e-4 to train the model. A learning rate scheduler with a cosine decay is used to decrease the learning rate from an initial value of 0.03. In addition, we train the model for 30 epochs to warm up before applying our proposed distribution con-

Method	CIFAR10,$\|D_x\|$=40			CIFAR10,$\|D_x\|$=250			CIFAR100, $\|D_x\|$=2500		MiniImageNet, $\|D_x\|$=1000	
	$\gamma_u = 50$	100	200	$\gamma_u = 50$	100	200	$\gamma_u = 100$	200	$\gamma_u = 100$	200
FixMatch	57.54	54.82	50.67	76.54	73.51	70.89	52.46	50.24	25.52	21.65
FixMatch+DA	54.08	46.71	41.37	70.78	66.25	61.69	48.96	46.59	22.92	19.82
CoMatch	52.73	46.20	38.85	69.36	64.47	60.05	47.03	43.89	20.37	19.03
Ours (TF)	62.44	56.47	52.32	79.25	76.10	72.01	56.43	52.01	27.44	23.53
Ours (TB)	63.95	57.16	53.27	81.82	77.26	73.34	59.02	52.70	29.12	24.41

Table 1. Mean test accuracy (%) with mismatched class distribution: balanced labeled data and imbalanced unlabeled data. $|D_x|$ is the number of labeled samples. The higher the γ_u, the more the imbalance, and the more severe the distribution mismatch.

sistency.

4.2. Results for mismatched distribution

Imbalanced unlabeled samples. In Tab. 1, we test the performance in a mismatched distribution setting where we have balanced labeled data but imbalanced unlabeled data. It can be clearly seen that, as the γ_u gets larger , *i.e.*, the mismatch issue is more severe, the test accuracy decreases considerably on all SSL benchmarks across different amounts of labeled samples. The mismatched distribution in SSL is a very challenging problem indeed. Compared to other SOTA SSL methods, our methods with either TF or TB strategies can achieve a remarkable performance improvement. In all our tested cases, our TB strategy can boost the mean accuracy of FixMatch by around 3%, and the accuracy of CoMatch by around 11% on average. Interestingly, we find that CoMatch obtains the worst results in all the tests among different baselines. This is because CoMatch extensively exploits the label information carried on the labeled samples to modify the pseudo-labels of unlabeled samples. In addition to the standard DA technique, it maintains a large memory bank to smooth the pseudo-labels by aggregating information from nearby labeled samples in the embedding space. However, relying heavily on labeled samples can only be helpful when the labeled and unlabeled distributions are identical. In the mismatched distribution setting, closely dependending on label information can cause severe negative effects, as can be seen from the test results. Although our methods share a similar idea of the DA to improve pseudo-labels from a distribution perspective, our methods significantly outperform other DA-based baselines (*i.e.*, Fixmqtch+DA and Comatch) due to our proposed better RCD estimated directly on unlabeled samples.

Imbalanced labeled samples. We also investigate another mismatch setting in Tab. 3: imbalanced labeled data but balanced unlabeled data. It can be seen that our methods can effectively improve the performance by rectifying the pseudo-labels from a distribution perspective. The overall results further demonstrate the superiority of our methods, *e.g.*, TB strategy can obtain a mean accuracy of 40.13% on MiniImageNet with imbalanced 1000 labeled samples, against 36.20% of FixMatch and 30.24% of CoMatch.

Method	STL-10
	$\|D_x\|$=1000
FixMatch	65.38
FixMatch+DA	66.53
CoMatch	79.80
Ours (TF)	84.61
Ours (TB)	82.47

Table 2. Mean test accuracy (%) for STL-10 averaged on 5 different folds. All the related works are reported in CoMatch [18].

Observing the results from Tabs. 1 and 3, we can also find that, our TB strategy can mostly achieve better performance than our TF strategy at different degrees of distribution mismatch. This stems from their different levels of influence on the pseudo-labels. The TF strategy can pose strong effects on the pseudo-labels by directly modifying them with a ratio of RCD to PCD. Differently, the TB strategy does not directly adjust the pseudo-labels but indirectly improves the pseudo-labels by enforcing their aggregated distribution to gradually approach the RCD. That is, the TB strategy can take effects in a more moderate manner. In the mismatched case, as shown in Fig. 2, our estimated RCD may not be very accurate at the early stages of the training process, but can be gradually improved to approach the ground-truth distribution across the training process. Therefore, our TB strategy is more suitable for the mismatched cases and can gradually enhance the SSL performance along with the iteratively-improved RCD.

STL10. This dataset contains out-of-distribution images in the unlabeled set, where the distribution mismatch between labeled and unlabeled sets inherently exists. Following [18], we evaluate on the five pre-defined folds and Tab. 2 shows that DC-SSL with both strategies can consistently outperform the SOTA methods, with more than 15% average accuracy improvements against FixMatch and more than 3% improvements against CoMatch.

Method	CIFAR10, $\|D_x\|$=250		CIFAR100, $\|D_x\|$=2500		MiniImageNet, $\|D_x\|$=1000	
	$\Gamma_0 = 100$	200	$\Gamma_0 = 100$	200	$\Gamma_0 = 40$	80
FixMatch	69.76	46.53	61.31	41.38	36.20	28.33
FixMatch+DA	61.80	27.61	50.94	31.82	33.87	23.53
CoMatch	57.87	26.77	48.02	30.08	30.24	21.47
Ours (TF)	72.21	52.59	64.63	41.23	39.07	31.75
Ours (TB)	73.04	48.49	65.24	42.09	40.13	32.82

Table 3. Mean test accuracy (%) with mismatched class distribution: imbalanced labeled data and balanced unlabeled data. $|D_x|$ is the number of labeled samples. The higher the Γ_0, the more the imbalance, and thus the more severe the distribution mismatch.

4.3. Results for matched Distribution

Method	CIFAR10		CIFAR100		MiniImageNet
	$\|D_x\|$=40	250	400	2500	1000
MixMatch [5]	52.46	88.95	33.39	60.06	33.74*
FixMatch [26]	86.19	94.93	51.15	71.71	39.03*
AlphaMatch [11]	91.35	95.03	61.26	74.98	-
CoMatch [18]	93.21*	95.14*	60.71*	74.36*	43.72*
Ours (TF)	95.31	95.87	62.47	75.10	45.19
Ours (TB)	93.89	95.24	61.33	74.62	44.23

Table 4. Mean test accuracy (%) in conventional SSL settings with balanced and matched distributions, *i.e.*, $\Gamma_i = \frac{|D_x|}{N}$ and $\gamma_u = 1$. Results with * in baselines are provided by our own testings.

α	0.8	0.9	0.99	0.999
Accuracy (%)	93.14	94.82	95.38	95.07

Table 5. Effect of the EMA ratio in our TF strategy

In Tab. 4 we also compare our strategies with recent SOTA SSL methods on conventional SSL settings. Following AlphaMatch and CoMatch, we also exploit the pre-known GCD as RCD to test our proposed two strategies. Surprisingly, without introducing more advanced techniques like alpha-divergence or contrastive learning techniques, our two strategies can consistently achieve a higher test accuracy than these SOTA methods, especially when the labeled data is severely scarce. On CIFAR10 with only 40 labels, our TB strategy can obtain a high average accuracy of 95.31%, which is significantly better than 86.19% of FixMatch. It can also be seen from the table that AlphaMatch and CoMatch (both integrating the DA technique) can also achieve remarkable performance gains over FixMatch, demonstrating that modifying the pseudo-labels from a distribution perspective can effectively enhance the SSL performance. Comparing the results in Tab. 1, we further verify our claim that an accurate distribution of unlabeled samples is the key. Unsurprisingly, since we have the accurate distribution information in conventional SSL settings, directly modifying the pseudo-labels in our TB strategy can be more effective than our TF strategy that indirectly improves the pseudo-labels in a more moderate way.

4.4. Effects of Hyper-parameters

λ_d	1.0	3.0	5.0	7.0
TB (matched)	93.92	94.33	94.67	94.81
TB (mismatched)	64.01	62.79	59.03	61.55

Table 6. Effect of the loss weight of $\mathcal{L}_d$ in our TB strategy.

We first examine the effects of two hyper-parameters introduced in our proposed strategies using CIFAR10 with 40 labels in the conventional SSL setting. The momentum coefficient α affects how the class distribution is estimated from historical predictions. A larger value of α can involve more historical predictions and relatively weaken the importance of the current predictions, therefore leading to more stable results as shown in Tab. 5. Meanwhile, the effect of the loss weight λ_d can be seen from Tab. 6: different values of λ_d can slightly affect the accuracy in the matched case while a smaller λ_d can better favor the mismatched case (following the same mismatched distribution setting as in Figure 1b). It is simply because a lower weight can better fit the iteratively-improved RCD and improve the pseudo-labels smoothly. By default, we set $\lambda_d = 1$ in all tests.

5. Conclusion

In this paper, we carefully study how to improve SSL especially when there is a class distribution mismatch between the labeled and unlabeled sets. Our proposed DC-SSL method can improve the pseudo-labels from a distribution perspective and achieves the state-of-the-art performance across many SSL benchmarks under matched and mismatched class distribution scenarios. Thanks to its simplicity, DC-SSL can be easily applied to fundamental consistency-based methods with minor changes.

References

[1] Abulikemu Abuduweili, Xingjian Li, Humphrey Shi, Cheng-Zhong Xu, and Dejing Dou. Adaptive consistency regularization for semi-supervised transfer learning. In *CVPR*, pages 6923–6932, 2021. 2, 3

[2] Eric Arazo, Diego Ortego, Paul Albert, Noel E O'Connor, and Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *2020 International Joint Conference on Neural Networks*. IEEE, 2020. 4, 5

[3] Philip Bachman, Ouais Alsharif, and Doina Precup. Learning with pseudo-ensembles. *NIPS*, 27:3365–3373, 2014. 3

[4] David Berthelot, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. In *8th ICLR*, 2020. 2, 3, 4, 5, 6

[5] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *NIPS*, 2019. 3, 8

[6] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100, 1998. 2

[7] John S Bridle, Anthony JR Heading, and David JC MacKay. Unsupervised classifiers, mutual information and 'phantom targets'. In *NIPS*, 1992. 3

[8] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *arXiv preprint arXiv:1906.07413*, 2019. 6

[9] A. Coates, H. Lee, A. Y. Ng, A. Coates, H. Lee, and A. Y. Ng. An analysis of single-layer networks in unsupervised feature learning. In *Aistats*, 2011. 6

[10] Ekin D. Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V. Le. Randaugment: Practical automated data augmentation with a reduced search space. In *CVPRW*, 2020. 4

[11] Chengyue Gong, Dilin Wang, and Qiang Liu. Alphamatch: Improving consistency for semi-supervised learning with alpha-divergence. In *CVPR*, pages 13683–13692, 2021. 2, 3, 8

[12] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In *CAP*, pages 281–296, 2005. 3, 5

[13] Zijian Hu, Zhengyu Yang, Xuefeng Hu, and Ram Nevatia. Simple: Similar pseudo label exploitation for semi-supervised classification. *arXiv preprint arXiv:2103.16725*, 2021. 2, 4

[14] Byoungjip Kim, Jinho Choo, Yeong-Dae Kwon, Seongho Joe, Seungjai Min, and Youngjune Gwon. Selfmatch: Combining contrastive self-supervision and consistency for semi-supervised learning. *arXiv preprint arXiv:2101.06480*, 2021. 2

[15] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. *Handbook of Systemic Autoimmune Diseases*, 1(4), 2009. 6

[16] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016. 2, 3

[17] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, 2013. 2

[18] Junnan Li, Caiming Xiong, and Steven Hoi. Comatch: Semi-supervised learning with contrastive graph regularization. In *ICCV*, 2021. 1, 2, 3, 4, 6, 7, 8

[19] Geoffrey J McLachlan. Iterative reclassification procedure for constructing an asymptotically optimal rule of allocation in discriminant analysis. *Journal of the American Statistical Association*, 70(350):365–369, 1975. 2

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

[21] Avital Oliver, Augustus Odena, Colin Raffel, Ekin D Cubuk, and Ian J Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. *arXiv preprint arXiv:1804.09170*, 2018. 1

[22] Yassine Ouali, Céline Hudelot, and Myriam Tami. An overview of deep semi-supervised learning. *arXiv preprint arXiv:2006.05278*, 2020. 1

[23] Siyuan Qiao, Wei Shen, Zhishuai Zhang, Bo Wang, and Alan Yuille. Deep co-training for semi-supervised image recognition. In *ECCV*, pages 135–152, 2018. 2

[24] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *5th ICLR*, 2017. 6

[25] Mamshad Nayeem Rizve, Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. *arXiv preprint arXiv:2101.06329*, 2021. 3, 4

[26] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *arXiv preprint arXiv:2001.07685*, 2020. 1, 2, 3, 4, 5, 6, 8

[27] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *arXiv preprint arXiv:1703.01780*, 2017. 2, 3

[28] Chen Wei, Kihyuk Sohn, Clayton Mellina, Alan Yuille, and Fan Yang. Crest: A class-rebalancing self-training framework for imbalanced semi-supervised learning. *arXiv preprint arXiv:2102.09559*, 2021. 2, 3

[29] Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. Unsupervised data augmentation for consistency training. *arXiv preprint arXiv:1904.12848*, 2019. 2, 3

[30] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *CVPR*, pages 10687–10698, 2020. 2, 3

[31] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 3