AllMatch: Exploiting All Unlabeled Data for Semi-Supervised Learning

Zhiyu Wu , Jinshi Cui*

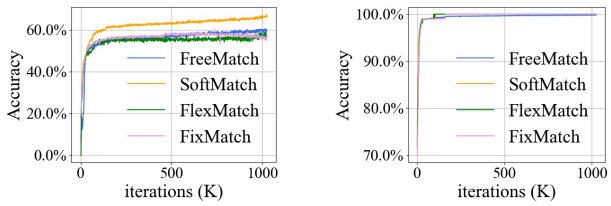
National Key Laboratory of General Artificial Intelligence,
School of Intelligence Science and Technology, Peking University
wuzhiyu@pku.edu.cn, cjs@cis.pku.edu.cn

Abstract

Existing semi-supervised learning algorithms adopt pseudo-labeling and consistency regulation techniques to introduce supervision signals for unlabeled samples. To overcome the inherent limitation of threshold-based pseudo-labeling, prior studies have attempted to align the confidence threshold with the evolving learning status of the model, which is estimated through the predictions made on the unlabeled data. In this paper, we further reveal that classifier weights can reflect the differentiated learning status across categories and consequently propose a class-specific adaptive threshold mechanism. Additionally, considering that even the optimal threshold scheme cannot resolve the problem of discarding unlabeled samples, a binary classification consistency regulation approach is designed to distinguish candidate classes from negative options for all unlabeled samples. By combining the above strategies, we present a novel SSL algorithm named AllMatch, which achieves improved pseudo-label accuracy and a 100% utilization ratio for the unlabeled data. We extensively evaluate our approach on multiple benchmarks, encompassing both balanced and imbalanced settings. The results demonstrate that AllMatch consistently outperforms existing state-of-the-art methods.

1 Introduction

Semi-supervised learning (SSL) [Zhu, 2005; Rosenberg *et al.*, 2005; Berthelot *et al.*, 2019b; Sohn *et al.*, 2020], a research topic that aims to boost the model's generalization performance by leveraging the potential of unlabeled data, has received extensive attention in recent years. Among the proposed techniques, the combination of pseudo-labeling [Lee and others, 2013; Arazo *et al.*, 2020] and consistency regulation [Sajjadi *et al.*, 2016; Laine and Aila, 2016], as introduced by FixMatch [Sohn *et al.*, 2020], has emerged as a predominant approach. Specifically, FixMatch first assigns a pseudo-label to each unlabeled sample based on the prediction of its weakly augmented view. Subsequently, pseudo-labels exceeding a predefined confidence threshold are used as supervision for corresponding strongly augmented views, while those below the threshold are discarded. To ensure high-quality pseudo-labels, FixMatch uses a high constant threshold throughout training. However, this strategy results in the underutilization of unlabeled data, presenting a central challenge in SSL: making efficient use of unlabeled data.



(a) Dropped pseudo-label *acc*. (b) Top-5 *acc* of pseudo-labels.

Figure 1: Pilot study for pseudo-label quality on CIFAR-10 with 40 labels. For SoftMatch, pseudo-labels with a confidence lower than $\mu_t - \sigma_t$ are assigned close-to-zero weights and thus considered dropped in the analysis. Here, μ_t/σ_t denotes the *mean/std* of the overall confidence on unlabeled data in SoftMatch.

To address the trade-off between the quality and quantity of pseudo-labels in threshold-based pseudo-labeling, previous studies introduce dynamic threshold strategies that align with the evolving learning status. For example, FlexMatch [Zhang *et al.*, 2021] utilizes the number of confident pseudo-labels to estimate the learning difficulty for each class, subsequently mapping the predefined threshold to class-specific thresholds based on the determined difficulty levels. Additionally, FreeMatch [Wang *et al.*, 2022] leverages the average confidence of unlabeled data to establish a dynamic global threshold. Besides, SoftMatch [Chen *et al.*, 2023] employs a Gaussian function representing the disparity between individual and global confidence to model sample weights, assigning a positive weight to each unlabeled sample. However, samples with confidence significantly lower than the global confidence receive close-to-zero weights, essentially treated as if they were discarded. This characteristic positions SoftMatch as a variant of the threshold-based method. While these algorithms effectively employ pseudo-labels to assess the learning status, biased data sampling and potential inter-class similarities can significantly influence the predictions of

*Corresponding author

unlabeled data. Consequently, a pivotal question arises: *Can we incorporate additional evidence along with pseudo-labels to achieve a more accurate estimation of the learning status?*

While an improved threshold scheme can enhance the utilization of unlabeled data, a portion of unlabeled samples still face exclusion. This raises another question: *Can pseudo-labels assigned lower confidence provide valuable semantic guidance?* To address this concern, we examine the pseudo-label quality of previous algorithms. In the case of CIFAR-10 [Krizhevsky *et al.*, 2009] with 40 labeled samples, more than half of the dropped pseudo-labels prove to be correct, as depicted in Figure 1(a). Furthermore, Figure 1(b) illustrates that the top-5 accuracy of the pseudo-labels reaches 100% within just a few thousand iterations. Accordingly, pseudo-labels with lower confidence can eliminate false options (*e.g.*, bottom-5 classes) and provide effective supervision signals.

Motivated by the aforementioned questions, this paper introduces AllMatch, a novel SSL model designed to enhance learning status estimation and provide semantic guidance for all unlabeled data. Specifically, AllMatch proposes a class-specific adaptive threshold (CAT) strategy, comprising global estimation and local adjustment steps, to achieve an improved characterization of the model's learning status. The global estimation step, similar to FreeMatch, employs the average confidence of unlabeled data as the global threshold. The ensuing local adjustment step utilizes the classifier weights to estimate the learning status of each class, adaptively decreasing thresholds for classes facing challenges. As illustrated in Figure 4(b, c) and Figure 4(f, g), CAT outperforms previous approaches in terms of the utilization ratio and pseudo-label accuracy of unlabeled samples. Besides, in response to the underutilization of unlabeled data resulting from the exclusion of low-confidence pseudo-labels, AllMatch introduces a binary classification consistency (BCC) regulation strategy to exploit the latent potential within such pseudo-labels. In essence, the BCC regulation divides the class space into candidate and negative classes, encouraging consistent candidate-negative division across diverse perturbed views of the same sample to eliminate negative options. The candidate class for each sample corresponds to its top-k predictions, considering the impressive top-k performance of various algorithms. Note that the parameter k is dynamically determined based on varying sample-specific learning status and evolving model performance. As depicted in Figure 4(c, d) and Figure 4(g, h), the BCC regulation effectively identifies candidate classes for unlabeled samples and achieves a 100% utilization ratio for the unlabeled data. Overall, our contributions can be summarized as follows:

(1) We revisit existing SSL algorithms and raise two questions: how to develop an effective threshold mechanism and how to utilize the low-confidence pseudo-labels.

(2) We propose the class-specific adaptive threshold mechanism, which employs pseudo-labels and classifier weights to estimate global and class-specific learning status respectively.

(3) We design the binary classification consistency regulation to provide supervision signals for all unlabeled samples.

(4) We conduct experiments on multiple benchmarks, considering both balanced and imbalanced settings. The results indicate that AllMatch achieves state-of-the-art performance.

2 Related Work

Consistency regulation represents the fundamental approach in SSL. The primary objective is to ensure consistent predictions across different perturbed views of the same sample. The original approaches [Laine and Aila, 2016; Sajjadi *et al.*, 2016] minimize the L2 distance between the predictions of two randomly augmented views. The Mean Teacher method [Tarvainen and Valpola, 2017] extends this approach by replacing one of the predictions with the output of an exponential moving average (EMA) model. Furthermore, FixMatch [Sohn *et al.*, 2020] combines consistency regulation and pseudo-labeling techniques, establishing an effective SSL paradigm. To avoid similar perturbed views, [Sohn *et al.*, 2020; Xie *et al.*, 2020] use the predictions of weakly augmented samples as pseudo-labels and minimize their divergence from the predictions of the corresponding strongly augmented views. Additionally, these methods use cross-entropy loss instead of the L2 distance for stable training.

To ensure high-quality pseudo-labels, FixMatch [Sohn *et al.*, 2020] utilizes a high constant threshold throughout training to filter out potentially incorrect pseudo-labels. However, this strategy results in the underutilization of unlabeled data. To address this issue, FlexMatch [Zhang *et al.*, 2021] draws inspiration from curriculum learning [Bengio *et al.*, 2009], mapping the predefined threshold to class-specific thresholds based on the learning status of each category. Dash [Xu *et al.*, 2021] defines the threshold based on the loss of labeled data, eliminating the empirical threshold parameter. FreeMatch [Wang *et al.*, 2022] employs the average confidence on unlabeled data as the adaptive global threshold. SoftMatch [Chen *et al.*, 2023] estimates sample weights by a dynamic Gaussian function, maintaining soft margins between unlabeled samples of different confidence levels. Notably, samples with confidence significantly lower than the global confidence are assigned close-to-zero weights in SoftMatch. Accordingly, such samples can be considered dropped, rendering SoftMatch a variant of the threshold-based method. In addition to the threshold-based algorithms, CoMatch [Li *et al.*, 2021] and SimMatch [Zheng *et al.*, 2022] leverage contrastive loss to impose sample-level constraints on all unlabeled data. In contrast, AllMatch combines the advantages of both threshold-based and contrastive-based methods. The approach introduces CAT, a learning-status-aware threshold strategy, and BCC regulation, the semantic-level supervision for the entire unlabeled set, to maximize the utilization of unlabeled data and enhance the overall performance.

In addition to consistency regulation and pseudo-labeling, entropy-based regulation is another widely adopted strategy. Entropy minimization [Grandvalet and Bengio, 2004] promotes high-confidence predictions during training. Maximizing the entropy of the expectation over all samples [Krause *et al.*, 2010; Arazo *et al.*, 2020; Zhao *et al.*, 2022] introduces the concept of *fairness*, which encourages the model to predict each class with equal frequency. Specifically, distribution alignment (DA) [Berthelot *et al.*, 2019a] and uniform alignment (UA) [Chen *et al.*, 2023] are prevailing strategies for achieving fairness in SSL, which adjusts the pseudo-label based on the overall predictions on the unlabeled data.

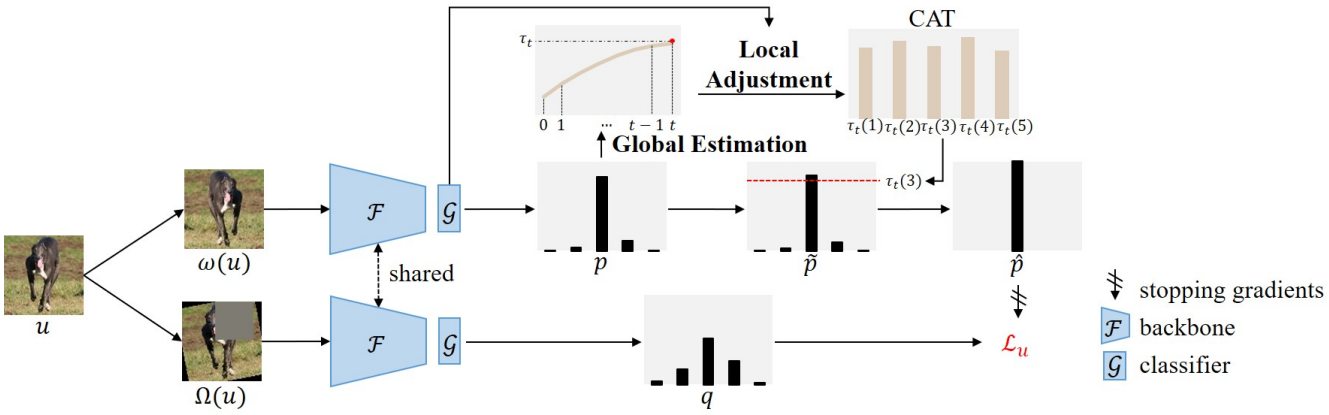


Figure 2: Pipeline of the class-specific adaptive threshold (CAT) mechanism. CAT employs the average confidence on unlabeled data as the global threshold and subsequently utilizes the classifier weights to establish class-specific thresholds.

3 Methodology

3.1 Preliminary

We begin by reviewing the widely adopted SSL framework. Let $\mathcal{D}_L = \{(x_i, y_i)\}_{i=1}^{N_L}$ and $\mathcal{D}_U = \{u_i\}_{i=1}^{N_U}$ represent the labeled and unlabeled datasets, respectively. Here, x_i and u_i denote the labeled and unlabeled training samples, and y_i represents the one-hot label for the labeled sample x_i. We denote the prediction of sample x as $p(y|x)$. Given a batch of labeled and unlabeled data, the model is optimized with the objective $\mathcal{L} = \mathcal{L}_s + \lambda_u \mathcal{L}_u$. Here, $\mathcal{L}_s$ represents the cross-entropy loss ($\mathcal{H}$) for the labeled batch of size B_L.

$$\mathcal{L}_s = \frac{1}{B_L} \sum_{i=1}^{B_L} \mathcal{H}(y_i, p(y|x_i)) \tag{1}$$

$\mathcal{L}_u$ indicates the consistency regulation between the prediction of the strongly augmented view $\Omega(u)$ and the pseudo-label derived from the corresponding weakly augmented view $\omega(u)$. To filter out incorrect pseudo-labels, FixMatch [Sohn *et al.*, 2020] introduces a predefined threshold τ. Specifically, $\mathcal{L}_u$ is defined as follows.

$$\mathcal{L}_u = \frac{1}{B_U} \sum_{i=1}^{B_U} \lambda(\tilde{p}_i) \mathcal{H}(\hat{p}_i, q_i) \tag{2}$$

$$\lambda(p) = \begin{cases} 1 & \textit{if } max(p) \geq \tau \\ 0 & \textit{otherwise} \end{cases} \tag{3}$$

Here, $\tilde{p}_i$ is the abbreviation for $DA(p(y|\omega(u_i)))$, where DA indicates the distribution alignment strategy [Berthelot *et al.*, 2019a]. $\hat{p}_i$ represents the one-hot pseudo-label obtained from $argmax(\tilde{p}_i)$. Moreover, q_i is the abbreviation for $p(y|\Omega(u_i))$. Lastly, B_U corresponds to the batch size of unlabeled data.

3.2 Class-Specific Adapative Threshold

Previous studies [Zhang *et al.*, 2021; Wang *et al.*, 2022] have demonstrated that the threshold should be aligned with the evolving learning status of the model. To achieve this, these approaches leverage predictions on unlabeled data to establish the dynamic threshold. In this paper, we unveil the ability of the classifier weights to differentiate the learning status of each class. By combining pseudo-labels and classifier weights, we introduce a class-specific adaptive threshold (CAT) mechanism. As depicted in Figure 2, CAT comprises the global estimation and local adjustment steps. The following parts provide a detailed description of these two steps.

Global Estimation. The global estimation step learns from FreeMatch [Wang *et al.*, 2022] and evaluates the overall learning status of the model. Given that deep neural networks tend to prioritize fitting easier samples before memorizing harder and noisier ones, a lower threshold is necessary during early training stages to incorporate more correct pseudo-labels. Conversely, as training progresses, a higher threshold is required to filter out incorrect pseudo-labels. Given that the cross-entropy loss encourages confident predictions, the average confidence of the unlabeled set captures information from all unlabeled data and steadily increases throughout training, thereby reflecting the overall learning status. However, making predictions for the entire unlabeled set at each time step incurs significant computational costs. Accordingly, we employ the mean confidence of the current batch as an estimation and update it using exponential moving average (EMA). Specifically, the global learning status estimation at t-th iteration, denoted as τ_t, can be computed as follows:

$$\tau_t = \begin{cases} \frac{1}{C} & \textit{if } t = 0 \\ m\tau_{t-1} + (1-m)\frac{1}{B_U} \sum_{i=1}^{B_U} max(p_i) & \textit{otherwise} \end{cases} \tag{4}$$

Here, p_i represents $p(y|\omega(u_i))$, m denotes the momentum decay, and C corresponds to the number of classes.

Local Adjustment. Due to the inherent variations in learning difficulty among different classes and the stochastic nature of parameter initialization, the model's learning status varies across categories. To address this issue, we introduce the local adjustment step, which makes the model pay more attention to the underfitting classes by decreasing their thresholds. Specifically, our study reveals that the L2 norm of the classifier weights provides insights into class-specific learning status. The reasons are explained as follows.

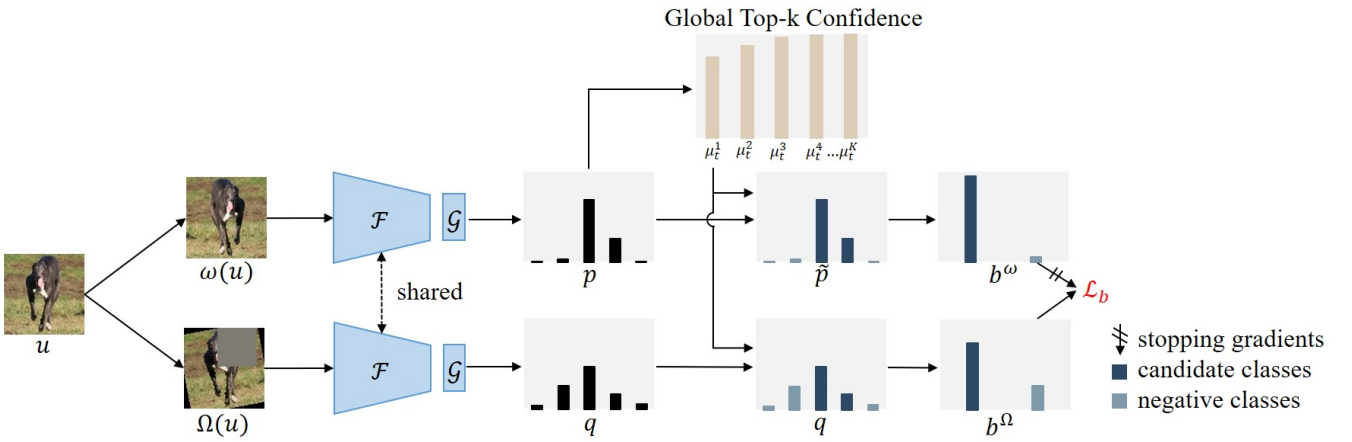


Figure 3: Pipeline of the binary classification consistency (BCC) regulation. The module compares global and local top-k confidence to identify the candidate and negative classes for each unlabeled sample. Moreover, it encourages consistent candidate-negative division between different perturbed views of the same sample, thereby introducing supervision signals for all unlabeled samples.

Firstly, let $\mathcal{M} = \mathcal{G} \circ \mathcal{F}$ denote the model, where $\mathcal{F}$ and $\mathcal{G}$ correspond to the encoder and the one-layer classifier, respectively. Given the feature vector $f = \mathcal{F}(u) \in R^d$ of the unlabeled sample u, the predicted logits $z \in R^C$ can be computed as $z = \mathcal{G}(f) = fW^T$, where $W \in R^{C \times d}$ denotes the weight matrix of the classifier. The bias term is dropped as it typically has a negligible impact on the results. Consequently, the logits of class c can be expressed as $z_c = ||f|| \cdot ||W_c|| \cdot cos(\theta)$, where $||\cdot||$ denotes the L2 norm, and W_c corresponds to the weight of class c. Accordingly, the model tends to produce logits with larger absolute values for categories with larger weight norms, implying that classes with larger weight norms exhibit a preferred learning status.

Secondly, [Kang *et al.*, 2019] reveals a positive correlation between the weight norm $||W_c||$ and n_c, the number of samples in class c. Given the abundant unlabeled data and limited labeled data, n_c can be approximated as the number of unlabeled samples classified into class c with confidence scores surpassing the threshold. Hence, a larger weight norm suggests that more samples are confidently categorized into the class, signifying a better learning status.

According to the above analysis, the L2 norm of classifier weights characterizes the learning status of each class. Consequently, the local adjustment step leverages this indicator to establish the mapping from the global threshold to class-specific thresholds. Specifically, we linearly scale the threshold for each class based on the deviation of its learning status from the optimal learning status. As such, the threshold for class c at t-th iteration, denoted as $\rho_t(c)$, can be computed as follows.

$$\rho_t(c) = \tau_t \cdot \frac{||W_c||}{max\{||W_c|| : c \in [1, \cdots, C]\}} \tag{5}$$

Moreover, to ensure a stable estimation of the learning status, we employ the classifier weights obtained from the EMA model. Notably, in contrast to FlexMatch, which maintains an additional list for recording the selected pseudo-label of each sample, the proposed CAT refrains from storing any sample-specific information during training. This eliminates the indexing budget concerns on large-scale datasets. With CAT incorporated, the mask for unlabeled samples in $\mathcal{L}_u$ can be expressed as follows.

$$\lambda(p) = \begin{cases} 1 & \textit{if } max(p) \geq \rho_t(argmax(p)) \\ 0 & \textit{otherwise} \end{cases} \tag{6}$$

3.3 Binary Classification Consistency Regulation

While the proposed class-specific adaptive threshold alleviates the underutilization of unlabeled data, a substantial number of pseudo-labels continue to be discarded. As illustrated in Figure 1, in the case of CIFAR-10 with 40 labeled samples, the top-5 accuracy of pseudo-labels effortlessly achieves 100% regardless of the adopted algorithm. In other words, pseudo-labels assigned lower confidence contribute to identifying candidate classes (*e.g.,* top-k predictions) and excluding negative options (*e.g.,* classes not included in top-k predictions). Motivated by these observations and the consistency regulation technique, we propose the binary classification consistency (BCC) regulation, whose overview is shown in Figure 3. In a nutshell, the strategy introduces semantic supervision for all unlabeled data by encouraging consistent candidate-negative division across diverse perturbed views of the same sample. The details are described as follows.

Given the impressive top-k pseudo-label accuracy obtained by numerous algorithms, the BCC regulation adopts the top-k predictions of each unlabeled sample as its candidate classes and the rest as the negative options. Thus, the candidate-negative division is simplified to the selection of the parameter k. Moreover, considering the variations in learning difficulty among different samples and the evolving performance of the model, the candidate-negative division for each sample should be determined based on both individual and global learning status. To achieve this, the BCC regulation first computes sample-specific top-k confidence and the global top-k confidence of the entire unlabeled set. Specifically, let p_i^k denote the top-k probability of sample u_i, and μ_t^k represent the

global top-k probability at t-th iteration. The global top-k confidence can be estimated by the exponential moving average (EMA) of the average top-k confidence at each time step.

$$p_i^k = \sum_{j=1}^{k} p_{i,c_j} \quad (p_{i,c_1} \geq p_{i,c_2} \geq \cdots) \tag{7}$$

$$\mu_t^k = \begin{cases} \frac{k}{C} & \textit{if } t = 0 \\ m\mu_{t-1}^k + (1-m)\frac{1}{B_U}\sum_{i=1}^{B_U} p_i^k & \textit{otherwise} \end{cases} \tag{8}$$

Here, $c_1, \ldots, c_k$ represent the k classes assigned the highest probability in p_i. With the global top-k confidence determined, the number of candidate classes for each unlabeled sample is defined as the minimum value that makes individual top-k confidence higher than global top-k confidence. Particularly, the candidate class for confident unlabeled samples is defined as the pseudo-label. Accordingly, the number of candidate classes k_i for sample u_i can be expressed as follows.

$$k_i = \begin{cases} 1 & \textit{if } \lambda(\tilde{p}_i) = 1 \\ \textit{min}(\textit{min}\{k : \tilde{p}_i^k \geq \mu_t^k\}, K) & \textit{otherwise} \end{cases} \tag{9}$$

where K is the upper bound for the number of candidate classes to prevent trivial candidate-negative division. With the division obtained, the candidate and negative probabilities for the weakly (b_i^ω) and strongly (b_i^Ω) perturbed views of the unlabeled sample u_i can be calculated as follows.

$$b_i^\omega = [\sum_{j=1}^{k_i} \tilde{p}_{i,c_j}, \sum_{j=k_i+1}^{C} \tilde{p}_{i,c_j}] \quad (\tilde{p}_{i,c_1} \geq \tilde{p}_{i,c_2} \geq \cdots) \tag{10}$$

$$b_i^\Omega = [\sum_{j=1}^{k_i} q_{i,c_j}, \sum_{j=k_i+1}^{C} q_{i,c_j}] \tag{11}$$

Here, $c_1, \ldots, c_{k_i}$ represents the k_i classes assigned the highest probability in $\tilde{p}_i$. Finally, the BCC regulation for a batch of unlabeled data can be calculated as follows:

$$\mathcal{L}_b = \frac{1}{B_U}\sum_{i=1}^{B_U} \mathcal{H}(b_i^\omega, b_i^\Omega) \tag{12}$$

3.4 Overall Objective

The overall objective of AllMatch is defined as the weighted sum of all semantic-level supervision.

$$\mathcal{L} = \mathcal{L}_s + \lambda_u \mathcal{L}_u + \lambda_b \mathcal{L}_b \tag{13}$$

λ_u and λ_b denote the weights to balance different supervision signals. For all experiments, we set both λ_u and λ_b to 1.0.

4 Experiments

4.1 Balanced Semi-Supervised Learning

Settings. For balanced image classification, we conduct experiments on CIFAR-10/100 [Krizhevsky *et al.*, 2009], SVHN [Netzer *et al.*, 2011], STL-10 [Coates *et al.*, 2011], and ImageNet [Deng *et al.*, 2009] with various numbers of labeled data, where the class distribution of the labeled data is balanced. To ensure fair comparisons, we employ the unified codebase TorchSSL [Zhang *et al.*, 2021] to evaluate all methods. Regarding the backbone architecture, we follow previous studies and use specific models for different datasets: WRN-28-2 [Zagoruyko and Komodakis, 2016] for CIFAR-10 and SVHN, WRN-28-8 for CIFAR-100, WRN-37-2 [Zhou *et al.*, 2020] for STL-10, and ResNet-50 [He *et al.*, 2016] for ImageNet. The batch sizes B_L and B_U are set to 128 and 128 for ImageNet and 64 and 448 for the remaining datasets. AllMatch is trained using the SGD optimizer with an initial learning rate of 0.03 and a momentum decay of 0.9. The learning rate is adjusted by a cosine decay scheduler over a total of 2^{20} iterations. We set m to 0.999 and generate the EMA model with a momentum decay of 0.999 for inference. The upper bound K is set to 20 for ImageNet and 10 for the other datasets. For SVHN, CIFAR-10 with 10 labels, and STL-10 with 40 labels, we constrain the threshold within the range of [0.9, 1.0] to prevent overfitting noisy pseudo-labels in the early training stages. To account for randomness, we repeat each experiment three times and report the mean and standard deviation of the top-1 accuracy.

Performance. Table 1 presents the top-1 accuracy on CIFAR-10/100, SVHN, and STL-10 with various numbers of labeled samples. The performance on ImageNet is reported in Table 2. The experimental results demonstrate that AllMatch achieves state-of-the-art performance on all datasets, except for CIFAR-100 with 10000 labels and SVHN with 40 labels, where our model's performance remains comparable to the best competitor. Moreover, regarding CIFAR-100, AllMatch outperforms ReMixMatch when only 400 or 2500 labels are available, while the latter achieves better performance when 10000 labels are available. The competitive results obtained by ReMixMatch mainly stem from the Mixup technique [Zhang *et al.*, 2017] and the additional self-supervised learning part. Furthermore, AllMatch exhibits substantial advantages over previous algorithms when dealing with extremely limited labeled samples. Specifically, the approach surpasses the second-best counterpart by 1.87% on CIFAR-10 with 10 labels, 0.66% on CIFAR-100 with 400 labels, and 2.86% on STL-10 with 40 labels. Particularly, STL-10 poses significant challenges due to its large unlabeled set that comprises 100k images. Accordingly, the impressive improvement obtained on STL-10 highlights the potential of AllMatch to be deployed in real-world applications.

4.2 Imbalanced Semi-Supervised Learning

Settings. We evaluate AllMatch in the context of imbalanced SSL, where both labeled and unlabeled data exhibit a long-tailed distribution. All experiments are conducted on the TorchSSL codebase. Following prior studies [Lee *et al.*, 2021; Oh *et al.*, 2022; Wei *et al.*, 2021; Lai *et al.*, 2022; Fan *et al.*, 2022; Chen *et al.*, 2023], we generate the labeled and unlabeled sets using the configurations of $N_c = N_1 \cdot \gamma^{-\frac{c-1}{C-1}}$ and $M_c = M_1 \cdot \gamma^{-\frac{c-1}{C-1}}$. Specifically, for CIFAR-10-LT, we set N_1 to 1500, M_1 to 3000, and γ to range from 50 to 150. For CIFAR-100-LT, we set N_1 to 150, M_1 to 300, and γ to range from 20 to 100. In all experiments, we employ WRN-28-2 as the backbone and utilize the Adam optimizer [Kingma and Ba, 2014] with the weight decay of 4e-5.

Datasets	CIFAR-10				CIFAR-100			SVHN		STL-10	
# Label	10	40	250	4000	400	2500	10000	40	1000	40	1000
MeanTeacher [Tarvainen and Valpola, 2017]	23.63	29.91	62.54	91.90	18.89	54.83	68.25	63.91	96.73	28.28	66.10
MixMatch [Berthelot *et al.*, 2019b]	34.24	63.81	86.37	93.34	32.41	60.24	72.22	69.40	96.31	45.07	78.30
ReMixMatch [Berthelot *et al.*, 2019a]	79.23	90.12	93.70	95.16	57.25	73.97	**79.98**	75.96	94.84	67.88	93.26
UDA [Xie *et al.*, 2020]	65.47	89.38	94.84	95.71	53.61	72.27	77.51	94.88	98.11	62.58	93.36
FixMatch [Sohn *et al.*, 2020]	82.09	92.53	95.14	95.79	53.58	71.97	77.80	96.19	98.04	64.03	93.75
Dash [Xu *et al.*, 2021]	73.72	91.07	94.84	95.64	55.18	72.85	78.12	97.81	98.03	65.48	93.61
MPL [Pham *et al.*, 2021]	76.45	93.38	94.24	95.45	53.74	72.29	78.26	90.67	97.72	64.24	93.34
FlexMatch [Zhang *et al.*, 2021]	90.02	95.03	95.02	95.81	60.06	73.51	78.10	91.81	93.28	76.65	94.23
SoftMatch [Chen *et al.*, 2023]	93.04	95.09	95.18	95.96	62.90	73.34	77.97	97.67	97.99	85.28	94.27
FreeMatch [Wang *et al.*, 2022]	92.09	95.10	95.12	95.90	62.02	73.53	78.32	**98.03**	98.04	84.68	94.37
AllMatch	**94.91**	**95.20**	**95.28**	**96.14**	**63.56**	**74.16**	78.66	97.56	**98.15**	**88.14**	**94.91**

Table 1: Top-1 accuracy (%) on CIFAR-10, CIFAR-100, SVHN, and STL-10 with varying number of labeled samples. **Bold** indicates the best performance, and underline denotes the second best performance.

Datasets	ImageNet	
# Label	100k	400k
FixMatch [Sohn *et al.*, 2020]	56.34	67.72
FlexMatch [Zhang *et al.*, 2021]	58.15	68.69
SoftMatch [Chen *et al.*, 2023]	59.48	70.51
FreeMatch [Wang *et al.*, 2022]	59.43	69.48
AllMatch	**59.99**	**70.82**

Table 2: Top-1 accuracy (%) on ImageNet.

Dataset	CIFAR-10-LT			CIFAR-100-LT		
γ	50	100	150	20	50	100
FixMatch	81.54	74.89	70.38	49.58	42.11	37.60
FlexMatch	81.87	74.49	70.20	50.89	42.80	37.30
SoftMatch	83.45	77.07	72.60	51.91	43.76	38.92
FreeMatch	83.30	75.96	71.20	51.60	42.95	37.50
AllMatch	**84.21**	**78.76**	**74.25**	**52.52**	**44.10**	**39.09**

Table 3: Performance (%) on CIFAR-10-LT and CIFAR-100-LT.

GE	LA	BCC	CIFAR-10-10	CIFAR-100-400	STL-10-40
			82.09	53.58	64.03
✓			89.44	61.58	80.48
✓	✓		92.66	63.04	87.43
✓	✓	✓	94.91	63.56	88.14

Table 4: Ablation study (%). (GE: global estimation in CAT. LA: local adjustment in CAT. BCC: binary classification consistency.)

Threshold	BCC	CIFAR-10-10	CIFAR-100-400	STL-10-40
FlexMatch		90.02	60.06	76.65
FreeMatch		92.09	62.02	84.68
SoftMatch		**93.04**	62.90	85.28
AllMatch		92.66	**63.04**	**87.43**
FlexMatch	✓	92.04	61.18	80.05
FreeMatch	✓	94.19	62.75	86.49
SoftMatch	✓	**94.99**	63.21	87.10
AllMatch	✓	94.91	**63.56**	**88.14**

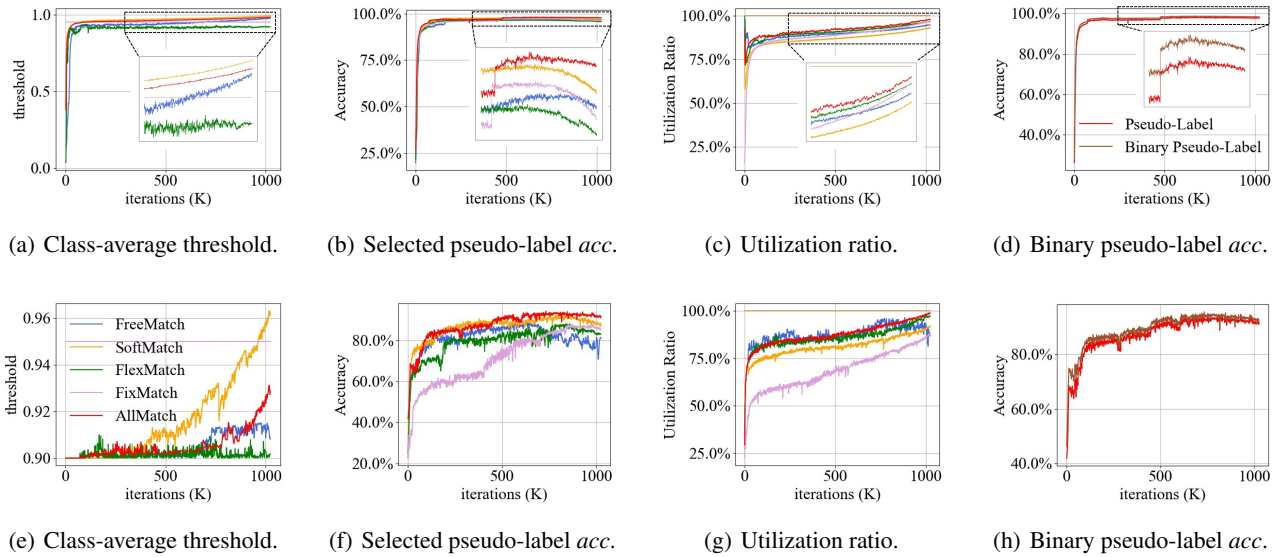
Table 5: Threshold comparison study (%). SoftMatch assigns trivial weights to samples with confidence significantly lower than the global confidence, approximating itself as a threshold-based model.

The batch sizes B_L and B_U are set to 64 and 128, respectively. The learning rate is initially set to 2e-3 and adjusted by a cosine decay scheduler during training. We repeat each experiment three times and report the overall performance.

Performance. In the context of imbalanced SSL, we compare AllMatch with several strong baselines, including FixMatch, FlexMatch, SoftMatch, and FreeMatch. The results in Table 3 demonstrate that AllMatch achieves state-of-the-art performance on all benchmarks. It is particularly noteworthy that AllMatch outperforms the second-best approach by 1.69% and 1.65% at γ=100 and γ=150 on CIFAR-10-LT, respectively, highlighting its robustness in handling significant imbalances. The consistently impressive performance observed in imbalanced SSL suggests that AllMatch can effectively address real-world challenges.

4.3 Ablation Study

Component analysis. We conduct an ablation study on four challenging datasets: CIFAR-10 with 10 labels, CIFAR-100 with 400 labels, STL-10 with 40 labels, and CIFAR-10-LT with an imbalance ratio of 150. For simplicity, we refer to the performance on the four benchmarks as (a, b, c, d) in subsequent analysis. As shown in Table 4, the global estimation step in CAT (line 2) promotes the performance by (7.35%, 8.00%, 16.45%, 1.20%) compared to the baseline model in line 1. The significant improvement highlights the crucial role of aligning the threshold with the model's global learning status. Furthermore, the local adjustment step in CAT (line 3) yields additional gains of (3.22%, 1.46%, 6.95%, 1.90%), suggesting that it effectively captures class-specific learning difficulties and facilitates the learning of classes facing challenges. Additionally, the BCC regulation enables a 100% utilization ratio of the unlabeled data and achieves the improvement of (2.25%, 0.52%, 0.71%, 0.77%). The substantial improvement observed on CIFAR-10 with 10 labels indicates the potential of the BCC regulation when dealing with ex-



(a) Class-average threshold. (b) Selected pseudo-label *acc*. (c) Utilization ratio. (d) Binary pseudo-label *acc*.

(e) Class-average threshold. (f) Selected pseudo-label *acc*. (g) Utilization ratio. (h) Binary pseudo-label *acc*.

Figure 4: Learning process visualization on CIFAR-10-40 (a-d) and STL-10-40 (e-h). The threshold for STL-10-40 is restricted within [0.9, 1.0] to mitigate the adverse effects of noisy pseudo-labels. In the analysis of SoftMatch, we employ $\mu_t - \sigma_t$ as its threshold. Samples with a confidence lower than $\mu_t - \sigma_t$ are assigned negligible weights, essentially treated as if they were discarded. Consequently, the class-average threshold of SoftMatch is $\mu_t - \sigma_t$, and its selected pseudo-label *acc* and utilization ratio can be computed like other threshold-based models. Here, μ_t/σ_t denotes the *mean/std* of the overall confidence on unlabeled data.

tremely limited labeled data. Overall, the results in Table 4 demonstrate the effectiveness of the proposed modules.

Comparative study on threshold strategies. We conduct a comparative analysis of existing threshold mechanisms in two aspects. Firstly, we directly compare the proposed CAT with the threshold strategies adopted in previous models. The results are presented in line 1-4 of Table 5. Secondly, we assess the threshold strategies within the AllMatch framework, *i.e.,* combining existing threshold schemes with the BCC regulation, and the results are provided in line 5-8 of Table 5. From both perspectives, AllMatch outperforms previous models in most cases, indicating the effectiveness of the proposed CAT. Moreover, the BCC regulation further boosts the performance of prior methods, suggesting its impressive compatibility and contribution to eliminating false options.

4.4 Quantitative Analysis

To gain further insights into AllMatch, we present various training indicators on CIFAR-10 with 40 labels and STL-10 with 40 labels, as illustrated in Figure 4. From Figure 4(a) and Figure 4(e), it can be observed that the threshold exhibits the expected behavior, starting with a small value and gradually increasing thereafter. Moreover, AllMatch demonstrates a smoother threshold evolution in contrast to other class-specific threshold-based models, implying a preferred learning status estimation. Additionally, in comparison to previous algorithms, Figure 4(b) and Figure 4(c), as well as Figure 4(f) and Figure 4(g), indicate that AllMatch achieves improved pseudo-label accuracy and a higher utilization ratio for the unlabeled data. Notably, Figure 4(b) and Figure 4(f) demonstrate that previous models consistently suffer from overfitting noisy pseudo-labels in the later training stages, whereas AllMatch successfully mitigates this issue. Furthermore, as depicted in Figure 4(d) and Figure 4(h), the accuracy of the candidate-negative division consistently outperforms pseudo-label accuracy, suggesting that the BCC regulation effectively identifies the candidate classes for all unlabeled data. Overall, the CAT precisely reflects the learning status of the model and the BCC regulation provides accurate supervision signals for all unlabeled samples.

5 Conclusion

This paper revisits prior SSL algorithms, focusing on two crucial questions: how to design an effective threshold mechanism and how to utilize the pseudo-labels assigned lower confidence. To address these challenges, we introduce two strategies named class-specific adaptive threshold (CAT) and binary classification consistency (BCC) regulation. CAT leverages predictions on unlabeled data and classifier weights to establish a threshold mechanism that aligns with the evolving learning status of each class. BCC regulation identifies candidate classes for each unlabeled sample and encourages consistent candidate-negative divisions across diverse perturbed views of the same sample. With these two modules incorporated, the proposed AllMatch maximizes the utilization of unlabeled data and achieves impressive pseudo-label accuracy. We conduct extensive experiments on multiple benchmarks, including both balanced and imbalanced settings. The results demonstrate that AllMatch achieves state-of-the-art performance and is capable of dealing with real-world challenges.

References

[Arazo *et al.*, 2020] 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 (IJCNN)*, pages 1–8. IEEE, 2020.

[Bengio *et al.*, 2009] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48, 2009.

[Berthelot *et al.*, 2019a] 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. *arXiv preprint arXiv:1911.09785*, 2019.

[Berthelot *et al.*, 2019b] 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.

[Chen *et al.*, 2023] Hao Chen, Ran Tao, Yue Fan, Yidong Wang, Jindong Wang, Bernt Schiele, Xing Xie, Bhiksha Raj, and Marios Savvides. Softmatch: Addressing the quantity-quality trade-off in semi-supervised learning. *arXiv preprint arXiv:2301.10921*, 2023.

[Coates *et al.*, 2011] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 215–223. JMLR Workshop and Conference Proceedings, 2011.

[Deng *et al.*, 2009] 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.

[Fan *et al.*, 2022] Yue Fan, Dengxin Dai, Anna Kukleva, and Bernt Schiele. 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*, pages 14574–14584, 2022.

[Grandvalet and Bengio, 2004] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. *Advances in neural information processing systems*, 17, 2004.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[Kang *et al.*, 2019] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. *arXiv preprint arXiv:1910.09217*, 2019.

[Kingma and Ba, 2014] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[Krause *et al.*, 2010] Andreas Krause, Pietro Perona, and Ryan Gomes. Discriminative clustering by regularized information maximization. *Advances in neural information processing systems*, 23, 2010.

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

[Lai *et al.*, 2022] Zhengfeng Lai, Chao Wang, Henrry Gunawan, Sen-Ching S Cheung, and Chen-Nee Chuah. Smoothed adaptive weighting for imbalanced semi-supervised learning: Improve reliability against unknown distribution data. In *International Conference on Machine Learning*, pages 11828–11843. PMLR, 2022.

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

[Lee and others, 2013] 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*, volume 3, page 896, 2013.

[Lee *et al.*, 2021] Hyuck Lee, Seungjae Shin, and Heeyoung Kim. Abc: Auxiliary balanced classifier for class-imbalanced semi-supervised learning. *Advances in Neural Information Processing Systems*, 34:7082–7094, 2021.

[Li *et al.*, 2021] Junnan Li, Caiming Xiong, and Steven CH Hoi. Comatch: Semi-supervised learning with contrastive graph regularization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9475–9484, 2021.

[Netzer *et al.*, 2011] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.

[Oh *et al.*, 2022] Youngtaek Oh, Dong-Jin Kim, and In So Kweon. Daso: Distribution-aware semantics-oriented pseudo-label for imbalanced semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9786–9796, 2022.

[Pham *et al.*, 2021] Hieu Pham, Zihang Dai, Qizhe Xie, and Quoc V Le. Meta pseudo labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11557–11568, 2021.

[Rosenberg *et al.*, 2005] Chuck Rosenberg, Martial Hebert, and Henry Schneiderman. Semi-supervised self-training of object detection models. 2005.

[Sajjadi *et al.*, 2016] Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. *Advances in neural information processing systems*, 29, 2016.

[Sohn *et al.*, 2020] 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*, 33:596–608, 2020.

[Tarvainen and Valpola, 2017] 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.

[Wang *et al.*, 2022] Yidong Wang, Hao Chen, Qiang Heng, Wenxin Hou, Marios Savvides, Takahiro Shinozaki, Bhiksha Raj, Zhen Wu, and Jindong Wang. Freematch: Self-adaptive thresholding for semi-supervised learning. *arXiv preprint arXiv:2205.07246*, 2022.

[Wei *et al.*, 2021] 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.

[Xie *et al.*, 2020] Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. *Advances in neural information processing systems*, 33:6256–6268, 2020.

[Xu *et al.*, 2021] Yi Xu, Lei Shang, Jinxing Ye, Qi Qian, Yu-Feng Li, Baigui Sun, Hao Li, and Rong Jin. Dash: Semi-supervised learning with dynamic thresholding. In *International Conference on Machine Learning*, pages 11525–11536. PMLR, 2021.

[Zagoruyko and Komodakis, 2016] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

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

[Zhang *et al.*, 2021] Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. *Advances in Neural Information Processing Systems*, 34:18408–18419, 2021.

[Zhao *et al.*, 2022] Zhen Zhao, Luping Zhou, Yue Duan, Lei Wang, Lei Qi, and Yinghuan Shi. Dc-ssl: Addressing mismatched class distribution in semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9757–9765, 2022.

[Zheng *et al.*, 2022] Mingkai Zheng, Shan You, Lang Huang, Fei Wang, Chen Qian, and Chang Xu. Simmatch: Semi-supervised learning with similarity matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14471–14481, 2022.

[Zhou *et al.*, 2020] Tianyi Zhou, Shengjie Wang, and Jeff Bilmes. Time-consistent self-supervision for semi-supervised learning. In *International Conference on Machine Learning*, pages 11523–11533. PMLR, 2020.

[Zhu, 2005] Xiaojin Jerry Zhu. Semi-supervised learning literature survey. 2005.