# FlexMatch: Boosting Semi-Supervised Learning with Curriculum Pseudo Labeling

**Anonymous Author(s)**
Affiliation
Address
email

## Abstract

The recently proposed FixMatch achieved state-of-the-art results on most semi-supervised learning (SSL) benchmarks. However, like other modern SSL algorithms, FixMatch uses a pre-defined constant threshold for all classes to select unlabeled data that contribute to the training, thus failing to consider different learning status and learning difficulties of different classes. To address this issue, we propose Curriculum Pseudo Labeling (CPL), a curriculum learning approach to leverage unlabeled data according to the model's learning status. The core of CPL is to flexibly adjust thresholds for different classes at each time step to let pass informative unlabeled data and their pseudo labels. CPL does not introduce additional parameters or computations (forward or backward propagation). We apply CPL to FixMatch and call our improved algorithm *FlexMatch*. FlexMatch achieves state-of-the-art performance on a variety of SSL benchmarks, with especially strong performances when the labeled data are extremely limited or when the task is challenging. For example, FlexMatch outperforms FixMatch by **14.32**% and **24.55**% on CIFAR-100 and STL-10 datasets respectively, when there are only 4 labels per class. CPL also significantly boosts the convergence speed, e.g., FlexMatch can use only 1/5 training time of FixMatch to achieve even better performance. Furthermore, we show that CPL can be easily adapted to other SSL algorithms and remarkably improve their performances. We also build a unified PyTorch-based library, named TorchSSL for convenient and fair study of SSL algorithms.

## 1 Introduction

Semi-supervised learning (SSL) has attracted increasing attention in recent years due to its superiority in leveraging a large amount of unlabeled data. This is particularly advantageous when the labeled data are limited in quantity or laborious to obtain. Consistency regularization [1, 2, 3] and pseudo labeling [4, 5, 6, 7, 8] are two powerful techniques for utilizing unlabeled data and have been widely used in modern SSL algorithms [9, 10, 11, 12, 13]. The recently proposed FixMatch [14] achieves competitive results by combining these techniques with weak and strong data augmentations and using cross-entropy loss as the consistency regularization criterion.

However, a drawback of FixMatch and other popular SSL algorithms such as Pseudo-Labeling [4] and Unsupervised Data Augmentation (UDA) [11] is that they rely on a *fixed* threshold to compute the unsupervised loss, using only unlabeled data whose prediction confidence is above the threshold. While this strategy can make sure that only high-quality unlabeled data contribute to the model training, it ignores a considerable amount of other unlabeled data, especially at the early stage of the training process, where only a few unlabeled data have their prediction confidence above the threshold. Moreover, modern SSL algorithms handle all classes *equally* without considering their different learning difficulties.

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

To address these issues, we propose Curriculum Pseudo Labeling (CPL), a curriculum learning [15] strategy to take into account the learning status of each class for semi-supervised learning. CPL substitutes the pre-defined thresholds with *flexible* thresholds that are dynamically adjusted for each class according to the current learning status. Notably, this process does not introduce any additional parameter (hyperparameter or trainable parameter) or extra computation (forward or back propagation). We apply this curriculum learning strategy directly to FixMatch and call the improved algorithm *FlexMatch*.

While the training speed remains as efficient as that of FixMatch, FlexMatch converges significantly faster and achieves state-of-the-art performances on most SSL image classification benchmarks. The benefit of introducing CPL is particularly remarkable when the labels are scarce or when the task is challenging. For instance, on the CIFAR-100 dataset, FlexMatch surpasses the performance of FixMatch by 14.32%, 4.30%, and 2.55% when the label amount is 400, 2500, and 10000 respectively. Moreover, CPL further shows its superiority by boosting the convergence speed – with CPL, FlexMatch takes less than 1/5 training time of FixMatch to reach its final accuracy. Adapting CPL to other modern SSL algorithms also leads to improvements in accuracy and convergence speed.

To sum up, this paper makes the following three contributions:

- We propose Curriculum Pseudo Labeling (CPL), a curriculum learning approach of dynamically leveraging unlabeled data for SSL. It is almost cost-free and can be easily integrated to other SSL methods.
- CPL significantly boosts the accuracy and convergence performance of several popular SSL algorithms on common benchmarks. Specifically, FlexMatch, the integration of FixMatch and CPL, achieves state-of-the-art results.
- We build TorchSSL, a unified PyTorch-based semi-supervised learning codebase for the fair study of SSL algorithms. TorchSSL includes implementations of popular SSL algorithms and their corresponding training strategies, and is easy to use and customize.

## 2 Background

Consistency regularization follows the continuity assumption of SSL [1, 2]. The most basic consistency loss in SSL, such as in $\Pi$ Model [9], Mean Teacher [10] and MixMatch [12], is the $\ell$-2 loss:

$$\sum_{b=1}^{\mu B} ||p_m(y|\omega(u_b)) - p_m(y|\omega(u_b))||_2^2, \tag{1}$$

where $B$ is the batch size of labeled data, $\mu$ is the ratio of unlabeled data to labeled data, $\omega$ is a stochastic data augmentation function (thus the two terms in Eq.(1) are different), $u_b$ denotes a piece of unlabeled data, and $p_m$ represents the output probability of the model. With the introduction of pseudo labeling techniques [5, 7], the consistency regularization is converted to an entropy minimization process [16], which is more suitable for the classification task. The improved consistency loss with pseudo labeling can be represented as:

$$\frac{1}{\mu B}\sum_{b=1}^{\mu B} \mathbb{1}(\max(p_m(y|\omega(u_b))) > \tau) H(\hat{p}_m(y|\omega(u_b)), p_m(y|\omega(u_b))), \tag{2}$$

where $H$ is cross-entropy, $\tau$ is the pre-defined threshold and $\hat{p}_m(y|\omega(u_b))$ is the pseudo label that can either be a 'hard' one-hot label [4, 14] or a sharpened 'soft' one [11]. The intention of using a threshold is to mask out noisy unlabeled data that have low prediction confidence.

FixMatch utilizes such consistency regularization with strong augmentation to achieve competitive performance. For unlabeled data, FixMatch first uses weak augmentation to generate artificial labels. These labels are then used as the target of strongly-augmented data. The unsupervised loss term in FixMatch thereby has the form:

$$\frac{1}{\mu B}\sum_{b=1}^{\mu B} \mathbb{1}(\max(p_m(y|\omega(u_b))) > \tau) H(\hat{p}_m(y|\omega(u_b)), p_m(y|\Omega(u_b))), \tag{3}$$

where $\Omega$ is a strong augmentation function instead of weak augmentation $\omega$.

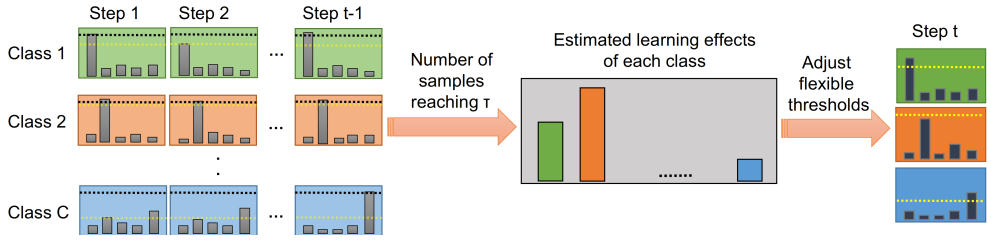



Figure 1: Illustration of Curriculum Pseudo Label (CPL). The estimated learning effects of each class are decided by the number of unlabeled data samples falling into this class and above the fixed threshold. They are then used to adjust the flexible thresholds to let pass the optimal unlabeled data. Note that the estimated learning effects do not always grow – they may also decrease if the predictions of the unlabeled data fall into other classes in later iterations.

Of the aforementioned works, the pre-defined threshold ($\tau$) is constant. We believe this can be improved because the data of some classes may be inherently more difficult to learn than others. Curriculum learning [15] is a learning strategy where learning samples are gradually introduced according to the model's learning process. In such a way, the model is always optimally challenged. This technique is widely employed in deep learning research [17, 18, 19, 20, 21].

# 3 FlexMatch

## 3.1 Curriculum Pseudo Labeling

While current SSL algorithms render pseudo labels of only high-confidence unlabeled data cut off by a pre-defined threshold, CPL renders the pseudo labels *to different classes* and *at different time steps*. Such a process is realized by adjusting the thresholds according to the model's learning status of each class.

However, it is non-trivial to dynamically determine the thresholds according to the learning status. The most ideal approach would be calculating evaluation accuracies for each class and use them to scale the threshold, as:

$$\mathcal{T}_t(c) = a_t(c) \cdot \tau, \tag{4}$$

where $\mathcal{T}_t(c)$ is the flexible threshold for class $c$ at time step $t$ and $a_t(c)$ is the corresponding evaluation accuracy. In this way, lower accuracy that indicates a less satisfactory learning status of the class will lead to a lower threshold that encourages more samples of this class to be learned. Since we cannot use the evaluation set in the model learning process, one may have to separate an extra validation set from the training set for such accuracy evaluations. However, this practice show two fatal problems: First, such a *labeled* validation set separated from the training set is expensive under SSL scenario as the labeled data are already scarce. Second, to dynamically adjust the thresholds in the training process, accuracy evaluations must be done continually at each time step $t$, which will considerably slow down the training speed.

In this work, we propose Curriculum Pseudo Labeling (CPL) for semi-supervised learning. Our CPL uses an alternative way to estimate the learning status, which does not introduce additional inference processes, nor needs an extra validation set. As believed in [14], a high threshold that filters out noisy pseudo labels and leaves only high-quality ones can considerably reduce the confirmation bias [22]. Therefore, our key assumption is that when the threshold is high, the learning effect of a class can be reflected by the number of samples whose predictions fall into this class and above the threshold. Namely, the class with fewer samples having their prediction confidence reach the threshold is considered to have a greater learning difficulty or a worse learning status, formulated as:

$$\sigma_t(c) = \sum_{n=1}^{N} \mathbb{1}(\max(p_{m,t}(y|u_n)) > \tau) \cdot \mathbb{1}(\arg\max(p_{m,t}(y|u_n) = c). \tag{5}$$

where $\sigma_t(c)$ reflects the learning effect of class $c$ at time step $t$. $p_{m,t}(y|u_n)$ is the model's prediction for unlabeled data $u_n$ at time step $t$, and $N$ is the total number of unlabeled data. When the unlabeled

dataset is balanced (i.e., the number of unlabeled data belonging to different classes are equal or close), larger $\sigma_t(c)$ indicates a better estimated learning effect. By applying the following normalization to $\sigma_t(c)$ to make its range between 0 to 1, it can then be used to scale the fixed threshold $\tau$:

$$\beta_t(c) = \frac{\sigma_t(c)}{\max \sigma_t(c)}, \tag{6}$$

$$\mathcal{T}_t(c) = \beta_t(c) \cdot \tau. \tag{7}$$

One characteristic of such a normalization approach is that the best-learned class has its $\beta_t(c)$ equal to 1, causing its flexible threshold equal to $\tau$. This is desirable. For classes that are hard to learn, the thresholds are lowered down, encouraging more training samples in these classes to be learned. This also improves the data utilization ratio. As learning proceeds, the threshold of a well-learned class is raised higher to selectively pick up higher-quality samples. Eventually, when all classes have reached reliable accuracies, the thresholds will all approach $\tau$. Note that the thresholds do not always grow, it may also decrease if the unlabeled data is classified into a different class in later iterations. This new threshold is used for calculating the unsupervised loss in FlexMatch, which can be formulated as:

$$\mathcal{L}_{u,t} = \frac{1}{\mu B} \sum_{b=1}^{\mu B} \mathbb{1}(\max(q_b) > \mathcal{T}_t(\arg\max(q_b))) H(\hat{q}_b, p_m(y|\Omega(u_b))), \tag{8}$$

where $q_b = p_m(y|\omega(u_b))$. The flexible thresholds are updated at each iteration. Finally, we can formulate the loss in FlexMatch as the weighted combination (by $\lambda$) of supervised and unsupervised loss:

$$\mathcal{L}_t = \mathcal{L}_s + \lambda \mathcal{L}_{u,t}, \tag{9}$$

where $\mathcal{L}_s$ is the supervised loss on labeled data:

$$\mathcal{L}_s = \frac{1}{B} \sum_{b=1}^{B} H(y_b, p_m(y|\omega(x_b))). \tag{10}$$

Note that the cost of introducing CPL is almost free. Practically, every time the prediction confidence of an unlabeled data $u_n$ is above the fixed threshold $\tau$, the data, and its predicted class are marked and will be used for calculating $\beta_t(c)$ at the next time step. Such marking actions are bonus actions each time the consistency loss is computed. Therefore, FlexMatch does not introduce additional forward propagation processes for evaluating the model's learning status, nor new parameters.

### 3.2 Threshold warm-up

We noticed in our experiments that at the early stage of the training, the model may blindly predict most unlabeled samples into a certain class depending on the parameter initialization(i.e., more likely to have confirmation bias). Hence, the estimated learning status may not be reliable at this stage. Therefore, we introduce a warm-up process by rewriting the denominator in Eq. (6) as:

$$\beta_t(c) = \frac{\sigma_t(c)}{\max\left\{\max \sigma_t(c), N - \sum_{c=1}^{C} \sigma_t(c)\right\}}, \tag{11}$$

where the term $N - \sum_{c=1}^{C} \sigma_t(c)$ can be regarded as the number of unlabeled data that have not been used. This ensures that at the beginning of the training, all estimated learning effects gradually rise from 0 until the number of unused unlabeled data is no longer predominant. The duration of such a period depends on the unlabeled data amount (ref. $N$ in Eq. (11)) and the learning difficulty (ref. the growing speed of $\sigma_t(c)$ in Eq. (11)) of the dataset. In practice, such a warm-up process is very easy to implement as we can add an extra class to denote the unused unlabeled data. Thus calculating the denominator of Eq. (11) is simply converted to finding the maximum among $c + 1$ classes.

### 3.3 Non-linear mapping function

The flexible threshold in Eq. (7) is determined by the normalized estimated learning effects via a linear mapping. However, it may not be the most suitable mapping in the real training process, where

**Algorithm 1** FlexMatch algorithm.

**Input:** $\mathcal{X} = \{(x_m, y_m) : m \in (1, \dots, M)\}$, $\mathcal{U} = \{u_n : n \in (1, \dots, N)\}$ {M labeled data and N unlabeled data.}
$\hat{u}_n = -1 : n \in (1, \dots, N)$ {Initialize predictions of all unlabeled data as -1 indicating unused.}
**while** not reach the maximum iteration **do**
  **for** $c = 1$ to $C$ **do**
    $\sigma(c) = \sum_{n=1}^{N} \mathbb{1}(\hat{u}_n = c)$ {Compute estimated learning effect.}
    **if** $\max \sigma(c) < \sum_{n=1}^{N} \mathbb{1}(\hat{u}_n = -1)$ **then**
      Calculate $\beta(c)$ using Eq. (11) {Threshold warms up when unused data dominate.}
    **else**
      Calculate $\beta(c)$ using Eq. (6) {Compute normalized estimated learning effect.}
    **end if**
    Calculate $\mathcal{T}(c)$ using Eq. (7) {Determine the flexible threshold for class $c$.}
  **end for**
  **for** $b = 1$ to $\mu B$ **do**
    **if** $p_m(y|\omega(u_b)) > \tau$ **then**
      $\hat{u}_b = \arg\max q_b$ {Update the prediction of unlabeled data $u_b$.}
    **end if**
  **end for**
  Compute the loss via Eq. (8), (10) and (9).
**end while**
**Return:** Model parameters.

the increase or decrease of $\beta_t(c)$ may make big jumps in the early phase where the predictions of the model are still unstable; and only make small fluctuations after the class is well-learned in the mid and late training stage. Therefore, it is preferable if the flexible thresholds can be more sensitive when $\beta_t(c)$ is large and vice versa.

We propose a non-linear mapping function to enable the thresholds to have a non-linear increasing curve when $\beta_t(c)$ ranges uniformly from 0 to 1, as formulated below:

$$\mathcal{T}_t(c) = \mathcal{M}(\beta_t(c)) \cdot \tau, \tag{12}$$

where $\mathcal{M}(\cdot)$ is a non-linear mapping function. It is clear that Eq. (7) can be seen as a special case by setting $\mathcal{M}$ to the identity function. The mapping function $\mathcal{M}$ should be monotonically increasing and have a maximum no larger than $1/\tau$ (otherwise the flexible threshold can be larger than 1 and filter out all samples). To avoid introducing additional hyperparameters (e.g. lower limits of the flexible thresholds), we consider the mapping function to have a range from 0 to 1 so that the flexible thresholds range from 0 to $\tau$.

A monotone increasing convex function lets the thresholds grow slowly when $\beta_t(c)$ is small, and become more sensitive as $\beta_t(c)$ gets larger. Hence, we intuitively choose a convex function with the above-mentioned properties $\mathcal{M}(x) = \frac{x}{2-x}$ for our experiments. We also conduct an ablation study to compare among mapping functions with different convexity and concavity in Sec. 4.4. The full algorithm of FlexMatch is shown in Algorithm 1.

# 4 Experiments

We evaluate FlexMatch and other CPL-enabled algorithms on common SSL datasets: CIFAR-10/100 [23], SVHN [24], STL-10 [25] and ImageNet [26], and extensively investigate the performance under various labeled data amounts. We mainly compare our method with Pseudo-Labeling [4], UDA [11] and FixMatch [14], since they all involve a pre-defined threshold. The results of other popular SSL algorithms are in the appendix. We also add a fully-supervised result for each dataset to better understand the results of SSL algorithms. Our fully-supervised experiments use only weak data augmentations for labeled data according to Eq. (10). Following SSL evaluation standards [27], we re-implement all baselines using our PyTorch [28] codebase: TorchSSL, which will be introduced in the next section and made publicly available.

For a fair comparison, we use the same hyperparameters following FixMatch [14]. Concretely, the optimizer for all experiments is standard stochastic gradient descent (SGD) with a momentum of 0.9

Table 1: Error rates on CIFAR-10/100, SVHN, and STL-10 datasets. The 'Flex' prefix denotes applying CPL to the algorithm, and 'PL' is an abbreviation of Pseudo-Labeling. STL-10 dataset does not have label information for unlabeled data, thus its fully-supervised result is unavailable.

| Dataset | CIFAR-10 | | | CIFAR-100 | | | STL-10 | | | SVHN | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Label Amount | 40 | 250 | 4000 | 400 | 2500 | 10000 | 40 | 250 | 1000 | 40 | 1000 |
| PL | 69.51±4.55 | 41.02±3.56 | 13.15±1.84 | 86.10±1.50 | 58.00±0.38 | 36.48±0.13 | 74.48±1.48 | 55.63±5.38 | 31.80±0.29 | 60.32±2.46 | 9.56±0.25 |
| Flex-PL | **65.41**±1.35 | **36.37**±1.57 | **10.82**±0.04 | **74.85**±1.53 | **44.15**±0.19 | **29.13**±0.26 | **69.26**±0.60 | **41.28**±0.46 | **24.63**±0.14 | **36.90**±1.19 | **8.64**±0.08 |
| UDA | 7.33±2.03 | 5.11±0.07 | 4.20±0.12 | 44.99±2.28 | 27.59±0.24 | 22.09±0.19 | 37.31±3.03 | 12.07±1.50 | 6.65±0.25 | 4.40±2.31 | **1.93**±0.01 |
| Flex-UDA | **5.33**±0.13 | **5.05**±0.02 | **4.07**±0.06 | **33.64**±0.92 | **24.34**±0.20 | **20.07**±0.13 | **12.84**±2.60 | **8.05**±0.21 | **5.77**±0.08 | **3.78**±1.67 | 1.97±0.06 |
| FixMatch | 6.78±0.50 | 4.95±0.07 | 4.09±0.02 | 46.76±0.79 | 28.15±0.81 | 22.47±0.66 | 35.42±6.43 | 10.49±1.03 | 6.20±0.20 | **4.36**±2.16 | **1.97**±0.03 |
| FlexMatch | **4.99**±0.16 | **4.80**±0.06 | **3.95**±0.03 | **32.44**±1.99 | **23.85**±0.23 | **19.92**±0.06 | **10.87**±1.15 | **7.71**±0.14 | **5.56**±0.22 | 5.36±2.38 | 2.86±0.91 |
| Fully-Supervised | 4.45± 0.12 | | | 19.07± 0.18 | | | - | | | 2.14± 0.02 | |

[29, 30, 31]. For all datasets, we use an initial learning rate of $0.03$ with a cosine learning rate decay schedule [32] as $\eta = \eta_0 \cos(\frac{7\pi k}{16K})$, where $\eta_0$ is the initial learning rate, $k$ is the current training step and $K$ is the total training step that is set to $2^{20}$. We also perform an exponential moving average with the momentum of $0.999$. The batch size of labeled data is $64$ except for ImageNet, which uses a batch size of 32. $\mu$ is set to be $1$ for Pseudo-Label and $7$ for UDA, FixMatch, and FlexMatch. $\tau$ is set to $0.8$ for UDA and $0.95$ for Pseudo Label, FixMatch, and FlexMatch. These setups follow the original papers. The strong augmentation function used in our experiments is RandAugment [33]. Detailed hyperparameters are listed in the appendix.

We adopt two evaluation metrics: (1) the median error rate of the last 20 checkpoints following [12, 14], and (2) the best error rate in all checkpoints. We argue that the median approach is not suitable when the convergence speeds of the algorithms show significant differences – the large number of redundant iterations may result in over-fitting for the fast-converge algorithms. Therefore, we report the best error rates for all algorithms, while the results of the median approach are also provided in the appendix, showing that our FlexMatch still achieves the best performance. We run each task three times using distinct random seeds to obtain the error bars.

### 4.1 Main results

The classification error rates on CIFAR-10/100, STL-10 and SVHN datasets are in Table 1, and the results on ImageNet are in Sec. 4.2. Note that the SVHN dataset used in our experiment also includes the extra set that contains 531,131 additional samples. Results demonstrate that FlexMatch achieves the best performance under all label conditions on all datasets except for SVHN where Flex-UDA (i.e., UDA with CPL) and UDA have the lowest error rate on the 40-label split and 1000-label split, respectively. We also provide the detailed precision, recall, F1, and AUC results in the appendix. Our CPL (FlexMatch) has the following advantages:

**CPL achieves better performance on tasks with extremely limited labeled data.** Our FlexMatch significantly outperforms other methods When the amount of labels is extremely small. For instance, on the CIFAR-100 dataset with 400 labels (i.e., only 4 label samples per class), FlexMatch achieves an average error rate of **32.44**%, which significantly outperforms FixMatch by **14.32**%.

**CPL improves the performance of existing SSL algorithms.** Other than FixMatch, the error rate of Pseudo-Labeling and UDA are also dramatically reduced by employing CPL. For instance, the error rate is reduced by **24.53**% for Pseudo-Labeling on SVHN (40 labels), and by **24.47**% for UDA on STL-10 (40 labels). These results further prove the effectiveness of CPL in better leveraging unlabeled data. Figure 2 shows the average running time of a single iteration with or without adding our CPL, it is clear that while improving the performance of existing SSL algorithms, our CPL *does not* introduce additional computational burden.

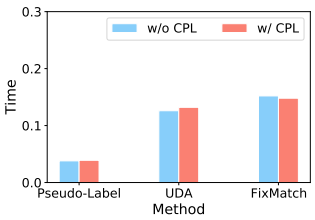
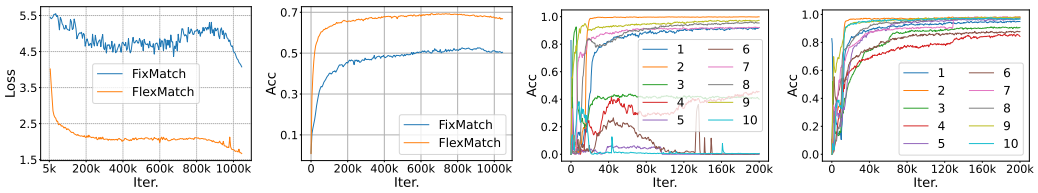



Figure 2: Average running time of one iteration on a single GeForce RTX 3090 GPU.

**CPL achieves better performance on complicated tasks.** The STL-10 dataset contains unlabeled data from a similar but broader distribution of images than its labeled set. The existence of new types of objects in the unlabeled dataset makes STL-10 a more challenging




(a) Loss (b) Top-1 acc. (c) FixMatch: 56.4% (d) FlexMatch: 94.3%

Figure 3: Convergence analysis of FixMatch and FlexMatch. (a) and (b) depict the loss and top-1-accuracy on CIFAR-100 with 400 labels. Evaluations are done every 5K iterations. (c) and (d) demonstrate the class-wise accuracy within the first 200K iterations on CIFAR-10 dataset. The numbers in legend correspond to the ten classes in the dataset.

and realistic task. Our FlexMatch achieves even greater performance improvement under such a challenging situation. To the best of our knowledge, we are the first to use only 40 labeled data for STL-10, and obtain a *noteworthy* error rate of **10.87**%, which is substantially better than FixMatch (35.42%). Similar strong improvements are also observed on CIFAR-100 dataset, which has as many as one hundred classes.

We also analyze the reason why FlexMatch fails to surpass FixMatch on SVHN. This is probably because the data of each class in SVHN are highly unbalanced. This leads to the classes with fewer samples never have their estimated learning effects close to 1 according to Eq. (6), even if they are already well-learned. These low upper thresholds allow noisy data to be used even in the late stage of training. We conclude this as a limitation of our approach of estimating the learning effects. Nevertheless, CPL still improves Pseudo-Labeling under both label conditions and UDA under the 40-label condition despite the above issue. The reason behind this is worth studying in future work.

### 4.2 Results on ImageNet

We test FlexMatch on ImageNet-1K dataset and compare it with FixMatch. We randomly choose the same 100K labeled data (i.e., 100 labels per class). This label amount is less than 6% of the total labels. As shown in Table 2, the top-1 and top-5 error rate of FlexMatch are 35.21% and 13.96%, which are significantly lower than FixMatch (43.08% and 19.55%).

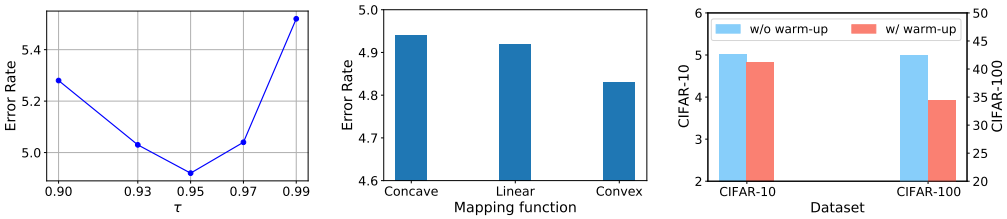
Table 2: Results on ImageNet.

| Method | Top-1 | Top-5 |
|---|---|---|
| FixMatch | 43.08 | 19.55 |
| FlexMatch | **35.21** | **13.96** |

### 4.3 Convergence speed analysis

Another strong advantage of FlexMatch is its superior convergence speed. Figure 3(a) and 3(b) shows the comparison between FlexMatch and FixMatch with respect to the loss and top-1-accuracy on CIFAR-100 400-label split. The loss of FlexMatch decreases much faster and smoother than FixMatch, demonstrating its superior convergence speed. The major fluctuations of the loss in FixMatch may due to the pre-defined threshold that lets pass most unlabeled data belonging to certain classes, whereas with CPL a larger batch of unlabeled data containing samples from various classes enables the gradient to more directly head toward the global optimum. As a result, with only 50K iterations, FlexMatch has already surpassed the final results of FixMatch. After 800K iterations, however, we observe a further decrease in loss and accuracy. This is likely due to over-fitting, which also occurs in FixMatch after 900K iterations. Thus, we believe it is not fair to use the median results of the last few checkpoints for evaluating algorithms with different convergence speeds.

We further compare the class-wise accuracy of FixMatch and FlexMatch on CIFAR-10 in their early training stages. As shown in Figure 3(c) and 3(d), at iteration 200K, FixMatch only hits an overall accuracy of 56.35% as half of the classes are still learned unsatisfactorily, whereas FlexMatch has already achieved an overall accuracy of 94.29% which is even higher than the final accuracy reached by FixMatch after 1M iterations. It is manifest that the introduction of CPL successfully encourages the model to proactively learn those difficult classes thereby improving the overall learning effect.




(a) Threshold $\tau$ (b) Mapping function (c) Threshold warm-up

Figure 4: Ablation study of FlexMatch.

### 4.4 Ablation study

We conduct experiments to evaluate three components of FlexMatch: the upper limit of thresholds $\tau$, mapping functions $\mathcal{M}(x)$, and threshold warm-up.

**Threshold upper bound.** We investigate 5 different $\tau$ values and 3 different mapping functions on CIFAR-10 dataset with 40 labels. As shown in Figure 4(a), the optimal choice of $\tau$ is around $0.95$, either increasing or decreasing this value results in a performance decay. Note that in FlexMatch, tuning $\tau$ does not only affect the upper limit of the threshold but also the estimated learning effects because they are determined by the number of samples that fall above $\tau$.

**Mapping function.** We explore three different mapping functions in Figure 4(b): (1) concave: $\mathcal{M}(x) = \ln(x+1)/\ln 2$, (2) linear: $\mathcal{M}(x) = x$, and (3) convex: $\mathcal{M}(x) = x/(2-x)$ . We see that the convex function shows the best performance and the concave function shows the worst. Although tweaking the degree of convexity may probably lead to further improvement, we do not make further investigation in this paper. It is noteworthy that all these functions have their outputs grow from $0$ to $1$ when the inputs go from $0$ to $1$. One may also design a function with a different range, for instance, from $0.5$ to $1$. In this case, it is equivalent to setting a lower limit to the flexible threshold so that even at the beginning of the training, only samples with prediction confidence higher than this limit will contribute to the unsupervised loss. We do not include such a lower limit in FlexMatch since it will introduce a new hyperparameter. However, we did find that setting a lower limit at $0.5$ can slightly improve the performance. A possible reason is that the lower threshold prevents noisy training caused by incorrect pseudo labels at the early stage [34].

**Threshold warm-up.** We analyze the performance of threshold warm-up on both CIFAR-10 (40 labels) and CIFAR-100 (400 labels) datasets. As shown in Figure 4(c), the effect of threshold warm-up is mediocre on CIFAR-10 but is measurable on CIFAR-100. This is reasonable since CIFAR-100 is a more complicated dataset with more classes compared to CIFAR-10. At the beginning of the training without the threshold warm-up, the flexible thresholds may go through heavy fluctuations because the denominator in Eq.(6) is small. In the meantime, there will always be some classes whose flexible thresholds reach or approach $\tau$, thereby filtering out most unlabeled data in the batch. The threshold warm-up solves this issue by gradually raising the thresholds of all classes from zero – it creates a learning boom at the early training stage where most of the unlabeled data can be utilized.

## 5 TorchSSL: A PyTorch-based SSL Codebase

The PyTorch [28] framework has gained increasing attention in the deep learning research community. However, the main existing SSL codebase [35] is based on TensorFlow. For the convenience and customizability, we re-implement and open source a PyTorch-based SSL toolbox, named *TorchSSL* as shown in Figure 5. TorchSSL contains eight popular semi-supervised learning methods: Π-Model [9], Pseudo-Labeling [4], VAT [36], Mean Teacher [10], MixMatch [12], ReMixMatch [13], UDA [11], and FixMatch [14], along with our proposed method FlexMatch. Most of our implementation details are based on [35]. More importantly, in addition to the basic SSL methods and components, we implement several techniques to make the results stable under PyTorch framework. For instance, we add synchronized batch normalization [37] to avoid the performance degradation caused by multi-GPU training with small batch size, and a batch norm controller to prevent performance crashes

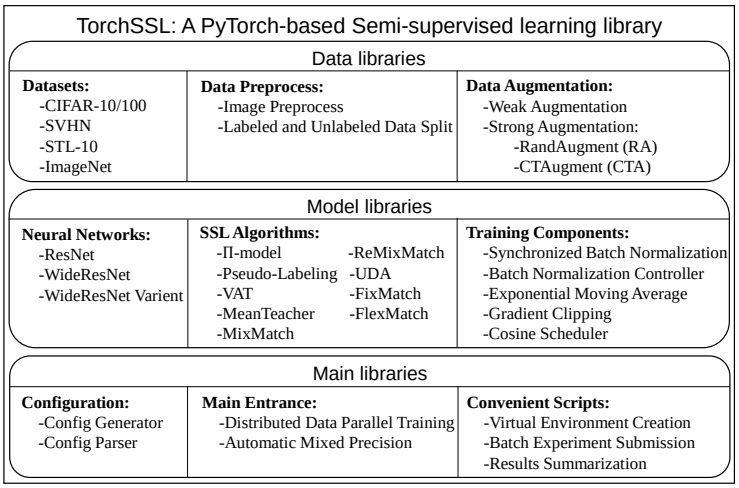



Figure 5: Components of TorchSSL.

for some algorithms, which is not officially supported in PyTorch. Detailed information of TorchSSL is presented in the appendix.

# 6 Related Work

Pseudo-Labeling [4] is a pioneer SSL method that uses hard artificial labels converted from model predictions. A confidence-based strategy was used in [6] along with pseudo labeling so that the unlabeled data are used only when the predictions are sufficiently confident. Such confidence-based thresholding also presents in recently proposed UDA [11] and FixMatch [14] with the difference being that UDA used sharpened 'soft' pseudo labels with a temperature whereas Fixmatch adopted one-hot 'hard' labels. The success of UDA and FixMatch, however, relies heavily on the usage of strong data augmentations to improve the consistency regularization. ReMixMatch [13] also leveraged such strong augmentations.

The combination of curriculum learning and semi-supervised learning is popular in recent years [38, 39, 40]. For multi-model image classification task, [38] optimized the learning process of unlabeled images by judging their reliability and discriminability. In [39], the easy image-level properties are learned first and then used to facilitate segmentation via constrained CNNs. Curriculum learning is also used to alleviate out-of-distribution problems by picking up in-distribution samples from unlabeled data according to the out-of-distribution scores [40].

Several researches have investigated on dynamic threshold in related fields such as sentiment analysis [41] and semantic segmentation [42]. In [41], the threshold was gradually reduced to make high-quality data selected into labeled data set in the early stage and large-quantity in the later stage. An extra classifier is added to automate the threshold to deal with domain inconsistency in [42].

# 7 Conclusion and Future Work

In this paper, we introduce Curriculum Pseudo Labeling (CPL), a curriculum learning approach of leveraging unlabeled data for SSL. CPL dramatically improves the performance and convergence speed of SSL algorithms that involve thresholds while being extremely simple and almost cost-free. FlexMatch, our improved algorithm of FixMatch, achieves state-of-the-art performance on a variety of SSL benchmarks. We hope that CPL can attract more future attention to explore the effectiveness of utilizing unlabeled data according to the model's learning status. In future work, we would like to improve our method under the long-tail scenario where the unlabeled data belonging to each class are extremely unbalanced.

# References


[1] Philip Bachman, Ouais Alsharif, and Doina Precup. Learning with pseudo-ensembles. In *NeurIPS*, pages 3365–3373, 2014.

[2] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *ICLR*, 2017.

[3] Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. In *NeurIPS*, pages 1171–1179, 2016.

[4] 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, 2013.

[5] 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.

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

[7] Henry Scudder. Probability of error of some adaptive pattern-recognition machines. *IEEE Transactions on Information Theory*, 11(3):363–371, 1965.

[8] 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.

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

[10] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIPS*, pages 1195–1204, 2017.

[11] Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. *NeurIPS*, 33, 2020.

[12] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. Mixmatch: A holistic approach to semi-supervised learning. *NeurIPS*, page 5050–5060, 2019.

[13] David Berthelot, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. Remixmatch: Semi-supervised learning with distribution matching and augmentation anchoring. In *ICLR*, 2019.

[14] 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. *NeurIPS*, 33, 2020.

[15] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *ICML*, pages 41–48, 2009.

[16] Yves Grandvalet, Yoshua Bengio, et al. Semi-supervised learning by entropy minimization. In *CAP*, pages 281–296, 2005.

[17] Anastasia Pentina, Viktoriia Sharmanska, and Christoph H Lampert. Curriculum learning of multiple tasks. In *CVPR*, pages 5492–5500, 2015.

[18] Lu Jiang, Deyu Meng, Qian Zhao, Shiguang Shan, and Alexander Hauptmann. Self-paced curriculum learning. In *AAAI*, volume 29, 2015.

[19] Alex Graves, Marc G Bellemare, Jacob Menick, Remi Munos, and Koray Kavukcuoglu. Automated curriculum learning for neural networks. In *ICML*, pages 1311–1320. PMLR, 2017.

[20] Guy Hacohen and Daphna Weinshall. On the power of curriculum learning in training deep networks. In *ICML*, pages 2535–2544. PMLR, 2019.

[21] Daphna Weinshall, Gad Cohen, and Dan Amir. Curriculum learning by transfer learning: Theory and experiments with deep networks. In *ICML*, pages 5238–5246. PMLR, 2018.

[22] Eric Arazo, Diego Ortego, Paul Albert, Noel E O'Connor, and Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *IJCNN*, pages 1–8, 2020.

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

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

[25] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *AISTATS*, pages 215–223, 2011.

[26] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee, 2009.

[27] Avital Oliver, Augustus Odena, Colin Raffel, Ekin D Cubuk, and Ian J Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. In *NeurIPS*, pages 3239–3250, 2018.

[28] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *NeurIPS*, 32:8026–8037, 2019.

[29] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *ICML*, pages 1139–1147. PMLR, 2013.

[30] Boris T Polyak. Some methods of speeding up the convergence of iteration methods. *Ussr computational mathematics and mathematical physics*, 4(5):1–17, 1964.

[31] Yu Nesterov. A method of solving a convex programming problem with convergence rate o (1/kˆ 2) o (1/k2). In *Sov. Math. Dokl*, volume 27.

[32] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *ICLR*, 2016.

[33] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 702–703, 2020.

[34] 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.

[35] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. fixmatch. `https://github.com/google-research/fixmatch`, 2020.

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

[37] Hang Zhang, Kristin Dana, Jianping Shi, Zhongyue Zhang, Xiaogang Wang, Ambrish Tyagi, and Amit Agrawal. Context encoding for semantic segmentation. In *CVPR*, pages 7151–7160, 2018.

[38] Chen Gong, Dacheng Tao, Stephen J Maybank, Wei Liu, Guoliang Kang, and Jie Yang. Multi-modal curriculum learning for semi-supervised image classification. *IEEE Transactions on Image Processing*, 25(7):3249–3260, 2016.

[39] Hoel Kervadec, Jose Dolz, Éric Granger, and Ismail Ben Ayed. Curriculum semi-supervised segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 568–576. Springer, 2019.

[40] Qing Yu, Daiki Ikami, Go Irie, and Kiyoharu Aizawa. Multi-task curriculum framework for open-set semi-supervised learning. In *ECCV*, pages 438–454. Springer, 2020.

[41] Yue Han, Yuhong Liu, and Zhigang Jin. Sentiment analysis via semi-supervised learning: a model based on dynamic threshold and multi-classifiers. *Neural Computing and Applications*, 32(9):5117–5129, 2020.

[42] Zhedong Zheng and Yi Yang. Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation. *IJCV*, 129(4):1106–1120, 2021.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 4.1
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] The code will be sent in the supplemental material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 4
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Table 1
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] This will be included in the appendix
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]