# Shrinking Class Space for Enhanced Certainty in Semi-Supervised Learning

Lihe Yang[1] Zhen Zhao[4] Lei Qi[5] Yu Qiao[3] Yinghuan Shi[2*] Hengshuang Zhao[1*]
[1]The University of Hong Kong [2]Nanjing University
[3]Shanghai AI Laboratory [4]The University of Sydney [5]Southeast University
https://github.com/LiheYoung/ShrinkMatch

## Abstract

*Semi-supervised learning is attracting blooming attention, due to its success in combining unlabeled data. To mitigate potentially incorrect pseudo labels, recent frameworks mostly set a fixed confidence threshold to discard uncertain samples. This practice ensures high-quality pseudo labels, but incurs a relatively low utilization of the whole unlabeled set. In this work, our key insight is that these uncertain samples can be turned into certain ones, as long as the confusion classes for the top-1 class are detected and removed. Invoked by this, we propose a novel method dubbed ShrinkMatch to learn uncertain samples. For each uncertain sample, it adaptively seeks a shrunk class space, which merely contains the original top-1 class, as well as remaining less likely classes. Since the confusion ones are removed in this space, the re-calculated top-1 confidence can satisfy the pre-defined threshold. We then impose a consistency regularization between a pair of strongly and weakly augmented samples in the shrunk space to strive for discriminative representations. Furthermore, considering the varied reliability among uncertain samples and the gradually improved model during training, we correspondingly design two reweighting principles for our uncertain loss. Our method exhibits impressive performance on widely adopted benchmarks.*

## 1. Introduction

In the last decade, our computer vision community has witnessed inspiring progress, thanks to large-scale datasets [21, 10]. Nevertheless, it is laborious and costly to annotate massive images, hindering the progress to benefit a broader range of real-world scenarios. Inspired by this, semi-supervised learning (SSL) was proposed to utilize the unlabeled data under the assistance of limited labeled data.

The frameworks in SSL are typically based on the strategy of pseudo labeling. Briefly, the model acquires knowledge from the labeled data, and then assigns predictions on the unlabeled data. The two sources of data are finally combined to train a better model. During this process, it is obvious that predictions on unlabeled data are not reliable. If the model is iteratively trained with incorrect pseudo labels, it will suffer the confirmation bias issue [1]. To address this dilemma, recent works [28] simply set a fixed confidence threshold to discard potentially unreliable samples. This simple strategy effectively retains high-quality pseudo labels, however, it also incurs a low utilization of the whole unlabeled set. As evidenced by our pilot study on CIFAR-100 [17], nearly 20% unlabeled images are filtered out for not satisfying the threshold of 0.95. Instead of blindly throwing them away, we believe there should exist a more promising approach. This work is just aimed to fully leverage previously uncertain samples in an *informative* but also *safe* manner.

So first, why are these samples uncertain? According to our observations on CIFAR-100 and ImageNet, although the top-1 accuracy could be low, the top-5 accuracy is much higher. This indicates in most cases, the model struggles to discriminate among a small portion of classes. As illustrated

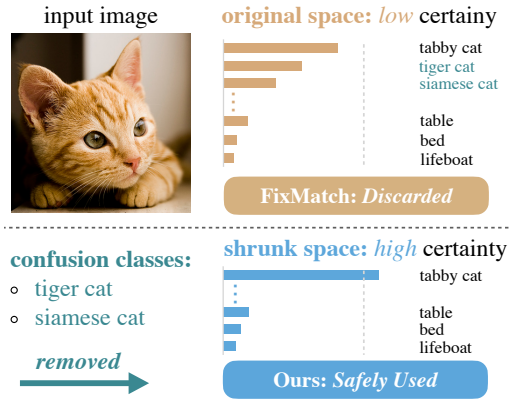



Figure 1: Illustration of our motivation. Due to confusion classes for the top-1 class, the certainty fails to reach the pre-defined threshold (gray dotted line). FixMatch discards such uncertain samples. Our method, however, detects and removes confusion classes to enhance certainty, then enjoying full and safe utilization of all unlabeled images.

*Corresponding authors

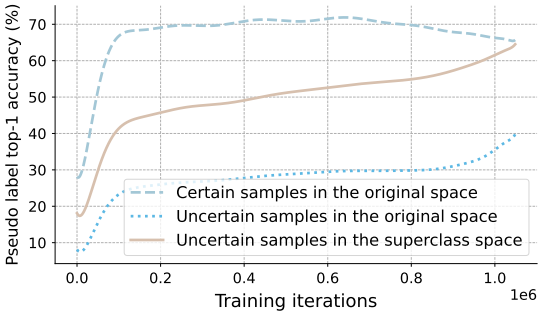



Figure 2: Pseudo label accuracy on CIFAR-100 with 400 labels. We highlight that even for uncertain samples, their top-1 predictions are of high accuracy *in the superclass space* (20 classes). This accuracy can even be comparable to the delicately selected certain samples in the original class space.

in Fig. 1, given a cat image, the model is not sure whether it belongs to `tabby cat`, `tiger cat`, or other cats. On the other hand, however, it is absolutely certain that the object is more like a `tabby cat`, rather than a `table` or anything else. In other words, it is reliable for the model to distinguish the top class from the remaining less likely classes.

Invoked by these, we propose a novel method dubbed ShrinkMatch to learn uncertain samples. The prediction fails to satisfy the pre-defined threshold due to the existence of confusion classes for the top-1 class. Hence, our approach adaptively seeks a shrunk class space where the confusion classes are removed, to enable the re-calculated confidence to reach the original threshold. Moreover, the obtained shrunk class space is also required to be the largest among those that can satisfy the threshold. In a word, we seek a *certain and largest shrunk space* for uncertain samples. Then, logits of the strongly augmented image are correspondingly gathered in the new space. And a consistency regularization is imposed between the shrunk weak-strong predictions.

Note that, the confusion classes are detected and removed in a fully automatic and instance-adaptive fashion. Moreover, even if the predicted top-1 class is not fully in line with the groundtruth label, they mostly belong to the same superclass. To prove this, as shown in Fig. 2, uncertain samples exhibit much higher pseudo label accuracy in the superclass space than the original space. This accuracy is even comparable to that of certain samples in the original class space. Therefore, contrasting these groundtruth-related classes against remaining unlikely classes is still highly beneficial, yielding more discriminative representations of our model. To avoid affecting the main classifier, we further adopt an auxiliary classifier to disentangle the learning in the shrunk space.

Despite the effectiveness, there still exist two main drawbacks to the above optimization target. (1) First, it treats all uncertain samples equally. The truth, however, is that in the original class space, the top-1 confidence of different uncertain samples can vary dramatically. And it is clear that samples with larger confidence should be attached more importance. To this end, we propose to balance different uncertain samples by their confidence in the original space. (2) Moreover, the regularization term also overlooks the gradually improved model state during training. At the start of training, there are abundant uncertain samples, but their predictions are extremely noisy, or even random. So even the highest scored class may share no relationship with the true class. Then as the training proceeds, the top classes become reliable. Considering this, we further propose to adaptively reweight the uncertain loss according to the model state. The model state is tracked and approximately estimated via performing exponential moving average on the proportion of certain samples in each mini-batch. With the two reweighting principles, the model turns out more stable, and avoids accumulating much noise from uncertain samples, especially at early training iterations.

To summarize, our contributions lie in three aspects:

- We first point out that low certainty is typically caused by a small portion of confusion classes. To enhance the certainty, we propose to shrink the original class space by adaptively detecting and removing confusion ones for the top-1 class to turn it certain in the new space.
- We manage to reweight the uncertain loss from two perspectives: the image-based varied reliability among different uncertain samples, and the model-based gradually improved state as the training proceeds.
- Our proposed ShrinkMatch establishes new state-of-the-art results on widely acknowledged benchmarks.

## 2. Related Work

**Semi-supervised learning (SSL).** The primary concern in SSL [19, 28, 36, 20, 25, 16, 37, 35, 41, 43, 14, 24, 29, 23, 4, 30, 11, 33, 2, 42] is to design effective constraints for unlabeled samples. Dating back to decades ago, pioneering works [19, 38] integrate unlabeled data via assigning pseudo labels to them, with the knowledge acquired from labeled data. In the era of deep learning, subsequent methods mainly follow such a bootstrapping fashion, but greatly boost it with some key components. Specifically, to enhance the quality of pseudo labels, Π-model [18] and Mean Teacher [31] ensemble model predictions and model parameters respectively. Later works start to exploit the role of perturbations. During this trend, [27] proposes to apply stochastic perturbations on inputs or features, and enforce consistency across these predictions. Then, UDA [34] emphasizes the necessity of strengthening the perturbation pool. It also follows VAT [22] and MixMatch [3] to supervise the prediction under strong perturbations with that under weak perturbations. Since then, weak-to-strong consistency regularization has become a standard practice in SSL. Eventually, the milestone

work FixMatch [28] presents a simplified framework using a fixed confidence threshold to discard uncertain samples. Our ShrinkMatch is built upon FixMatch. But, we highlight the value of previously neglected uncertain samples, and leverage them in an informative but also safe manner.

More recently, DST [7] decouples the generation and utilization of pseudo labels with a main and an auxiliary head respectively. Besides, SimMatch [45] explores instance-level relationships with labeled embeddings to supplement original class-level relations. Compared with them, our ShrinkMatch achieves larger improvements.

**Defining uncertain samples.** Earlier works estimate the uncertainty with Bayesian Neural Networks [15], or its faster approximation, *e.g.*, Monte Carlo Dropout [12]. Some other works measure the prediction disagreement among multiple randomly augmented inputs [32]. The latest trend is to directly use the entropy of predictions [39], cross entropy [36], or softmax confidence [28] as a measurement for uncertainty. Our work is not aimed at the optimal uncertainty estimation strategy, so we adopt the simplest solution from FixMatch, *i.e.*, using the maximum softmax output as the certainty.

**Utilizing uncertain samples.** UPS [26] leverages negative class labels whose confidence is below a pre-defined threshold, from a reversed multi-label classification perspective. In comparison, our model is enforced to tell the most likely class without being cheated by the less likely ones. So our supervision on uncertain samples is more informative and produces more discriminative representations. Moreover, we do not introduce any extra hyper-parameters, *e.g.*, the lower threshold in [26], into our framework.

# 3. Method

We primarily provide some notations and review a common practice in semi-supervised learning (SSL) (Sec. 3.1). Next, we present our ShrinkMatch in detail (Sec. 3.2 and Sec. 3.3). Finally, we summarize our approach and provide a further discussion (Sec. 3.4 and Sec. 3.5).

## 3.1. Preliminaries

Semi-supervised learning aims to learn a model with limited labeled images $\mathcal{D}^l = \{(x_k, y_k)\}$, aided by a large number of unlabeled images $D^u = \{u_k\}$. Recent frameworks commonly follow the FixMatch practice. Concretely, an unlabeled image $u$ is first transformed by a weak augmentation pool $\mathcal{A}^w$ and a strong augmentation pool $\mathcal{A}^s$ to yield a pair of weakly and strongly augmented images $(u^w, u^s)$. Then, they are fed into the model together to produce corresponding predictions $(p^w, p^s)$. Typically, $p^w$ is of much higher accuracy than $p^s$, while $p^s$ is beneficial to learning. Therefore, $p^w$ serves as the pseudo label for $p^s$. Moreover, a core practice introduced by FixMatch is that, to improve the quality of selected pseudo labels, a fixed confidence threshold is pre-defined to abandon uncertain ones. So the unsupervised loss $\mathcal{L}_u$ can be formulated as:

$$\mathcal{L}_u = \frac{1}{B_u} \sum_{k=1}^{B_u} \mathbb{1}(\xi(p_k^w) \geq \tau) \cdot \mathrm{H}(p_k^w, p_k^s), \tag{1}$$

where $B^u$ is the batch size of unlabeled images and $\tau$ is the pre-defined threshold. $\xi(p_k^w)$ computes the confidence of logits $p_k^w$ by $\xi(\cdot) = \max(\sigma(\cdot))$, where $\sigma$ is the softmax function. The H denotes the consistency regularization between the two distributions. It is typically the cross entropy loss.

In addition, the labeled images are learned with a regular cross entropy loss to obtain the supervised loss $\mathcal{L}_x$. The overall loss in each mini-batch then will be:

$$\mathcal{L} = \mathcal{L}_x + \lambda_u \cdot \mathcal{L}_u, \tag{2}$$

where $\lambda_u$ acts as a trade-off term between the two losses.

## 3.2. Shrinking the Class Space for Certainty

**Our motivation.** As reviewed above, FixMatch discards the samples whose confidence is lower than a pre-defined threshold, because their pseudo labels are empirically found relatively noisier. These samples are named uncertain samples in this work. Although this practice retains high-quality pseudo labels, it incurs a low utilization of the whole unlabeled set, especially when the scenario is challenging and the selection criterion is strict. Take the CIFAR-100 dataset as an instance, with a common threshold of 0.95 and 4 labels per class, there will be nearly 20% unlabeled samples being ignored due to their low certainty. We argue that, these uncertain samples can still benefit the model optimization, as long as we can design appropriate constraints (loss functions) on them. So we first investigate the cause of low certainty to gain some better intuitions.

The reason for the low certainty of an unlabeled image is that, the model tends to be confused among some top classes. For example, given a cat image, the score of the class `tabby cat` and class `tiger cat` may be both high, so the model is not absolutely certain what the concrete class is. Motivated by this observation, we propose to shrink the original class space via adaptively detecting and removing the confusion classes for the top-1 class. Then the shrunk space is only composed of the original top-1 class, as well as the remaining less likely classes. After this process, re-calculated confidence of the top-1 class will satisfy the pre-defined threshold. Thereby, we can enforce the model to learn the previously uncertain samples in this new certain space, as shown in Fig. 3. Following the previous cat example, if the top-1 class is `tabby cat`, our method will scan scores of all other classes, and remove confusion ones (*e.g.*, `tiger cat` and `siamese cat`) to construct a confident shrunk space, where the model is sure that the image is a `tabby cat`, rather than a `table`. Since we do not ask the model to discriminate among several top classes, it will avoid suffering from the noise when it makes a wrong judgment in the original space.

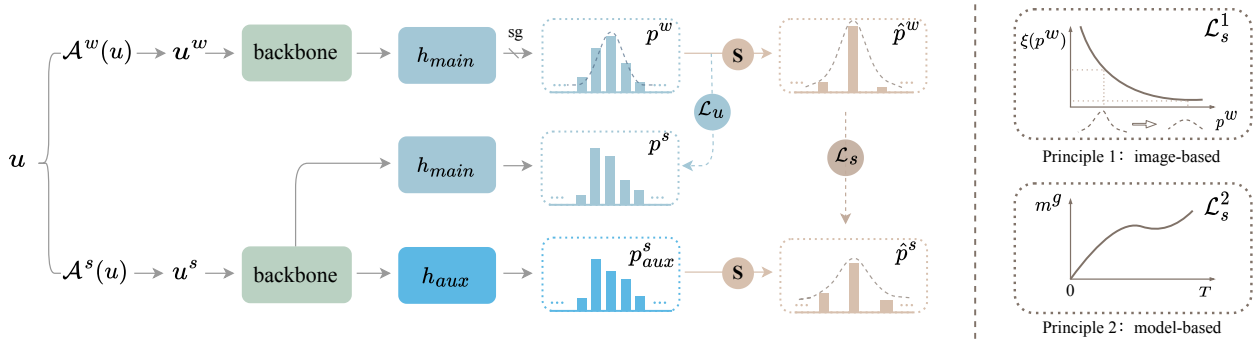



Figure 3: An overview of our proposed ShrinkMatch. Our motivation is to fully leverage the originally uncertain samples. "**S**" denotes **shrinking the class space**. The confusion classes for the top-1 class are detected and removed *in a fully automatic and instance-adaptive fashion*, to construct a shrunk space where the top-1 class is turned certain. $\mathcal{L}_u$ is the original certain loss, while $\mathcal{L}_s$ calculates the uncertain loss in the shrunk class space. We add an auxiliary head $h_{aux}$ to learn in the new space. On the right, we further reweight $\mathcal{L}_s$ based on two principles. Principle 1 (**image-based**): image predictions with larger reliability are attached more importance. Principle 2 (**model-based**): We track the model state during training for reweighting.

**How to seek the shrunk class space?** Now, how can we enable our method to *automatically* seek the optimal shrunk space of an uncertain unlabeled image? Ideally, we hope this seeking process is free from any prior knowledge from humans, *e.g.*, class relationships, and also does not require any extra hyper-parameters. Considering these, we opt to inherit the pre-defined confidence threshold as a criterion. To be specific, we first *sort* the predicted logits in a descending order to obtain $p^w = \{s^w_{n_i}\}_{i=1}^C$ for classes $\{n_i\}_{i=1}^C$, where $s^w_{n_i} \geq s^w_{n_{i+1}}$. In the shrunk space, we will retain the original top-1 class, because it is still the most likely one to be true. Then, we find a set of less likely classes $\{n_i\}_{i=K}^C$ by enforcing two constraints on the $K$:

$$\xi(\{s^w_{n_1}\} \cup \{s^w_{n_i}\}_{i=K}^C) \geq \tau, \tag{3}$$

$$\xi(\{s^w_{n_1}\} \cup \{s^w_{n_i}\}_{i=K-1}^C) < \tau, \tag{4}$$

where $\xi$ is defined the same as that in Eq. (1), calculating the confidence of the re-assembled logits. The final shrunk space is composed of the re-assembled classes $\{n_1\} \cup \{n_i\}_{i=K}^C$. The two constraints on $K$ not only ensure the top-1 class is turned certain in the new space (Eq. (3)), but also select the largest space among all candidates (Eq. (4)).

**How to learn in the shrunk class space?** For a *weakly* augmented uncertain image $x^w$, the model is certain about its top-1 class in the shrunk class space. To learn effectively in this space, we follow the popular practice of weak-to-strong consistency regularization. The *correspondingly shrunk* prediction on the *strongly* augmented image is enforced to match that on the *weakly* augmented one. Concretely, for clarity, the re-assembled logits from $p^w$ is denoted as $\hat{p}^w$, which means $\hat{p}^w = \{s^w_{n_1}\} \cup \{s^w_{n_i}\}_K^C$. We use the re-assembled classes $\{n_1\} \cup \{n_i\}_K^C$ in the shrunk space to correspondingly gather the logits $p^s$ on $x^s$, yielding $\hat{p}^s = \{s^s_{n_1}\} \cup \{s^s_{n_i}\}_K^C$. Then we can regularize the consistency between the two shrunk distributions $\hat{p}^s$ and $\hat{p}^w$, similar to that in Eq. (1):

$$\mathcal{L}_s = \frac{1}{B_u} \sum_{k=1}^{B_u} \mathbb{1}(\xi(p^w_k) < \tau) \cdot \hat{\mathrm{H}}(\hat{p}^w_k, \hat{p}^s_k), \tag{5}$$

where the indicator function is to find uncertain samples.

Empirically, we observe that if we use the original linear head $h_{main}$ to learn this auxiliary supervision, it will make the confidence of our model increase aggressively. Most *noisy* unlabeled samples are blindly judged as certain ones. We conjecture that it is because the $\mathcal{L}_s$ strengthens weights of the classes that are frequently uncertain, then these classes will be incorrectly turned certain. With this in mind, our solution is simple. We adopt an auxiliary MLP head $h_{aux}$ that shares the backbone with $h_{main}$, to deal with this auxiliary optimization target, as shown in Fig. 3. So the $\hat{p}^s$ is indeed gathered from predictions of $h_{aux}$ ($\hat{p}^w$ is still from $h_{main}$). This modification enables our feature extractor to acquire more discriminative representations, and meantime does not affect predictions of the main head. Note that $h_{aux}$ is only applied for training, bringing no burden to the test stage.

### 3.3. Reweighting the Uncertain Loss

Despite the effectiveness of the above uncertain loss, there still exist two main drawbacks. (1) On one hand, it overlooks the varied reliability of the top-1 class among different uncertain images. For example, suppose two uncertain images $u_1$ and $u_2$ with softmax predictions [0.8, 0.1, 0.1] and [0.5, 0.3, 0.2] in the original space, they should not be treated equally in the shrunk space. The $u_1$ with top-1 confidence of 0.8 is more likely to be true than $u_2$, and thereby should be attached more attention to. (2) On the other, it ignores the gradually improved model performance as the training proceeds. To be concrete, at the very start of training, the predictions are extremely noisy or even random. And then

at later stages, the predictions become more and more reliable. So the uncertain predictions at different training stages should not be treated equally. Therefore, we further design two reweighting principles for the two concerns.

**Principle 1: Reweighting with image-based varied reliability.** According to the above intuition, we directly reweight the uncertain loss of each uncertain image by its top-1 confidence $\xi(p_k^w)$ in the *original* class space, which is:

$$\mathcal{L}_s^1 = \frac{1}{B_u}\sum_{k=1}^{B_u} \mathbb{1}(\xi(p_k^w) < \tau) \cdot \hat{\mathrm{H}}(\hat{p}_k^w, \hat{p}_k^s) \cdot \xi(p_k^w). \quad (6)$$

We do not use the top-1 confidence $\xi(\hat{p}_k^w)$ in the *shrunk* space as the weight, because after shrinking, this value of different predictions is very close to each other. So generally, $\xi(p_k^w)$ is more discriminative than $\xi(\hat{p}_k^w)$ as the weight.

**Principle 2: Reweighting with model-based gradually improved state.** One naïve solution is to linearly increase the loss weight of $\mathcal{L}_s$ from 0 at the beginning to $\mu$ at iteration $T$, and then keep $\mu$ until the end. However, this practice has two severe disadvantages. First, the two additional hyper-parameters $\mu$ and $T$ are not easy to determine, and could be sensitive. More importantly, the linear scheduling criterion simply assumes the model state also improves linearly. Indeed, it can not be true. Thus, we here present a more promising principle to perform reweighting, that is free from any extra hyper-parameters. It can adjust the loss weight according to the model state in a fully adaptive fashion. To be specific, we use the certain ratio of the unlabeled set as an indicator for the model state. The certain ratio is traced at each iteration and accumulated globally in an exponential moving average (EMA) manner. Formally, the certain ratio in a single mini-batch is given by:

$$m = \frac{1}{B_u}\sum_{k=1}^{B_u} \mathbb{1}(\xi(p_k^w) \geq \tau). \quad (7)$$

The *global* certain ratio $m^g$ is initialized as 0, and accumulated at each training iteration by:

$$m^g \leftarrow \gamma \cdot m^g + (1-\gamma) \cdot m, \quad (8)$$

where $\gamma$ is the momentum coefficient. It is a hyper-parameter already defined in FixMatch (baseline), where it is used to update the teacher parameters for final evaluation.

Obviously, $m^g$ falls between 0 and 1. And it will approximately increase from 0 to a nearly saturated value. Then, the reweighted uncertain loss is given by:

$$\mathcal{L}_s^2 = m^g \cdot \frac{1}{B_u}\sum_{k=1}^{B_u} \mathbb{1}(\xi(p_k^w) < \tau) \cdot \hat{\mathrm{H}}(\hat{p}_k^w, \hat{p}_k^s). \quad (9)$$

Integrating the above two intuitions and principles, the final reweighted uncertain loss will be:

$$\mathcal{L}_s = m^g \cdot \frac{1}{B_u}\sum_{k=1}^{B_u} \mathbb{1}(\xi(p_k^w) < \tau) \cdot \hat{\mathrm{H}}(\hat{p}_k^w, \hat{p}_k^s) \cdot \xi(p_k^w). \quad (10)$$

### 3.4. Summary

To summarize, the final loss in a mini-batch is a combination of the supervised loss ($\mathcal{L}_x$), certain loss ($\mathcal{L}_u$, Eq. (1)), and uncertain loss in the shrunk class space ($\mathcal{L}_s$, Eq. (10)):

$$\mathcal{L} = \mathcal{L}_x + \lambda_u \cdot (\mathcal{L}_u + \mathcal{L}_s). \quad (11)$$

We do not carefully fine-tune the fusion weight between $\mathcal{L}_u$ and $\mathcal{L}_s$ , but use 1:1 by default to avoid hyper-parameters.

### 3.5. Discussions

Our uncertain loss in the shrunk space owns two properties: **informative** and also **safe**. The first property is because we manage to find the largest shrunk space that reaches the confidence threshold. Besides, we also adopt the weak-to-strong consistency regularization to pose a challenging optimization target. Both constraints ensure the learning in the shrunk space is not trivial and still quite informative. Next, we wish to explain the second property "safe", especially about *how we avoid noise in the shrunk space*.

Noises in pseudo labels distinguish the semi-supervised paradigm from the fully-supervised one. So designing a safe optimization target for unlabeled data is crucial. Typically, the cross entropy loss will maximize the softmax probability $\exp(s_t)/\sum_{i=1}^{C}\exp(s_i) \rightarrow 1$ for the target class $t$. It inevitably suppresses scores of all other classes except $t$. However, the true class could not be $t$, and may be the 2$^{\text{nd}}$ or 3$^{\text{rd}}$ largest class, *etc*., which is wrongly restrained. This is frequently observed when the confidence of class $t$ is not high enough. As a milder alternative, the soft labeling still encounters a similar problem. In contrast, our shrunk target directly *excludes* these confusion classes. So only the almost unlikely classes are suppressed.

**It is worth noting**, even if the predicted top-1 class is not in line with the human label, it is probably one of the closest semantics, *e.g*., belonging to the same superclass as the groundtruth label, as shown in Fig. 2. Thus, contrasting these relevant classes against other less likely classes with our noise-robust shrunk loss is still beneficial. The model is encouraged to make discriminative predictions closer to top classes compared to tail classes. In addition, our disentangled auxiliary head can effectively leverage such supervision while not affecting the main classification tasks [7].

## 4. Experiment

In this section, we first describe the implementation details of our framework. Then, we compare our ShrinkMatch with previous state-of-the-art methods (SOTAs) under extensive evaluation protocols. Lastly, we conduct comprehensive ablation studies on each component to validate the necessity.

| Seed | 0 | 1 | 2 | 3 | 4 | Mean |
|---|---|---|---|---|---|---|
| SimMatch [45] | 95.34 | 95.16 | 92.63 | 93.76 | 95.10 | 94.39 |
| **ShrinkMatch** \| 40 | 95.09 | 94.66 | 95.12 | 94.78 | 94.95 | **94.92** |
| SimMatch [45] | 95.58 | 95.50 | 95.34 | 94.06 | 95.26 | 95.15 |
| **ShrinkMatch** \| 250 | 95.39 | 95.44 | 95.36 | 94.76 | 95.35 | **95.26** |

Table 1: Comparison with SOTAs on **CIFAR-10**. The same seed ensures exactly the same data split. The 400 or 2500 denotes the number of labels.

| Seed | 0 | 1 | 2 | 3 | 4 | Mean |
|---|---|---|---|---|---|---|
| SimMatch [45] | 62.06 | 60.19 | 59.89 | 64.88 | 63.92 | 62.19 |
| **ShrinkMatch** \| 400 | 65.00 | 63.47 | 63.77 | 66.42 | 64.52 | **64.64** |
| SimMatch [45] | 74.64 | 75.19 | 74.53 | 75.03 | 75.24 | **74.93** |
| **ShrinkMatch** \| 2500 | 75.00 | 75.11 | 74.54 | 74.78 | 74.72 | 74.83 |

Table 2: Comparison with SOTAs on **CIFAR-100**.

### 4.1. Implementation Details

**Baselines.** We use FixMatch + distribution alignment (DA) as our baseline on all datasets except ImageNet and SVHN. On ImageNet, we build our method on SimMatch. On SVHN, we discard DA. To be more convincing, we adopt the same codebase as our compared methods.

**Hyper-parameters.** Following prior arts, Wide ResNet-28-2 [40], WRN-28-8, WRN-37-2, and WRN-28-2 are used for CIFAR-10, CIFAR-100, STL-10, and SVHN respectively. A ResNet-50 [13] is used for ImageNet. The auxiliary head $h_{aux}$ for uncertain samples is a 3-layer MLP. On the ImageNet, we set $B_u = 64 \times 5$, but for other datasets, $B_u = 64 \times 7$. The labeled batch size is 64 for all datasets. On the ImageNet, the model is trained for 400 epochs, while on the others, it is trained for $2^{20}$ iterations. The initial learning rate is set as 0.03 for all datasets with a cosine scheduler. Specially, on the ImageNet, it is warmed up for 5 epochs. The consistency regularization H in $\mathcal{L}_u$ on STL-10 and SVHN is a hard cross entropy (CE) loss, while on the CIFAR-10/100 and ImageNet, following the SimMatch and CoMatch [20], it is modified to a soft CE loss. The $\hat{\text{H}}$ in our proposed $\mathcal{L}_s$ is simply a hard CE loss. The weight $\lambda_u$ for the two unsupervised losses is set as 10 on the ImageNet and 1 on others. The confidence threshold $\tau$ is 0.7 on the ImageNet and 0.95 for others. The shared momentum coefficient $\gamma$ between our global certain ratio $m^g$ and the teacher parameters is 0.999. Following the common practice, the teacher model is only maintained for final evaluation.

### 4.2. CIFAR-10 and CIFAR-100

The CIFAR dataset is composed of 50000/10000 training/validation images of size $32 \times 32$. The CIFAR-10 contains 10 balanced classes to recognize, while CIFAR-100 is much more challenging, containing 100 classes. We run our ShrinkMatch on five different seeds and report each result as well as the average result. On the CIFAR-10 of Tab. 1 where performance is almost saturated, we still obtain a non-trivial improvement of 0.53% (94.39% $\rightarrow$ 94.92%) with 4 labels per class (400 labels in total). In addition, notably, as shown in Tab. 2, with only 4 labels per class on the challenging CIFAR-100, our ShrinkMatch remarkably outperforms SimMatch by 2.45% (62.19% $\rightarrow$ 64.64%) on average.

| Seed | 0 | 1 | 2 | Mean |
|---|---|---|---|---|
| FixMatch [28] | 65.85 | 67.94 | 58.30 | 64.03 |
| FlexMatch [42] | 76.71 | 68.28 | 67.55 | 70.85 |
| **ShrinkMatch** \| 40 | 85.75 | 85.64 | 86.55 | **85.98** |
| FixMatch [28] | 90.91 | 88.71 | 90.94 | 90.19 |
| FlexMatch [42] | 91.35 | 92.29 | 91.66 | **91.77** |
| **ShrinkMatch** \| 250 | 91.13 | 92.43 | 91.10 | 91.55 |

Table 3: Comparison with SOTAs on **STL-10**.

| Seed | 0 | 1 | 2 | Mean |
|---|---|---|---|---|
| FixMatch [28] | 94.53 | 96.90 | 97.14 | 96.19 |
| FlexMatch [42] | 89.19 | 89.93 | 96.32 | 91.81 |
| **ShrinkMatch** \| 40 | 97.96 | 97.81 | 96.70 | **97.49** |
| FixMatch [28] | 98.00 | 97.99 | 97.94 | 97.98 |
| FlexMatch [42] | 91.76 | 91.83 | 96.65 | 93.41 |
| **ShrinkMatch** \| 250 | 98.08 | 98.06 | 97.98 | **98.04** |

Table 4: Comparison with SOTAs on **SVHN**.

### 4.3. STL-10 and SVHN

The STL-10 dataset originally consists of 5K/100K/8K labeled/unlabeled/validation images. We follow FlexMatch to only select a subset of 40/250 labeled images, while the unlabeled set remains unchanged. As shown in Tab. 3, with 4 labels per class, our ShrinkMatch surpasses FlexMatch tremendously from 70.85% to 85.98% (+15.13%).

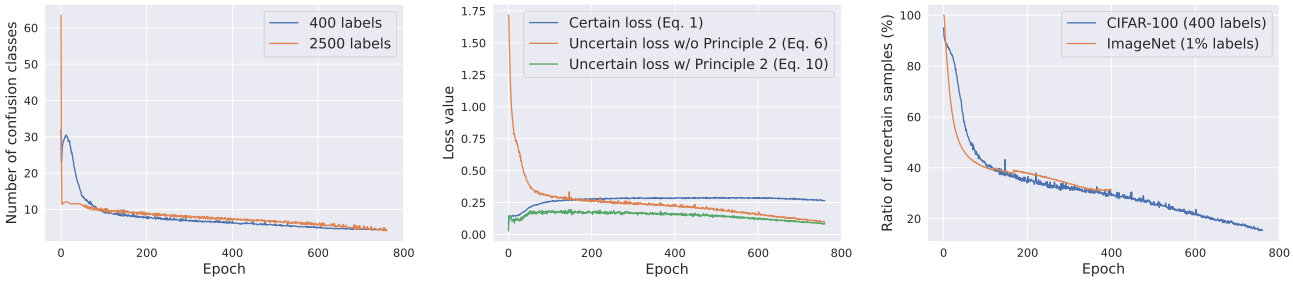
The SVHN dataset is suitable to reveal the ability of different semi-supervised algorithms in the presence of class imbalance issue. As demonstrated in Tab. 4, although the performance of the original FixMatch (our baseline) has almost touched the upper bound, we still further boost it from 96.19% to 97.49% (+1.30%) with 4 labels per class.

### 4.4. ImageNet-1K

The ImageNet dataset is rather challenging, containing 1.28M/50K training/validation images of 1000 classes. We exactly follow the codebase of SimMatch [45]. As shown in Tab. 5, our ShrinkMatch further boosts previous best results on both settings of 1% and 10% labeled images.

| Pre-training | Method | Epochs | Params (train / test) | Top-1 1% | Top-1 10% | Top-5 1% | Top-5 10% |
|---|---|---|---|---|---|---|---|
| SimCLR v2 [8] | Fine-tune | 800 | 34.2M / 25.6M | 57.9 | 68.4 | 82.5 | 89.2 |
| SwAV [6] | | 800 | 30.4M / 25.6M | 53.9 | 70.2 | 78.5 | 89.9 |
| WCL [44] | | 800 | 34.2M / 25.6M | 65.0 | 72.0 | 86.3 | 91.2 |
| MoCo v2 [9] | Fine-tune | 800 | 30.0M / 25.6M | 49.8 | 66.1 | 77.2 | 87.9 |
| | CoMatch [20] | 1200 | 30.0M / 25.6M | 67.1 | 73.7 | 87.1 | 91.4 |
| MoCo-EMAN [5] | FixMatch-EMAN [5] | 1100 | 30.0M / 25.6M | 63.0 | 74.0 | 83.4 | 90.9 |
| None | CoMatch [20] | 400 | 30.0M / 25.6M | 66.0 | 73.6 | 86.4 | 91.6 |
| | SimMatch† [45] | 400 | 30.0M / 25.6M | 67.0 | 74.1 | 86.9 | 91.5 |
| | **ShrinkMatch** | 400 | 31.8M / 25.6M | **67.5** | **74.5** | **87.4** | **91.9** |

Table 5: Accuracy (%) on the **ImageNet-1K** with 1% and 10% labeled images. †: Reproduced in our environment.




(a) Number of removed confusion classes. (b) Certain loss $\mathcal{L}_u$ and uncertain loss $\mathcal{L}_s$. (c) Ratio of uncertain samples.

Figure 4: (a) The number of removed confusion classes for each uncertain image as the training proceeds. (b) Value of different loss functions, *i.e.*, Eq. (1), Eq. (6), and Eq. (10). (c) The ratio of uncertain samples (certainty $< \tau$) in each mini-batch.

| Seed | 0 | 1 | 2 | 3 | 4 | Mean |
|---|---|---|---|---|---|---|
| Baseline | 64.40 | 57.51 | 63.07 | 64.77 | 62.53 | 62.46 |
| **ShrinkMatch** | 65.00 | 63.47 | 63.77 | 66.42 | 64.52 | **64.64** |

Table 6: Ablation studies of ShrinkMatch on **CIFAR-100**.

| Seed | 0 | 1 | 2 | Mean |
|---|---|---|---|---|
| Baseline | 84.86 | 84.21 | 85.47 | 84.85 |
| **ShrinkMatch** | 85.75 | 85.64 | 86.55 | **85.98** |

Table 7: Ablation studies of ShrinkMatch on **STL-10**.

### 4.5. Ablation Studies

Unless otherwise specified, we conduct our ablation studies on CIFAR-100 with 4 labels per class.

**Effectiveness of our holistic ShrinkMatch.** We first view our ShrinkMatch as a holistic component added to our baseline. As displayed in Tab. 6 for CIFAR-100, our ShrinkMatch boosts the strong baseline significantly by 2.18% (62.46% $\rightarrow$ 64.64%). And on the STL-10 of Tab. 7, the baseline is also improved evidently from 84.85% to 85.98% (+1.13%).

**Effectiveness of two reweighting principles.** In Sec. 3.3, considering the varied prediction reliability of different uncertain samples and different training stages, we propose two principles to reweight our uncertain loss $\mathcal{L}_s$. So here we carefully examine their necessity in Tab. 8. It can be observed that the two principles are both beneficial, jointly improving the basic shrinking practice from 63.16% to 64.24%. We also attempt a simple alternative of linearly increasing the uncertain loss weight from 0 (start) to 1 (end). But as evidenced in the Exp 3 & 4 of Tab. 8, it is obviously inferior to our designed reweighting strategy. We visualize our $m^g$ as training proceeds on ImageNet in Fig. 5. It looks essentially different from linear scheduling. Moreover, we visualize the uncertain loss value with or without the second reweighting principle in Fig. 4b. After reweighting, the uncertain loss is of a similar magnitude as the certain loss, which can avoid dominating the gradient at early training stages.

**The number of removed confusion classes.** As introduced before, our approach detects and removes the confusion classes for the top-1 class in a fully automatic and instance-adaptive fashion. So we visualize the number of removed confusion classes for each uncertain sample in Fig. 4a. At the very start of training, model predictions are almost uniform, so abundant classes are removed for an uncertain sample. But as the training proceeds, much fewer confusion classes need to be removed to form a confident shrunk class space.

**Hard label or soft label in the shrunk class space?** By

| Exp | Principle 1 | Principle 2 | LS | Seed 0 | Seed 1 | Mean |
|---|---|---|---|---|---|---|
| 1 | | | | 65.62 | 60.70 | 63.16 |
| 2 | ✓ | | | 65.68 | 62.08 | 63.88 |
| 3 | | ✓ | | 65.09 | 62.90 | 64.00 |
| 4 | | | ✓ | 63.01 | 59.45 | 61.23 |
| 5 | ✓ | ✓ | | 65.00 | 63.47 | **64.24** |

Table 8: Ablation studies on the effectiveness of the two reweighting principles. Due to randomness, the "Mean" results are more convincing. **LS** is short for linear scheduling, which linearly increases the loss weight.

| Exp | Shrink | Aux Head | Label | Seed 0 | Seed 1 | Mean |
|---|---|---|---|---|---|---|
| 1 | Our Strong Baseline | | | 64.40 | 57.51 | 60.96 |
| 2 | ✓ | ✓ | H | 65.00 | 63.47 | **64.24** |
| 3 | ✓ | ✓ | S | 62.92 | 59.76 | 61.34 |
| 4 | ✓ | | H | 62.43 | 58.86 | 60.65 |
| 5 | ✓ | | S | 60.92 | 60.22 | 60.57 |
| 6 | | ✓ | H | 64.21 | 59.17 | 61.69 |
| 7 | | ✓ | S | 64.75 | 60.89 | 62.82 |
| 8 | | | H | 52.90 | 51.88 | 52.39 |
| 9 | | | S | 54.10 | 55.75 | 54.93 |

Table 9: Ablation studies on different options to learn uncertain sample. (1) whether to shrink the class space (**Shrink**), (2) whether to use an auxiliary head (**Aux Head**), and (3) use hard (**H**) or soft label (**S**). Exp 1 is our baseline without using uncertain images. Exp 8 & 9 are experiments of directly learning uncertain samples, which are quite noisy.

default, we use the hard label in the shrunk space, following FixMatch. But we also ablate the choice of using the soft label as supervision. As shown in Exp 2 & 3 of Tab. 9, the hard label is much better than the soft label. We conjecture that this is because we have shrunk the class space to a safe one. The soft label will incur some unnecessary noise.

**Effects of the auxiliary head.** To prevent the main classifier from being affected, we use an auxiliary head to learn the uncertain samples in the shrunk class space only for discriminative representations. So we first provide the results when still using the main head for the shrunk space. As shown in Exp 2 & 4 of Tab. 9, the auxiliary head is indispensable for our shrinking practice. So we continue to figure out whether the improvement of our ShrinkMatch merely comes from the auxiliary head. We attempt to use the auxiliary head to directly learn the uncertain samples in the original space without shrinking. It can be concluded from Exp 2 & 6 & 7 of Tab. 9, the original space is indeed much inferior to our shrunk space for uncertain samples, although the auxiliary head still helps the baseline to some extent. Lastly, in Tab. 9, we also include two additional results for readers, *i.e.*, directly learning uncertain samples without shrinking or

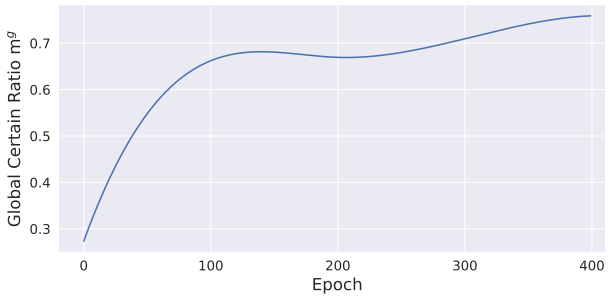



Figure 5: Visualization of our global certain ratio $m^g$.

| Threshold | 0.98 | 0.95 | 0.90 | 0.80 | 0.70 |
|---|---|---|---|---|---|
| ShrinkMatch | **65.80** | 65.00 | 64.55 | 62.74 | 59.58 |

Table 10: Ablation studies of the confidence threshold.

auxiliary head (Exp 8 & 9), which are quite terrible because of introducing the abundant noise to our main classifier.

**The ratio of uncertain samples.** To further justify our motivation, in Fig. 4c, we display the ratio of uncertain samples in each mini-batch. It can be seen that even in the middle of the whole training course, there are still around 40% and 30% uncertain samples on ImageNet and CIFAR-100 respectively. Therefore, an appropriate approach to utilizing these samples is necessary and definitely beneficial.

**Different confidence thresholds $\tau$.** We also try different thresholds for uncertain samples, as shown in Tab. 10. When increasing $\tau$ from widely adopted 0.95 to 0.98 (more uncertain samples), our ShrinkMatch can be further improved. These results clearly highlight the effective and safe utilization of uncertain samples with our ShrinkMatch.

## 5. Conclusion

In this work, we aim to fully leverage the uncertain samples in semi-supervised learning. We point out that the low certainty is typically caused by a small portion of confusion classes. Invoked by this, we propose a novel method dubbed ShrinkMatch to automatically detect and remove the confusion classes to construct a shrunk class space, where the top-1 class is turned certain. A weak-to-strong consistency regularization is enforced in the confident new space. Furthermore, we design two reweighting principles for the auxiliary uncertain loss, according to the reliability of different uncertain samples and the gradually improved state of the model. Consequently, our method establishes new state-of-the-art results on widely acknowledged benchmarks.

**Acknowledgements.** This work is supported in part by the National Natural Science Foundation of China (62201484, 62222604, 62192783, 62206052), HKU Startup Fund, and HKU Seed Fund for Basic Research.

## References


[1] 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*, 2020. 1

[2] Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Armand Joulin, Nicolas Ballas, and Michael Rabbat. Semi-supervised learning of visual features by non-parametrically predicting view assignments with support samples. In *ICCV*, 2021. 2

[3] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *NeurIPS*, 2019. 2

[4] Zhaowei Cai, Avinash Ravichandran, Paolo Favaro, Manchen Wang, Davide Modolo, Rahul Bhotika, Zhuowen Tu, and Stefano Soatto. Semi-supervised vision transformers at scale. In *NeurIPS*, 2022. 2

[5] Zhaowei Cai, Avinash Ravichandran, Subhransu Maji, Charless Fowlkes, Zhuowen Tu, and Stefano Soatto. Exponential moving average normalization for self-supervised and semi-supervised learning. In *CVPR*, 2021. 7

[6] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *NeurIPS*, 2020. 7

[7] Baixu Chen, Junguang Jiang, Ximei Wang, Jianmin Wang, and Mingsheng Long. Debiased pseudo labeling in self-training. In *NeurIPS*, 2022. 3, 5

[8] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. In *NeurIPS*, 2020. 7

[9] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv:2003.04297*, 2020. 7

[10] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 1

[11] Yue Fan, Dengxin Dai, Anna Kukleva, and Bernt Schiele. Cossl: Co-learning of representation and classifier for imbalanced semi-supervised learning. In *CVPR*, 2022. 2

[12] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *ICML*, 2016. 3

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6

[14] Zijian Hu, Zhengyu Yang, Xuefeng Hu, and Ram Nevatia. Simple: similar pseudo label exploitation for semi-supervised classification. In *CVPR*, 2021. 2

[15] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *NeurIPS*, 2017. 3

[16] Jiwon Kim, Youngjo Min, Daehwan Kim, Gyuseong Lee, Junyoung Seo, Kwangrok Ryoo, and Seungryong Kim. Conmatch: Semi-supervised learning with confidence-guided consistency regularization. In *ECCV*, 2022. 2

[17] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 1

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

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

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

[21] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 1

[22] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *TPAMI*, 2018. 2

[23] Islam Nassar, Samitha Herath, Ehsan Abbasnejad, Wray Buntine, and Gholamreza Haffari. All labels are not created equal: Enhancing semi-supervision via label grouping and co-training. In *CVPR*, 2021. 2

[24] Yassine Ouali, Céline Hudelot, and Myriam Tami. An overview of deep semi-supervised learning. *arXiv:2006.05278*, 2020. 2

[25] Hieu Pham, Zihang Dai, Qizhe Xie, and Quoc V Le. Meta pseudo labels. In *CVPR*, 2021. 2

[26] 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. In *ICLR*, 2021. 3

[27] Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. In *NeurIPS*, 2016. 2

[28] 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. In *NeurIPS*, 2020. 1, 2, 3, 6

[29] Fariborz Taherkhani, Ali Dabouei, Sobhan Soleymani, Jeremy Dawson, and Nasser M Nasrabadi. Self-supervised wasserstein pseudo-labeling for semi-supervised image classification. In *CVPR*, 2021. 2

[30] Hui Tang and Kui Jia. Towards discovering the effectiveness of moderately confident samples for semi-supervised learning. In *CVPR*, 2022. 2

[31] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIPS*, 2017. 2

[32] Guotai Wang, Wenqi Li, Michael Aertsen, Jan Deprest, Sébastien Ourselin, and Tom Vercauteren. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing*, 2019. 3

[33] Xudong Wang, Long Lian, and Stella X Yu. Unsupervised selective labeling for more effective semi-supervised learning. In *ECCV*, 2022. 2

[34] Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. Unsupervised data augmentation for consistency training. In *NeurIPS*, 2020. 2

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

[36] 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 *ICML*, 2021. 2, 3

[37] Fan Yang, Kai Wu, Shuyi Zhang, Guannan Jiang, Yong Liu, Feng Zheng, Wei Zhang, Chengjie Wang, and Long Zeng. Class-aware contrastive semi-supervised learning. In *CVPR*, 2022. 2

[38] David Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *ACL*, 1995. 2

[39] Lequan Yu, Shujun Wang, Xiaomeng Li, Chi-Wing Fu, and Pheng-Ann Heng. Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation. In *MICCAI*, 2019. 3

[40] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *BMVC*, 2016. 6

[41] Xiaohua Zhai, Avital Oliver, Alexander Kolesnikov, and Lucas Beyer. S4l: Self-supervised semi-supervised learning. In *ICCV*, 2019. 2

[42] Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. In *NeurIPS*, 2021. 2, 6

[43] Zhen Zhao, Luping Zhou, Lei Wang, Yinghuan Shi, and Yang Gao. Lassl: Label-guided self-training for semi-supervised learning. In *AAAI*, 2022. 2

[44] Mingkai Zheng, Fei Wang, Shan You, Chen Qian, Changshui Zhang, Xiaogang Wang, and Chang Xu. Weakly supervised contrastive learning. In *ICCV*, 2021. 7

[45] Mingkai Zheng, Shan You, Lang Huang, Fei Wang, Chen Qian, and Chang Xu. Simmatch: Semi-supervised learning with similarity matching. In *CVPR*, 2022. 3, 6, 7