Safe-Student for Safe Deep Semi-Supervised Learning with Unseen-Class Unlabeled Data

Rundong He[*]
Shandong University
rundong_he@mail.sdu.edu.cn

Zhongyi Han[*]
Shandong University
hanzhongyicn@gmail.com

Xiankai Lu
Shandong University
carrierlxk@gmail.com

Yilong Yin[†]
Shandong University
ylyin@sdu.edu.cn

Abstract

Deep semi-supervised learning (SSL) methods aim to take advantage of abundant unlabeled data to improve the algorithm performance. In this paper, we consider the problem of safe SSL scenario where unseen-class instances appear in the unlabeled data. This setting is essential and commonly appears in a variety of real applications. One intuitive solution is removing these unseen-class instances after detecting them during the SSL process. Nevertheless, the performance of unseen-class identification is limited by the small number of labeled data and ignoring the availability of unlabeled data. To take advantage of these unseen-class data and ensure performance, we propose a safe SSL method called ***SAFE-STUDENT*** *from the teacher-student view. Firstly, a new scoring function called energy-discrepancy (ED) is proposed to help the teacher model improve the security of instances selection. Then, a novel unseen-class label distribution learning mechanism mitigates the unseen-class perturbation by calibrating the unseen-class label distribution. Finally, we propose an iterative optimization strategy to facilitate teacher-student network learning. Extensive studies on several representative datasets show that SAFE-STUDENT remarkably outperforms the state-of-the-art, verifying the feasibility and robustness of our method in the under-explored problem.*

1. Introduction

The recent remarkable success of deep learning methods attributes to the advancements of learning algorithms and the availability of large-scale labeled data [17, 30]. However, large-scale annotated data is scarce and costly to collect for various real-world applications, such as medical image analysis [12, 29] and image classification [14, 27]. In contrast, unlabeled data is plentiful and cheap to collect, stimulating the great research interest of SSL. Currently, the research of deep SSL methods can be categorized into three main branches: consistency regularization [18, 31, 33], pseudo-labeling [3, 30, 36], and hybrid methods [1, 2, 32].

Although current deep SSL methods can enhance the performance of learning models, the prerequisite is that the unlabeled set derives from the identical distribution with the labeled set. Once this condition is not satisfied in real scenarios, the performance of SSL models degenerates significantly [4, 9, 27]. Many real-world tasks involve the situation where unlabeled data contains some unseen-class instances. For example, at the beginning stage of the outbreak of COVID-19, the unlabeled data inevitably contains some imperceptible COVID-19 instances in the deep SSL pneumonia classification [11]. These unseen-class instances heavily hinder the safety of the pneumonia classification model. We define this case as the problem of *Safe Deep semi-supervised learning with Unseen-class unlabeled data* (SDU), which accommodates a variety of real-world applications but is rarely considered in the literature.

Considering the unseen-class instances contained in unlabeled data hurt the performance of seen-class classification easily, one intuitive way is to detect these unseen-class instances by unseen-class identification methods, remove them, and then use any existing SSL method to obtain promising performance. Nevertheless, the performance of unseen-class identification is limited by the small number of labeled data. Several deep safe SSL methods are proposed to alleviate the limitations [3, 4, 9, 37]. These deep safe SSL methods select reliable seen-class instances from unlabeled data to increase the number of labeled seen-class instances. However, the selected seen-class instances are inaccurate due to limited identification ability and the lack of suit-

[*] Contribute to this work equally
[†] Corresponding author

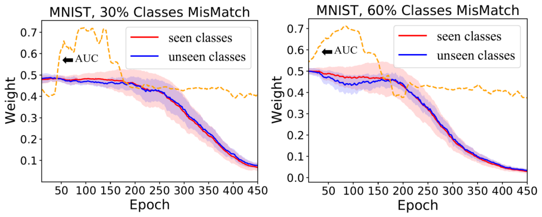


Figure 1. With the increase of training iterations of DS^{3}L, the learned weights of seen-class and unseen-class instances tend to be consistent and negligible. Shaded regions indicate the range of the weights over unlabeled instances. The orange denotes the AUC of unseen-class identification.

able scoring function. For example, in the pioneering work DS^{3}L [9], as training progresses (see Fig. 1), the learned weights of seen-class and unseen-class instances tend to be consistent and small, which causes the poor performance of unseen-class identification. Moreover, the valuable unseen-class data in unlabeled data set are not used effectively, which indeed contribute to unseen-class identification.

To improve unseen-class identification, we propose a novel SAFE-STUDENT from the teacher-student view. Specifically, we propose a novel scoring function called energy-discrepancy (ED) that addresses the maximum-logit dominance problem of the existing energy function. This new scoring function provides a better indication for the teacher model to capture unseen-class data in the unlabeled set. Then, we propose the unseen-class label distribution learning module to help the student model calibrate the unseen-class label distribution. Therefore, the overconfidence problem of the student model to unseen-class data is mitigated. Finally, we further propose an iterative optimization strategy to boost the performance of the whole teacher-student models.

We summarize our contributions as follows:

- We investigate a limited-explored problem, safe deep SSL learning with unseen-class unlabeled data, and propose a novel safe deep SSL framework called SAFE-STUDENT.
- We raise a new scoring function called ED to help the teacher model improve the performance of unseen-class identification, and verify the consistency between the optimization objective and ED.
- We propose an unseen-class label distribution learning module that adequately calibrates the unseen-class probability distribution into a uniform distribution.
- We carry out extensive experiments on four benchmark datasets. Extensive results verify the learnability of this new problem and the robustness of SAFE-STUDENT.

2. Related Work

Deep Semi-Supervised Learning. Deep SSL has made revolutionary advances to diverse machine learning problems. The remarkable success attributes to the advancements of learning algorithms and the utilization of abundant unlabeled data. Deep SSL methods mainly contain three categories: consistency regularization methods, pseudo-labeling methods, and hybrid methods. Consistency regularization methods [18, 31, 33] take advantage of unlabeled data by assuming that the model should output similar predictions for any image and its perturbed version. Pseudo-labeling methods [3, 21, 28, 30, 36] use the model itself to obtain artificial labels of unlabeled data. Hybrid methods [1, 2, 32] combine both consistency regularization and pseudo-labeling, and use data augmentation [5, 7, 35]. However, the effectiveness of these methods is based on that all labeled and unlabeled data come from the same distribution. Once the assumption is broken, the performance of seen-class classification would degrade and even fall below the performance of supervised learning methods [4, 9, 27].

Safe Semi-Supervised Learning. Safe SSL ensures the performance of SSL methods is no worse than a simple supervised learning model. The study about safe SSL problem mainly refer to three aspects [22]: data quality [9, 10, 39], model uncertainty [23], and measure diversity [22]. In this paper, we focus on the data quality that the unlabeled data contains some unseen-class instances. One intuitive solution is removing these unseen-class instances after detecting them during the SSL process. Thus, the performance of unseen-class identification is the key to mitigating the side effects of unseen-class unlabeled data.

Guo *et al.* [9] assigned soft weights to each unlabeled instance by a weighting function, which can be viewed as an unseen-class detector. Then, Chen *et al.* [4] used the model to identify unseen-class instances at the beginning of the training time but which is unstable. Yu *et al.* [37] identified unseen-class by considering labeled data as in-distribution data and unlabeled data as out-of-distribution data. However, the performance of unseen-class identification is limited by the small number of in-distribution data and noisy out-of-distribution data. In [3], labeled data is used to train a supervised model and then identify unseen classes based on the model confidence. However, as the training process progresses, all unlabeled instances are absorbed into the seen-class set. The above representative works suffer from the poor performance due to the inaccurate unseen-class identification and the lack of suitable scoring function. Moreover, the valuable unseen-class data in unlabeled data set are not used effectively, which indeed contribute to unseen-class identification. Different from these works, this paper provides a novel and elegant approach by improving the unseen-class identification so as to reduce the perturbation on the seen-classifier.

3. SAFE-STUDENT

3.1. Learning Set-up

Definition 1 *(Distribution for Safe SSL Scenario.) Given a feature space $\mathcal{X} \subset \mathbb{R}^d$ and the label space $\mathcal{Y}$, the labeled and unlabeled data have different joint distributions $P(X^l, Y^l)$ and $P(X^u, Y^u)$, where the feature space $X^l, X^u \subset \mathcal{X}$ and the label space $Y^l, Y^u \subset \mathcal{Y}$.*

Definition 2 *(Safe Deep semi-supervised learning with Unseen-class unlabeled data (SDU).) Let $D_L = \{(\boldsymbol{x}_i^l, y_i^l)\}_{i=1}^m$ denote the labeled data set, where m denotes the number of labeled data, $\boldsymbol{x}_i^l \in X^l$, $y_i^l \in Y^l$. Let $D_U = \{\boldsymbol{x}_i^u\}_{i=1}^n$ denote the unlabeled data set, where n denotes the number of unlabeled data, $\boldsymbol{x}_i^u \in X^u$, and $m \ll n$. $Y^l \subset Y^u$ and $Y^{new} = Y^u \backslash Y^l$, where Y^{new} denotes unseen classes that only emerge in the unlabeled set D_U. Let $K = |Y^l|$ denote the number of seen classes.*

When $Y^{new} \neq \emptyset$, unseen classes lead to class distribution mismatch, which damages the performance of seen-class classification. SAFE-STUDENT is to keep the seen-class classification model safe.

As shown in Fig. 2, SAFE-STUDENT mainly contains four modules: *teacher pre-training module* (Sec. 3.2) to obtain a teacher model that serves as a mentor for the student model, *seen and unseen classes identification module* (Sec. 3.3) to select reliable seen-class and unseen-class instances, *seen-class learning module* (Sec. 3.4) to achieve the seen-class classification, *unseen-class label distribution learning module* (Sec. 3.5) to mitigate the adverse effects of unseen classes. Moreover, *iterative optimization strategy* (Sec. 3.6) helps the teacher model improve the identification of unseen classes and helps the student model improve the performance of seen-class classification.

3.2. Teacher Pre-Training

A good teacher model lays solid foundations for assigning accurate pseudo-labels to seen-class unlabeled data, estimating the uncertainties of unseen-class unlabeled data, and initializing the student model. Accordingly, we use the data augmentation (aug) strategy to pre-train a reliable teacher model. For any image $\boldsymbol{x}_i^l$ with label y_i^l from D_L, we perform the augmentation strategy on $\boldsymbol{x}_i^l$ denoted by $\text{aug}(\boldsymbol{x}_i^l)$. The augmentation strategy combines several carefully selected data augmentation manners to prevent the semantic information change of labeled data. The teacher model is optimized by the cross-entropy loss as follows:

$$\mathcal{L}_{CE} = -\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{K} y_{ij}^l \log\left(\frac{e^{f_j(\text{aug}(\boldsymbol{x}_i^l))}}{\sum_{j=1}^{K} e^{f_j(\text{aug}(\boldsymbol{x}_i^l))}}\right), \quad (1)$$

where $f_j(\text{aug}(\boldsymbol{x}_i^l)) = [\eta \circ \Phi\left(\text{aug}(\boldsymbol{x}_i^l)\right)]_j$ denotes the j^{th} index of logits passed trough a classifier head η connected to a feature extractor Φ, K denotes the number of seen classes, and y_{ij}^l corresponds to the j^{th} element of one-hot encoded ground-truth label of the instance $\boldsymbol{x}_i^l$.

3.3. Seen and Unseen Classes Identification

Unlike previous works that only select reliable seen-class instances [3, 4, 37], we believe that reliable unseen-class instances also contribute to solving the SDU problem. We propose to utilize reliable unseen-class instances for the unseen-class label distribution learning such that we can easily detect them (elaborated in Sec. 3.5). Accordingly, the objective of this seen and unseen classes identification (SUCI) module is to select out reliable seen-class instances and unseen-class instances simultaneously.

We propose a novel scoring function called energy-discrepancy (ED) to measure the teacher model outputs. Denote by D_{sc} and D_{uc} the reliable seen-class set and unseen-class set, respectively. Both sets are selected by:

$$\begin{aligned} D_{sc} &= \{\boldsymbol{x}_i^u | (\text{ED}(\boldsymbol{x}_i^u) > \tau_1), i = 1, \dots, n\}, \\ D_{uc} &= \{\boldsymbol{x}_i^u | (\text{ED}(\boldsymbol{x}_i^u) < \tau_2), i = 1, \dots, n\}, \end{aligned} \quad (2)$$

where $\tau_1 > \tau_2$ are thresholds, τ_1 is set according to the mismatch ratio in D_U, and $\tau_2 = \max\{\tau_1 - 0.1, 0\}$. If the mismatch ratio is not known in advance, we can consider the unseen-class instances in D_U as open set noises and then easily infer it by [24].

Energy-Discrepancy. In the previous study of out-of-distribution (OOD) detection [25], energy is used to measure the physical properties and the outlier degrees of OOD data. Unlike confidence, energy is theoretically aligned with the probability density of the inputs and is less susceptible to the overconfidence issue [25]:

$$\text{E}(\boldsymbol{x}_i^u) = -T' \cdot \log \sum_{j=1}^{K} e^{f_j(\boldsymbol{x}_i^u)/T'}, \quad (3)$$

where T' denotes the temperature parameter, $\text{E}(\cdot)$ represents energy. The higher negative energy of $\boldsymbol{x}_i^u$ means the $\boldsymbol{x}_i^u$ more likely belongs to the seen classes [25]. Energy is effective to detect unseen classes in many open tasks, like open world object detection [15].

However, we find that the energy cannot accurately identify seen and unseen classes instances because the energy is always dominated by the maximum logit. For example, Fig. 3 illustrates four instances' logit distributions and corresponding negative energy values. The third instance (c) intuitively belongs to the unseen class according to the logit distribution; however, it has the largest negative energy value, which breaks our intuition. In contrast, compared to the first and fourth instances, the second instance is more likely to belong to the seen class; however, it has a lower negative energy value, which again breaks the agreement on the energy function. Therefore, it is not accurate to identify seen and unseen classes by energy.

We propose a novel scoring function called energy-discrepancy to give full play to the role of non-maximum logits such that we can mitigate the dominance problem of

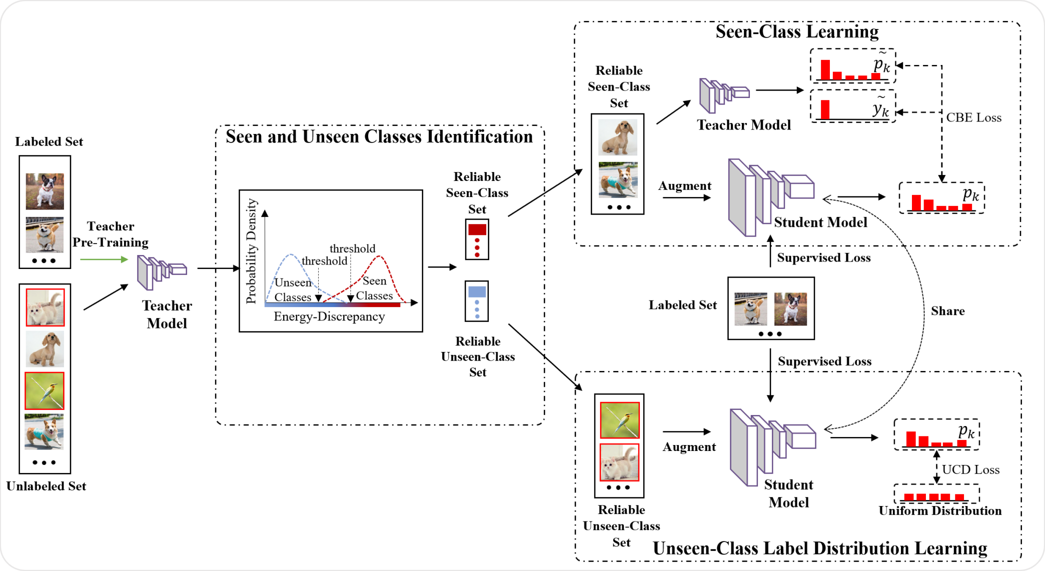


Figure 2. The overview of SAFE-STUDENT for deep safe SSL with unseen-class unlabeled data. It mainly consists of four modules. Especially, SAFE-STUDENT uses a new scoring function to achieve seen-class and unseen-class identification, and uses an unseen-class label distribution learning module to improve unseen-class identification.

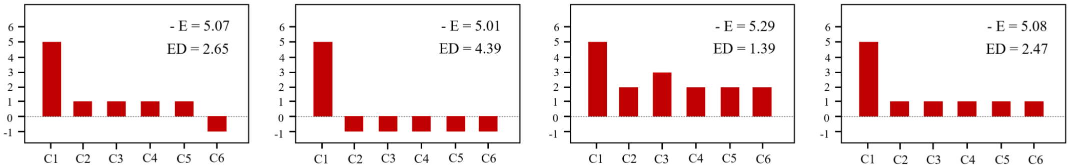


Figure 3. Logit distributions of four instances with negative Energy (-E) and Energy-Discrepancy (ED).

maximum logit. The key novelty of ED lies in the measurement of the negative energy discrepancy between the original negative energy and the updated negative energy after removing the maximum logit. Specificity, we define ED by:

$$\begin{aligned}\mathrm{ED}(\boldsymbol{x}_i^u) &= T^{'} \cdot \log \sum_{j=1}^{K} e^{f_j(\boldsymbol{x}_i^u)/T^{'}} \\ &- T^{'} \cdot \log \left[\sum_{j=1}^{K} e^{f_j(\boldsymbol{x}_i^u)/T^{'}} - \max\{e^{f_j(\boldsymbol{x}_i^u)/T^{'}}\}_{j=1}^{K} \right].\end{aligned} \tag{4}$$

The first item in Eq. (4) denotes the original negative energy of instance $\boldsymbol{x}_i^u$. Since the energy is dominated by the maximum logit, it is not accurate enough to select reliable seen-class instances and unseen-class instances by using negative energy directly. Accordingly, we recalculate the negative energy of instance $\boldsymbol{x}_i^u$ after removing the maximum logit. The updated negative energy is denoted by the second item in Eq. (4). Then, we utilize the discrepancy between the original negative energy and the updated negative energy to select reliable samples, renamed energy-discrepancy (ED).

Compared to Energy, ED improves the selection accuracy from the perspective of discrepancy and gives full play to the role of non-maximum logits. It considers the whole distribution of logits and can reflect the uncertainty change brought by each logit (see Figs. 3(a) and (d)). Meanwhile, ED retains the essential characteristics of energy, which is less susceptible to the overconfidence issue. The higher ED of $\boldsymbol{x}_i^u$ means the lower the uncertainty of $\tilde{y}_i^u$. Correspondingly, $\boldsymbol{x}_i^u$ more likely belongs to the seen classes. The ED values of the four instances in Fig. 3 accurately represent their uncertainties, which demonstrates that ED contributes to selecting more reliable instances. ED is a very general method and can be flexibly embedded into other frameworks, such as adversarial defense frameworks, etc.

Energy-Discrepancy vs Energy. Then, we perform an in-depth analysis of energy and ED. Let T' equal 1 in Eq. (3). Suppose the ground truth label of sample $\boldsymbol{x}$ is y. Then we can get

$$\begin{aligned}\mathrm{E}(\boldsymbol{x}) &= -\log \left[e^{f_y(\boldsymbol{x})} \cdot \left(e^{f_1(\boldsymbol{x}) - f_y(\boldsymbol{x})} + \cdots + e^{f_K(\boldsymbol{x}) - f_y(\boldsymbol{x})} \right) \right] \\ &= -\log e^{f_y(\boldsymbol{x})} - \log \left(e^{f_1(\boldsymbol{x}) - f_y(\boldsymbol{x})} + \cdots + e^{f_K(\boldsymbol{x}) - f_y(\boldsymbol{x})} \right) \\ &\approx -f_y(\boldsymbol{x}) .\end{aligned} \tag{5}$$

Formally speaking, for each $\boldsymbol{x}$, we define the sub-optimal label by $y' = \arg\max_{\bar{y}\in\mathcal{Y}\setminus\{y\}} f_{\bar{y}}(\boldsymbol{x})$. After removing the logit corresponding to the optimal label, the new energy score is $\mathrm{E}^{'}(\boldsymbol{x})$, which is

$$\begin{aligned}\mathrm{E}^{'}(\boldsymbol{x}) &= -f_{y'}(\boldsymbol{x}) - \log\left(e^{f_1(\boldsymbol{x})-f_{y'}(\boldsymbol{x})} + \cdots + e^{f_K(\boldsymbol{x})-f_{y'}(\boldsymbol{x})}\right) \\ &\geq -f_{y'}(\boldsymbol{x}) - \log(K-1)\,. \end{aligned} \tag{6}$$

And due to $\log\left(e^{f_1(\boldsymbol{x})-f_{y'}(\boldsymbol{x})} + \cdots + e^{f_K(\boldsymbol{x})-f_{y'}(\boldsymbol{x})}\right) \geq 0$, we can know that $\mathrm{E}^{'}(\boldsymbol{x}) \leq -f_{y'}(\boldsymbol{x})$. So we can know $-f_{y'}(\boldsymbol{x}) - \log(K-1) \leq \mathrm{E}^{'}(\boldsymbol{x}) \leq -f_{y'}(\boldsymbol{x})$. Let $\mathrm{E}^{'}(\boldsymbol{x}) = -f_{y'}(\boldsymbol{x})$, and the bias is less than $\log(K-1)$. Then, we can get $\mathrm{ED}(\boldsymbol{x})$ as follows,

$$\mathrm{ED}(\boldsymbol{x}) = \left|\mathrm{E}(\boldsymbol{x}) - \mathrm{E}^{'}(\boldsymbol{x})\right| \approx f_y(\boldsymbol{x}) - f_{y'}(\boldsymbol{x})\,. \tag{7}$$

According to Eq. (7), we can know ED is approximately the discrepancy between the logit of the optimal label and the logit of the sub-optimal label. Moreover, the final loss function Eq. (11) is aimed at enlarging ED of seen classes by the first two items and minimizing ED of unseen classes by the last item, which verifies the consistency between the optimization objective and our proposed scoring function ED. On the contrary, energy is inconsistent with optimization objective, which causes inaccurate unseen-class identification. Please see Appendix for specific analysis.

In summary, SUCI uses the newly proposed ED that contributes to the accurate selection of seen-class and unseen-class data from the view of discrepancy property. To ensure the safe SUCI, we design the following learning paradigm.

3.4. Seen-Class Learning

In practice, the student's performance on seen-class classification is restricted when learning from inaccurate pseudo-labels due to confirmation bias of the teacher model. When only using the vanilla cross-entropy loss function, the student model will be easily confused by false pseudo-labels. The reason is that the cross-entropy loss function focuses on learning a hyperplane for discriminating each class from the other classes, causing unsatisfactory performance on the test set. Therefore, vanilla cross-entropy loss is not a suitable optimization objective when confirmation bias exists. To overcome the confirmation bias, a new and effective loss function is designed, which considers both probability distribution and pseudo label.

Specifically, we propose a confirmation bias elimination (CBE) loss function to mitigate the reliance on pseudo-labels and boost the tolerance to inaccurate pseudo-labels. Different from the vanilla cross-entropy loss function, the CBE loss function attempts to measure the probability distributions between an original instance $\boldsymbol{x}_i^u$ and its augmented version $\mathrm{aug}(\boldsymbol{x}_i^u)$. Note that the probability distribution of $\boldsymbol{x}_i^u$ is taken from the output of the teacher model. In contrast, the probability distribution of $\mathrm{aug}(\boldsymbol{x}_i^u)$ is taken from the student model. Such a way mitigates the local optima in which the network might get stuck due to weak guidance. Formally, we define the confirmation bias elimination (CBE) loss function to optimize the student model by

$$\begin{aligned}\mathcal{L}_{CBE} = &\frac{1}{|D_{sc}|}\sum_{\boldsymbol{x}_i^u\in D_{sc}} \left[-\tilde{y}_i^u \log\left(\tilde{\eta}\circ\Phi\left(\mathrm{aug}(\boldsymbol{x}_i^u)\right)\right)\right. \\ &\left.+\tilde{p}_i^u \log\frac{\tilde{p}_i^u}{\tilde{\eta}\circ\Phi\left(\mathrm{aug}(\boldsymbol{x}_i^u)\right)}\right]\,,\end{aligned} \tag{8}$$

where $|D_{sc}|$ denotes the number of instances in D_{sc}, $\tilde{y}_i^u$ denotes hard pseudo label of instance $\boldsymbol{x}_i^u$ from D_{sc} by teacher model, $\tilde{p}_i^u$ denotes the probability distribution of instance $\boldsymbol{x}_i^u$ by teacher model, and $\tilde{\eta}$ denotes the classifier head η followed by softmax function.

To further mitigate the confirmation bias, we also incorporate Eq. (1) to use the labeled set D_L that includes a small amount of valuable seen-class data with ground-truth labels.

3.5. Unseen-Class Label Distribution Learning

To some degree, the seen and unseen classes identification module mitigates the overconfidence problem on unseen classes by using ED. The strategy is based on the predictions of the network, but the overconfidence problem of the network itself still exists. Accordingly, there is still a part of hard unseen-class instances that make the teacher model overconfidence, which confuses the seen and unseen classes identification module. Overconfident predictions with low uncertainties cause the seen-class set to contain dangerous unseen-class instances. These dangerous instances disrupt the seen-class distribution, resulting in degradation of seen-class classification performance [19].

Inspired by label distribution learning [8, 34], we propose the unseen-class label distribution learning module to increase the unseen-class uncertainty by calibrating the unseen-class probability distribution. Based on the unseen instances set D_{uc} selected by the SUCI module, the unseen-class label distribution learning module attempts to make the probability distribution of D_{uc} converge to the uniform distribution $\mathcal{U}$. In practice, we optimize the Kullback-Leibler (KL) divergence to minimize the difference between the probability distribution of D_{uc} and $\mathcal{U}$ by

$$\mathcal{L}_{UCD} = \frac{1}{|D_{uc}|}\sum_{\boldsymbol{x}_i^u\in D_{uc}} \omega(\boldsymbol{x}_i^u)\mathrm{KL}(\mathcal{U}(y)\|\tilde{\eta}\circ\Phi((\mathrm{aug}(\boldsymbol{x}_i^u)))\,, \tag{9}$$

where $|D_{uc}|$ denotes the number of instances in D_{uc}, $\omega(\boldsymbol{x}_i^u)$ denotes the weight of sample $\boldsymbol{x}_i^u$,

$$\omega(\boldsymbol{x}_i^u) = \exp\left(\frac{\max_{\boldsymbol{x}_j^u\in D_{uc}}\mathrm{ED}(\boldsymbol{x}_j^u) - \mathrm{ED}(\boldsymbol{x}_i^u)}{\max_{\boldsymbol{x}_j^u\in D_{uc}}\mathrm{ED}(\boldsymbol{x}_j^u)}\right)\,. \tag{10}$$

The mined unseen-class data have different reliability levels, so we perform exponential weighting on these data according to Eq. (10), and $1 \leq \omega(\boldsymbol{x}_i^u) \leq e$. The exponential weighting can amplify the role of reliable unseen-class data. Since KL divergence measures the distribution discrepancy, its optimization can adequately calibrate the unseen-class probability distribution. The calibration strategy mitigates the overconfidence of the student model to unseen-class instances. The teacher model, after iterative optimization, achieves more accurate predictions, which help the SUCI module identify seen-class and unseen-class instances more easily. Therefore, we could mitigate the perturbation of unseen-class instances to the seen-class distribution.

3.6. Iterative Optimization

Considering Eqs. (1), (8), and (9), we state the overall loss as:

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda_1 \mathcal{L}_{CBE} + \lambda_2 \mathcal{L}_{UCD}\,, \tag{11}$$

where λ_1 and λ_2 are coefficients. We adopt an iterative training way to optimize SAFE-STUDENT as follows. We first fully pre-train the teacher model by Eq. (1), then share its parameters to a new student model. After that, we fix the teacher model to select reliable seen-class and unseen-class sets used to train the student model. Once this round of training is completed, the student model performs better than the teacher model. Then, to improve the teacher model's ability to guide the next round of student training, we update the teacher model with the technology of exponential moving average (EMA) [33] and the student model's parameters at the current round. Next, with the guidance of the updated teacher model, SAFE-STUDENT trains the student model with a large learning rate and keeps decreasing the learning rate until the end of the current training round. Utilizing a large learning rate aims to prevent the model from falling into local optimal points. By using the iterative strategy, the seen-class classification performance of the student model can be continuously improved. The iterative training of these modules provides a guarantee for the student model. Algorithm 1 summarizes the detailed optimization procedure, which can be seen in Appendix.

3.7. Convergence Analysis

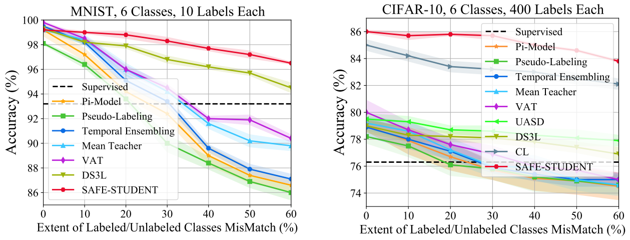
We further analyze the convergence of optimization process in SAFE-STUDENT. Proof can be seen in Appendix.

Theorem 1 *(Convergence.) Suppose the final loss function $\mathcal{L}$ is Lipschitz-smooth with constant $L \leq 2$ and ρ-bounded gradients, then by following our optimization algorithm, the loss $\mathcal{L}$ always decreases along with the iteration t, and step size $\eta_{\theta_s} \leq 1$, $\mathcal{L}\left(\theta_s^{t+1}\right) \leq \mathcal{L}\left(\theta_s^t\right)$.*

4. Experiments

4.1. Experimental Setup

Datasets. We evaluate SAFE-STUDENT on image classification datasets: MNIST [20], CIFAR-10 [16], CIFAR-100 [16], and TinyImageNet (a subset of ImageNet [6]).



(a) MNIST

(b) CIFAR-10

Figure 4. Seen-class classification accuracy (%) of SAFE-STUDENT and compared deep SSL methods on MNIST and CIFAR-10 with different class mismatch ratios between labeled and unlabeled data. Shaded regions indicate standard deviation over five runs.

Baselines. We compare **SAFE-STUDENT** on test data that only contain seen-class instances with SSL baselines: **Pseudo-Labeling** [21], **Pi-Model** [31], **Temporal Ensembling** [18], **Mean Teacher** [33], Virtual Adversarial Training (**VAT**) [26], **FixMatch** [32], **UASD** [4], **DS^{3}L** [9], Multi-Task Curriculum (**MTC**) [37], and Curriculum Labeling (**CL**) [3]. Moreover, we make the **supervised method** trained on D_L as another baseline.

4.2. Results

4.2.1 MNIST

Due to the small size of images, we adopt a simple two-layer CNN model as a backbone network [9]. The network is trained using stochastic gradient descent (SGD) with a learning rate of $1e^{-3}$. We train the model for 500 epochs with a batch size of 100. Half of the batch size comes from labeled data, and the other half comes from unlabeled data. By optimizing Eq. (11), we obtain a safe student model.

Fig. 4(a) reports the average accuracy of different deep SSL methods over five runs with different class mismatch ratios. From Fig. 4(a), we can find that the performance of all the deep SSL methods, including our proposal, is better than the supervised learning method when the extent of labeled/unlabeled class mismatch ratio is 0%. However, as the class distribution mismatch increases, the performance of the existing deep SSL method drops rapidly. Many deep SSL methods even under-perform the supervised learning method when the extent of labeled/unlabeled class mismatch ratio is 40%. Although these safe SSL methods still exceed the supervised learning method when the mismatch ratio is 60%, our proposed SAFE-STUDENT is significantly better and can reach **96.5%** averaged accuracy, which is about **3.3%** higher than the supervised method.

4.2.2 CIFAR-10

As for CIFAR-10, we use the standard Wide ResNet [38], *i.e.*, WRN-28-2, as the base network for training. Both the teacher model and the student model are based on the same

Table 1. Seen-class classification accuracy (%) of different methods on the four datasets.

Method	MNIST		CIFAR-10		CIFAR-100		TinyImagenet	
	ratio=0.3	ratio=0.6	ratio=0.3	ratio=0.6	ratio=0.3	ratio=0.6	ratio=0.3	ratio=0.6
Supervised	93.2±0.3	93.2±0.3	76.3±0.4	76.3±0.4	58.6±0.5	58.6±0.5	36.5±0.5	36.5±0.5
Pi-Model [31]	92.4±0.6	86.6±0.5	75.7±0.7	74.5±1.0	59.4±0.3	57.9±0.3	36.9±0.4	36.4±0.5
PL [21]	90.0±0.7	86.0±0.6	75.8±0.8	74.6±0.7	60.2±0.3	57.5±0.6	36.6±0.6	35.8±0.4
VAT [26]	94.5±0.3	90.4±0.3	76.9±0.6	75.0±0.5	61.8±0.4	59.6±0.6	36.7±0.5	36.3±0.6
FixMatch [32]	-	-	81.5±0.2	80.9±0.3	65.9±0.3	65.2±0.3	-	-
DS^{3}L [9]	96.8±0.3	94.5±0.4	78.1±0.4	76.9±0.5	-	-	-	-
UASD [4]	96.2±0.6	94.3±0.8	77.6±0.4	76.0±0.4	61.8±0.4	58.4±0.5	37.1±0.7	36.9±0.6
MTC [37]	93.7±0.5	88.5±0.3	85.5±0.6	81.7±0.5	63.1±0.6	61.1±0.3	37.0±0.5	36.6±0.4
CL [3]	96.9±0.1	95.6±0.4	83.2±0.4	82.1±0.4	63.6±0.4	61.5±0.5	37.3±0.7	36.7±0.8
SAFE-STUDENT	**98.3±0.3**	**96.5±0.1**	**85.7±0.3**	**83.8±0.1**	**68.4±0.2**	**68.2±0.1**	**37.7±0.3**	**37.1±0.3**

Table 2. Seen-class classification accuracy (%) of ablation studies on CIFAR-10 with the 60% ratio of class distribution mismatch under the different numbers of labeled data.

Method	m=100×6	m=200×6	m=400×6
SAFE-STUDENT w/o ED	67.0±0.5	75.1±0.2	82.0±0.2
SAFE-STUDENT w/o AUG	67.2±0.3	76.2±0.2	82.3±0.2
SAFE-STUDENT w/o CBE	66.9±0.8	75.3±0.4	82.6±0.2
SAFE-STUDENT w/o UCDL	65.4±1.1	75.1±0.7	81.5±0.3
SAFE-STUDENT w/o IO	67.5±0.9	75.8±0.6	81.8±0.4
SAFE-STUDENT w/o EW	69.5±0.7	76.9±0.5	82.9±0.3
SAFE-STUDENT	**69.9±0.3**	**78.3±0.4**	**83.8±0.1**

network. The network is trained using SGD with a learning rate of 0.128 like [36], a momentum of 0.9, and a weight decay of $5e^{-4}$. We train the model for 400 epochs with a batch size of 512, where half of the batch comes from labeled data, and the other half comes from unlabeled data. The number of iterative optimization rounds is 3.

Fig. 4(b) reports the averaged accuracy on CIFAR-10 over five runs with the different class mismatches. Firstly, our proposed SAFE-STUDENT significantly outperforms existing deep SSL methods. For example, when the mismatch ratio is 60%, our method achieves **83.8%** averaged accuracy, about **7.5%** higher than the supervised learning method, about **6.9%** higher than DS^{3}L. These results verify that our method achieves the best performance compared with the safe deep SSL methods on the SDU problem. Note that our method still outperforms all the compared methods when the class mismatch ratio is 0%.

4.2.3 CIFAR-100 and TinyImageNet

Similar to CIFAR-10, we use WRN-28-2 as the base network for training on CIFAR-100 and TinyImageNet. The rest experimental setups are the same as CIFAR-10. The third and fourth columns of Table 1 show the results on CIFAR-100 and TinyImageNet with the 30% and 60% of unseen-class instances in unlabeled data. SAFE-STUDENT outperforms the compared methods. For example, when the class distribution mismatch is 60% on CIFAR-100, SAFE-STUDENT achieves **68.2%** averaged accuracy, about **6.7%** higher than CL, and **9.6%** higher than the supervised learning method. These results indicate the effectiveness and robustness of SAFE-STUDENT.

Table 3. AUC(%) for unseen-class identification on MNIST with 40% of unseen-class data in the test data.

Method	ratio=0	ratio=0.1	ratio=0.2	ratio=0.3	ratio=0.4	ratio=0.5	ratio=0.6	Avg
Probabilities [13]	84.3±0.9	84.3±0.9	84.3±0.9	84.3±0.9	84.3±0.9	84.3±0.9	84.3±0.9	84.3
DS^{3}L [9]	95.9±0.8	93.1±0.4	91.7±0.2	90.6±0.1	90.5±0.5	89.1±0.2	85.1±0.8	90.9
SAFE-STUDENT	**98.1±0.1**	**97.3±0.2**	**96.5±0.1**	**96.0±0.9**	**94.6±0.9**	**93.5±0.3**	**91.4±0.2**	**95.3↑**

4.3. Ablation Analysis

We validate the effectiveness of the components in SAFE-STUDENT by ablating them and measuring the performance on CIFAR-10 with 100 labeled instances per class, 200 labeled instances per class, and 400 labeled instances per class. Table 2 reports the results of ablation studies that contain SAFE-STUDENT without energy-discrepancy (ED), without augment (AUG), without confirmation bias elimination (CBE) loss, without unseen-class label distribution learning (UCDL), without iterative optimization (IO), and without exponential weighting (EW). We can see that all components have a significant effect as removing any of them causes a decline in performance.

4.4. Sensitivity of Hyperparameters

τ_1, τ_2 are two important hyperparameters that work as thresholds to select seen-class data and unseen-class data from unlabeled data in the SUCI module. Fig. 5 shows the results on CIFAR-10 with the 40% ratio of class distribution mismatch under the wider values of τ_1, τ_2, which verifies seen-class classification is not sensitive to τ_1, τ_2.

4.5. Unseen-Class Identification

To further measure the potential to identify unseen classes, we compare SAFE-STUDENT with probability estimation method [13] and DS^{3}L [9] on MNIST. AUC can measure the identification ability by treating the unseen-class data as a negative class and the others as a positive

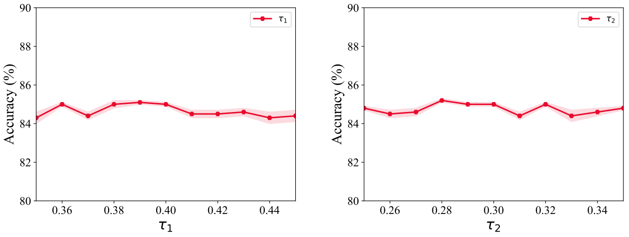


Figure 5. An analysis of τ_1, τ_2 under different values.

Table 4. AUC(%) for unseen-class identification on MNIST with 40% of unseen-class instances in the test data by different scores.

Method	Ratio	confidence	entropy	energy	**ED**
Probabilities [13]	–	84.3±0.9	85.1±0.7	86.3±0.7	**88.9±0.6**
DS^{3}L [9]	0	95.9±0.8	96.1±0.7	97.4±0.8	**97.5±0.7**
	0.1	93.1±0.4	93.4±0.5	90.5±0.5	**94.3±0.4**
	0.2	91.7±0.2	92.1±0.3	92.6±0.2	**93.1±0.1**
	0.3	90.6±0.1	91.4±0.2	88.3±0.1	**92.9±0.1**
	0.4	90.5±0.5	91.0±0.5	89.9±0.4	**92.7±0.3**
	0.5	89.1±0.2	90.0±0.2	92.8±0.2	**93.2±0.2**
	0.6	85.1±0.8	86.7±0.6	88.6±0.5	**89.7±0.5**
SAFE-STUDENT	0	98.1±0.1	98.2±0.1	97.8±0.1	**99.0±0.1**
	0.1	97.3±0.2	97.4±0.2	96.9±0.2	**97.9±0.1**
	0.2	96.5±0.1	96.6±0.2	96.1±0.1	**97.3±0.1**
	0.3	96.0±0.9	96.2±0.8	95.7±0.7	**97.0±0.6**
	0.4	94.6±0.9	94.8±0.6	94.7±0.6	**95.1±0.4**
	0.5	93.5±0.3	93.6±0.4	93.4±0.4	**94.3±0.3**
	0.6	91.4±0.2	91.7±0.3	92.6±0.2	**93.8±0.2**

one. Table 3 shows the results on MNIST with 40% of unseen-class data in the test data under different mismatch ratios in unlabeled data, respectively. By these results, we can discover that SAFE-STUDENT outperforms probability estimation method [13] and DS^{3}L [9] by a large margin.

To further verify the newly proposed ED for unseen-class identification, we compare ED with confidence, entropy, and energy. We apply the four different scoring functions for probability estimation, DS^{3}L, and SAFE-STUDENT. As shown in Table 4, ED comprehensively outperforms confidence, entropy, and energy [25], which can prove the universality and effectiveness of ED for various out-of-distribution detection methods. We can find that the performance of unseen-class identification shows an upward trend as the ratios of class mismatch in D_U decreases. Under the same ratio of class mismatch and the same scoring function, SAFE-STUDENT performs better than DS^{3}L. Fig. 6 exhibits the distributions of ED and energy on the MNIST under the 60% ratio of class mismatch in D_U by DS^{3}L, which proves that ED outperforms energy and owns the greater ability to identify unseen classes.

5. Conclusion

We presented a new analysis of safe deep SSL with unseen-class unlabeled data, a limited-explored but more realistic scenario. We also proposed a practical framework called SAFE-STUDENT guaranteed by the new ED scoring function, the unseen-class label distribution learning module, and the iterative optimization strategy. Empirical studies show that, unlike the compared deep SSL methods, SAFE-STUDENT still achieves stable performance gain when the class distribution mismatch proportion exceeds 60% and outperforms the existing safe SSL methods with a large margin. This work also provides a unified analysis work for this new problem. One can also extend our work into other safe SSL environments.

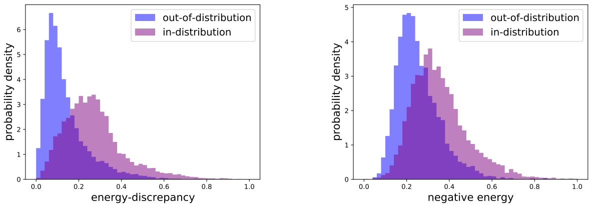


Figure 6. The distribution of the ED and negative energy under the 60% ratio of class mismatch in D_U by DS^{3}L.

Broader Impact

In this work, we study the problem of safe deep semi-supervised learning with unseen-class unlabeled data, a less explored task in the literature. Our project aims to improve the safeness and robustness of deep SSL models. We are aware that numerous uses of the technique can pose ethical issues and that best practice. Our approach should not be used in mission-critical applications or to make essential decisions without human oversight.

Limitations

While we successfully identify seen classes and unseen classes and perform safe semi-supervised learning, the thresholds to select seen-class data and unseen-class data from unlabeled data in the SUCI module require manual setting, but we have proved its insensitivity. Furthermore, since our method is based on deep neural networks, it is susceptible to adversarial attacks, lack of interpretability.

Acknowledgment

This work is supported by the National Natural Science Foundation of China (62176139, 61876098, 62106128), the Major Basic Research Project of Natural Science Foundation of Shandong Province (ZR2021ZD15), the Natural Science Foundation of Shandong Province (ZR202102240155), the Young Elite Scientists Sponsorship Program by CAST.

References

[1] 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*, 2020. 1, 2

[2] David Berthelot, Nicholas Carlini, Ian J. Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *NeurIPS*, 2019. 1, 2

[3] Paola Cascante-Bonilla, Fuwen Tan, Yanjun Qi, and Vicente Ordonez. Curriculum labeling: Revisiting pseudo-labeling for semi-supervised learning. In *AAAI*, 2021. 1, 2, 3, 6, 7

[4] Yanbei Chen, Xiatian Zhu, Wei Li, and Shaogang Gong. Semi-supervised learning under class distribution mismatch. In *AAAI*, 2020. 1, 2, 3, 6, 7

[5] Ekin D. Cubuk, Barret Zoph, Dandelion Mané, Vijay Vasudevan, and Quoc V. Le. Autoaugment: Learning augmentation strategies from data. In *CVPR*, 2019. 2

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 6

[7] Terrance Devries and Graham W. Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. 2

[8] Xin Geng. Label distribution learning. *IEEE Transactions on Knowledge and Data Engineering*, 2016. 5

[9] Lan-Zhe Guo, Zhenyu Zhang, Yuan Jiang, Yu-Feng Li, and Zhi-Hua Zhou. Safe deep semi-supervised learning for unseen-class unlabeled data. In *ICML*, 2020. 1, 2, 6, 7, 8

[10] Zhongyi Han, Xian-Jin Gui, Chaoran Cui, and Yilong Yin. Towards accurate and robust domain adaptation under noisy environments. In *IJCAI*, 2020. 2

[11] Zhongyi Han, Benzheng Wei, Yanfei Hong, Tianyang Li, Jinyu Cong, Xue Zhu, Haifeng Wei, and Wei Zhang. Accurate screening of COVID-19 using attention-based deep 3d multiple instance learning. *TMI*, 2020. 1

[12] Zhongyi Han, Benzheng Wei, Xiaoming Xi, Bo Chen, Yilong Yin, and Shuo Li. Unifying neural learning and symbolic reasoning for spinal medical report generation. *Medical Image Anal.*, 2021. 1

[13] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *ICLR*, 2017. 7, 8

[14] Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondrej Chum. Label propagation for deep semi-supervised learning. In *CVPR*, 2019. 1

[15] K. J. Joseph, Salman Khan, Fahad Shahbaz Khan, and Vineeth N. Balasubramanian. Towards open world object detection. In *CVPR*, 2021. 3

[16] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 6

[17] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, 2012. 1

[18] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. In *ICLR*, 2017. 1, 2, 6

[19] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *NeurIPS*, 2017. 5

[20] Yann LeCun. The mnist database of handwritten digits. 1998. 6

[21] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *ICML Workshop*, 2013. 2, 6, 7

[22] Yu-Feng Li and De-Ming Liang. Safe semi-supervised learning: a brief introduction. *FCS*, 2019. 2

[23] Yu-Feng Li and Zhi-Hua Zhou. Towards making unlabeled data never hurt. *TPAMI*, 2015. 2

[24] Tongliang Liu and Dacheng Tao. Classification with noisy labels by importance reweighting. *TPAMI*, 2015. 3

[25] Weitang Liu, Xiaoyun Wang, John D. Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *NeurIPS*, 2020. 3, 8

[26] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *TPAMI*, 2019. 6, 7

[27] Avital Oliver, Augustus Odena, Colin Raffel, Ekin Dogus Cubuk, and Ian J. Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. In *NeurIPS*, 2018. 1, 2

[28] Hieu Pham, Qizhe Xie, Zihang Dai, and Quoc V. Le. Meta pseudo labels. In *CVPR*, 2020. 2

[29] Zhongzheng Ren, Raymond A. Yeh, and Alexander G. Schwing. Not all unlabeled data are equal: Learning to weight data in semi-supervised learning. In *NeurIPS*, 2020. 1

[30] Mamshad Nayeem Rizve, Kevin Duarte, Yogesh Singh Rawat, and Mubarak Shah. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. In *ICLR*, 2021. 1, 2

[31] Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. In *NeurIPS*, 2016. 1, 2, 6, 7

[32] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *NeurIPS*, 2020. 1, 2, 6, 7

[33] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIPS*, 2017. 1, 2, 6

[34] Jing Wang and Xin Geng. Label distribution learning by exploiting label distribution manifold. *IEEE Transactions on Neural Networks and Learning Systems*, 2021. 5

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

[36] Qizhe Xie, Minh-Thang Luong, Eduard H. Hovy, and Quoc V. Le. Self-training with noisy student improves imagenet classification. In *CVPR*, 2020. 1, 2, 7

[37] Qing Yu, Daiki Ikami, Go Irie, and Kiyoharu Aizawa. Multi-task curriculum framework for open-set semi-supervised learning. In *ECCV*, 2020. 1, 2, 3, 6, 7

[38] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016. 6

[39] Dengyong Zhou, Olivier Bousquet, Thomas Navin Lal, Jason Weston, and Bernhard Schölkopf. Learning with local and global consistency. In *NeurIPS*, 2003. 2