# CaliMatch: Adaptive Calibration for Improving Safe Semi-supervised Learning

Jinsoo Bae
Korea University
Seoul, Republic of Korea

Seoung Bum Kim*
Korea University
Seoul, Republic of Korea

Hyungrok Do
NYU Grossman School of Medicine
New York, NY, USA

## Abstract

*Semi-supervised learning (SSL) uses unlabeled data to improve the performance of machine learning models when labeled data is scarce. However, its real-world applications often face the label distribution mismatch problem, in which the unlabeled dataset includes instances whose ground-truth labels are absent from the labeled training dataset. Recent studies, referred to as safe SSL, have addressed this issue by using both classification and out-of-distribution (OOD) detection. However, the existing methods may suffer from overconfidence in deep neural networks, leading to increased SSL errors because of high confidence in incorrect pseudo-labels or OOD detection. To address this, we propose a novel method, CaliMatch, which calibrates both the classifier and the OOD detector to foster safe SSL. CaliMatch presents adaptive label smoothing and temperature scaling, which eliminates the need to manually tune the smoothing degree for effective calibration. We give a theoretical justification for why improving the calibration of both the classifier and the OOD detector is crucial in safe SSL. Extensive evaluations on CIFAR-10, CIFAR-100, SVHN, TinyImageNet, and ImageNet demonstrate that CaliMatch outperforms the existing methods in safe SSL tasks.*

## 1. Introduction

Semi-supervised learning (SSL) is a useful approach to improving machine learning model performance, particularly when labeled data is limited, but plenty of unlabeled data are available [5, 10, 24, 30, 31]. However, most existing SSL methods are based on the assumption that the labeled and unlabeled data have the same set of class labels. In reality, there could be unlabeled instances whose labels do not belong to any classes in the labeled dataset, called unseen-class data. Such unseen-class data can adversely affect the decision boundary of the classifier trained using labeled data and unlabeled seen-class data through SSL methods, thereby decreasing the effectiveness of SSL approaches. To address this issue, several deep learning-based SSL methods, known as *safe semi-supervised learning*, have been proposed [1, 4, 9, 14, 24, 32]. Most safe SSL methods use out-of-distribution (OOD) detection techniques to identify whether the labels of unlabeled data belong to the set of labels in the labeled training dataset. That is, they categorize the unlabeled data into "seen" and "unseen" classes and use labeled data and unlabeled seen class data to improve the classifier or unlabeled unseen class data to improve the OOD detection performance.

Most safe SSL studies for image classification rely on deep convolutional neural networks (CNNs). However, recent studies have reported that deep CNNs show poor calibration performances despite their effectiveness measured in classification accuracies [7, 16, 17, 22, 28]. Specifically, deep CNNs are overconfident in their decisions and struggle to accurately assess the likelihood of error based on their confidence levels. In this study, we investigate the overconfidence issue of deep CNNs in the context of safe SSL. The overconfidence in safe SSL can cause the classification model to learn from incorrect pseudo-labels with unreasonably high confidence in their decisions, thereby amplifying the model's error. Moreover, because of the overconfidence issue, OOD detection in safe SSL may erroneously label a considerable portion of unseen class data in the unlabeled dataset as a seen class, consequently undermining the effectiveness of safe SSL.

We propose a new safe SSL method, CaliMatch, to mitigate overconfidence-related issues in SSL by improving the calibration performance. CaliMatch effectively rejects unseen-class data and accurately identifies pseudo-labeled samples from the unlabeled dataset using a well-calibrated classifier and an OOD detector, thereby improving the efficacy of SSL. To achieve this, CaliMatch uses adaptive label smoothing and logit scaling to calibrate both the multiclass classifier and the OOD detector. Unlike conventional smoothing-based calibration methods, which require manual adjustment of the smoothing degree for effective calibration, CaliMatch dynamically adjusts the degree of label smoothing based on the accuracy distribution of the vali-

[*] Seoung Bum Kim is the corresponding author: sbkim@korea.ac.kr.

dation dataset, ensuring appropriate smoothing degrees relative to current confidence levels. Moreover, incorporating learnable scaling parameters for the multiclass classifier and OOD detector helps the models learn adaptively smoothed labels and mitigates overconfidence in both classification and OOD detection. This approach can alleviate the adverse training effects of incorrectly pseudo-labeled examples in safe SSL.

To the best of our knowledge, this is the first study to highlight the importance of improving the calibration performance of both multiclass classifiers and OOD detectors in safe SSL. The key contributions are as follows:

- We investigate the importance of calibration in both classification and OOD detection in safe SSL, supported by extensive experiments and theoretical justification.
- We propose CaliMatch, a new safe SSL method that calibrates both the confidence of classifier and OOD scores to improve the quality of pseudo-label and effectively exclude instances from unseen classes within the unlabeled training dataset.
- CaliMatch outperforms existing safe SSL methods across five benchmark datasets, including a large-scale ImageNet dataset. Furthermore, our adaptive calibration approach is more effective than popular calibration methods in helping the safe SSL task.

## 2. Related Works

**Semi-supervised Learning.** Hybrid SSL methods integrate data augmentation and pseudo-labeling techniques to use large amounts of unlabeled data effectively [2, 3, 13, 26]. Notably, FixMatch has achieved state-of-the-art performance on several SSL benchmarks by using a confidence-based thresholding approach that considers pseudo-labels in training only when predictions meet a high confidence threshold. However, even with high confidence, pseudo-label quality can suffer because of overconfidence issues common in deep CNNs. To address this reliability concern, we propose calibrating the neural networks to improve the confidence-based thresholding process for accurately pseudo-labeled data, thus improving the quality of safe SSL.

**Safe Semi-supervised Learning.** Multi-task curriculum (MTC) trains an OOD detector alongside MixMatch using a joint optimization framework that alternates between updating neural network parameters and adjusting the OOD score [32]. SafeStudent uses a novel scoring function, known as energy discrepancy, to detect OOD instances [9]. For better OOD detection, SafeStudent calibrates the probability distribution of detected unseen-class instances to a uniform distribution. OpenMatch, built on the framework of FixMatch, is an advanced, safe SSL method that incorporates an OOD rejection mechanism. IOMatch is based on OpenMatch and proposes to use multiclass and one-versus-rest (OvR) classifiers with a projection head to create an open-set classifier [14]. Similarly, SCOMatch presents an open-set classifier and an OOD memory queue for selecting reliable OOD samples as new labeled data [29]. However, the overconfident neural network-based OOD scores may decrease the model's capability of correctly rejecting OOD instances with reliability during SSL. To address the negative issue in safe SSL, we propose calibrating an OOD detector and defining an OOD score by leveraging well-calibrated predictions of both the multiclass classifier and the OOD detector.

**Improving Calibration Performance of Deep Learning.** Calibration methods generally fall into two categories: (i) post-hoc techniques, such as temperature scaling [7, 15], which calibrate models after training, and (ii) real-time techniques that perform calibration during training. Our focus is on real-time calibration techniques because post-hoc calibration cannot correct for miscalibrated confidence that has already polluted the safe SSL process. Although smoothing-based calibration methods, such as mixup and label smoothing, can calibrate neural networks during training, achieving optimal calibration requires careful hyperparameter tuning to determine an appropriate degree of smoothing [18, 27]. To achieve efficient calibration for safe SSL, CaliMatch uses adaptive label smoothing based on the accuracy distribution over different confidence levels.

**Long-tailed Semi-supervised Learning.** Recent long-tailed SSL (LTSSL) methods are designed to address not only class imbalance in both labeled and unlabeled data, but also different imbalance ratios between them [25, 30]. Specifically, these methods aim to correct biased pseudo-labels for unlabeled data, which are skewed toward majority classes. A recent LTSSL method, ADELLO [25], further incorporates the concept of calibration to mitigate overconfidence in majority classes and enhance the quality of pseudo-labels for minority classes. In contrast, CaliMatch addresses the presence of unlabeled samples from unknown classes, where calibration is critical for preventing overconfident pseudo-labeling of OOD data.

## 3. Proposed Method

### 3.1. Safe Semi-supervised Learning

In our problem setting, we are given a labeled dataset $D_\ell = \{(x_i^\ell, y_i^\ell) \in \mathcal{X} \times \mathcal{Y} : i = 1, \cdots, n_\ell\}$, where $\mathcal{X}$ is the input feature space and $\mathcal{Y} = \{1, \cdots, K\}$ is the output variable space. Additionally, a set of unlabeled data $D_u = \{x_i^u \in \mathcal{X} : i = 1, \cdots, n_u\}$ is given. We assume the presence of label distribution mismatch, where the true labels of the unlabeled instances in $D_u$ may not belong to $\mathcal{Y}$. Our primary goal is to build an accurate multiclass classification model using pseudo-labeling-based SSL techniques with both $D_\ell$ and $D_u$. However, to address the mismatch

**Algorithm 1:** CaliMatch

```
Require: $D_\ell$ and $D_u$: labeled and unlabeled datasets; $\phi_\theta$, $f_\theta$, and $g_\theta$: encoder, multiclass classifier, and OOD
         detector; $T_M$ and $T_O$: learnable parameters; $\mathcal{T}_w$ and $\mathcal{T}_s$: weak and strong augmentations; $\tau_1$ and $\tau_2$:
         thresholds for consistency regularization and OOD rejection; $\lambda_{\text{O}}$, $\lambda_{\text{OCal}}$, and $\lambda_{\text{S}}$: coefficients; $E_{max}$ and
         $I_{max}$: epoch and iteration; $E_{\text{warm-up}}$: warm-up epoch; $\eta$: learning rate
for $Epoch = 1$ to $E_{max}$ do
    for $Iteration = 1$ to $I_{max}$ do
        Draw minibatches $B_\ell$ and $B_u$ from $D_\ell$ and $D_u$, respectively
        $\mathcal{L} = \mathcal{L}_{\text{CE}}(B_\ell) + \lambda_{\text{O}}\mathcal{L}_{\text{OOD}}(B_\ell) + \lambda_{\text{S}}\mathcal{L}_{\text{SC}}(B_u; \mathcal{T}_w)$    ▷ Equations (1), (2), (3)
        if $Epoch \geq E_{warm\text{-}up}$ then
            $B_u^t = \{x_i^u \in B_u : (s_i^u > \tau_1) \wedge (c_i^u > \tau_2)\}$    ▷ Select unlabeled data with Equation (7)
            $\mathcal{L} = \mathcal{L} + \mathcal{L}_{\text{MCal}}(B_\ell; \Gamma) + \lambda_{\text{OCal}}\mathcal{L}_{\text{OCal}}(B_\ell; \Delta) + \mathcal{L}_{\text{Fix}}(B_u^t; \mathcal{T}_w, \mathcal{T}_s)$    ▷ Equations (4), (5), (6)
        Update $[\theta, T_M, T_O]$ at every iteration with stochastic gradient descent: $[\theta, T_M, T_O] \leftarrow [\theta, T_M, T_O] - \eta\nabla\mathcal{L}$
    Update $\Gamma$ and $\Delta$ using validation dataset at every epoch    ▷ For adaptive label smoothing.
Output: Trained networks for CaliMatch.
```

of the label distribution, we also require an OOD detector to distinguish between known classes (those in $\mathcal{Y}$) and unknown classes (those not in $\mathcal{Y}$). Therefore, following the safe SSL setting of [24], we consider a neural network consisting of an encoder $\phi_\theta : \mathcal{X} \rightarrow \mathbb{R}^d$, which takes the input variable and generates a $d$-dimensional embedding vector that summarizes the features of each instance, a multiclass classifier $f_\theta : \mathbb{R}^d \rightarrow \mathbb{R}^K$, which takes an embedding vector as input and generates logit vectors for the classification task, and an OOD detector $g_\theta : \mathbb{R}^d \rightarrow \mathbb{R}^K$, implemented as a set of OvR binary classifiers, which generate logit vectors to distinguish unseen from seen classes.

Training a safe SSL model involves two key strategies: first, selecting a reliable subset of unlabeled data, which does not include instances of unseen classes from $D_u$ using the OOD detector, and second, applying consistency regularization techniques to the reliable subset. This reliable subset consists of unlabeled samples whose labels belong to the known classes, and their pseudo labels, derived from the model's predictions (i.e., $\text{argmax}_{k\in\mathcal{Y}} p_k(x_i^u)$), align with their true labels. To construct a reliable dataset, the existing safe SSL methods use OOD scores to discard unseen-class instances and use confidence scores in multiclass classification to correctly identify pseudo-labeled samples. However, because of the overconfidence issue of deep CNNs used in the existing safe SSL methods, the existing OOD and the confidence scores-based approach often fail to reject unlabeled instances successfully, thereby compromising the effectiveness of SSL techniques. Unexpectedly included unseen-class instances and incorrectly pseudo-labeled data can worsen a model's confirmation bias, resulting in poor classification performance. Therefore, addressing the overconfidence issue of deep CNN is crucial for the safe SSL task.

### 3.2. CaliMatch: Safe Semi-supervised Learning with Improved Calibration

To address the overconfident predictions of deep CNN and improve the efficacy of SSL, we propose an approach to improve the calibration performance of the multiclass classifier and the OOD detector by presenting adaptive label smoothing and temperature scaling techniques using labeled instances. Two scaling parameters $T_M$ and $T_O$ are initialized at 1.5 and optimized during training to allow the multiclass classifier and the OOD detector to learn adaptively smoothed labels efficiently. Our safe SSL method, CaliMatch, builds upon FixMatch, a highly effective consistency regularization SSL technique. CaliMatch incorporates an OOD detector to reject instances with unseen labels effectively and is trained to be well-calibrated, thereby enhancing the efficacy of consistency regularization.

Algorithm 1 outlines CaliMatch. The encoder, multiclass classifier, and OOD detector are trained using a minibatch stochastic gradient descent. Given two sampled minibatches $B_\ell$ and $B_u$ from $D_\ell$ and $D_u$, respectively, in the warm-up phase, we update the model parameters to minimize three losses: cross-entropy loss for classification ($\mathcal{L}_{\text{CE}}(B_\ell)$), binary cross-entropy loss for OOD ($\mathcal{L}_{\text{OOD}}(B_\ell)$), and soft consistency regularization loss ($\mathcal{L}_{\text{SC}}(B_u; \mathcal{T}_w)$) for the OOD detector. These three loss functions are given as follows:

$$-\sum_{i=1}^{|B_\ell|}\sum_{k=1}^{K} y_{ik}^\ell \log p_k(x_i^\ell), \tag{1}$$

$$-\sum_{i=1}^{|B_\ell|}\Big[\sum_{k=1}^{K}\Big(y_{ik}^\ell \log q_k(x_i^\ell)\Big) + \min_{l\neq k}\log(1 - q_l(x_i^\ell))\Big], \tag{2}$$

$$\sum_{i=1}^{|B_u|}\sum_{k=1}^{K}\Big(q_k\big(\mathcal{T}_w^{(1)}(x_i^u)\big) - q_k\big(\mathcal{T}_w^{(2)}(x_i^u)\big)\Big)^2, \tag{3}$$

where $p(x_i^\ell)$ is defined as $\text{softmax}((f_\theta \circ \phi_\theta)(x_i^\ell))$ and $p_k(x_i^\ell)$

denotes the $k$-th element of $p(x_i^\ell)$, which represents the predicted probability of the instance $x_i^\ell$ being in class $k$. Note that $y_{ik}^\ell$ is equivalent to $\mathbb{I}(y_i^\ell = k)$. $q(x_i^\ell)$ refers to $\text{sigmoid}((g_\theta \circ \phi_\theta)(x_i^\ell))$, and $q_k(x_i^\ell)$ denotes the $k$-th element of $q(x_i^\ell)$, representing the predicted probability that $x_i^\ell$ is in class $k$. The distinction between the multiclass classifier and the OOD detector lies in how they interpret probabilities. Specifically, $p_k(x_i^\ell)$ is the probability of being in class $k$, given that $x_i^\ell$ belongs to the in-distribution data. In contrast, $q_k(x_i^\ell)$ indicates whether the instance belongs to class $k$ or not without conforming to the distribution of any known classes. $\mathcal{T}_w^{(1)}$ and $\mathcal{T}_w^{(2)}$ are transformations randomly selected from weak augmentation functions. The soft consistency regularization loss encourages the OOD detector to generate similar outputs for differently augmented input images, thereby improving OOD detection.

After the warm-up period, we add the following two loss functions $\mathcal{L}_{\text{MCal}}(B_\ell;\Gamma)$ and $\mathcal{L}_{\text{OCal}}(B_\ell;\Delta)$ to improve calibration performance of the multiclass classifier and the OOD detector on the labeled data. These two loss functions are formulated as follows:

$$-\sum_{i=1}^{|B_\ell|}\sum_{k=1}^{K}\Big(\gamma_i^\ell y_{ik}^\ell + \frac{1-\gamma_i^\ell}{K-1}(1-y_{ik}^\ell)\Big)\log p_k^s(x_i^\ell), \quad (4)$$

$$\begin{aligned}-\sum_{i=1}^{|B_\ell|}\Big[&\sum_{k=1}^{K}\Big(\big(\delta_i^\ell y_{ik}^\ell + (1-\delta_i^\ell)(1-y_{ik}^\ell)\big)\log q_k^s(x_i^\ell)\Big)\\ &+\min_{k\in\mathcal{Y}}\Big(\big((1-\delta_i^\ell)y_{ik}^\ell + \delta_i^\ell(1-y_{ik}^\ell)\big)\log(1-q_k^s(x_i^\ell))\Big)\Big], \quad (5)\end{aligned}$$

where $\Gamma$ and $\Delta$ are sets of reference confidence values for the multiclass classifier and the OOD detector, respectively. Specifically, $\gamma_i^\ell$ and $\delta_i^\ell$ are the reference confidence values assigned to the instance $x_i^\ell$ to determine the degrees of label smoothing based on the current model's confidence level. Unlike classic label smoothing, which applies a fixed smoothing degree to every sample, CaliMatch dynamically adjusts the smoothing levels for each instance to $1-\gamma_i^\ell$ for the classifier and $1-\delta_i^\ell$ for the OOD detector. Here, $p_k^s(x_i^\ell)$ and $q_k^s(x_i^\ell)$ denote the $k$-th elements of scaled-logit-based probabilities, $\text{softmax}((f_\theta \circ \phi_\theta)(x_i^\ell)/T_M)$ and $\text{softmax}((g_\theta \circ \phi_\theta)(x_i^\ell)/T_O)$, respectively. To assign reference value $\gamma_i^\ell$ for the multiclass classifier, we first divide the segment $[0,1]$ into $M$ equally spaced bins and define a set of validation instances whose confidences belong to each segment: $B_m^p = \{x_i \in D_{\text{Val}} : \max_k p_k(x_i) \in (\frac{m-1}{M}, \frac{m}{M}]\}$ for $m = 1, \cdots, M$. Subsequently, we calculate the accuracy of the multiclass classifier for each bin $B_m^p$ as $\gamma_m = |B_m^p|^{-1}\sum_{x_i\in B_m^p}\mathbb{I}(\text{argmax}_k p_k(x_i) = y_i)$. Lastly, when we suppose the confidence value $\text{max}_k p_k(x_i^\ell)$ for a labeled example $x_i^\ell$ falls into $B_{m_1}^p$, CaliMatch considers the accuracy value $\gamma_{m_1}$ as the well-calibrated reference confidence score $\gamma_i^\ell$ for the labeled data $x_i^\ell$. The same method calculates $\delta_i^\ell$ for the OOD detector using $B_m^q$. Through minimizing our calibration loss functions, the multiclass classifier and the OOD detector with two scaling parameters $T_M$ and $T_O$ learn the smoothed labels, thereby aligning their confidences $p_{y_i^\ell}^s(x_i^\ell)$ and $q_{y_i^\ell}^s(x_i^\ell)$ with the current model's accuracy $\gamma_i^\ell$ and $\delta_i^\ell$. Note that our smoothed labels take the reference values $\gamma_i^\ell$ and $\delta_i^\ell$ for the true class, and $(1-\gamma_i^\ell)/(K-1)$ and $1-\delta_i^\ell$ for the remaining classes.

Next, we add FixMatch loss $\mathcal{L}_{\text{Fix}}(B_u^t;\mathcal{T}_w,\mathcal{T}_s)$ based on the set of reliable unlabeled instances $B_u^t$ for consistency regularization. The FixMatch loss is:

$$-\sum_{i=1}^{|B_u^t|}\sum_{k=1}^{K}\mathbb{I}\big(\text{argmax}_l p_l(\mathcal{T}_w(x_i^u)) = k\big)\log p_k(\mathcal{T}_s(x_i^u)), \quad (6)$$

where $\mathcal{T}_w$ and $\mathcal{T}_s$ are weak and strong augmentations applied to input variables, and they can be used to improve the classifier's consistency, thereby increasing the performance of the classifier, as used in FixMatch [26]. To identify the reliable set of unlabeled instances $B_u^t$ with classes in $\mathcal{Y}$ while ensuring the correctness of their pseudo-labels, we propose two selection criteria based on predictions of the model calibrated with the labeled dataset. The predictions of the model on the data $x_i^u \in B_u^t$ satisfy the following condition with two thresholds, $\tau_1$ and $\tau_2$:

$$s_i^u = \sum_{k\in\mathcal{Y}} p_k^s(x_i^u)q_k^s(x_i^u) > \tau_1 \quad \wedge \quad c_i^u = \max_{k\in\mathcal{Y}} p_k^s(x_i^u) > \tau_2. \quad (7)$$

Our proposed seen-class score $s_i^u$ consists of predicted probabilities $p_k^s(x_i^u)$ and $q_k^s(x_i^u)$, estimating the likelihood that the unlabeled data point $x_i^u$ belongs to one of the seen classes. This is based on the fact that the two well-calibrated probabilities, namely $p_k^s(x_i^u)$ and $q_k^s(x_i^u)$, account for the likelihood of $x_i^u$ being $k$-th seen class better than $p_k(x_i^u)$ and $q_k(x_i^u)$, which are likely to be overconfident. On the other hand, $u_i^u$, calculated as $1-s_i^u$, serves as a CaliMatch OOD score implying the likelihood that $x_i^u$ belongs to unseen classes not included in $\mathcal{Y}$. CaliMatch considers an unlabeled sample as reliable seen-class data if $s_i^u$ exceeds $\tau_1$. Furthermore, we propose using better-calibrated confidence $c_i^u$ to select unlabeled samples whose prediction confidence is greater than $\tau_2$ among those classified as unlabeled seen-class data, instead of using the overconfident score $\max_{k\in\mathcal{Y}} p_k(x_i^u)$ used in other safe SSL methods. Unless otherwise noted, $\tau_1$ and $\tau_2$ are fixed at 0.5 and 0.95, which are the same values from previous studies [24, 26].

In Supplementary Section S-3, we present a theoretical justification demonstrating how better-calibrated models improve safe SSL performance. Specifically, we demonstrate that *mitigating overconfidence in both classification and OOD detection can bring the gradient of*

$\mathcal{L}_{Fix}(B_u^t; \mathcal{T}_w, \mathcal{T}_s)$*, closer to that of the ideal scenario (no incorrect pseudo-labels and no OOD samples) in safe SSL settings.*

## 4. Experiments

**Setups.** We evaluated CaliMatch on a comprehensive list of benchmark datasets, including SVHN [20], CIFAR-10, CIFAR-100 [11], TinyImageNet [12], and ImageNet [6], in the presence of label distribution mismatch between labeled and unlabeled datasets. We compared CaliMatch with conventional SSL, safe SSL methods, and a recent calibration-based LTSSL method. These include MTC [32], FixMatch [26], OpenMatch [24], SafeStudent [9], IOMatch [14], SCOMatch [29], and ADELLO [25]. For fair evaluation, we followed the SSL evaluation protocol of [23], which outlines the SSL training procedure, including the number of training iterations, CNN architecture, batch size, optimizer, learning rate, and scheduler. In experiments on SVHN, CIFAR-10, CIFAR-100, and TinyImageNet, we used a Wide ResNet 28-2 [33] as the CNN backbone, while ResNet 50 [8] was used for experiments on ImageNet. All other implementation details, such as the settings of label distribution mismatch and hyperparameters in each dataset, are summarized in Supplementary Section S-1. Note that we also compared computational complexity of a supervised learning baseline and all SSL methods, and the results are presented in Supplementary Section S-2. To ensure robustness, we conducted experiments on SVHN, CIFAR-10, CIFAR-100, and TinyImageNet using the same evaluation protocols on five different random seeds, reporting average metrics and standard deviation. In the case of the ImageNet dataset, we limited the number of repeated experiments to three runs to alleviate extensive training costs. Lastly, we evaluated the calibration performance by calculating expected calibration error (ECE) with 15 bins [19].

### 4.1. Multiclass Classification

Table 1 presents the classification performance (top-1 accuracy) across four image benchmark datasets, with unseen-label data proportion ($\kappa$) set to 30% or 60%. Across all eight combinations of datasets and proportions, five SSL methods improved performance compared to the supervised learning approach using only labeled data. Notably, the proposed CaliMatch outperformed most SSL competitors in both 30% and 60% proportions of unseen classes, demonstrating its efficacy and robustness in the presence of unlabeled OOD instances. Furthermore, a Friedman test conducted across the eight SSL scenarios yielded a statistically significant result ($p$-value $< 0.001$), with CaliMatch ranked first based on the Friedman scores. We also evaluated the calibration performance of the methods, as presented in Supplementary Section S-2 (Table S-3). CaliMatch provides not only improved accuracy but also decent calibration performance.

The proposed CaliMatch outperformed other methods by providing more accurate pseudo-labels for high-confidence instances (confidence score greater than 0.95), which are selected for consistency regularization. Additionally, it includes a lower proportion of unseen-label instances in the set of high-confidence and low-OOD instances (OOD score less than 0.5). To demonstrate this, we investigated the learning curves of the SSL methods. Figure 1 shows the learning curves for CIFAR-100 with $\kappa$ set to 60%. Learning curves for other datasets are described in Supplementary Section S-2 (Figure S-1). First, as shown in the top-left figure, CaliMatch provides better pseudo-label accuracy for seen-class instances with confidence scores above the threshold. That is, the quality of the unlabeled instances selected by CaliMatch for consistency regularization is better than that of others, while the other SSL methods provide less accurate pseudo-labels. This advantage is because of CaliMatch's better calibration, whereas other methods suffer from overconfident predictions, as shown in Table S-3. This significantly improves the effectiveness of consistency regularization and leads CaliMatch to better classification performance.

The top-right figure depicts the ratio of high-confidence unlabeled samples with seen classes among all unlabeled samples with seen classes. While SCOMatch, ADELLO, and IOMatch showed high confidence in most of the training unlabeled data, they could not ensure the quality of pseudo-labels for high-confidence unlabeled data compared to CaliMatch. The bottom-left figure shows the ratio of unseen-label data among all high-confidence unlabeled samples. CaliMatch has the lowest percentage compared to all other comparison methods, indicating that it also improves OOD detection calibration. This leads to better detection of unseen classes in the high-confidence set, which is used for consistency regularization. Furthermore, CaliMatch's lower percentage in the bottom-right figure, indicating the proportion of unseen-label data among samples with low OOD scores and high confidence, highlights its superior capability to filter out unseen-label data in $D_u$.

### 4.2. Out-of-distribution Detection

We evaluated the OOD detection performance of safe SSL methods using the F1 score, as shown in Supplementary Section S-2 (Table S-4). CaliMatch and OpenMatch outperformed other SSL methods in unseen-class detection across four datasets. In this section, we contrast the differences between CaliMatch and OpenMatch. To compare their unseen-class detection details, we used reliability diagrams [21] to visualize the calibration performance, as shown in Figure 2, with ECE metrics for each diagram. Our evaluation of unseen-label data detection calibration is based on considering OOD detection as the binary classification

Table 1. Evaluation of multiclass classification using the averaged accuracy and standard deviation (in parentheses) on four image benchmark datasets under two different existence rates ($\kappa$%) of unseen-label data. The best results are in **bold**, and the second-best results are underlined.

| Dataset | $\kappa$ | Method | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Supervised | MTC | FixMatch | OpenMatch | SafeStudent | IOMatch | SCOMatch | ADELLO | **CaliMatch** |
| SVHN | 30% | 85.82<br>(1.24) | 89.34<br>(2.09) | 95.33<br>(0.38) | 96.62<br>(0.23) | 90.94<br>(0.62) | 95.72<br>(0.43) | 95.80<br>(0.48) | 95.70<br>(0.51) | **96.81**<br>**(0.23)** |
| | 60% | | 86.77<br>(2.76) | 94.20<br>(0.32) | 95.67<br>(0.14) | 90.34<br>(0.49) | 94.52<br>(0.32) | 94.59<br>(0.38) | 94.30<br>(0.43) | **96.56**<br>**(0.11)** |
| CIFAR-10 | 30% | 76.96<br>(0.39) | 82.22<br>(0.88) | 88.12<br>(0.39) | 88.04<br>(0.24) | 81.88<br>(0.65) | 89.12<br>(0.31) | **90.30**<br>**(0.35)** | 88.81<br>(0.31) | 90.25<br>(0.27) |
| | 60% | | 79.74<br>(0.42) | 85.52<br>(0.54) | 86.19<br>(0.74) | 80.64<br>(0.33) | 86.23<br>(0.92) | 86.80<br>(1.00) | 86.73<br>(0.78) | **87.62**<br>**(0.36)** |
| CIFAR-100 | 30% | 60.24<br>(0.53) | 66.88<br>(0.68) | 68.74<br>(0.48) | 69.65<br>(0.93) | 64.38<br>(1.00) | 69.35<br>(0.78) | 66.33<br>(0.64) | 69.00<br>(0.96) | **72.32**<br>**(0.34)** |
| | 60% | | 63.02<br>(0.64) | 64.66<br>(0.53) | 65.65<br>(1.02) | 62.72<br>(1.19) | 65.59<br>(0.81) | 63.78<br>(0.91) | 65.18<br>(0.28) | **68.56**<br>**(1.08)** |
| TinyImageNet | 30% | 39.68<br>(0.68) | 44.50<br>(0.97) | 46.06<br>(0.67) | 46.90<br>(0.63) | 42.44<br>(0.38) | 46.33<br>(0.89) | 44.11<br>(0.89) | 44.72<br>(0.81) | **47.46**<br>**(0.69)** |
| | 60% | | 41.04<br>(0.89) | 42.90<br>(0.54) | 44.02<br>(0.63) | 40.76<br>(0.83) | 43.15<br>(0.95) | 42.05<br>(0.83) | 42.80<br>(0.61) | **44.66**<br>**(0.37)** |

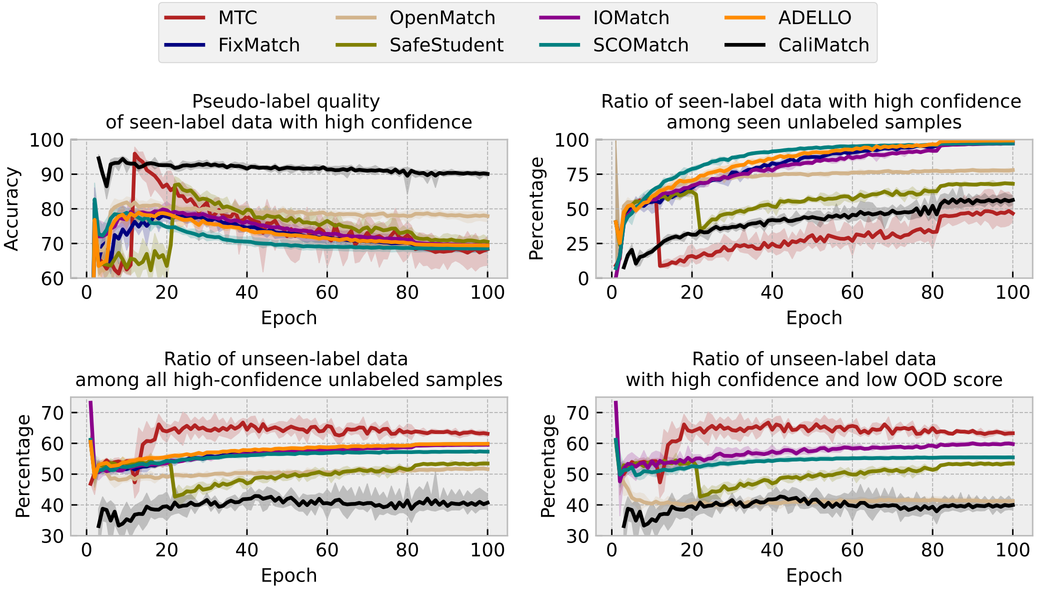

Figure 1. Learning curves averaged over five runs on CIFAR-100 for CaliMatch and other SSL methods. The shaded region indicates standard deviations calculated from five runs.

where the unseen class is recognized as a positive class.

The blue bar represents the average confidence scores provided by the models in each bin, while the red bar represents the accuracy of the OOD detection in each bin. If the blue bar is above the red one, the model overestimates the confidence in OOD detection. Therefore, OpenMatch's OOD detector generally overestimates the probabilities in OOD detection. This means that OpenMatch is more likely to accept unseen classes as seen classes or seen classes as unseen classes with unreasonably high confidence in OOD detection. On the other hand, CaliMatch provides better calibration performance across all four datasets, demonstrating that our approach to improving calibration performance worked as expected. Our proposed method's consistently lower ECE scores compared to those of OpenMatch across all datasets also suggest that the seen-class and OOD scores generated by our well-calibrated model are more effective in reliably selecting unlabeled seen-label samples during safe SSL.

To further investigate the importance of a well-calibrated OOD detector in safe SSL, we investigated how the performance of CaliMatch and OpenMatch varies as the decision threshold changes, as shown in Figure 3. We increased the threshold $\tau_1$ from 0.5 to 0.8. The first graph in Figure 3 shows how the accuracy of the multiclass classification changes as a function of $\tau_1$. The accuracy of CaliMatch is

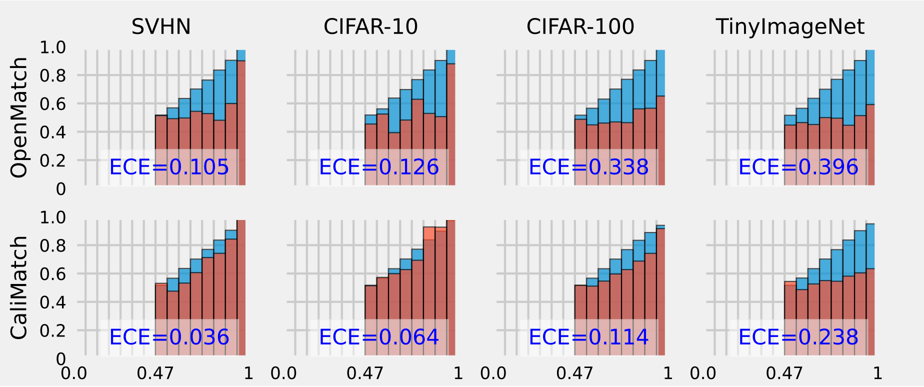

Figure 2. Reliability diagrams of OpenMatch and CaliMatch for unseen-label data detection on four datasets. The blue and red bars represent the average confidence score and sample accuracy of the confidence bin, respectively.

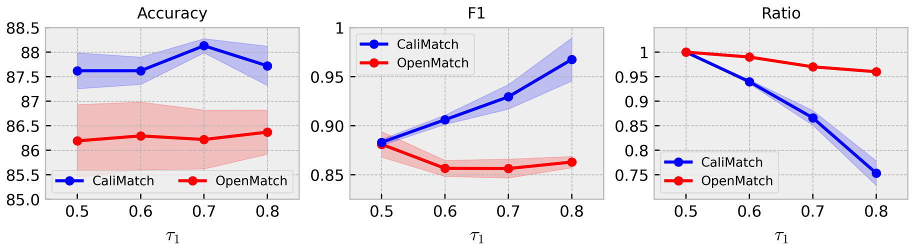

Figure 3. Performance variations in CaliMatch and OpenMatch with threshold $\tau_1$ on CIFAR-10. We marked the averaged metrics and standard deviation as straight lines and shaded areas over five runs.

higher than that of OpenMatch in all $\tau_1$ values, demonstrating that CaliMatch consistently outperforms OpenMatch with different OOD detection thresholds because of its superior calibration performance. The second graph provides the variations in the F1 score for unseen-class detection, selecting only observations with high-confidence values greater than $\tau_1$. Here, confidence refers to the higher value between the seen-class score and the OOD score. The last graph represents changes in the ratio of selected samples with high confidence among the seen-label data. As more uncertain predictions are removed, the F1 score increases, indicating that CaliMatch effectively avoids incorrect decisions and has a positive impact on safe SSL, thereby increasing accuracy as well, except when the value of $\tau_1$ changes from 0.7 to 0.8. This exception comes from the trade-off between the quality and quantity of selected seen-label samples. However, OpenMatch did not show any significant performance improvement in Figure 3, even though it filtered out the uncertain decision results. Note that a more detailed analysis on $\tau_1$ and $\tau_2$ is also provided in Supplementary Section S-2.

## 4.3. Further Analyses

Under $\kappa$ set to 60% on four benchmark datasets, we conducted extensive ablation studies to demonstrate the effectiveness of calibration loss terms for the multiclass classifier and the OOD detector. We also compared OpenMatch with ablated CaliMatch approaches because OpenMatch is identical to CaliMatch when not calibrating all classifiers except for calculating the seen-class and OOD scores.

Table 2 presents that for multiclass classification, multiclass calibration had a more significant impact than OOD calibration across all datasets regarding both accuracy and ECE. In other words, our multiclass calibration is especially useful in mitigating overconfidence-based negative issues on SSL, such as low-quality pseudo-labels. Note that removing OOD calibration also resulted in slightly reduced classification and calibration performance for SVHN, CIFAR-10, and TinyImageNet while simultaneously causing a significant change in accuracy for CIFAR-100. It represents that the shared encoder $\phi_\theta$ of two models, $f_\theta \circ \phi_\theta$ and $g_\theta \circ \phi_\theta$, can also benefit from learning our adaptive smoothed labels of OOD detector regarding both accuracy and calibration improvements. Note that a Friedman test conducted across the four datasets also yielded statistically significant results ($p$-values $< 0.001$), with the full CaliMatch configuration ranking first in both classification and calibration performance.

Table 3 shows that, for unseen-label detection, particularly in terms of calibration, OOD calibration had a greater

Table 2. Multiclass classification results of the ablation study on four datasets. The best results are in **bold**, and the second-best results are underlined. (ACC: Accuracy, w/o: without)

| Method | SVHN | | CIFAR-10 | | CIFAR-100 | | TinyImageNet | |
|---|---|---|---|---|---|---|---|---|
| | ACC | ECE | ACC | ECE | ACC | ECE | ACC | ECE |
| OpenMatch | 95.67 | 0.026 | 86.19 | 0.115 | 65.65 | 0.256 | 44.02 | 0.413 |
| (=w/o both calibration) | (0.14) | (0.002) | (0.74) | (0.007) | (1.02) | (0.006) | (0.63) | (0.008) |
| without | 96.08 | 0.022 | 86.47 | 0.108 | 66.16 | 0.223 | 43.68 | 0.341 |
| multiclass calibration | (0.02) | (0.002) | (0.87) | (0.008) | (0.50) | (0.006) | (0.63) | (0.020) |
| without | 96.23 | 0.009 | 87.57 | 0.038 | 66.96 | 0.033 | 44.65 | 0.250 |
| OOD calibration | (0.09) | (0.004) | (0.22) | (0.010) | (0.30) | (0.008) | (0.15) | (0.003) |
| **CaliMatch** | **96.56** | **0.006** | **87.62** | **0.029** | **68.56** | **0.025** | **44.66** | **0.233** |
| | **(0.11)** | **(0.002)** | **(0.36)** | **(0.003)** | **(1.08)** | **(0.007)** | **(0.37)** | **(0.006)** |

Table 3. Unseen-label data detection results of the ablation study on four datasets. The best results are in **bold**, and the second-best results are underlined. (w/o: without)

| Method | SVHN | | CIFAR-10 | | CIFAR-100 | | TinyImageNet | |
|---|---|---|---|---|---|---|---|---|
| | F1 | ECE | F1 | ECE | F1 | ECE | F1 | ECE |
| OpenMatch | 0.858 | 0.105 | 0.881 | 0.126 | 0.696 | 0.338 | 0.688 | 0.396 |
| (=w/o both calibration) | (0.009) | (0.004) | (0.013) | (0.006) | (0.002) | (0.010) | (0.002) | (0.006) |
| without | **0.890** | 0.037 | **0.909** | 0.080 | 0.693 | 0.149 | 0.683 | **0.189** |
| multiclass calibration | **(0.004)** | (0.009) | **(0.004)** | (0.010) | (0.005) | (0.006) | (0.004) | **(0.008)** |
| without | 0.865 | 0.085 | 0.858 | 0.086 | **0.709** | 0.320 | 0.686 | 0.366 |
| OOD calibration | (0.005) | (0.005) | (0.001) | (0.003) | **(0.001)** | (0.006) | (0.003) | (0.010) |
| **CaliMatch** | 0.889 | **0.036** | 0.883 | **0.064** | 0.687 | **0.114** | **0.691** | 0.238 |
| | (0.028) | **(0.018)** | (0.003) | **(0.005)** | (0.006) | **(0.008)** | **(0.001)** | (0.005) |

impact than multiclass calibration across all datasets. It is noteworthy that not calibrating the multiclass classifier also resulted in significant increases in ECE across three datasets: CIFAR-10, CIFAR-100, and TinyImageNet. These results demonstrate the effectiveness of our calibration techniques in improving the calibration of unseen-label detection. Such calibration is a valuable asset in reliably selecting seen-label data, thereby positively impacting safe SSL, as illustrated in Figure 3. In the case of TinyImageNet, only calibrating OOD detector resulted in the lowest calibration error for unseen-label detection compared to the original CaliMatch. This suggests that when the accuracies of both multiclass classifier and OOD detector are relatively poor, combining their predictions in CaliMatch's $s_i^u$ and $u_i^u$ may lead to more calibration errors in unseen-label detection compared to solely calibrating OOD detector, although CaliMatch already outperformed OpenMatch in terms of calibration. Additionally, there were minor fluctuations in the F1 score with $\tau_1$ set to 0.5 across all datasets, indicating that finding a proper threshold $\tau_1$ during calibration would be an additional task in safe SSL, as depicted in Figure 3.

To highlight CaliMatch's effectiveness on a more complex dataset, we compared it with OpenMatch and the baseline (supervised learning) on ImageNet with $\kappa$ set to 60%. CaliMatch achieved a top-1 accuracy of 63.04±0.43%, outperforming OpenMatch by 1.24% and the baseline by 6.34%. In terms of calibration in both classification and OOD detection, CaliMatch demonstrated ECE scores of 0.028 and 0.070, outperforming OpenMatch's 0.041 and 0.121, respectively. These results highlight the critical role of calibration in improving SSL performance and CaliMatch's effectiveness for large-scale safe SSL scenarios.

To further examine the robustness and efficiency of CaliMatch, we performed two additional studies on CIFAR-10 with $\kappa$ set to 60%. Detailed discussions are in Supplementary Section S-2 (Tables S-5 and S-6). We conducted a sensitivity analysis of $\lambda_{\text{O}}$ and $\lambda_{\text{OCal}}$, in which we did not observe a significant failure of our method in the range of hyperparameters considered. Moreover, we compared our adaptive calibration with other calibration methods in improving safe SSL. We found that while all of the calibration methods improved calibration when combined with OpenMatch, our method showed the highest improvement in calibration, ultimately leading to the best safe SSL performance.

## 5. Conclusion

This study proposed CaliMatch to address overconfidence-based negative issues in safe SSL scenarios. CaliMatch improved calibration of classification and OOD detection through adaptive label smoothing and scaling logit for multiclass and OvR classifiers. Our extensive experimental results with theoretical justification demonstrated that well-calibrated classifiers could enhance the safety and robustness of thresholding-based SSL methods by reducing negative training effects from incorrectly pseudo-labeled data whose confidence is high.

## Acknowledgements

The authors thank the ICCV reviewers for their careful evaluation and constructive suggestions. This research, conducted by Jinsoo Bae and Seoung Bum Kim, was supported by Brain Korea 21 FOUR, the Ministry of Science and ICT (MSIT) in Korea under the ITRC support program supervised by the Institute for Information Communication Technology Planning and Evaluation (IITP-2020-0-01749), and the National Research Foundation of Korea grant funded by the Korea government (RS-2022-00144190).

## References

[1] Jinsoo Bae, Minjung Lee, and Seoung Bum Kim. Safe semi-supervised learning using a bayesian neural network. *Information Sciences*, 612:453–464, 2022. 1

[2] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[3] David Berthelot, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. In *International Conference on Learning Representations*, 2020. 2

[4] Yanbei Chen, Xiatian Zhu, Wei Li, and Shaogang Gong. Semi-supervised learning under class distribution mismatch. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3569–3576, 2020. 1

[5] Yuhao Chen, Xin Tan, Borui Zhao, Zhaowei Chen, Renjie Song, Jiajun Liang, and Xuequan Lu. Boosting semi-supervised learning by exploiting all unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7548–7557, 2023. 1

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5

[7] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017. 1, 2

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[9] Rundong He, Zhongyi Han, Xiankai Lu, and Yilong Yin. Safe-student for safe deep semi-supervised learning with unseen-class unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14585–14594, 2022. 1, 2, 5

[10] Durk P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. *Advances in Neural Information Processing Systems*, 27, 2014. 1

[11] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 5

[12] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015. 5

[13] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on Challenges in Representation Learning, ICML*, page 896. Atlanta, 2013. 2

[14] Zekun Li, Lei Qi, Yinghuan Shi, and Yang Gao. Iomatch: Simplifying open-set semi-supervised learning with joint inliers and outliers utilization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15870–15879, 2023. 1, 2, 5

[15] Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*, 2018. 2

[16] Bingyuan Liu, Ismail Ben Ayed, Adrian Galdran, and Jose Dolz. The devil is in the margin: Margin-based label smoothing for network calibration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 80–88, 2022. 1

[17] Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. Revisiting the calibration of modern neural networks. *Advances in Neural Information Processing Systems*, 34:15682–15694, 2021. 1

[18] Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? *Advances in Neural Information Processing Systems*, 32, 2019. 2

[19] Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well-calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, 2015. 5

[20] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, Andrew Y Ng, et al. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, page 7. Granada, Spain, 2011. 5

[21] Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *International Conference on Learning Representations*, pages 625–632, 2005. 5

[22] Jongyoun Noh, Hyekang Park, Junghyup Lee, and Bumsub Ham. Rankmixup: Ranking-based mixup training for network calibration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1358–1368, 2023. 1

[23] Avital Oliver, Augustus Odena, Colin A Raffel, Ekin Dogus Cubuk, and Ian Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. *Advances in Neural Information Processing Systems*, 31, 2018. 5

[24] Kuniaki Saito, Donghyun Kim, and Kate Saenko. Openmatch: Open-set semi-supervised learning with open-set consistency regularization. *Advances in Neural Information Processing Systems*, 34:25956–25967, 2021. 1, 3, 4, 5

[25] Emanuel Sanchez Aimar, Nathaniel Helgesen, Yonghao Xu, Marco Kuhlmann, and Michael Felsberg. Flexible distribution alignment: Towards long-tailed semi-supervised learn-

ing with proper calibration. In *European Conference on Computer Vision*, pages 307–327. Springer, 2024. 2, 5

[26] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in Neural Information Processing Systems*, 33:596–608, 2020. 2, 4, 5

[27] Sunil Thulasidasan, Gopinath Chennupati, Jeff A Bilmes, Tanmoy Bhattacharya, and Sarah Michalak. On mixup training: Improved calibration and predictive uncertainty for deep neural networks. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[28] Christian Tomani, Sebastian Gruber, Muhammed Ebrar Erdem, Daniel Cremers, and Florian Buettner. Post-hoc uncertainty calibration for domain drift scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10124–10132, 2021. 1

[29] Zerun Wang, Liuyu Xiang, Lang Huang, Jiafeng Mao, Ling Xiao, and Toshihiko Yamasaki. Scomatch: Alleviating overtrusting in open-set semi-supervised learning. In *European Conference on Computer Vision*, pages 217–233. Springer, 2024. 2, 5

[30] Tong Wei and Kai Gan. Towards realistic long-tailed semi-supervised learning: Consistency is all you need. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3469–3478, 2023. 1, 2

[31] Jinsung Yoon, Yao Zhang, James Jordon, and Mihaela Van der Schaar. Vime: Extending the success of self-and semi-supervised learning to tabular domain. *Advances in Neural Information Processing Systems*, 33:11033–11043, 2020. 1

[32] Qing Yu, Daiki Ikami, Go Irie, and Kiyoharu Aizawa. Multi-task curriculum framework for open-set semi-supervised learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, pages 438–454. Springer, 2020. 1, 2, 5

[33] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *Proceedings of the British Machine Vision Conference*, pages 87.1–87.12, 2016. 5