Improving Open-Set Semi-Supervised Learning with Self-Supervision

Erik Wallin[1,2], Lennart Svensson[2], Fredrik Kahl[2], Lars Hammarstrand[2]
[1]Saab AB, [2]Chalmers University of Technology
{walline,lennart.svensson,fredrik.kahl,lars.hammarstrand}@chalmers.se

Abstract

Open-set semi-supervised learning (OSSL) embodies a practical scenario within semi-supervised learning, wherein the unlabeled training set encompasses classes absent from the labeled set. Many existing OSSL methods assume that these out-of-distribution data are harmful and put effort into excluding data belonging to unknown classes from the training objective. In contrast, we propose an OSSL framework that facilitates learning from all unlabeled data through self-supervision. Additionally, we utilize an energy-based score to accurately recognize data belonging to the known classes, making our method well-suited for handling uncurated data in deployment. We show through extensive experimental evaluations that our method yields state-of-the-art results on many of the evaluated benchmark problems in terms of closed-set accuracy and open-set recognition when compared with existing methods for OSSL. Our code is available at https://github.com/walline/ssl-tf2-sefoss.

1. Introduction

Using a combination of labeled and unlabeled data for training a model, so-called *semi-supervised learning* (SSL), is a well-studied field of machine learning [23,36,37,39,48], motivated by exploiting the extensive amounts of cheap and readily available unlabeled data. However, semi-supervised learning is typically studied in a *closed-set context*, where labeled and unlabeled data are assumed to follow the same distribution. In practice though, one can expect that the labeled set is of a much more curated character, *e.g.*, hand-picked examples from known classes, compared to its unlabeled counterpart, which may contain outliers or corrupted data. Semi-supervised learning where the unlabeled set contains more classes than the labeled set (see Fig. 1) is referred to as *open-set semi-supervised learning* (OSSL), where classes present in the labeled set are considered in-distribution (ID), whereas other classes are recognized as out-of-distribution (OOD).

A common approach for training models in OSSL is to use a standard SSL objective but only include unlabeled data that are predicted as ID [5,9,12,49], while the rest are, *e.g.*, discarded [4] or given less importance [9]. This approach is motivated by an assumption that the training signal from OOD data is harmful. Consequently, by restricting learning to samples predicted as ID they are not efficiently using all available data. Some algorithms take steps towards better employment of unlabeled data in OSSL. For example, [33] uses consistency regularization and [17] introduces a self-supervised rotation loss on all unlabeled data, while [19] identifies OOD data semantically similar to ID data that can be "recycled" as such.

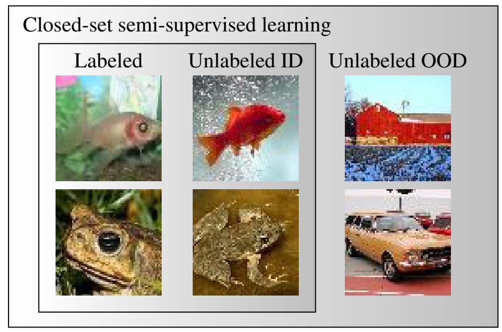

Figure 1. Comparison between open- and closed-set semi-supervised learning.

In this work, we propose a **Se**lf-supervision-centric **F**ramework for **O**pen-set **S**emi-**S**upervised learning (SeFOSS). Our framework unconditionally promotes learning from all unlabeled data by utilizing a self-supervised consistency loss, which is effective in the traditional closed-set SSL setting [41]. Compared to previous methods focusing on predicted inliers through an SSL objective, we argue that making self-supervision the primary source of learning is key, both for data efficiency and robustness in OSSL. The impact of classifying unlabeled data incorrectly as ID, or as the incorrect class, is intuitively much less significant for a self-supervised training objective. To our knowledge, we are the first to utilize self-supervised feature consistency of the type used in [41] for OSSL. Moreover, we prioritize the final model's capacity for open-set recognition (OSR) (predicting data as ID or OOD). To this end, we resort to the recently popularized *free-energy score* [26]. This theoretically founded approach for OSR is well-performing with the added benefit of not requiring any architectural modifications to the classification model.

We assess our framework and compare it with existing approaches through extensive experiments across diverse datasets. The results show that SeFOSS reaches state-of-the-art performance for closed-set accuracy and OSR on many of the evaluated benchmark problems. Although SeFOSS does not reach SOTA on all problems, it consistently displays strong results across the range of evaluated scenarios. We also show that the training of SeFOSS is stable, avoiding the need for additional validation data for early stopping. Furthermore, our experiments reveal that the previous assumption [5, 9, 12] that OOD data significantly hurt the closed-set accuracy of traditional SSL is not always valid. On the contrary, the seminal SSL method FixMatch [37] outperforms all evaluated OSSL methods on closed-set accuracy. Thus, we suggest that designated OSSL methods are mainly important when OSR performance is vital.

Our main contributions are:

- a framework for OSSL that incorporates self-supervised feature consistency for accurate closed-set performance and OSR
- an extensive evaluation across a wide range of open-set scenarios showing that our framework achieves strong results compared to existing methods.
- a challenge to the prior assumption that OOD data in the unlabeled set significantly harms the closed-set performance of traditional SSL methods, indicating that research efforts should be put elsewhere.

2. Related work

Semi-supervised learning: Research in SSL has a long history [1, 8, 23, 32, 36, 48]. Recently, frameworks such as FixMatch [37] and UDA [44] introduced a new paradigm for SSL combining pseudo-labeling with consistency regularization using data augmentations. These works emphasize the importance of strong domain-specific data augmentations for high performance in SSL. In the image domain, these augmentations are, *e.g.*, RandAugment [6] and Cutout [7]. The effectiveness of FixMatch and UDA sparked a new wave of research trying to extend or improve these frameworks. For example, FlexMatch [51], Dash [46], SimMatch [54], CCSSL [47], DP-SSL [45], and DoubleMatch [41] all propose ways to improve the strategies for pseudo-labeling and consistency regularization of UDA and FixMatch.

For this work, we take particular inspiration from DoubleMatch [41], which highlighted the effectiveness of enabling learning from *all* unlabeled data. DoubleMatch is motivated by the fact that UDA and FixMatch restrict learning from unlabeled data to samples for which the model produces confident predictions. In this regard, DoubleMatch adds a self-supervised cosine-similarity loss on the feature predictions across augmentations of all unlabeled data. By promoting prediction consistency also for uncertain unlabeled data, DoubleMatch sees improvements in terms of both final accuracy and training speed. In the context of OSSL, we suggest that this ability to safely learn from all unlabeled data, without inferring class predictions, is of particular interest as it can enable the model to learn from outlier data. For this reason, we include the self-supervision proposed by DoubleMatch as a core part of SeFOSS.

Open-set recognition: The ability to identify previously unseen classes is an important safety feature in many machine learning applications and the task of predicting if data belongs to a pre-defined set of classes or not is often referred to as open-set recognition. This problem is widely studied [2, 14, 15, 22, 24, 25, 35] where existing solutions are based on, *e.g.*, modeling class conditional feature distributions [24], using ensembles of models [22], or analyzing the model predictions under perturbations of input data [25].

Recently, Li *et al.* [26] proposed to use the *free-energy score* for OSR, which for a data point $\mathbf{x}$ is obtained by interpreting the logits $f_y(\mathbf{x})$ for each class y as negative energies: $E(\mathbf{x}, y) = -f_y(\mathbf{x})$. The free-energy score is then given by

$$F(\mathbf{x}) = -\frac{1}{\beta} \log \sum_{y'=1}^{C} e^{-\beta E(\mathbf{x}, y')}, \tag{1}$$

where β is a hyperparameter and C is the number of classes. The free-energy score is theoretically aligned with the marginal distribution for ID data, $p(\mathbf{x})$, so that we can expect $F(\mathbf{x}_\mathrm{a}) < F(\mathbf{x}_\mathrm{b})$ for an $\mathbf{x}_\mathrm{a}$ that is ID and an $\mathbf{x}_\mathrm{b}$ that is OOD. It's worth noting that, for large β, we get $F(\mathbf{x}) \approx \min_{y'} E(\mathbf{x}, y')$, *i.e.*, the maximum logit score, which also has been used successfully for OSR [40].

For SeFOSS, we utilize the free-energy score to determine which samples are ID or OOD. Our main motivation is its simplicity, *i.e.*, not requiring architectural modifications or significant computational complexity, while still being a powerful discriminant. Another benefit of using a method that can be "plugged in" to any existing model is that it allows for easy and fair comparisons of results.

Open-set semi-supervised learning: In OSSL [5, 9, 10, 12, 13, 17–19, 27, 31, 33, 43, 49, 53], we study the case where the unlabeled training set, and sometimes also the test set, contains additional classes not present in the labeled set. Compared to (closed-set) SSL, this is a more general and much less studied problem, and compared to *open-set domain adaptation* (OSDA) [30, 34], there are no assumptions regarding domain shifts in the training sets.

Many proposed methods for OSSL classify the unlabeled data as either ID or OOD and use only confident ID samples for the training objective in a traditional SSL scheme.

One such example is UASD [5], where unlabeled data are classified as inliers based on thresholding the maximum of predicted softmax distribution. The method averages multiple predictions of the same unlabeled data from different time steps during training for increased predictive calibration. The framework MTCF [49] uses a similar strategy but employs a separate OSR head for predicting the probability of a sample being ID. The OSR head trains in an SSL fashion where the OSR head uses model predictions as optimization targets. In SAFE-STUDENT [12], the model is trained using both pseudo-in- and outliers, which are predicted from unlabeled data by using energy discrepancy.

OpenMatch [33] and T2T [17] use a slightly different strategy built around a one-vs-all framework. In both cases, we have one head for each class predicting the probability of a sample belonging to the corresponding class. Predicted inliers are used for SSL objectives based on FixMatch (OpenMatch) or UDA (T2T). Both T2T and OpenMatch take steps towards utilizing all unlabeled data: OpenMatch through a consistency loss for its one-vs-all predictions, and T2T through a self-supervised rotation loss.

In the DS^{3}L method [9], the focus is on preserving closed-set performance by solving the OSSL problem via a bi-level optimization. The inner task is to learn the classifier based on a standard two-term SSL optimization but scaling the loss for unlabeled samples using a data-dependent weight function. The outer task is thus to learn the weight function that minimizes the *labeled* loss term of the inner task. This way, the outer optimization steers the training by using the labeled set as a proxy so that the model never drops in closed-set performance.

Lastly, TOOR [19] proposes to identify "recyclable" OOD data, *i.e.*, semantically close to one of the ID classes. The method projects the recyclable OOD data on the space of ID data by domain adaptation and uses them in conjunction with unlabeled data predicted as ID for training through an SSL objective.

In summary, most existing contributions for OSSL are based on the assumption that OOD data are "harmful" and focus on detecting ID data from the unlabeled set to use in an SSL training objective. In SeFOSS, we instead facilitate learning from all unlabeled data while also learning to distinguish between ID and OOD. Additionally, in contrast to many prior works, we do not require extra model heads for classifying data as ID or OOD. Using the free energy score, we avoid additional model parameters or heavy computational complexity for solving the OSR task.

3. Method

This section describes SeFOSS, our proposed method for OSSL. The main philosophy behind this method is that we encourage learning from *all* unlabeled data, whether it is ID or OOD. Our proposed method achieves this by applying the self-supervised loss of DoubleMatch [41] on all unlabeled samples in each training batch.

We complement the self-supervision with losses on unlabeled data confidently predicted as ID or OOD to improve OSR performance. For samples confidently predicted as ID, we apply a pseudo-labeling loss similar to those of FixMatch [37] and UDA [44]. For unlabeled data confidently predicted as outliers, we instead use energy regularization to increase the model's confidence that these are OOD. Figure 2 illustrates how SeFOSS treats unlabeled data.

Following standard SSL practice, losses on unlabeled data are combined with supervised cross-entropy on labeled data. The components of SeFOSS are detailed below.

3.1. Self-supervision on all unlabeled data

The central source of learning from unlabeled data in SeFOSS is self-supervision. To this end, we use the loss proposed by [41], *i.e.*, a cosine similarity between feature predictions for different augmentations of unlabeled data:

$$l_s = -\frac{1}{\mu B}\sum_{i=1}^{\mu B}\frac{h(\mathbf{v}_i)\cdot\mathbf{z}_i}{\|h(\mathbf{v}_i)\|\|\mathbf{z_i}\|}, \tag{2}$$

where μB is the unlabeled batch size, $\mathbf{v}_i$ and $\mathbf{z}_i$ are d-sized feature vectors from the penultimate network layer for weak and strong augmentations of sample i, respectively. The operator $\|\cdot\|$ is the l_2 norm. The mapping $h:\mathbb{R}^d\to\mathbb{R}^d$ is a trainable linear projection to allow for differences in feature predictions for weak and strong augmentations. In gradient evaluations of l_s, we treat $\mathbf{z}_i$ as constant.

A principal difference between the self-supervision of (2) and the losses for all unlabeled data in T2T [17] and OpenMatch [33], is that (2) makes use of strong data-augmentation whereas the corresponding losses of T2T and OpenMatch use weak data-augmentation only.

3.2. Pseudo-labeling loss for pseudo-inliers

SeFOSS uses the free-energy score [26] as defined in (1) for OSR. For convenience, we define the equivalent function $s:\mathbb{R}^C\to\mathbb{R}$ that operates on the logits $\boldsymbol{\sigma}$ as

$$s(\boldsymbol{\sigma}) = -\log\sum_{y'=1}^{C}e^{\sigma_{y'}}, \tag{3}$$

where $\sigma_{y'} = f_{y'}(\mathbf{x})$ is the logit associated with class y' for data point $\mathbf{x}$. For data that are confidently ID (pseudo-inliers), we expect $s(\boldsymbol{\sigma})$ to be low. To amplify this confidence, we apply the pseudo-labeling loss

$$\begin{aligned} l_p = \frac{1}{\mu B}\sum_{i=1}^{\mu B}&\mathbb{1}\{s(\mathbf{w}_i)<\tau_{\text{id}}\} \\ &\times H(\operatorname{argmax}(\mathbf{w}_i),\operatorname{softmax}(\mathbf{q}_i)), \end{aligned} \tag{4}$$

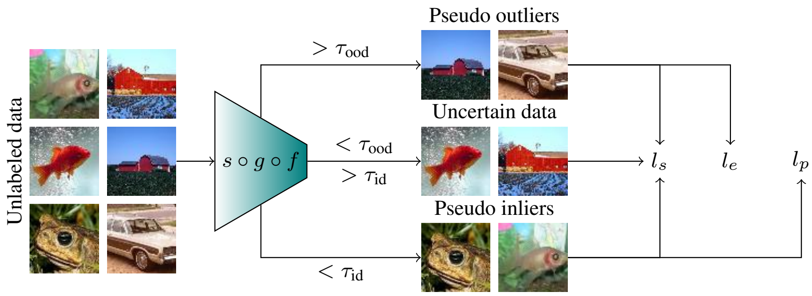

Figure 2. Sorting of *unlabeled data* in SeFOSS. Predicted outliers are used for energy regularization in l_e. Predicted inliers are used for a pseudo-labeling loss in l_p. All unlabeled data are used for self-supervision in l_s.

where τ_{id} is a confidence threshold. We let $\text{argmax} : \mathbb{R}^C \rightarrow \mathbb{R}^C$ so that it returns a one-hot vector. Similarly, we define $\text{softmax} : \mathbb{R}^C \rightarrow \mathbb{R}^C$ so that the j:th element is

$$\text{softmax}(\boldsymbol{\sigma})_j = \frac{e^{\sigma_j}}{\sum_{y'=1}^{C} e^{\sigma_{y'}}}, \text{ for } j = 1, \ldots, C. \tag{5}$$

The inputs $\mathbf{w}_i$ and $\mathbf{q}_i$ in (4) are the predicted logits for unlabeled sample i for weak and strong data augmentations, respectively. Lastly, $\mathbb{1}\{\cdot\}$ is the indicator function and $H(\cdot, \cdot)$ is the cross entropy between two discrete probability distributions $\mathbf{p}^a$, $\mathbf{p}^b$ given by

$$H(\mathbf{p}^a, \mathbf{p}^b) = -\sum_{i=1}^{C} p_i^a \log p_i^b, \tag{6}$$

where p_i^a and p_i^b are the i:th elements of $\mathbf{p}^a \in \mathbb{R}^C$ and $\mathbf{p}^b \in \mathbb{R}^C$, respectively. As for l_s from (2), we consider the predictions on weakly augmented data as constant when evaluating the gradient of l_p.

Note that l_p is equivalent to the pseudo-labeling loss in FixMatch [37] with the exception that FixMatch selects pseudo-labels by thresholding the maximum value of the predicted softmax distributions. We choose pseudo-labels by thresholding the free-energy score, which is a better score for OSR than the maximum softmax probability [26].

3.3. Energy regularization for pseudo-outliers

Many existing methods for OSSL [5, 9, 49] focus on identifying unlabeled data that are confidently ID. To boost the separation of OSR predictions between ID and OOD, we suggest including a loss for unlabeled data confidently predicted as OOD, *i.e.*, pseudo-outliers. Inspired by [26], we employ a hinge loss on the free-energy score of the pseudo-outliers to stimulate the model to raise the free-energy score to a given margin. This energy regularization is given by

$$l_e = \frac{\sum_{i=1}^{\mu B} \mathbb{1}\{s(\mathbf{w}_i) > \tau_{\text{ood}}\} \max(0, m_{\text{ood}} - s(\mathbf{w}_i))^2}{\sum_{i=1}^{\mu B} \mathbb{1}\{s(\mathbf{w}_i) > \tau_{\text{ood}}\}}, \tag{7}$$

where $\tau_{\text{ood}} \in \mathbb{R}$ is the confidence threshold for OOD data and $m_{\text{ood}} \in \mathbb{R}$ is the margin for the hinge loss. Note that l_e only uses the predictions for weakly augmented data because we empirically found that using strongly augmented data in l_e led to instabilities during training.

3.4. Adaptive confidence thresholds

In SeFOSS, we select pseudo-in- and outliers from unlabeled data, cf. (4) and (7), based on thresholding the free-energy score s defined in (3). The free-energy score is non-probabilistic and thus unbounded, so selecting these thresholds is not trivial. Thus, we propose adaptively calculating $\tau_{\text{id}}, \tau_{\text{ood}}$ and m_{ood} based on the distribution of s given our labeled training set at the end of a pre-training phase. During this pre-training phase, the model is trained only with a supervised loss on labeled data and the self-supervised loss given by (2) on unlabeled data. Following the pre-training phase, we compute s for the complete (unaugmented) labeled training set. The energy scores are evaluated using an exponential moving average of the model parameters for stability. Given the set of energy scores $\{S_i : i = 1, \ldots, M\}$, where M is the number of labeled training data, we compute the median S_m and the interquartile range S_{iqr}. The confidence thresholds and the margin for the energy regularization are then set as

$$\begin{aligned} \tau_{\text{id}} &\leftarrow S_m - S_{iqr} \cdot \zeta_{\text{id}}, \\ \tau_{\text{ood}} &\leftarrow S_m + S_{iqr} \cdot \zeta_{\text{ood}} \\ m_{\text{ood}} &\leftarrow S_m + S_{iqr} \cdot \xi_{\text{ood}}, \end{aligned} \tag{8}$$

where ζ_{id}, ζ_{ood}, and ξ_{ood} are scalar hyperparameters. By using these adaptive metrics, we expect a tuned set of the hyperparameters ζ_{id}, ζ_{ood}, and ξ_{ood} to work for a wider set of problems, compared to direct tuning of τ_{id}, τ_{ood}, and m_{ood}. Compared to the Otsu method [29] used for adaptive thresholding on unlabeled data in OSSL [17, 43], our method has no requirement for clearly bi-modal situations [20] since we only consider ID training data. Moreover, in contrast to the adaptive thresholds for pseudo-labeling used in closed-set SSL [42, 46, 51], their purpose is to lower thresholds to assign more pseudo-labels, which is not desired in OSSL.

Algorithm 1 SeFOSS training loop

Require: Trainable models f, g, and h, labeled training data $\{(\mathbf{x}_1, \mathbf{y}_1), \dots, (\mathbf{x}_M, \mathbf{y}_M)\}$, unlabeled training data $\{\tilde{\mathbf{x}}_1, \dots, \tilde{\mathbf{x}}_{\tilde{M}}\}$, scaling parameters w_p, w_s, w_e, and w_w, learning rate scheme $\eta(k)$

▷ Pretraining loop
for $i = 1, \dots, K_\text{p}$ **do**
  Compute loss according to Algorithm 2 setting $w_e = w_p = 0$
  Update parameters of f, g, and h using SGD
end for
▷ Compute confidence thresholds by looping through all labeled data
for $i = 1, \dots, M$ **do**
  $S_i = s(g(f(\mathbf{x}_i)))$
end for
Compute thresholds and margin τ_id, τ_ood, and m_ood from Eq. (8)
▷ Training loop
for $i = K_\text{p} + 1, \dots, K$ **do**
  Compute loss using Algorithm 2 using $w_p = 1$ and a tuned w_e
  Update parameters of f, g, and h using SGD
end for
return trained classification model $g \circ f$

3.5. Full training objective

The full training loss is a weighted sum of five terms:

$$l = l_l + w_p l_p + w_s l_s + w_e l_e + w_w l_w, \tag{9}$$

where w_p, w_s, w_e, and w_w are scalar hyperparameters for controlling the relative importance of each of the terms. Here, l_s is the self-supervision loss in (2), l_p is the pseudo-labeling loss in (4), and l_e is the energy regularization term in (7). We also use a supervised loss for the labeled data:

$$l_l = \frac{1}{B} \sum_{i=1}^{B} H(\mathbf{y}_i, \text{softmax}(\mathbf{o}_i)), \tag{10}$$

where B is the number of labeled data in each batch, $\mathbf{y}_i \in \mathbb{R}^C$ is the one-hot label vector for labeled sample i, and $\mathbf{o}_i \in \mathbb{R}^C$ is the predicted logits for weakly augmented labeled sample i. Lastly, we add weight-regularization

$$l_w = \frac{1}{2} \|\boldsymbol{\theta}\|^2, \tag{11}$$

$\boldsymbol{\theta}$ being a vector of all trainable weights (excluding biases).

3.6. Data augmentation and optimization

Our method utilizes weak or strong data augmentation for labeled and unlabeled data during training, where we follow one of the proposed augmentation strategies from FixMatch [37]. Weak augmentations involve a stochastic flip-and-shift strategy. Strong augmentation stacks the weak augmentation with two randomly selected transformations from RandAugment [6], followed by Cutout [7].

For optimization, we use stochastic gradient descent with Nesterov momentum [38]. We define a scheme for the learning rate η where the learning rate stays constant in the pre-training phase and follows a cosine decay in the subsequent training phase:

Algorithm 2 SeFOSS training step

Require: Strong augmentation β, weak augmentation α, labeled batch $\{(\mathbf{x}_1, \mathbf{y}_1), \dots, (\mathbf{x}_B, \mathbf{y}_B)\}$, unlabeled batch $\{\tilde{\mathbf{x}}_1, \dots, \tilde{\mathbf{x}}_{\mu B}\}$, scaling parameters w_p, w_s, w_e, and w_w, thresholds τ_id, τ_ood, and m_ood, backbone model f, prediction layer g, projection layer h

▷ Cross-entropy loss for (weakly augmented) labeled data
for $i = 1, \cdots, B$ **do**
  $\mathbf{o}_i = g(f(\alpha(\mathbf{x}_i)))$
end for
$l_l = \frac{1}{B} \sum_{i=1}^{B} H(\mathbf{y}_i, \text{softmax}(\mathbf{o}_i))$
▷ Predictions on unlabeled data
for $i = 1, \cdots, \mu B$ **do**
  $\mathbf{z}_i = f(\alpha(\tilde{\mathbf{x}}_i))$ ▷ Weak augmentation
  $\mathbf{v}_i = f(\beta(\tilde{\mathbf{x}}_i))$ ▷ Strong augmentation
  $\mathbf{q}_i = g(\mathbf{v}_i)$
  $\mathbf{w}_i = g(\mathbf{z}_i)$
end for
Compute l_s, l_p, l_e, and l_w according to (2), (4), (7), and (11)
return $l_l + w_p l_p + w_s l_s + w_e l_e + w_w l_w$

$$\eta(k) = \begin{cases} \eta_0 & \text{for} \quad k < K_p \\ \eta_0 \cos\left(\gamma \frac{\pi(k-K_p)}{2(K-K_p)}\right) & \text{otherwise} \end{cases}, \tag{12}$$

where η_0 is the initial learning rate, γ is a hyperparameter that controls the decay rate, k is the current training step, K_p is the number of pre-training steps, and K is the total number of training steps.

Our training procedure is summarized in Algorithm 1, and each training step is detailed in Algorithm 2, where we denote the trainable parts of our model as f, g, and h. The backbone model $f : \mathbb{R}^D \rightarrow \mathbb{R}^d$ maps the input images of dimension D to the feature predictions of dimension d. The classification head, $g : \mathbb{R}^d \rightarrow \mathbb{R}^C$ is a linear head that predicts logits given feature predictions. Finally, $h : \mathbb{R}^d \rightarrow \mathbb{R}^d$ is a projection head that performs a linear transformation on the feature predictions of strongly augmented data to the feature space of weakly augmented data, cf. (2).

4. Experiments

We present experimental results for SeFOSS alongside other SSL and OSSL methods across various datasets. The OSSL methods that we compare with, MTCF [49], OpenMatch [33], and T2T [17], all seek to produce well-performing models in terms of both closed-set accuracy and OSR. For comparison, we include closed-set SSL method FixMatch [37] and a fully supervised baseline trained using only the labeled subset. We evaluate closed-set accuracy and OSR performance for each method, where OSR is measured as the AUROC for classifying data as ID or OOD. For our method, FixMatch, and fully supervised, we use the free-energy score (3) to evaluate AUROC, whereas for the others we use the scores proposed in the respective paper.

Figure 3. A few representative examples from the datasets used in our experiments.

4.1. Datasets

We use several datasets with different characteristics for a complete performance assessment. For ID data, we use CIFAR-10, CIFAR-100 [21] and ImageNet-30 [16]. The CIFAR sets are of size 32^2 and comprise 10 and 100 classes, respectively. Both have training sets of 50,000 images and test sets of size 10,000. When CIFAR-10/100 is used as ID, we use SVHN [28], uniform noise, and the corresponding CIFAR set as OOD. SVHN consists of images showing house numbers divided into 73,257 images in the training set and 26,032 for the test set. The uniform noise dataset has 50,000 training images and 10,000 test images. ImageNet-30 is a 30-class subset of ImageNet, selected such that there is no overlap between the classes. Following [33], we use the first 20 classes as ID and the last 10 classes as OOD. Each class has 1,300 training images and 100 test images. The images of ImageNet-30 are first resized so that the shortest side gets a length of 256 while keeping the aspect ratio. Each image is then center-cropped to size 224^2.

The different open-set scenarios presented above pose distinct challenges. CIFAR-10 and CIFAR-100 originate from the same source set [21] and contain semantically similar classes, *e.g.*, CIFAR-10 contains *dogs* whereas CIFAR-100 has images of *wolves*, making for a challenging OSR problem. On the other hand, the similarities could potentially increase the possibility of learning useful features from the OOD set. Conversely, we have OOD data constituting pure noise images containing no semantic content or learnable features. However, they could cause unexpected behavior if misclassified and thus generate unwanted training signals. The in-between scenario is the SVHN set containing real images with learnable features but is semantically very different from CIFAR-10 and CIFAR-100. ImageNet-30 showcases how the methods perform on more realistic data similar to that used in practical applications. Example images are shown in Fig. 3.

4.2. Limitations

The experiments in this work only consider OSSL problems where the OOD data in training and testing follow the same distributions. Furthermore, we do not evaluate our method for very low-label regimes (*i.e.*, only a few labeled training samples per class). Lastly, we only use ID sets that are balanced in terms of classes.

4.3. Implementation details

The architectures used for the experiments are WRN-28-2 [50] (ID: CIFAR-10), WRN-28-8 [50] (ID: CIFAR-100), and ResNet-18 [11] (ImageNet-30). In SeFOSS, when CIFAR-10 is ID, we use $w_e = 10^{-4}$, $w_s = 5.0$, $\eta_0 = 0.03, \gamma = 7/8, B = 64, \mu = 7, \xi_{\text{id}} = 0.2, \xi_{\text{ood}} = 1.3, \zeta_{\text{ood}} = 1.9, K_p = 5 \cdot 10^4, K = 4 \cdot 10^5, w_d = 5 \cdot 10^{-4}$, and SGD momentum 0.9. When CIFAR-100 is ID we use $w_s = 15.0, w_d = 10^{-4}$ and $\gamma = 5/8$ (following [41]), keeping the other hyperparameters the same. For ImageNet-30, we use the same hyperparameters as for CIFAR-10 except that we use $K = 2 \cdot 10^5$ because of the more expensive training steps. We evaluate SeFOSS using an exponential moving average of the model parameters (with momentum 0.999). For T2T and OpenMatch, we use the original authors' implementations and hyperparameters. For a fair comparison, we implement MTCF with the FixMatch backbone (with original FixMatch hyperparameters). Our experiments with FixMatch use hyperparameters from the original work. The fully supervised baseline is trained for 50,000 steps, uses batch size 256, and a learning rate from (12) with $\gamma = 7/8$ and $\eta_0 = 0.03$.

4.4. OSSL performance

CIFAR-10/100: The results from using CIFAR-10 and CIFAR-100 as ID data are shown in Tab. 1. For CIFAR-10, we use labeled sets of sizes 1,000 and 4,000, whereas, for CIFAR-100, the sets have sizes 2,500 and 10,000. The top row for each method shows closed-set accuracy in %, and the bottom row shows AUROC for OSR. We evaluate each combination of ID and OOD datasets for each method using five different sub-samplings of the complete labeled data. The reported numbers are the mean and standard deviation from these five training sessions. The results from each session are evaluated by taking median performance values from the last five model evaluations during training.

Tab. 1 shows that SeFOSS is the only method that reaches good performances for both closed-set accuracy and AUROC across all scenarios. MTCF shows overall good AUROC but generally performs worse than SeFOSS on closed-set accuracy. T2T reaches good results on a few scenarios (*e.g.* CIFAR-100 with 10,000 labels with noise as OOD) but does not consistently perform well across scenarios. A drawback of T2T is that it displays high variance in AUROC for many scenarios. OpenMatch shows very good results in terms of closed-set accuracy when CIFAR-10 is

	CIFAR-10: 1,000 labels			CIFAR-10: 4,000 labels			CIFAR-100: 2,500 labels			CIFAR-100: 10,000 labels		
	CIFAR-100	SVHN	Noise	CIFAR-100	SVHN	Noise	CIFAR-10	SVHN	Noise	CIFAR-10	SVHN	Noise
Fully supervised	54.51±1.82	54.03±2.05	55.39±2.67	75.57±2.88	76.70±2.27	77.62±1.79	34.62±1.43	33.19±1.80	34.03±1.50	59.12±0.91	60.32±0.54	59.40±2.11
	0.62±0.01	0.61±0.04	0.56±0.22	0.74±0.02	0.80±0.03	0.78±0.15	0.61±0.01	0.57±0.05	0.34±0.19	0.71±0.01	0.76±0.08	0.77±0.17
FixMatch [37]	92.70±0.14	94.80±0.19	95.02±0.10	94.07±0.15	94.93±0.22	95.38±0.07	71.95±0.49	69.39±0.14	70.89±0.42	77.72±0.32	75.89±0.39	77.04±0.24
	0.66±0.00	0.67±0.03	0.71±0.05	0.69±0.01	0.67±0.02	0.73±0.01	0.46±0.01	0.49±0.03	0.67±0.17	0.51±0.01	0.48±0.03	0.65±0.04
MTCF [49]	82.96±1.08	90.49±0.79	89.32±0.65	89.87±0.21	92.72±0.14	92.01±0.32	40.46±1.49	53.55±1.24	46.56±0.66	62.88±0.92	66.10±0.63	63.80±0.75
	0.81±0.00	$\underline{1.00}$±0.00	$\underline{1.00}$±0.00	0.84±0.00	$\underline{1.00}$±0.00	$\underline{1.00}$±0.00	$\underline{0.82}$±0.01	$\underline{1.00}$±0.00	$\underline{1.00}$±0.00	0.80±0.01	$\underline{1.00}$±0.00	$\underline{1.00}$±0.00
T2T [17]	86.99±1.09	91.83±1.20	91.13±1.12	86.11±1.91	92.16±1.00	92.91±0.57	38.30±9.72	58.44±18.14	51.33±9.59	62.02±3.73	70.93±4.38	73.01±0.37
	0.57±0.02	0.96±0.07	0.72±0.26	0.57±0.04	0.80±0.24	0.90±0.19	0.63±0.08	0.80±0.40	$\underline{1.00}$±0.00	0.59±0.08	0.66±0.42	$\underline{1.00}$±0.00
OpenMatch [33]	**92.20**±0.15	**94.12**±0.34	**94.07**±0.08	**94.82**±0.21	**94.73**±0.10	**94.76**±0.15	20.84±8.65	18.66±2.59	16.10±5.70	40.95±20.44	32.69±9.68	21.19±8.55
	$\underline{0.93}$±0.00	0.98±0.03	0.68±0.40	$\underline{0.96}$±0.00	$\underline{1.00}$±0.00	0.58±0.24	0.66±0.05	0.69±0.10	0.85±0.18	0.77±0.11	0.68±0.19	0.50±0.32
SeFOSS (*ours*)	91.49±0.16	91.16±0.27	92.78±1.00	93.73±0.27	92.60±0.40	94.14±0.09	**68.48**±0.26	**62.99**±0.39	**64.54**±1.00	**77.63**±0.21	**73.60**±0.20	**75.25**±0.34
	0.90±0.01	0.99±0.01	$\underline{1.00}$±0.00	0.92±0.00	$\underline{1.00}$±0.00	$\underline{1.00}$±0.00	0.79±0.01	$\underline{1.00}$±0.00	$\underline{1.00}$±0.00	$\underline{0.83}$±0.00	$\underline{1.00}$±0.00	$\underline{1.00}$±0.00

Table 1. Closed-set accuracy (top rows) and AUROC (bottom rows) for SSL and OSSL methods when using CIFAR-10/100 as ID data. Boldface marks the best closed-set accuracies and underline marks the best AUROCs among OSSL methods.

	MTCF [49]	OpenMatch [33]	SeFOSS *(ours)*
Accuracy	86.4±0.7	89.6±1.0	**92.53**±0.10
AUROC	0.94±0.00	0.96±0.00	$\underline{0.97}$±0.00

Table 2. Closed-set accuracy and AUROC on ImageNet-30 dataset when using 2,600 labels. Baseline results are taken from [33].

	CIFAR-100: 2,500 labels			CIFAR-100: 10,000 labels		
	CIFAR-10	SVHN	Noise	CIFAR-10	SVHN	Noise
OpenMatch	63.33±0.86	63.41±1.32	58.97±0.52	75.89±0.23	75.56±0.17	75.08±0.28
	$\underline{0.86}$±0.01	$\underline{1.00}$±0.00	0.42±0.47	$\underline{0.92}$±0.01	$\underline{1.00}$±0.00	0.24±0.20
SeFOSS	**76.16**±0.39	**71.78**±0.27	**73.26**±0.69	**79.37**±0.21	**76.11**±0.28	**77.53**±0.17
	0.83±0.01	$\underline{1.00}$±0.00	$\underline{1.00}$±0.00	0.83±0.01	$\underline{1.00}$±0.00	$\underline{1.00}$±0.00

Table 3. Closed-set accuracy and AUROC with CIFAR-100 as ID using an additional 50 labels/class as validation data. OpenMatch uses the validation data for selecting the best model during training, and SeFOSS uses the data as extra labeled training data.

ID. OpenMatch however seems unable to handle noise as OOD data since it displays poor and high-variance AUROC for these scenarios. OpenMatch also drops in performance in the CIFAR-100 experiments for both closed-set accuracy and AUROC, where it is outperformed by the fully supervised lower bound for some scenarios. Slightly surprisingly, the closed-set SSL method FixMatch obtains the highest closed-set accuracy in nearly all scenarios. More expected, however, is that FixMatch consistently gives poor AUROC since it can freely assign pseudo-labels to OOD samples, leading to erroneous and overconfident predictions on these data. Existing works [17,33] have reported comparably worse performance for FixMatch, the reason for this is likely the use of hyperparameters that are different from the original work (*e.g.*, fewer training steps).

ImageNet-30: For the results on ImageNet-30, we compare with numbers reported by [33]. The number of labeled data used here is 2,600. To make our results comparable with [33], we report the test performance also for SeFOSS at the point of best validation performance given a labeled validation set of 1,000 images. The reported numbers are means and standard deviations over three runs. The results in Tab. 2 show that SeFOSS reaches better results than MTCF and OpenMatch in terms of both AUROC and closed-set accuracy. Note also that the hyperparameters used for SeFOSS are the same as for CIFAR-10, indicating that SeFOSS scales well to high-resolution data.

Avoiding collapse using validation data: To present the fairest possible evaluation of OpenMatch, we note that the poor results displayed in Tab. 1 when CIFAR-100 is ID are many times a result of a training collapse from a much better performance. This collapse can be avoided by using a labeled validation set and selecting the model that yields the best performance on the validation set during training. The official code of [33] uses 50 images per class for this purpose. The results from evaluating OpenMatch using this procedure are shown in Tab. 3, where we see that OpenMatch displays much better results, although it does not solve the poor AUROC for noise OOD. However, for fairness, as SeFOSS does not suffer from training collapse and, thus, has no need for a validation set, it is free to use the additional data during training instead, resulting in a significant boost to closed-set accuracy. As SSL methods are meant for situations where labeled data are scarce or expensive, assuming the presence of a sufficiently large labeled validation set during training goes against this philosophy.

4.5. Influence of OOD data on SSL methods

From Tab. 1, we see that FixMatch displays high closed-set accuracies for all datasets. These results contradict prior works where OOD data in SSL are assumed to significantly harm the closed-set performance [9, 12]. To investigate this further, we study how a few SSL methods perform when trained using unlabeled data containing different fractions of OOD data. The SSL methods that we evaluate are Fix-

Match, UDA [44], and MixMatch [3]. We also include SeFOSS for comparison. The dataset used consists of CIFAR-10 with 4,000 labels as ID data and CIFAR-100 as OOD data. The unlabeled datasets with different fractions of OOD data are created by adding CIFAR-100 data (up to 0.5) or removing CIFAR-10 data (above 0.5). For OSR, the SSL methods are evaluated using the free-energy score (3). The results are shown in Fig. 4.

Most notable from the results is that FixMatch and UDA show no significant drop in closed-set accuracy when the fraction of OOD data is below 0.4. MixMatch loses closed-set accuracy faster, likely due to its use of mixup [52] augmentations, which intuitively should not handle OOD data well. For AUROC, we see a quick drop in performance for all SSL methods as OOD data gets added to the unlabeled set. It is also here we see a significant difference in the performance of SeFOSS and the traditional SSL methods. Our framework consistently shows high AUROC (around 0.9) when the fraction of OOD data is below 0.7.

4.6. Ablation

SeFOSS uses three loss functions to learn from unlabeled data, see Fig. 2. To study the importance of these terms, we conduct experiments using 1) only self-supervision on unlabeled data, 2) self-supervision and energy regularization, and 3) self-supervision and pseudo-labeling. Additionally, we evaluate the OSR performance of case 1) using the maximum softmax confidence to confirm that the free-energy score gives better performance.

The results in Tab. 4 show that using only the self-supervision gives nearly as good results as using the entire framework. Adding pseudo-labeling and energy regularization barely affects the closed-set accuracy but gives better AUROC by a couple of percentage points. This indicates that self-supervision alone, at least under these conditions, is a strong and safe training signal for OSSL.

Moreover, we study the effect of manually adjusting τ_{id}. In Tab. 5 we see an increase in accuracy at the cost of a reduction in AUROC when lowering τ_{id}. The increase in accuracy can be explained by the model assigning pseudo-labels to more data. However, when τ_{id} is low, many predicted inliers are likely outliers, causing weaker OSR performance. These experiments are done with $w_e = 0$ to isolate the effect of τ_{id}.

CIFAR-10: 4,000 labels - OOD: CIFAR-100					
l_s	l_e	l_p	OSR-score	Accuracy	AUROC
✓			conf	$93.87_{\pm 0.17}$	$0.89_{\pm 0.00}$
✓			energy	$93.87_{\pm 0.17}$	$0.90_{\pm 0.01}$
✓		✓	energy	$93.74_{\pm 0.17}$	$0.90_{\pm 0.01}$
✓	✓		energy	$93.85_{\pm 0.12}$	$0.91_{\pm 0.01}$
✓	✓	✓	energy	$93.73_{\pm 0.27}$	$0.92_{\pm 0.00}$

Table 4. Evaluating different modifications of SeFOSS.

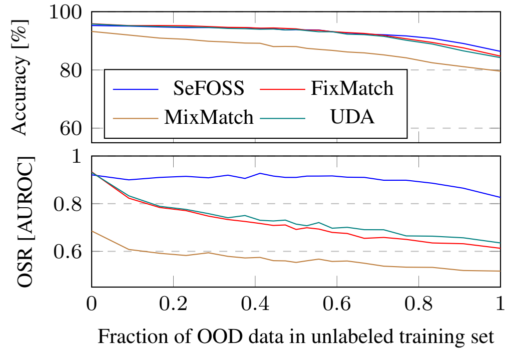

Figure 4. Closed-set accuracy and AUROC using different fractions of OOD data in the unlabeled set. ID data are CIFAR-10 with 4,000 labels. The OOD set is CIFAR-100.

CIFAR-100: 2,500 labels, OOD: CIFAR-10						
τ_{id}	-30	-25	-20	-15	-10	-5
Accuracy	68.93	68.44	68.51	68.35	70.21	69.25
AUROC	0.78	0.77	0.79	0.78	0.68	0.57

Table 5. Evaluating SeFOSS when manually adjusting τ_{id}.

5. Conclusion

This paper shows that self-supervision on all unlabeled data combined with OSR predictions using the free-energy score can be successfully applied in an OSSL context. Our proposed framework reaches overall strong and consistent results on a wide range of OSSL problems when considering both closed-set accuracy and OSR. In particular, it displays SOTA performance on the more challenging tests when CIFAR-100 is ID and the more realistic scenarios using ImageNet-30. However, we should note that the performance is marginally below SOTA when CIFAR-10 is ID. Moreover, if focusing solely on closed-set accuracy, we show that SSL methods can perform equal to or better than designated OSSL methods even when unlabeled data contain OOD data, indicating that the challenge with OSSL lies in ensuring OSR performance during employment.

Acknowledgment

This work was supported by Saab AB, the Swedish Foundation for Strategic Research, and Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. The experiments were enabled by resources provided by the Swedish National Infrastructure for Computing (SNIC) at Chalmers Centre for Computational Science and Engineering (C3SE), and National Supercomputer Centre (NSC) at Linköping University.

References

[1] A Agrawala. Learning with a probabilistic teacher. *IEEE Transactions on Information Theory*, 1970. 2

[2] Abhijit Bendale and Terrance E Boult. Towards open set deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016. 2

[3] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. MixMatch: A holistic approach to semi-supervised learning. In *Advances in Neural Information Processing Systems*, 2019. 8

[4] Guangyao Chen, Limeng Qiao, Yemin Shi, Peixi Peng, Jia Li, Tiejun Huang, Shiliang Pu, and Yonghong Tian. Learning open set network with discriminative reciprocal points. In *European Conference on Computer Vision*, 2020. 1

[5] 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*, 2020. 1, 2, 3, 4

[6] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. RandAugment: Practical automated data augmentation with a reduced search space. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020. 2, 5

[7] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. 2, 5

[8] S Fralick. Learning to recognize patterns without a teacher. *IEEE Transactions on Information Theory*, 1967. 2

[9] Lan-Zhe Guo, Zhen-Yu Zhang, Yuan Jiang, Yu-Feng Li, and Zhi-Hua Zhou. Safe deep semi-supervised learning for unseen-class unlabeled data. In *International Conference on Machine Learning*, 2020. 1, 2, 3, 4, 7

[10] Lu Han, Han-Jia Ye, and De-Chuan Zhan. On pseudo-labeling for class-mismatch semi-supervised learning. *Transactions on Machine Learning Research*, 2022. 2

[11] 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. 6

[12] 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*, 2022. 1, 2, 3, 7

[13] Rundong He, Zhongyi Han, Yang Yang, and Yilong Yin. Not all parameters should be treated equally: Deep safe semi-supervised learning under class distribution mismatch. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022. 2

[14] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *The International Conference on Learning Representations*, 2017. 2

[15] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *The International Conference on Learning Representations*, 2019. 2

[16] Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. Using self-supervised learning can improve model robustness and uncertainty. *Advances in neural information processing systems*, 32, 2019. 6

[17] Junkai Huang, Chaowei Fang, Weikai Chen, Zhenhua Chai, Xiaolin Wei, Pengxu Wei, Liang Lin, and Guanbin Li. Trash to treasure: Harvesting ood data with cross-modal matching for open-set semi-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 1, 2, 3, 4, 5, 7

[18] Zhe Huang, Mary-Joy Sidhom, Benjamin Wessler, and Michael C Hughes. Fix-a-step: Semi-supervised learning from uncurated unlabeled data. In *International Conference on Artificial Intelligence and Statistics*, 2023. 2

[19] Zhuo Huang, Jian Yang, and Chen Gong. They are not completely useless: Towards recycling transferable unlabeled data for class-mismatched semi-supervised learning. *IEEE Transactions on Multimedia*, 2022. 1, 2, 3

[20] J. Kittler and J. Illingworth. On threshold selection using clustering criteria. *IEEE Transactions on Systems, Man, and Cybernetics*, 1985. 4

[21] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. 6

[22] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 2017. 2

[23] 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*, 2013. 1, 2

[24] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 2018. 2

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

[26] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2020. 2, 3, 4

[27] Sangwoo Mo, Jong-Chyi Su, Chih-Yao Ma, Mido Assran, Ishan Misra, Licheng Yu, and Sean Bell. Ropaws: Robust semi-supervised representation learning from uncurated data. In *International Conference on Learning Representations*, 2023. 2

[28] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011. 6

[29] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, 1979. 4

[30] Pau Panareda Busto and Juergen Gall. Open set domain adaptation. In *Proceedings of the IEEE international conference on computer vision*, 2017. 2

[31] Jongjin Park, Sukmin Yun, Jongheon Jeong, and Jinwoo Shin. Opencos: Contrastive semi-supervised learning for handling open-set unlabeled data. In *European Conference on Computer Vision*, 2022. 2

[32] Antti Rasmus, Mathias Berglund, Mikko Honkala, Harri Valpola, and Tapani Raiko. Semi-supervised learning with ladder networks. In *Advances in Neural Information Processing Systems*, 2015. 2

[33] Kuniaki Saito, Donghyun Kim, and Kate Saenko. Openmatch: Open-set semi-supervised learning with open-set consistency regularization. In *Advances in Neural Information Processing Systems*, 2021. 1, 2, 3, 5, 6, 7

[34] Kuniaki Saito, Shohei Yamamoto, Yoshitaka Ushiku, and Tatsuya Harada. Open set domain adaptation by backpropagation. In *Proceedings of the European conference on computer vision (ECCV)*, 2018. 2

[35] Walter J Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E Boult. Toward open set recognition. *IEEE transactions on pattern analysis and machine intelligence*, 2012. 2

[36] Henry Scudder. Probability of error of some adaptive pattern-recognition machines. *IEEE Transactions on Information Theory*, 1965. 1, 2

[37] 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. In *Advances in Neural Information Processing Systems*, 2020. 1, 2, 3, 4, 5, 7

[38] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, 2013. 5

[39] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems*, 2017. 1

[40] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: A good closed-set classifier is all you need. In *The International Conference on Learning Representations*, 2022. 2

[41] Erik Wallin, Lennart Svensson, Fredrik Kahl, and Lars Hammarstrand. DoubleMatch: Improving semi-supervised learning with self-supervision. In *International Conference on Pattern Recognition*, 2022. 1, 2, 3, 6

[42] Yidong Wang, Hao Chen, Qiang Heng, Wenxin Hou, Marios Savvides, Takahiro Shinozaki, Bhiksha Raj, Zhen Wu, and Jindong Wang. Freematch: Self-adaptive thresholding for semi-supervised learning. In *International Conference on Learning Representations*, 2023. 4

[43] Yu Wang, Pengchong Qiao, Chang Liu, Guoli Song, Xiawu Zheng, and Jie Chen. Out-of-distributed semantic pruning for robust semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2, 4

[44] Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. In *Advances in Neural Information Processing Systems*, 2020. 2, 3, 8

[45] Yi Xu, Jiandong Ding, Lu Zhang, and Shuigeng Zhou. DP-SSL: Towards robust semi-supervised learning with a few labeled samples. In *Advances in Neural Information Processing Systems*, 2021. 2

[46] Yi Xu, Lei Shang, Jinxing Ye, Qi Qian, Yu-Feng Li, Baigui Sun, Hao Li, and Rong Jin. Dash: Semi-supervised learning with dynamic thresholding. In *International Conference on Machine Learning*, 2021. 2, 4

[47] Fan Yang, Kai Wu, Shuyi Zhang, Guannan Jiang, Yong Liu, Feng Zheng, Wei Zhang, Chengjie Wang, and Long Zeng. Class-aware contrastive semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2

[48] David Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *33rd annual meeting of the association for computational linguistics*, 1995. 1, 2

[49] Qing Yu, Daiki Ikami, Go Irie, and Kiyoharu Aizawa. Multi-task curriculum framework for open-set semi-supervised learning. In *European Conference on Computer Vision*, 2020. 1, 2, 3, 4, 5, 7

[50] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2016. 6

[51] Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. FlexMatch: Boosting semi-supervised learning with curriculum pseudo labeling. In *Advances in Neural Information Processing Systems*, 2021. 2, 4

[52] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018. 8

[53] Xujiang Zhao, Killamsetty Krishnateja, Rishabh Iyer, and Feng Chen. How out-of-distribution data hurts semi-supervised learning. In *2022 IEEE International Conference on Data Mining (ICDM)*, 2022. 2

[54] Mingkai Zheng, Shan You, Lang Huang, Fei Wang, Chen Qian, and Chang Xu. SimMatch: Semi-supervised learning with similarity matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2