# Multi-Task Curriculum Framework for Open-Set Semi-Supervised Learning

Qing Yu[1], Daiki Ikami[1,2], Go Irie[2], and Kiyoharu Aizawa[1]

[1] The University of Tokyo, Japan
{yu,ikami,aizawa}@hal.t.u-tokyo.ac.jp
[2] NTT Corporation, Japan
goirie@ieee.org

**Abstract.** Semi-supervised learning (SSL) has been proposed to leverage unlabeled data for training powerful models when only limited labeled data is available. While existing SSL methods assume that samples in the labeled and unlabeled data share the classes of their samples, we address a more complex novel scenario named open-set SSL, where out-of-distribution (OOD) samples are contained in unlabeled data. Instead of training an OOD detector and SSL separately, we propose a multi-task curriculum learning framework. First, to detect the OOD samples in unlabeled data, we estimate the probability of the sample belonging to OOD. We use a joint optimization framework, which updates the network parameters and the OOD score alternately. Simultaneously, to achieve high performance on the classification of in-distribution (ID) data, we select ID samples in unlabeled data having small OOD scores, and use these data with labeled data for training the deep neural networks to classify ID samples in a semi-supervised manner. We conduct several experiments, and our method achieves state-of-the-art results by successfully eliminating the effect of OOD samples.

**Keywords:** Semi-supervised learning, out-of-distribution detection, multi-task learning

## 1 Introduction

After several breakthroughs in deep learning methods, deep neural networks (DNNs) have achieved impressive results and even outperformed humans on various machine perception tasks such as image classification [8][26], face recognition [18], and natural language processing [6] with large-scale, annotated training samples. However, creating these large datasets is typically time-consuming and expensive.

To solve this problem, semi-supervised learning (SSL) is proposed to leverage unlabeled data to improve the performance of a model when only limited labeled data is available. SSL is able to train large, powerful models when labeling data is expensive or inconvenient. There is a diverse collection of approaches to SSL. For example, one approach is consistency regularization [24][15][28], which encourages a model to produce the same prediction when the input is perturbed.

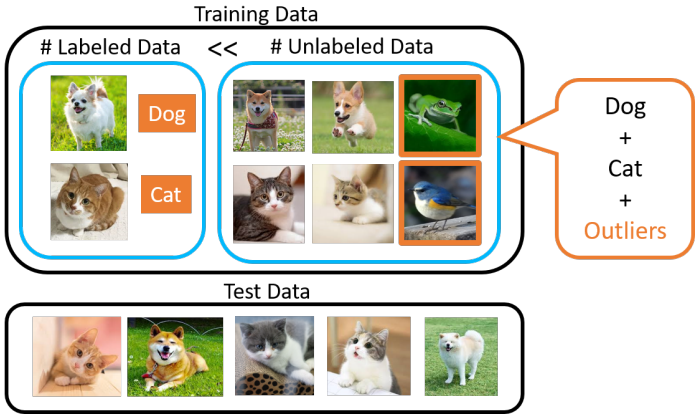


**Fig. 1.** Problem setting of open-set SSL. Outliers, which do not belong to any class of labeled data, exist in the unlabeled data.

Another approach, entropy minimization [7], encourages the model to produce high-confidence predictions. The recent state-of-the-art method, MixMatch [2], combines the aforementioned techniques in a unified loss function and achieves strong performance on a variety of image classification benchmarks.

These existing SSL methods assume that the labeled and unlabeled data have the same distribution, meaning that they share the classes of their samples, and there is no outlier sample in unlabeled data. However, in the real world, it is hard to ensure that the unlabeled data does not contain any out-of-distribution (OOD) sample that is drawn from different distributions. Oliver et al. [20] have shown that adding unlabeled data from a mismatched set of classes can actually damage the performance of SSL.

Hence, we consider a new, realistic setting called "Open-Set Semi-supervised Learning", as shown in Fig. 1. Outliers, which do not belong to the classes of labeled data, exist in the unlabeled data, and the model should be trained on labeled and unlabeled data by eliminating the effect of these outliers. To the best of our knowledge, our study is the first to tackle the problem of open-set SSL.

Although there are many algorithms for detecting OOD samples [9][16][17][29], these methods are trained on a large number of labeled in-distribution (ID) samples with class labels. In the setting of SSL, the number of labeled data is very limited. Hence, the previous methods cannot achieve high performance of detection and are not suitable for open-set SSL. Therefore, we propose a method that uses multi-task curriculum learning, which is a multi-task framework aiming to solve OOD detection and SSL simultaneously.

First, we detect OOD samples in the unlabeled data. We propose a new OOD detection method by a joint optimization framework, which can utilize the unlabeled data containing OOD data in the process of training an OOD detector. We train the network to estimate the probability of the sample belonging to

OOD. At the beginning of training, we treat all unlabeled samples as OOD and all labeled samples as ID by assigning an initial OOD score to each sample (0 for labeled data and 1 for unlabeled data). Next, we train the model to classify the sample as OOD or ID. Since unlabeled data also contains a reasonable amount of ID samples, treating all unlabeled samples as OOD samples would result in incorrect label assignments. Inspired by a solution of the noisy label problem [27], we update the network parameters and the OOD scores alternately as a joint optimization to clean the noisy OOD scores of unlabeled samples, which ranges from 0 to 1.

At the same time, while training the network for OOD detection, we also train the network to classify ID samples correctly, which forms multi-task learning. Since ID samples in the unlabeled data are expected to have smaller OOD scores than the real OOD samples, we use curriculum learning that excludes the samples with higher OOD scores in unlabeled data. Then we combine remaining ID unlabeled samples with labeled data for training the CNN to classify ID samples correctly by any SSL method, where MixMatch [2] is used in this paper.

We evaluate our method on a diverse set of open-set SSL settings. In many settings, our method outperforms existing methods by a large margin. We summarize the contributions of this paper as the following:

- We propose a novel experimental setting and training methodology for open-set SSL.
- We propose a multi-task curriculum learning framework that detects OOD samples by alternate optimization and classifies ID samples by applying SSL according to the results of OOD detection.
- We evaluate our method across several open-set SSL tasks and outperforms state-of-the-art by a considerable margin. Our approach successfully eliminates the effect of OOD samples in the unlabeled data.

## 2 Related Work

At present, there are several different methods of SSL and OOD detection. We will briefly explain some important studies in this section.

### 2.1 Semi-supervised Learning

Although there are many studies on SSL techniques, such as transductive models [10][11], graph-based methods [33] and generative modeling [13][23][25], we focus mainly on the recent state-of-the-art methods, based on consistency regularization [15][24][28].

In general supervised learning, data augmentation is a common regularization technique. In image classification, it is common to add some noise to an input image to change the pixel values of an image but keep its label [4], which means data augmentation is able to artificially increase the size of a training set by generating new modified data.

Consistency regularization is a method that applies data augmentation to SSL. It imposes a constraint in the form of regularization so that the classification result of each unlabeled sample does not change before and after augmentation.

In the simplest method, Laine and Aila [15] proposed $\pi$-model, which applies two different stochastic augmentations to an unlabeled data to generate two inputs and minimize the distance of the two network outputs of these two inputs.

Mean Teacher [28] used an exponential moving average of network parameter values to generate a more stable output on one of the two inputs, instead of generating two outputs for the two inputs by the same network, to improve the effectiveness of their method.

The state-of-the-art method for SSL is MixMatch [2], which works by guessing low-entropy labels for data-augmented unlabeled examples and mixing labeled and unlabeled data using MixUp [32]. MixMatch [2] then uses $\pi$-model to train a model using the mixed labeled and unlabeled data. We refer the reader to their paper [2] for further details.

However, these methods assume that the labeled and unlabeled data share the classes of their samples. When some OOD samples are contained in the unlabeled data, Oliver et al. [20] showed that the existing methods achieved bad performance, which is even lower than the performance of supervised learning trained only by limited labeled data in some cases. The motivation of this research is to solve this problem.

### 2.2 Out-of-distribution detection

There are also some methods for OOD detection. In the simplest method, Hendrycks & Gimpel [9] used the predicted softmax class probability to detect OOD samples. They observed that the prediction probability of incorrect and OOD samples tends to be lower than that of the correct samples. However, they also found that a pre-trained neural network can still classify some OOD samples overconfidently, which limits its performance.

To improve the effectiveness of Hendrycks & Gimpel's method [9], Liang et al. [17] applied temperature scaling and input preprocessing to detect OOD samples, called Out-of-DIstribution detector for Neural networks (ODIN). They found that the difference between the largest logit (the outputs of which are not normalized by softmax) and the remaining logits is larger for ID samples than for OOD samples if the logits are scaled by a large constant (temperature scaling). They showed that the separation of the softmax scores between the ID and OOD samples could be increased by temperature scaling. They also found that the addition of small perturbations to the input (through the loss gradient) increases the maximum predicted softmax score. As a result, the ID samples show a greater increase in score than the OOD samples. Using these techniques, their method outperformed the baseline method [9].

In another method using the predicted probability of the network, Bendale & Boult [1] calculated the score for an unknown class by taking the weighted average of all other classes obtained from a Weibull distribution, named openMax layer.

However, all the methods described earlier need a large number of labeled ID samples to achieve stable results and they are unable to utilize any unlabeled data. In open-set SSL, the number of labeled ID samples is small but we have access to a huge amount of unlabeled data containing some OOD samples. Our method aims at training a model to not only detect OOD samples with limited labeled and plenty of unlabeled data, but also achieve high recognition performance on the classification of ID samples.

## 3 Method

### 3.1 Problem Statement

We assume that an ID image-label pair, $\{\boldsymbol{x}_l, y_l\}$, drawn from a set of labeled ID images $\{X_l, Y_l\}$ , as well as an unlabeled image, $\boldsymbol{x}_{ul}$, drawn from a set of unlabeled images $X_{ul}$, is accessible. The labeled ID sample $\{\boldsymbol{x}_l, y_l\}$ can be classified into one of $K$ classes denoted by $\{c_1, \ldots, c_K\}$, meaning that $y_l \in \{c_1, \ldots, c_K\}$. Besides ID samples, outliers (OOD samples) also exist in the unlabeled data, which signifies that the true class of some unlabeled data $\boldsymbol{x}_{ul}$ is not in $\{c_1, \ldots, c_K\}$.

The goal of our method is to train a model that can correctly classify ID samples into $\{c_1, \ldots, c_K\}$ on a combination of labeled ID samples and unlabeled samples under semi-supervised setting. Our technique achieves this by distinguishing whether the image $\boldsymbol{x}_{ul}$ is from in-distribution to eliminate the negative effect of OOD samples during the training.

### 3.2 Overall Concept

The most challenging part of this task is the detection of OOD samples when only limited labeled ID samples are available. As mentioned in Section 2, traditional OOD detection methods [9, 17] assumed that there is a large number of labeled ID samples for training the recognition model and these methods did not utilize unlabeled data in training. Thus, these methods cannot achieve high performance OOD detection in SSL.

We propose a multi-task curriculum learning framework for open-set SSL, which aims to solve OOD detection and SSL simultaneously as a multi-task framework.

Since the number of labeled ID samples is limited in SSL, we use a joint optimization framework inspired by [27], which updates DNN parameters and estimates the probability of the sample belonging to OOD alternately. First, we assign an initial pseudo label representing the probability of the sample belonging to OOD, named OOD score, to all the data. For labeled samples, since they are ID, we initialize the OOD scores as 0 and for unlabeled samples, we initialize the OOD scores as 1. Since some amount of ID samples are present in unlabeled data, we can consider the binary classification of OOD as a noisy label problem. [27] showed that a DNN trained on noisy labeled datasets does not memorize

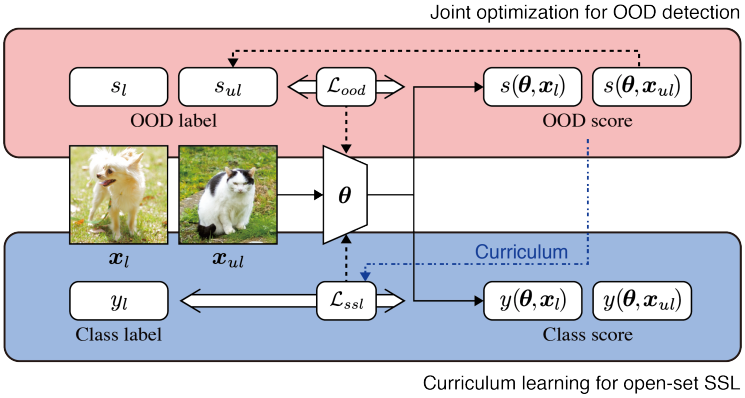


**Fig. 2.** An overview of our framework. Noisy OOD scores of unlabeled data are reassigned to the outputs of OOD scores by the DNN. The network parameters and OOD scores are alternately updated for each epoch. The unlabeled samples used to calculate semi-supervised loss are selected by their OOD scores.

noisy labels under a high learning rate. Thus, the noisy label of a sample can be corrected by reassigning the probability output of the DNN to the sample as a new label. We utilize this property to increase the number of ID samples during the training by cleaning the OOD scores of unlabeled data. To achieve this, the network parameters of the DNN and the OOD scores of unlabeled samples are alternately updated for each epoch and the OOD scores of unlabeled samples are reassigned to the estimation of OOD scores by DNN.

At the same time, to achieve high performance on the classification of ID samples, we select ID samples in unlabeled data having low OOD scores and combine them with labeled ID samples for training the DNN to classify ID samples correctly in a semi-supervised manner. Although our method can be applied to any SSL method, we choose the state-of-the-art SSL method, MixMatch [2], in this paper.

Instead of training OOD detection and ID classification separately, we further propose a multi-task training framework combining all the steps to formulate an end-to-end trainable network that can detect OOD samples and classify ID samples simultaneously. The overview of our framework is shown in Fig. 2.

### 3.3 Training Procedure

**Noisy Label Optimization for OOD Detection** First, we assign an initial OOD score $s$ to all the data, which makes an ID image-label pair, $\{\boldsymbol{x}_l, y_l\} \in \{X_l, Y_l\}$ become $\{\boldsymbol{x}_l, y_l, s_l\} \in \{X_l, Y_l, S_l\}$ , and an unlabeled image, $\boldsymbol{x}_{ul} \in X_{ul}$ become $\{\boldsymbol{x}_{ul}, s_{ul}\} \in \{X_{ul}, S_{ul}\}$. We initialize the OOD scores $S_l$ as 0 for the labeled (ID) samples, whereas the unlabeled samples are assigned the OOD score $S_{ul}$ of 1. This is denoted by:

$$s_l = 0 \, (\forall s_l \in S_l), \tag{1}$$

**Algorithm 1** Joint Optimization

**for** $t \leftarrow 0$ to $E$ **do**
  update $\boldsymbol{\theta}^{(t+1)}$ by Adam on $\mathcal{L}_{ood}(\boldsymbol{\theta}^t, S_{ul}^t | X_l, S_l, X_{ul})$
  update $S_{ul}^{(t+1)}$ by Eq. (6)
**end for**

$$s_{ul} = 1 \ (\forall s_{ul} \in S_{ul}). \tag{2}$$

As a binary classification of OOD, parameters of the network $\boldsymbol{\theta}$ can be optimized as follows:

$$\min_{\boldsymbol{\theta}} \mathcal{L}_{ood}(\boldsymbol{\theta} | X_l, S_l, X_{ul}, S_{ul}), \tag{3}$$

$$\begin{aligned} \mathcal{L}_{ood} = &-\frac{1}{|X_l|} \sum_{i=1}^{|X_l|} (s_{l_i} \log s(\boldsymbol{\theta}, \boldsymbol{x}_{l_i}) + (1 - s_{l_i}) \log(1 - s(\boldsymbol{\theta}, \boldsymbol{x}_{l_i}))) \\ &-\frac{1}{|X_{ul}|} \sum_{i=1}^{|X_{ul}|} (s_{ul_i} \log s(\boldsymbol{\theta}, \boldsymbol{x}_{ul_i}) + (1 - s_{ul_i}) \log(1 - s(\boldsymbol{\theta}, \boldsymbol{x}_{ul_i}))), \end{aligned} \tag{4}$$

where $\mathcal{L}_{ood}$ denotes the cross entropy loss under the supervision of OOD score and $s(\boldsymbol{\theta}, \boldsymbol{x}_{l_i})$ (or $s(\boldsymbol{\theta}, \boldsymbol{x}_{ul_i})$) denotes the predicted OOD score of the image $\boldsymbol{x}_{l_i}$ (or $\boldsymbol{x}_{ul_i}$) with the network parameters being $\boldsymbol{\theta}$.

However, although the OOD scores $S_{ul}$ are initialized to 1 for all unlabeled samples, the existence of some ID samples in the unlabeled data leads to the classification of OOD as a noisy label problem. Tanaka et al. [27] showed that a network trained with a high learning rate is less likely to overfit to noisy labels, which means the loss Eq. (4) is high for noisy labels and low for clean labels. So we obtain clean OOD scores by updating the OOD scores in the direction to decrease Eq. (4). Hence, we formulate the problem as the joint optimization of the network parameters and OOD scores as follows:

$$\min_{\boldsymbol{\theta}, S_{ul}} \mathcal{L}_{ood}(\boldsymbol{\theta}, S_{ul} | X_l, S_l, X_{ul}), \tag{5}$$

Alternately updating the network parameters $\boldsymbol{\theta}$ and OOD scores of unlabeled data $S_{ul}$ is achieved via joint optimization [27] by repeating the following two steps:

*Updating $\boldsymbol{\theta}$ with fixed $S_{ul}$:* Since all terms in the loss function Eq. (4) are differentiable with respect to $\boldsymbol{\theta}$, we update $\boldsymbol{\theta}$ by the Adam optimizer [12] on Eq. (4).

*Updating $S_{ul}$ with fixed $\boldsymbol{\theta}$:* Considering the update of $S_{ul}$, we need to minimize $\mathcal{L}_{ood}$ with fixed $\boldsymbol{\theta}$ to correct OOD scores. $\mathcal{L}_{ood}$ can be minimized when the predicted OOD scores of the network equals $S_{ul}$. As a result, $S_{ul}$ is updated as follows:

$$s_{ul_i} \leftarrow s(\boldsymbol{\theta}, \boldsymbol{x}_{ul_i}). \tag{6}$$

The whole algorithm of joint optimization is shown in Algorithm 1.

**Algorithm 2** Multi-task curriculum learning

**for** $t \leftarrow 0$ to $E$ **do**
  Select ID samples in unlabeled data $X_{ul}^{id}$ by Eq. (9)
  update $\boldsymbol{\theta}^{(t+1)}$ by Adam on $\mathcal{L}_{ssl}(\boldsymbol{\theta}^t|X_l, Y_l, X_{ul}^{id}) + \mathcal{L}_{ood}(\boldsymbol{\theta}^t, S_{ul}^t|X_l, S_l, X_{ul})$
  update $S_{ul}^{(t+1)}$ by Eq. (6)
**end for**

**Multi-task Curriculum Learning for Open-set SSL** As general SSL, the optimization problem of the network parameters $\boldsymbol{\theta}$ is formulated as follows:

$$\min_{\boldsymbol{\theta}} \mathcal{L}_{ssl}(\boldsymbol{\theta}|X_l, Y_l, X_{ul}), \tag{7}$$

where $\mathcal{L}_{ssl}$ denotes a loss function such as the sum of cross-entropy loss on labeled data and L2 loss on unlabeled data in [15].

During the training of the network for OOD detection, we also train the network to classify ID data by SSL. As a result, our method is a multi-task learning problem formulated as follows:

$$\min_{\boldsymbol{\theta}, S_{ul}} \mathcal{L}_{ssl}(\boldsymbol{\theta}|X_l, Y_l, X_{ul}) + \mathcal{L}_{ood}(\boldsymbol{\theta}, S_{ul}|X_l, S_l, X_{ul}). \tag{8}$$

The inclusion of the training for ID data classification is helpful to OOD detection because the network can learn more discriminative features. However, while optimizing $\boldsymbol{\theta}$ on the semi-supervised part $\mathcal{L}_{ssl}(\boldsymbol{\theta}^t|X_l, Y_l, X_{ul})$ in Eq. (8), the existence of OOD samples in $X_{ul}$ is detrimental to the training for the classification of ID samples. This problem is solved by using curriculum learning that picks up ID samples $X_{ul}^{id}$ from unlabeled data $X_{ul}$ according to the OOD scores of unlabeled data $S_{ul}$.

Although we can simply sample top $\eta\%$ samples from $X_{ul}$ in ascending order of OOD scores $S_{ul}$, we implement Otsu thresholding [21] to decide the threshold $th_{otsu}$ automatically, which reduces one hyper-parameter. Then, the selected unlabeled samples are denoted as follows:

$$X_{ul}^{id} = \{\boldsymbol{x}_{ul_i}|s(\boldsymbol{\theta}, \boldsymbol{x}_{ul_i}) < th_{otsu}, 1 \leq i \leq N\}, \tag{9}$$

which converts the total loss function to:

$$\min_{\boldsymbol{\theta}, S_{ul}} \mathcal{L}_{ssl}(\boldsymbol{\theta}|X_l, Y_l, X_{ul}^{id}) + \mathcal{L}_{ood}(\boldsymbol{\theta}, S_{ul}|X_l, S_l, X_{ul}). \tag{10}$$

The entire algorithm of our method is as shown in Algorithm 2. It is to be noted that the semi-supervised loss $\mathcal{L}_{ssl}(\boldsymbol{\theta}^t|X_l, Y_l, X'_{ul})$ is a general loss of SSL, which implies that our method is applicable to any SSL method. In this paper, we use MixMatch [2] as SSL.

## 4 Experiments

In this section, we discuss our experimental settings and results. We demonstrate the effectiveness of our method on a diverse set of in- and out-of-distribution

dataset pairs for open-set SSL. We found that our method outperformed the current state-of-the-art methods by a considerable margin. We used PyTorch 1.1.0 [22] to run all the experiments.

### 4.1 Neural Network Architecture

Following [2][20], we implemented our network based on Wide ResNet (WRN) [31]. We first trained the model only on the OOD classification loss for 100 epochs and the update of the OOD scores was set to start at the 10th epoch, to achieve stable performance. The model was then trained on the total loss function in Algorithm 2 for 1,024 epochs and 1,024 iterations of each epoch, which is the same as [2]. We used the Adam [12] optimizer and the learning rate was set as 0.002. 64 samples each from the labeled and unlabeled data are sampled for a batch. We report the average test accuracy of the last 10 checkpoints.

### 4.2 In-Distribution Datasets

CIFAR-10 [14] and SVHN [19] (each containing 10 classes) were used as in-distribution datasets. A total of 5,000 samples were split from the original training data as validation data, and all original test samples were used for testing. We further split the remaining training samples (45,000 for CIFAR-10 and 68,257 for SVHN) into labeled and unlabeled data. Following [20][2], we used {250, 1000, 4000} samples as labeled data and remaining samples as unlabeled data.

### 4.3 Out-of-Distribution Datasets

As the OOD data are mixed in the unlabeled data, we added 10,000 samples from the following four datasets for each setting:

1. **TinyImageNet (TIN).** The Tiny ImageNet dataset [5] contains 10,000 test images from 200 different classes, which are drawn from the original 1,000 classes of ImageNet [5]. All samples are downsampled by resizing the original image to a size of $32 \times 32$.
2. **LSUN.** The Large-scale Scene Understanding dataset (LSUN) consists of 10,000 test images from 10 different scene categories.[30]. Similar to TinyImageNet, all samples are downsampled by resizing the original image to a size of $32 \times 32$.
3. **Gaussian.** The synthetic Gaussian noise dataset contains 10,000 random 2D Gaussian noise images, where each RGB value of every pixel is sampled from an independent and identically distributed Gaussian distribution with mean 0.5 and unit variance. We further clip each pixel value into the range $[0, 1]$.
4. **Uniform.** The synthetic uniform noise dataset contains 10,000 images where each RGB value of every pixel is independently and identically sampled from a uniform distribution on $[0, 1]$.

Examples of these datasets are shown in Fig. 3. The experimental setting is summarized in Table 1.

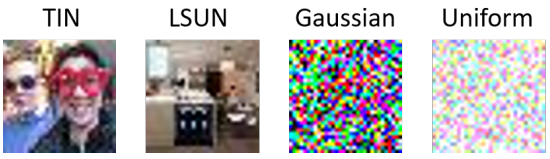


**Fig. 3.** Examples of Out-of-Distribution Datasets. The samples from TinyImageNet and LSUN are resized to $32 \times 32$.

**Table 1.** The number and type of labeled and unlabeled samples in the experimental setting.

**CIFAR-10**

| #Labeled | #Unlabeled | #Unlabeled Outlier | #Valid | #Test |
|---|---|---|---|---|
| 250 | 44,750 | 10,000 | 5,000 | 10,000 |
| 1,000 | 44,000 | | | |
| 4,000 | 41,000 | | | |

**SVHN**

| #Labeled | #Unlabeled | #Unlabeled Outlier | #Valid | #Test |
|---|---|---|---|---|
| 250 | 68,007 | 10,000 | 5,000 | 26,032 |
| 1,000 | 67,257 | | | |
| 4,000 | 64,257 | | | |

### 4.4 Results

The results for the CIFAR-10 dataset are summarized in the upper part of Table 2, which shows the comparison of our method and the baseline. We used the original MixMatch [2] without any OOD detection as the baseline method. Table 2 clearly shows that our approach significantly outperforms the baseline by eliminating the effect of OOD samples in unlabeled data. Compared to TIN and LSUN, which are natural images, synthetic datasets (Gaussian and Uniform) are more harmful to the performance of SSL. Our technique has successfully enabled SSL methods to achieve stable performance on these outliers by detecting them.

In Fig. 4, we show test accuracy vs. the number of epochs. We observe that our method continuously improves the performance of the model during the latter half of the training process and its performance is more stable compared to the baseline method during the training. At the beginning of the training process, our method is observed to converge slower than the baseline method. We consider this to be due to the multi-task learning, where our method also learns to detect OOD samples.

The lower part of Table 2 shows the comparison of our method and the baseline for the SVHN dataset. Compared to the case where CIFAR-10 is used as the ID dataset, the effect of outliers on the model is much smaller in this case, possibly because the classification of SVHN is a comparatively easier task. In this

**Table 2.** Accuracy (%) for CIFAR-10/SVHN and OOD dataset pairs. We report the averages and the standard deviations of the scores obtained from three trials. Bold values represent the highest accuracy in each setting. *Clean* shows the upper limit of the model when the unlabeled data contains no OOD data.

**CIFAR-10**

| OOD dataset | 250 labeled | | 1000 labeled | | 4000 labeled | |
|---|---|---|---|---|---|---|
| | Baseline | Ours | Baseline | Ours | Baseline | Ours |
| TIN | 82.42 ± 0.70 | **86.44 ± 0.64** | 88.03 ± 0.22 | **89.85 ± 0.11** | 91.25 ± 0.13 | **93.03 ± 0.05** |
| LSUN | 76.32 ± 4.19 | **86.65 ± 0.41** | 87.03 ± 0.41 | **90.19 ± 0.47** | 91.18 ± 0.33 | **92.91 ± 0.03** |
| Gaussian | 75.76 ± 3.49 | **87.34 ± 0.13** | 85.71 ± 1.14 | **89.80 ± 0.26** | 91.51 ± 0.35 | **92.53 ± 0.08** |
| Uniform | 72.90 ± 0.96 | **85.54 ± 0.11** | 84.49 ± 1.06 | **89.87 ± 0.08** | 90.47 ± 0.38 | **92.83 ± 0.04** |
| Clean | 87.65 ± 0.29 | | 90.67 ± 0.29 | | 93.30 ± 0.10 | |

**SVHN**

| OOD dataset | 250 labeled | | 1000 labeled | | 4000 labeled | |
|---|---|---|---|---|---|---|
| | Baseline | Ours | Baseline | Ours | Baseline | Ours |
| TIN | 94.66 ± 0.14 | **95.21 ± 0.27** | 95.58 ± 0.38 | **96.65 ± 0.14** | 96.73 ± 0.05 | **97.01 ± 0.03** |
| LSUN | 94.98 ± 0.23 | **95.40 ± 0.17** | 95.46 ± 0.05 | **96.51 ± 0.16** | 96.75 ± 0.01 | **97.15 ± 0.02** |
| Gaussian | 93.42 ± 1.09 | **95.23 ± 0.04** | 95.85 ± 0.33 | **96.50 ± 0.11** | 96.97 ± 0.02 | **97.07 ± 0.07** |
| Uniform | 94.78 ± 0.25 | **95.07 ± 0.12** | 95.62 ± 0.50 | **96.47 ± 0.24** | 96.86 ± 0.12 | **97.04 ± 0.02** |
| Clean | 96.04 ± 0.39 | | 96.84 ± 0.06 | | 97.23 ± 0.05 | |

situation, our method continues to exhibit a higher and more stable performance than the baseline.

We also studied the performance of OOD detection by the proposed method. Since we use the threshold calculated by Otsu thresholding to select ID samples in unlabeled data, we evaluate the performance of OOD detection by precision and recall. Precision is calculated by the percentage of ID samples in the selected samples by curriculum learning. Recall is calculated by the percentage of selected ID samples among all ID samples in unlabeled data. Higher precision and recall indicate better OOD detection – the precision and recall of a perfect detector are both 1. Table 3 shows the results. We find that our method achieves high precision and recall in all the cases, indicating that our method successfully in selecting ID samples from unlabeled data for semi-supervised training.

### 4.5 Ablation Studies

We further analyzed the effects of the following factors:

**The number of OOD samples in the unlabeled data.** We used LSUN as OOD and we changed the number of OOD samples in $X_{ul}$. The result is summarized in Table 4, which shows that our proposed method works under different OOD conditions except for a case with few outliers in unlabeled data. When there are more OOD samples in the unlabeled data, the performance of the baseline model is lower while our method can achieve stable performance.

**The performance of OOD detection compared to existing OOD detection methods.** As mentioned in Section 2, the existing OOD detection methods cannot achieve high performance when the labeled data is limited and

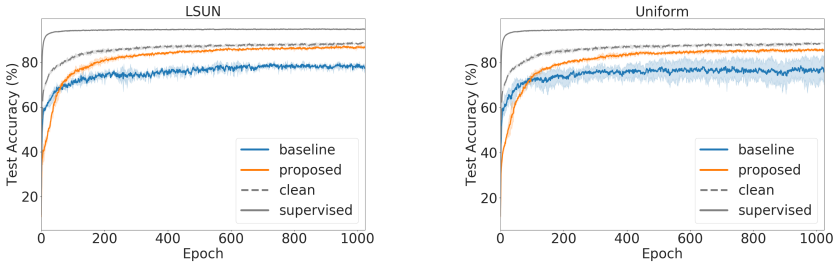


**Fig. 4.** Test accuracy vs. the number of epochs using CIFAR-10 as ID and other datasets as OOD when 250 labeled samples are used. *Clean* shows the upper limit of the model when the unlabeled data contains no OOD data and *supervised* shows the performance when all the samples are labeled and no OOD data is contained.

**Table 3.** Performance (%) of OOD detection by our proposed method. We report the averages of the scores obtained from three trials.

**CIFAR-10**

| OOD dataset | 250 labeled | | 1000 labeled | | 4000 labeled | |
|---|---|---|---|---|---|---|
| | Recall | Precision | Recall | Precision | Recall | Precision |
| TIN | 99.22 | 98.48 | 99.48 | 97.11 | 100.00 | 99.30 |
| LSUN | 99.48 | 99.38 | 99.95 | 98.95 | 100.00 | 99.64 |
| Gaussian | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| Uniform | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |

**SVHN**

| OOD dataset | 250 labeled | | 1000 labeled | | 4000 labeled | |
|---|---|---|---|---|---|---|
| | Recall | Precision | Recall | Precision | Recall | Precision |
| TIN | 84.28 | 99.93 | 98.4 | 99.83 | 99.59 | 99.87 |
| LSUN | 88.55 | 99.98 | 98.28 | 99.97 | 99.70 | 99.98 |
| Gaussian | 87.52 | 100.00 | 99.28 | 100.00 | 99.76 | 100.00 |
| Uniform | 82.67 | 100.00 | 99.21 | 100.00 | 99.78 | 100.00 |

the unlabeled data cannot be utilized in training. We show the results of applying the existing OOD detection method in the setting of open-set semi-supervised learning in Table 5. We choose the most challenging cases when only 250 labeled samples are available. We report the AUROC (the area under the false positive rate against the true positive rate curve) of the OOD detection score of each method. The AUROC of a perfect detector is 1. Table 5 shows that our approach significantly outperforms other OOD detection methods. It is also interesting that the AUROC of previous methods is less than 50% in some cases, which means the model shows more confidence on the predictions of OOD samples than those of ID samples (and this observation conflicts with the idea of [9][17]).

**The performance of Otsu thresholding.** We use Otsu thresholding to calculate the threshold for splitting ID and OOD samples for the curriculum learning of semi-supervised learning. Fig. 5 shows the histogram of OOD scores

**Table 4.** Accuracy (%) for CIFAR-10 as ID and LSUN as OOD on different numbers of OOD samples when 250 labeled samples are used.

| #OOD samples | 2000 | | 5000 | | 10000 | | 20000 | |
|---|---|---|---|---|---|---|---|---|
| | Baseline | Ours | Baseline | Ours | Baseline | Ours | Baseline | Ours |
| Accuracy (%) | **88.16** | 84.82 | 82.98 | **86.20** | 76.32 | **86.65** | 70.3 | **85.83** |

**Table 5.** The comparison of AUROC (%) in the task of OOD Detection.

| ID dataset | OOD dataset | Hendrycks & Gimpel [9] | ODIN [17] | Ours |
|---|---|---|---|---|
| CIFAR-10 | TIN | 50.92 | 54.54 | **98.86** |
| | LSUN | 54.34 | 58.02 | **99.82** |
| | Gaussian | 32.41 | 37.49 | **100.00** |
| | Uniform | 45.43 | 51.05 | **100.00** |
| SVHN | TIN | 50.48 | 57.09 | **99.57** |
| | LSUN | 51.44 | 53.68 | **99.84** |
| | Gaussian | 21.20 | 1.87 | **99.98** |
| | Uniform | 2.79 | 8.31 | **99.97** |

and the threshold calculated using Otsu thresholding. We find that both ID and OOD samples can be successfully separated by the threshold.

### 4.6 Discussion

**Limitations.** As shown in Table 4, our method fails to improve the baseline if there are few outliers in the unlabeled data. This failure mainly comes from the wrong threshold calculated by Otsu thresholding, since the number of ID samples and OOD samples is extremely imbalanced. This problem can be solved by changing the OOD threshold, which means we can introduce a new parameter to control the number of unlabeled samples selected as ID data.

**"Similar" outliers.** The outlier datasets used in Section 4 are still quite different from the original training datasets CIFAR-10 and SVHN. Including similar outliers in the unlabeled data is a more complicated scenario. We tried using the animal classes in CIFAR-10 as ID and other classes as OOD and found that these similar outliers are not as harmful as dissimilar outliers, which leads to a 3% decrease in test accuracy. Our method can still reach 2% higher than the test accuracy of the baseline with 94% precision and 98% recall of OOD detection.

**Additional comparisons.** Chen et al. [3] works on a close setting to our paper at around the same time as our paper, but there are two significant differences between [3] and our work. First, the experimental setting is different. [3] defines the mismatch of class distribution as some known classes are not contained in the unlabeled dataset and some unknown classes are contained in the unlabeled dataset. Second, the method of utilizing OOD samples in unlabeled data is different. In [3], the OOD samples are simply ignored by filtering the

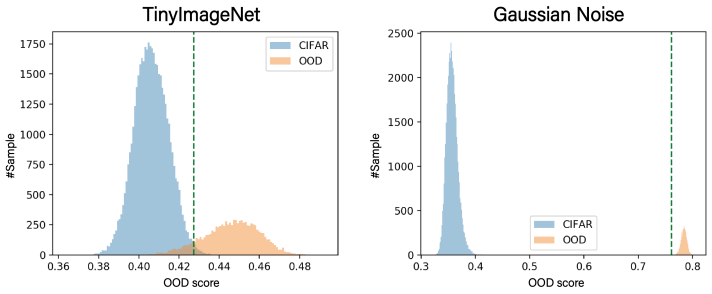


**Fig. 5.** The histogram of OOD scores and the threshold (the green line) calculated by Otsu thresholding when 250 labeled samples are used.

**Table 6.** Performance (%) of UASD [3] and our propsoed method.

| OOD dataset | Test Accuracy | | | Detection Recall | | Detection Precision | |
|---|---|---|---|---|---|---|---|
| | Baseline | UASD [3] | Ours | UASD [3] | Ours | UASD [3] | Ours |
| TIN | 82.42 | 83.53 | **86.44** | 66.47 | **99.22** | 96.84 | **98.48** |
| LSUN | 76.32 | 80.87 | **86.65** | 63.88 | **99.48** | 96.50 | **99.38** |

confidence score, which means they mainly train an SSL model and just use the output of the model directly.

For the comparison, we implemented [3] for filtering OOD in MixMatch and show the comparison in the setting of CIFAR-10 with 250 labeled samples. The results are summarized in Table 6 and it shows our method has better performance not only in SSL but also in OOD detection, because we explicitly train an OOD detector by unlabeled OOD samples together with the training of SSL model, which enables our method to achieve higher OOD detection performance.

## 5 Conclusion

In this paper, we have proposed a multi-task curriculum for open-set SSL, where the labeled data is limited and the unlabeled data contains some OOD samples. To detect these OOD samples, our method utilizes joint optimization framework to estimate the probability of the unlabeled sample belonging to OOD, which is achieved by updating the network parameters and the OOD score alternately. Simultaneously, we use curriculum learning to exclude these OOD samples from semi-supervised learning and utilize these data with labeled data for training the model to classify ID samples with high performance. We evaluated our method on several open-set semi-supervised benchmarks and proved that our method achieves state-of-the-art performance by detecting the OOD samples with high accuracy.

**ACKNOWLEDGEMENTS** This work was supported by JSPS KAKENHI Grant Number 18H03254 and JST CREST Grant Number JPMJCR1686, Japan.

## References


1. Bendale, A., Boult, T.E.: Towards open set deep networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition (2016)
2. Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., Raffel, C.: Mixmatch: A holistic approach to semi-supervised learning. In: Advances in Neural Information Processing Systems (2019)
3. Chen, Y., Zhu, X., Li, W., Gong, S.: Semi-supervised learning under class distribution mismatch. In: AAAI Conference on Artificial Intelligence (2020)
4. Cubuk, E.D., Zoph, B., Mane, D., Vasudevan, V., Le, Q.V.: Autoaugment: Learning augmentation policies from data. In: Proceedings of the IEEE conference on computer vision and pattern recognition (2019)
5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: Proceedings of the IEEE conference on computer vision and pattern recognition (2009)
6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (2019)
7. Grandvalet, Y., Bengio, Y.: Semi-supervised learning by entropy minimization. In: Advances in Neural Information Processing Systems (2005)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition (2016)
9. Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. In: International Conference on Learning Representations (2017)
10. Joachims, T.: Transductive inference for text classification using support vector machines. In: International Conference on Machine Learning (1999)
11. Joachims, T.: Transductive learning via spectral graph partitioning. In: International Conference on Machine Learning (2003)
12. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: International Conference on Learning Representations (2014)
13. Kingma, D.P., Mohamed, S., Rezende, D.J., Welling, M.: Semi-supervised learning with deep generative models. In: Advances in Neural Information Processing Systems (2014)
14. Krizhevsky, A.: Learning multiple layers of features from tiny images. Tech. rep. (2009)
15. Laine, S., Aila, T.: Temporal ensembling for semi-supervised learning. In: International Conference on Learning Representations (2017)
16. Lee, K., Lee, H., Lee, K., Shin, J.: Training confidence-calibrated classifiers for detecting out-of-distribution samples. In: International Conference on Learning Representations (2018)
17. Liang, S., Li, Y., Srikant, R.: Enhancing the reliability of out-of-distribution image detection in neural networks. In: International Conference on Learning Representations (2018)
18. Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., Song, L.: Sphereface: Deep hypersphere embedding for face recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition (2017)

19. Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., Ng, A.Y.: Reading digits in natural images with unsupervised feature learning. In: NeurIPS Workshop on Deep Learning and Unsupervised Feature Learning (2011)
20. Oliver, A., Odena, A., Raffel, C.A., Cubuk, E.D., Goodfellow, I.: Realistic evaluation of deep semi-supervised learning algorithms. In: Advances in Neural Information Processing Systems (2018)
21. Otsu, N.: A threshold selection method from gray-level histograms. IEEE transactions on systems, man, and cybernetics (1979)
22. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems (2019)
23. Pu, Y., Gan, Z., Henao, R., Yuan, X., Li, C., Stevens, A., Carin, L.: Variational autoencoder for deep learning of images, labels and captions. In: Advances in Neural Information Processing Systems (2016)
24. Sajjadi, M., Javanmardi, M., Tasdizen, T.: Regularization with stochastic transformations and perturbations for deep semi-supervised learning. In: Advances in Neural Information Processing Systems (2016)
25. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans. In: Advances in Neural Information Processing Systems (2016)
26. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning (2019)
27. Tanaka, D., Ikami, D., Yamasaki, T., Aizawa, K.: Joint optimization framework for learning with noisy labels. In: Proceedings of the IEEE conference on computer vision and pattern recognition (2018)
28. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In: Advances in Neural Information Processing Systems (2017)
29. Vyas, A., Jammalamadaka, N., Zhu, X., Das, D., Kaul, B., Willke, T.L.: Out-of-distribution detection using an ensemble of self supervised leave-out classifiers. In: Proceedings of the European Conference on Computer Vision (2018)
30. Yu, F., Seff, A., Zhang, Y., Song, S., Funkhouser, T., Xiao, J.: Construction of a large-scale image dataset using deep learning with humans in the loop. arXiv preprint arXiv: 1506.03365 (2015)
31. Zagoruyko, S., Komodakis, N.: Wide residual networks. In: Proceedings of the British Machine Vision Conference (2016)
32. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. In: International Conference on Learning Representations (2018)
33. Zhu, X., Ghahramani, Z., Lafferty, J.D.: Semi-supervised learning using gaussian fields and harmonic functions. In: International Conference on Machine Learning (2003)