Trash to Treasure: Harvesting OOD Data with Cross-Modal Matching for Open-Set Semi-Supervised Learning

Junkai Huang[1*] Chaowei Fang[2*] Weikai Chen[3] Zhenhua Chai[4]
Xiaolin Wei[4] Pengxu Wei[1] Liang Lin[1] Guanbin Li[1†]
[1]Sun Yat-sen University [2]Xidian University [3]Tencent America [4]Meituan

Abstract

Open-set semi-supervised learning (open-set SSL) investigates a challenging but practical scenario where out-of-distribution (OOD) samples are contained in the unlabeled data. While the mainstream technique seeks to completely filter out the OOD samples for semi-supervised learning (SSL), we propose a novel training mechanism that could effectively exploit the presence of OOD data for enhanced feature learning while avoiding its adverse impact on the SSL. We achieve this goal by first introducing a warm-up training that leverages all the unlabeled data, including both the in-distribution (ID) and OOD samples. Specifically, we perform a pretext task that enforces our feature extractor to obtain a high-level semantic understanding of the training images, leading to more discriminative features that can benefit the downstream tasks. Since the OOD samples are inevitably detrimental to SSL, we propose a novel cross-modal matching strategy to detect OOD samples. Instead of directly applying binary classification [39], we train the network to predict whether the data sample is matched to an assigned one-hot class label. The appeal of the proposed cross-modal matching over binary classification is the ability to generate a compatible feature space that aligns with the core classification task. Extensive experiments show that our approach substantially lifts the performance on open-set SSL and outperforms the state-of-the-art by a large margin.

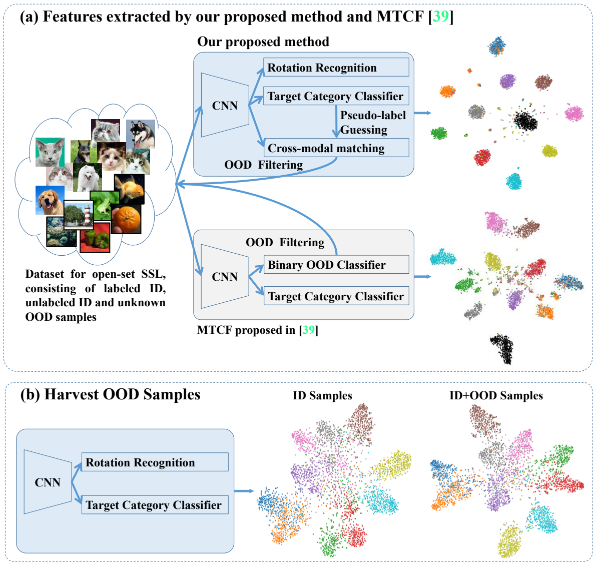

(a) Features extracted by our proposed method and MTCF [39]
Our proposed method
CNN
Rotation Recognition
Target Category Classifier
Pseudo-label Guessing
Cross-modal matching
OOD Filtering
Dataset for open-set SSL, consisting of labeled ID, unlabeled ID and unknown OOD samples
OOD Filtering
CNN
Binary OOD Classifier
Target Category Classifier
MTCF proposed in [39]
(b) Harvest OOD Samples
ID Samples
ID+OOD Samples
CNN
Rotation Recognition
Target Category Classifier

Figure 1: t-SNE [26] visualization of image features extracted from CIFAR100 [20]. Images of the same category are shown in the same color. (a) Features learned by our method are more compact and cleaner than those learned by MTCF [39]. (b) By leveraging OOD samples in the proposed pretext training, we have achieved more discriminative feature space than using ID samples only.

1. Introduction

> *"One man's trash is another man's treasure."*
>
> *– Hector Urquhart*

Semi-supervised learning (SSL) provides an effective way of leveraging massive unlabeled data to improve the performance of deep neural network when only limited labeled samples are available. Most existing SSL methods assume that labeled and unlabeled data share the same category space. However, this assumption is difficult to satisfy since it still requires tedious efforts to confirm the purity of the unlabeled data. Very recently, Yu *et al.* [39] proposed a more realistic setting called open-set semi-supervised learning (open-set SSL). Open-set SSL considers a more challenging but practical scenario where outliers, that do not belong to the categories of the labeled data, may exist in the unlabeled data. Resolving the open-set SSL problem has crucial practical meanings as it can significantly reduce the workload of data preparation in the actual applications.

A straightforward approach to cope with the out-of-distribution (OOD) samples is to completely remove them

[*] Equal contribution.

[†] Corresponding author.

from the SSL training, as the prior works [28] have shown that including OOD unlabeled data can severely impact the performance of SSL. While there exist a diverse collection of approaches for OOD sample detection, they typically require a large corpus of in-distribution (ID) data with class labels. However, due to the scarcity of the labeled data in SSL, the existing OOD detection methods would fail to achieve satisfactory performance and are hence not suitable to be deployed in the open-set SSL.

In order to eliminate the influence of OOD samples, [39] devises a multi-task curriculum framework (MTCF) with a binary OOD classification head that strives to filter out all the OOD samples. The classification of in-distribution (ID) samples and the detection of OOD samples are unified into a joint optimization framework, where unlabeled samples with lower OOD scores will be gradually added for semi-supervised training. However, the proposed binary OOD classification task and the ID classification have conflicting goals in terms of feature learning. Specifically, the training of OOD detection aims to cluster all ID samples (regardless of their categories) into one category (i.e., ID data) while the task of ID classification tends to enhance the category discrimination between ID samples. Unifying the contradictory optimization goals into one framework that shares a backbone network could compromise the final performance and increase the difficulty of training.

In this paper, we present a novel training framework for open-set SSL that can effectively exploit the presence of OOD data for enhanced feature learning while avoiding its adverse impact to the SSL. First, instead of completely discarding the OOD data, we introduce a warm-up training that makes full use of all the unlabeled data, including the OOD samples, to enhance the representation learning of our backbone network. Unlike the conventional pre-training, our warm-up training performs a pretext task that deviates from the target application. In particular, we ask the network to predict the rotation of the rotationally augmented data in a self-supervised manner. This enforces our backbone model to obtain high-level semantic understanding of the images and hence leads to more discriminative features that could benefit the downstream applications, e.g. the classification task. Particularly, as shown in Figure 1(b), OOD samples, which are outliers in SSL algorithms, turn out to be treasures that can enhance feature learning when fully utilized in the proposed self-supervised pretext training. The idea of leveraging self-supervised techniques for boosting the performance of semi-supervised learning has been shown effective in the previous work [41]. However, it is only verified in the traditional SSL, where the unlabeled data share the same category space with the labeled ones. We are the first to investigate this idea in the open-set setting and show that the self-supervised auxiliary task could be beneficial to open-set SSL with properly designed training strategy.

Second, we propose a more effective approach for detecting and filtering OOD samples based on a novel cross-modal matching mechanism. First of all, each unlabeled sample is assigned the category with the highest predicted probability of the model as a pseudo label. We then propose a cross-modal matching head to infer whether the embedding of the image and its pseudo-label are matched. Once trained, OOD samples can be screened out due to its low confidence with all the ID categories. Unlike the binary classification based OOD detection [39], the feature learning of cross-modal matching aligns well with that of the target ID classification task, as both strive to achieve better discrimination between the image features of different categories. We show in Figure 1(a) that our method can obtain features with much more compact and purified clusters than that of [39]. Furthermore, we can effectively detect ID samples with incorrect pseudo labels, also coded “hard” samples, via cross-modal matching. This helps to further improve the performance of the trained model since hard samples could harm the model training especially at the early stage when the prediction accuracy of pseudo labels is relatively low. We propose an adaptive training mechanism which gradually involves more hard samples as the model proceeds to achieve better performance.

Our proposed approach is a general training framework that can be easily implemented into existing SSL methods. We show that our method greatly improves the state-of-the-art performance in extensive open-set semi-supervised image recognition benchmarks including CIFAR-10 [20], Animals-10, CIFAR-100, and TinyImageNet [22]. We summarize our contributions as follows:

- A novel training pipeline for open-set SSL that leverages the presence of OOD samples for enhanced feature learning while avoiding their adverse impact.
- A specially tailored warm-up training method that uses self-supervised learning to boost the performance of open-set SSL.
- A novel OOD and hard sample detection algorithm based on cross-modal matching, which achieves compatible feature space with the target classification task.
- New state-of-the-art performance on open-set SSL over extensive benchmarks including CIFAR-10, CIFAR-100, TinyImageNet, and Animals-10.

2. Related Works

Semi-Supervised Learning. Semi-supervised learning (SSL) generally refers to a series of methods that aim to use both labeled and unlabeled data from the similar distribution for model training. There are a vast number of classic works on SSL across various disciplines [16, 5, 17]. Among them, semi-supervised image classification has been a long-term and extensive research topic. The most classic solution for the SSL problem is self-training, which iteratively enlarges the labeled set via guessing labels of unknown samples. We refer interested readers to [35] for a detailed survey. Other techniques including co-training [3], label propagation [31] and graphical model are also widely used in this task [43, 36, 1, 13].

Benefiting from deep learning, breakthrough achievements have been achieved in SSL. Traditional semi-supervised techniques are re-implemented with deep CNN, such as self-labeling [35], multi-view training [30, 8], label propagation [15] and graph-based method [18]. These methods focus on assigning pseudo hard/soft labels to unknown samples or clustering samples with similar semantics. Considering the sample density is low nearby the decision boundary, a training sample is assumed to share the same label with a synthesized sample close to it. Motivated by this intuition, consistency regularization can be exerted to spread labels of known samples and confident predictions of unknown samples. In [21], two models including π-model and temporal ensembling model are proposed to regularize the predictions on two different augmentations of a training sample. Unlike the temporal ensembling model which averages the prediction in every iteration, the mean teacher [34] aggregates model weights through the exponential moving average. MixMatch [2] introduces mixup [42] to explore inter-class relations in SSL. [37, 33] employ two variants of augmentations and propagate the prediction of the weakly augmented image to the strongly augmented counterpart. Most of the mainstream SSL algorithms mentioned above are based on the assumption that the labeled and unlabeled samples share the category space. We break this assumption and focus on the new setting of open-set SSL.

Self-supervised Learning. Self-supervised learning is an emerging technology which is widely considered to have the potential to initialize CNN with powerful representation capacity. It cleverly designs *pretext* tasks, which can be formulated using only unlabeled data but requiring high-level semantic understanding. As a result, the intermediate layers of convolutional neural networks trained for solving these *pretext* tasks encode high-level semantic representation that can be universally applied to downstream tasks, e.g., image classification. The most commonly-used pretext tasks include transformation based regularization [9], patch based jigsaw puzzle [27], relative location inference [7], and rotation recognition [10], etc. [19] revisits these self-supervised methods and provides comprehensive quantitative comparisons. Recently, Zhai *et al.* [41] proved that extra self-supervisions such as rotation recognition and transformation based regularization can benefit the classification performance in semi-supervised image classification. However, it is under explored whether the performance gains obtained by self-supervision will be exhausted by the interference caused by OOD samples in the process of open-set SSL. In this paper, rotation recognition is introduced as an auxiliary task to make full usage of all samples including OOD samples for feature representation enhancement.

Open-set Semi-supervised Learning. Recently, researchers have gradually focused on settling the open-set SSL problem. [6] proposes to solve the problem in which the categories of labeled and unlabeled samples are not completely matched. In [39], the concept of open-set SSL is put forward for the first time. With the help of a binary OOD classifier, a curriculum framework is proposed to solve this problem. We argue that the OOD classification does not benefit to learn discriminative feature representations for the core category recognition task but actually have conflicting targets with the target category classification. Based on this concern, we propose a novel method to filter out OOD samples based on self-labeling and cross-modal matching between images and labels. UASD [6] temporally accumulates the network predictions for self-distillation, and uses a simple threshold on the largest prediction score to detect OOD samples. However, its detection of OOD data is highly sensitive to the performance of the final classifier. Our proposed cross-modal matching strategy removes the dependency on the outcome of the classifier by inferring whether the embedding of the input image is matched to an assigned pseudo-label. It also helps detect ID samples with incorrect pseudo labels (“hard” samples) that cannot be handled by UASD. DS3L [12] introduces meta-learning to suppress the weight of OOD samples. Instead of directly filtering out OOD samples, [25] attempts to remove the distribution divergence between ID and OOD samples based on style transfer, and then explore the OOD samples during training via the unsupervised data augmentation [37]. The distribution gap between ID and OOD samples is caused by category difference. Style transfer can only change the style of OOD images but not the semantic content. Hence, it remains difficult for style transfer technique to fully eliminate the feature discrepancy between the ID and OOD samples.

3. Method

Similar to the SSL problem, the training dataset of the open-set SSL problem contains a small subset of labeled samples $\mathbb{D}^l = \left\{(\mathbf{x}_i^l, y_i^l)\right\}_{i=1}^{n=N}$ and a large subset of unlabeled samples $\mathbb{D}^u = \{\mathbf{x}_i^u\}_{i=1}^{M}$. Here, $\mathbf{x}_i^l$ or $\mathbf{x}_i^u$ represents a labeled or unlabeled image. y_i^l denotes the ground-truth category of $\mathbf{x}_i^l$. Assuming that there are K target categories, $y_i^l \in \{1, \cdots, K\}$. The ‘open-set’ setting indicates that there exists OOD samples in the unlabeled training set. Namely, $\mathbf{x}_i^u$ may not belong to any one of the K target categories. We assume that a single training batch is consisting of n labeled images and m unlabeled images. The overall framework of our method is illustrated in Figure 2.

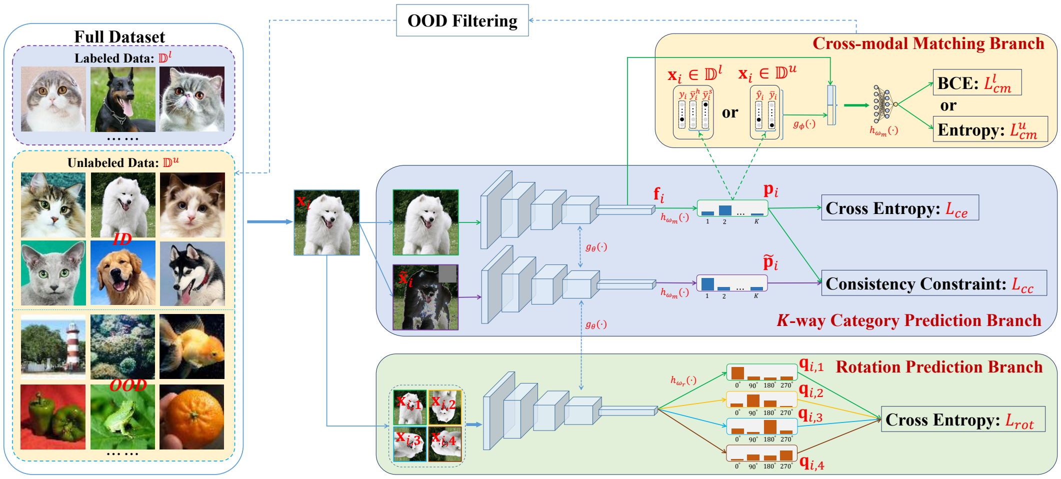


Figure 2: Overall architecture of our proposed method for open-set semi-supervised classification. It is composed of a multi-task framework, including a core category prediction branch, a rotation prediction branch for self-supervised feature learning, and a cross-modal matching branch for filtering out OOD samples within unlabeled data.

3.1. K-way Category Prediction

A convolutional neural network is chosen as the backbone for extracting feature representations from input images. It is used to extract a 128-dimensional feature $\mathbf{f}_i$ from an input image $\mathbf{x}_i$, namely $\mathbf{f}_i = g_\theta(\mathbf{x}_i)$. θ represents the network parameters, and $g_\theta(\cdot)$ denotes the calculation function of the backbone model. To obtain the prediction scores for the K target categories, a linear and a softmax layer is attached to the backbone, giving rise to a K-dimensional class probability vector $\mathbf{p}_i = h_{\omega_c}(\mathbf{f}_i) = h_{\omega_c}(g_\theta(\mathbf{x}_i))$. ω_c contains the weights and biases of the linear layer. During training, cross entropy is used to regularize the class probability vectors of labeled images,

$$L_{ce} = -\frac{1}{n}\sum_{i=1}^{n} \ln(p^l_i[y^l_i]). \tag{1}$$

Here, $p^l_i[k]$ indicates the k-th element of $\mathbf{p}^l_i$ which refers to the predicted probability vector of $\mathbf{x}^l_i$. Inspired by [37], the unsupervised consistency constraint is employed to pull close the distances between the predictions of every sample and its vicinal points. A strongly augmented counterpart $\tilde{\mathbf{x}}_i$ is synthesized for the training image $\mathbf{x}_i$. Denote the category prediction of $\mathbf{x}_i$ and $\tilde{\mathbf{x}}_i$ be $\mathbf{p}_i$ and $\tilde{\mathbf{p}}_i$ respectively. The KL divergence is adopted to calculate the distance between the two predictions, and the loss function of the consistency constraint is as below,

$$L_{cc} = \frac{1}{n+m}\sum_{i=1}^{n+m}\sum_{j=1}^{K} p_i[j]\ln(\frac{p_i[j]}{\tilde{p}_i[j]}). \tag{2}$$

3.2. Self-Supervised Representation Enhancement

For the sake of enhancing the representation capacity of the backbone with all training samples including both ID and OOD samples, the rotation recognition is introduced as an auxiliary task. In details, an extra 4-way rotation classification head consisting of 1 linear layer and 1 softmax function is attached to the backbone as shown in Figure 2. We denote the calculation process of the rotation classification head as $h_{w_r}(\cdot)$, where w_r represents related parameters. For every training image, four counterparts are generated through rotating it by 0°, 90°, 180° and 270°, respectively. We denote images generated via rotating $\mathbf{x}_i$ by $(j-1)*90°$ as $\mathbf{x}_{i,j}$, and its rotation prediction as $\mathbf{q}_{i,j} = h_{w_r}(g_\theta(\mathbf{x}_{i,j}))$. The following loss function is added during the training stage,

$$L_{rot} = -\frac{1}{4(n+m)}\sum_{i=1}^{n+m}\sum_{j=1}^{4}\ln(q_{i,j}[j]). \tag{3}$$

This rotation prediction branch is critical to take advantage of large amounts of unlabed samples, especially OOD samples, to improve the representation learning.

3.3. Cross-Modal Matching

In order to protect the learning of the K-way category recognition task from being distorted by OOD samples, a cross-modal matching branch is devised to purify unlabeled samples. For an unlabeled sample $\mathbf{x}^u_i$, we assume the predicted probability vector be $\mathbf{p}^u_i$. The category having the most confident probability value is allocated as the pseudo label of $\mathbf{x}^u_i$, namely, $\hat{y}^u_i = \arg\max_j p^u_i[j]$. The target of the cross-modal matching branch is trained to judge whether an ID image and a label from the set of target categories are matched. It can be used to identify out OOD samples since they are not belonging to any of target categories, namely they are not matched to any target category.

Given a pair of an input sample $\mathbf{x}$ and a category label y, we first extract a feature vector $\mathbf{f} = g_\theta(\mathbf{x})$ for $\mathbf{x}$. y is transformed into a one hot vector which is subsequently transformed into a 128-dimensional embedding vector via a linear layer, $\mathbf{e} = g_\phi(y)$. Afterwards, $\mathbf{e}$ and $\mathbf{f}$ are concatenated and fed into a multi-layer perceptron consisting of a hidden layer with 128-dimensional output followed by the ReLU function and a linear layer attached with the Sigmoid function, giving rise to a matching score $s(\mathbf{x}, y) = h_{\omega_m}(\mathbf{f}, \mathbf{e}) = h_{\omega_m}(g_\theta(\mathbf{x}), g_\phi(y))$. The matching score $s(\mathbf{x}, y)$ measures whether y is the correct category label of $\mathbf{x}$. When training the cross-modal matching head, positive samples can be easily collected from labeled images. Negative samples are synthesized through making pairs of images and categories which are not identical to the ground-truth label. For every training image, negative training samples are constructed in two manners: 1) Inspired by hard example mining [32], a so-called hardest label which is different from the ground-truth but having the largest prediction score is chosen; 2) The other relatively simple label is randomly selected from the category set excluding the ground-truth label and the mined hardest label. The following loss function is used for training the cross-modal matching head,

$$\begin{aligned} L^l_{cm} = \quad & -\tfrac{1}{n}\textstyle\sum_{i=1}^{n}[\ln(s(\mathbf{x}^l_i, y^l_i)) + \ln(1 - s(\mathbf{x}^l_i, \bar{y}^{l,h}_i)) \\ & \qquad + \ln(1 - s(\mathbf{x}^l_i, \bar{y}^{l,s}_i))]. \end{aligned} \tag{4}$$

Here, $\bar{y}^{l,h}_i$ and $\bar{y}^{l,s}_i$ indicates the hardest and a relatively simple negative label, respectively,

$$\bar{y}^{l,h}_i = \arg\max_{y \neq y^l_i} p_i[y], \tag{5}$$

$$\bar{y}^{l,s}_i = \text{rand}(\{y \in [1, K] \mid y \neq y^l_i;\ y \neq \bar{y}^{l,h}_i\}). \tag{6}$$

Considering the labeled samples are limited in open-set SSL, the unlabeled samples are employed to further strengthen the cross-modal matching head with entropy minimization [11],

$$\begin{aligned} L^u_{cm} = & -\frac{1}{m}\sum_{i=1}^{m}[s(\mathbf{x}^u_i, \hat{y}^u_i)\ln(s(\mathbf{x}^u_i, \hat{y}^u_i)) + \\ & (1 - s(\mathbf{x}^u_i, \hat{y}^u_i))\ln(1 - s(\mathbf{x}^u_i, \hat{y}^u_i)) + \\ & s(\mathbf{x}^u_i, \bar{y}^u_i)\ln(s(\mathbf{x}^u_i, \bar{y}^u_i)) + \\ & (1 - s(\mathbf{x}^u_i, \bar{y}^u_i))\ln(1 - s(\mathbf{x}^u_i, \bar{y}^u_i))], \end{aligned} \tag{7}$$

where, $\bar{y}^u_i = \text{rand}(\{y \in [1, K] \mid y \neq \hat{y}^u_i\})$.

Apart from identifying OOD samples, the other utility of the cross-modal matching head is to preclude parts of misclassified ID samples. This is essential to reduce the instability of training the core classification task, especially when the classification performance of the model is still unfavorable in the early iterations. The cross-modal matching branch is used to estimate the matching scores of all unlabeled samples and their pseudo labels inferred with the K-way category prediction branch. The Otsu algorithm [29] is used to select the threshold for cleaning away samples with relatively low matching scores w.r.t their pseudo labels.

3.4. Training Process

The training process consists of two stages. In the first stage, we take a warm-up training stage for optimizing the complete architecture with loss function $L = L_{ce} + L^l_{cm} + L_{rot}$. In the second stage, the cross-modal matching head is used to periodically clean unlabeled samples. The consistency constraint (2) and entropy minimization (7) are added to train the K-way category prediction branch and the cross-modal matching branch, respectively. The loss function of this stage is $L = L_{ce} + L^l_{cm} + L^u_{cm} + L_{rot} + L_{cc}$. In such a manner, OOD samples are precluded after they are fully exploited for feature enhancement in the first stage.

4. Experiments

4.1. Datasets

The following four public datasets are used to validate the performance of open-set SSL algorithms.

CIFAR-10 [20] consists of 60,000 images of size 32 × 32 which belong to 10 categories. Following the original split, 10,000 images are used for testing. The same splits in [39] are adopted for training and validating. The number of labeled training images varies in {250, 1,000, 4,000}.

Animals-10 which is obtained from Kaggle, contains 26,179 images of 10 animal categories. 500/1,000, 1,000 and 2,000 images are selected as labeled training, validating, and testing samples respectively. The remaining images are used as unlabeled samples.

CIFAR-100 has 100 classes and each class contains 600 images. We choose 25,000 images of the first 50 classes as in-distribution samples, forming the CIFAR-ID-50 dataset. They are split into 22,500 samples for training and 2,500 samples for validating. The number of labeled training images is 2,000 or 2,500.

TinyImageNet (TIN) [22] is composed of 120,000 images belonging to 200 classes. Similar to CIFAR-100, 27,500 images of the first 50 categories are regarded as in-distribution samples, which are separated into 22,500 samples for training, 2,500 samples for validating, and 2,500 samples for testing. We name the above subset of TIN as TIN-ID-50. In experiments below, 2,000/2,500 training images are selected as labeled samples.

The following inter-dataset and intra-dataset OOD settings are used in our experiments.

Inter-Dataset OOD Setting For the CIFAR-10 dataset, we follow [39] to synthesize OOD samples. 10,000 images are sampled from each of the TIN dataset, the Large-scale

Alg.	TIN			LSUN			Gaussian Noise			Uniform Noise		
	250	1000	4000	250	1000	4000	250	1000	4000	250	1000	4000
MixM	$82.42_{0.70}$	$88.03_{0.22}$	$91.25_{0.13}$	$76.32_{4.19}$	$87.03_{0.41}$	$91.18_{0.33}$	$75.76_{3.49}$	$85.71_{1.14}$	$91.51_{0.35}$	$72.90_{0.96}$	$84.49_{1.06}$	$90.47_{0.38}$
UDA	$88.83_{0.49}$	$91.75_{0.12}$	$93.63_{0.11}$	$88.54_{1.10}$	$91.12_{0.31}$	$93.25_{0.12}$	$88.93_{0.73}$	$89.23_{0.41}$	$92.35_{0.18}$	$88.69_{0.93}$	$89.74_{0.42}$	$92.74_{0.35}$
UASD	83.53	-	-	80.87	-	-	-	-	-	-	-	-
DS3L	-	$70.10_{0.47}$	$82.58_{0.14}$	-	$69.74_{0.08}$	$82.89_{0.69}$	-	$62.86_{0.67}$	$80.44_{0.01}$	-	$62.89_{1.65}$	$80.59_{0.03}$
MTCF	$86.44_{0.64}$	$89.85_{0.11}$	$93.03_{0.05}$	$86.65_{0.41}$	$90.19_{0.47}$	$92.91_{0.03}$	$87.34_{0.13}$	$89.80_{0.26}$	$92.53_{0.08}$	$85.54_{0.11}$	$89.87_{0.08}$	$92.83_{0.04}$
OTCT	-	$91.10_{0.65}$	$93.84_{0.10}$	-	$91.30_{0.36}$	$94.27_{0.21}$	-	$92.33_{0.59}$	$94.52_{0.07}$	-	$91.82_{0.04}$	$94.50_{0.13}$
Ours	$91.52_{0.11}$	$93.26_{0.14}$	$94.71_{0.06}$	$91.13_{0.21}$	$94.43_{0.10}$	$94.97_{0.10}$	$90.81_{0.12}$	$93.63_{0.06}$	$94.38_{0.12}$	$89.95_{0.18}$	$94.12_{0.14}$	$94.83_{0.15}$

Table 1: Comparisons with the conventional SSL and open-set SSL algorithms, including MixM (short for MixMatch) [2], UDA [37], UASD [6] DS3L [12], MTCF [39], and OTCT [25], on variants of CIFAR-10 which are respectively corrupted with two real-world OOD datasets (TIN and LSUN) and two synthetic OOD datasets (Gaussian Noise and Uniform Noise). Accuracy (%) is used for evaluating algorithms. The subscript of the accuracy value indicates its standard deviation. 250, 1000, and 4000 labeled images are used for training respectively. We use the reported results of [39] for UASD and that of [25] for DS3L and OTCT.

Scene Understanding (LSUN) dataset [38], Gaussian noise dataset, and uniform noise dataset, forming into 4 OOD settings. For the Animals-10 and CIFAR-ID-50 datasets, 10,000 images from TIN are used as OOD samples.

Intra-Dataset OOD Setting For CIFAR-ID-50, we select 100 images from each of the other 50 classes of CIFAR-100 as OOD images. For TIN-ID-50, 50 images from each of the other 150 classes of TIN are chosen as OOD images.

4.2. Implementation details

Existing methods including MixMatch [2], UDA [37] FixMatch [33], UASD [6], DS3L [12], OTCT [25] and MTCF [39], are used for comparison. For UDA, FixMatch and our method, SGD is used to optimize network weights. The learning rate is initially set to 0.03 and adjusted via the cosine decay strategy [37, 33]. The momentum is set to 0.9. In each training batch, $n = 64$, and $m = 320$. For our method, the first stage costs 50,000 iterations, and the second stage takes 200,000 iterations. Without specification, the cycle length of using the crossmodal matching head to clean unlabeled data is 2×10^4. For UDA and FixMatch, models are trained with 250,000 iterations for a fair comparison. When training MixMatch and MTCF, we follow the original settings of [39] in which models are trained with 1,024 epochs, and each epoch contains 1,024 iterations. During the training stage, network weights are saved every 1,000 iterations. The averaged classification accuracy of the last 20 copies is used to evaluate the performance of all methods. For all experiments, we use the Wide-ResNet28-2 [40] as the backbone model.

Method	Animals-10		CIFAR-ID-50	
	500	1000	2000	2500
MixMatch [2]	78.35	83.15	62.10	64.78
UDA [37]	83.30	84.74	64.34	66.65
MTCF [39]	73.50	75.83	63.22	65.10
MTCF+UDA	79.85	85.60	65.20	67.30
FixMatch [33]	89.06	91.00	68.98	72.92
Ours+UDA	87.86	89.70	71.58	73.19
Ours+FixMatch	**89.43**	**91.50**	**72.06**	**73.80**

Table 2: Accuracy (%) for Animals-10 and CIFAR-ID-50. Images of TIN are used as OOD samples. On Animals-10 dataset, 500 or 1000 labeled images are used for training. On CIFAR-ID-50 dataset, 2000 or 2500 labeled images are used for training.

Method	CIFAR-ID-50		TIN-ID-50	
	2000	2500	2000	2500
MixMatch [2]	60.20	66.17	48.12	50.52
UDA [37]	66.02	67.82	54.03	55.39
MTCF [39]	63.48	65.38	49.64	52.08
MTCF+UDA	60.06	63.08	47.32	51.20
FixMatch [33]	68.06	71.01	56.82	60.33
Ours+UDA	67.36	69.14	54.87	57.08
Ours+FixMatch	**68.65**	**73.14**	**57.48**	**62.64**

Table 3: Accuracy (%) for CIFAR-ID-50 and TIN-ID-50 under the intra-dataset OOD setting. On both datasets, 2000 or 2500 labeled images are used for training.

4.3. Experimental Results

4.3.1 Comparison with Other Methods

Inter-Dataset OOD Setting As introduced in Section 4.1, three datasets including CIFAR-10 (Table 1), Animals-10 (Table 2), and CIFAR-ID-50 (Table 2) are used to validate the classification performance under the inter-dataset OOD setting. For all inter-dataset OOD settings of CIFAR-10, our method surpasses all compared methods by significant margins. As shown in Table 1, when 250 labeled images are provided and images of TIN are used as OOD samples,

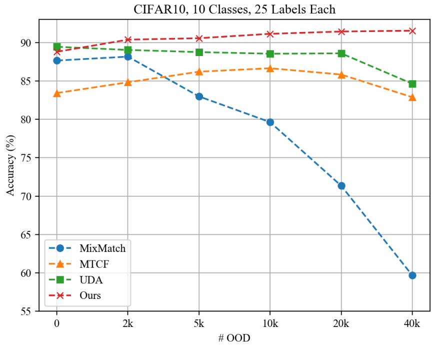


Figure 3: Impact of different number of OOD samples on baselines and our approach.

our method achieves an average accuracy of 91.52% which is 5.08% and 3.69% higher than MTCF and UDA, respectively. Compared with two synthetic noise OOD datasets, the rich semantic information of real-world datasets is beneficial to our proposed method. For instance, the accuracy of our method achieves 91.13% on the LSUN OOD setting, which is 1.28% higher than the accuracy on the uniform noise OOD setting.

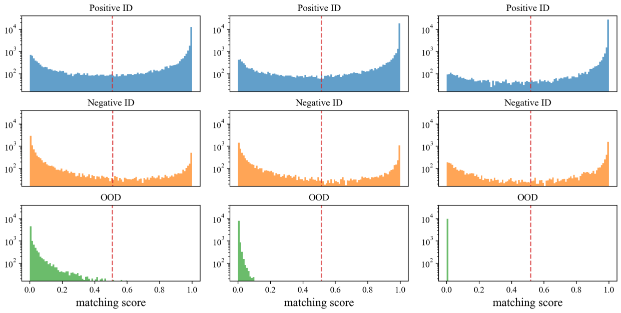
Table 2 reports the results of Animals-10 and CIFAR-ID-50 under the inter-dataset OOD setting in which images of TIN are added as OOD samples. Our method is capable of improving existing SSL methods, e.g. UDA and FixMatch, and performs much better than MTCF.

We further study the impact of the number of OOD samples on SSL and open-set SSL algorithms. In this experiment, we use CIFAR-10 as the ID dataset and LSUN as the OOD dataset. The number of OOD samples varies from 0 to 4×10^4. 25 labeled samples are provided for each class The results are presented in Figure 3. Our method achieves better performance with even more OOD samples, and consistently outperforms all the other methods in all settings.

Intra-Dataset OOD Setting The CIFAR-ID-50 and TIN-ID-50 datasets are adopted in this experiment. Images of the same dataset but having categories different from those ID categories are regarded as OOD samples. The experimental results are shown in Table 3. Our method consistently performs better than other methods.

ID	OOD	ODIN	SUF	MTCF	Ours
CIFAR-10	LSUN	98.47	99.03	99.82	**99.98**
Animals-10	TIN	76.35	90.01	92.59	**93.51**
CIFAR-ID-50	TIN	88.82	97.98	98.17	**99.85**
CIFAR-ID-50	CIFAR-50	69.47	72.32	69.75	**74.13**
TIN-ID-50	TIN-150	59.83	65.59	63.92	**65.67**

Table 4: The comparison of our method against ODIN [24], SUF [23], and MTCF [39] on the task of OOD detection. The evaluation metric is AUROC(%). ODIN and SUF are implemented based on the classifier model learnt by UDA. For CIFAR-10 and Animals-10, 250 labels and 500 labels are used during training respectively. For other datasets, 2500 labels are used during training.



1000-th iteration 19000-th iteration 28000-th iteration

Figure 4: Visualization of the matching scores in different iterations. The number of samples are showcased in logarithmic coordinates (vertical axis) with base of 10. ‘Positive ID’/‘Negative ID’ means ID samples having correct/incorrect classification results. The threshold value obtained via OTSU is indicated by the dashed vertical line.

We try to modify MTCF [39] via using UDA [37] for semi-supervised learning, which forms a new open-set SSL algorithm denoted by ‘UDA+MTCF’. In MTCF, the learning of the target category classifier is in conflict with the learning of the binary OOD classifier, thus the features learned by MTCF is not discriminative enough. Meanwhile, UDA depends on the label propagation from the weakly augmented image to the strongly augmented counterpart, which is sensitive to a confused feature space. The incorporation of MTCF only brings marginal improvement (e.g. CIFAR-ID-50 in Table 2) or even causes severe performance degradation (e.g. both datasets in Table 3) to UDA.

OOD Detection Performance In Table 4, we compare our method against ODIN [24], Mahalanobis [23] and MTCF [39] under extensive combinations of ID and OOD datasets, to validate the efficacy of the cross-modal matching branch in detecting OOD images. In our method, the matching score of an image and its pseudo label is regarded as the probability value of the image belonging to ID samples. The area under the receiver operating characteristic (AUROC) is used to measure the performance of OOD detection algorithms. Our method outperforms ODIN, Mahalanobis and MTCF under all settings.

4.3.2 Ablation Study

We carry out an extensive ablation study on the CIFAR-10 and CIFAR-ID-50 to tease apart the experimental factors that are most important to the success of our method, as

M-#	Settings	CIFAR-10	CIFAR-ID-50
1	Supervised	39.66	53.18
2	Supervised+SS	75.69	59.82
3	UDA	88.83	64.78
4	UDA+SS	90.69	70.10
5	UDA+SS+CT	90.81	71.16
6	UDA+SS+CMF-w/o-L^u_{cm}	91.09	72.38
7	UDA+SS+CMF-w/o-WU	89.45	71.81
8	UDA+SS+CMF-full	**91.52**	**73.19**

Table 5: Ablation Study on CIFAR-10 and CIFAR-ID-50. TIN is chosen as the OOD dataset. 250 and 2500 labeled images are for CIFAR-10 and CIFAR-ID-50, respectively. Abbreviations: unsupervised data augmentation (UDA) [37], crossmodal matching based filtering (CMF), self-supervision based on rotation recognition (SS), warming-up (WU), confidence-based thresholding (CT).

shown in Table 5.

Self-supervised Learning in Open-set SSL M-2 and M-4 in Table 5 represent the variant of the simple supervised learning and UDA respectively, which employs rotation recognition as the auxiliary task. The accuracy of M-2 and M-4 on CIFAR-ID-50 is 5.32% and 6.64% higher than that of M-1 and M-3 respectively. This indicates that the feature enhancement brought by the rotation recognition task can significantly improve the classification accuracy. This is the first time to prove that self-supervised representation learning based on rotation recognition is effective in open-set SSL. We have tried other self-supervised learning methods, like SimCLR [4] and MoCo [14], and empirically find that they perform much worse than the rotation prediction approach. Both SimCLR and MoCo strive to distinguish the features of different samples while our task requires samples of the same class to stay closer in the feature space. The conflicting goals make them unsuitable for our task.

Different Variants of OOD Sample Filtering The full version of our method (M-8 in Table 5) outperforms M-4 without cross-modal matching head by 3.09% and 0.83% on CIFAR-ID-50 and CIFAR-10 respectively. This validates that our proposed OOD sample filtering algorithm based on cross-modal matching is complementary to self-supervised learning, and it further benefits the classification performance of in-distribution categories. On the other hand, merely using the supervised loss function (4) to train the cross-modal matching branch (M-6) leads to performance degradation to a certain extent. The accuracy on CIFAR-ID-50 is decreased by 0.81%, compared to the results of M-8 in which both function (4) and (7) are adopted for training. We also attempt to replace the cross-modal matching based filtering with the confidence-based thresholding via precluding unlabeled samples with maximum prediction scores lower than the OTSU threshold . The resulted method (M-5) achieves the accuracy of 90.81% and 71.16% on CIFAR10 and CIFAR-ID-50 respectively, which is much inferior to that of our method. This indicates the superiority of the proposed cross-modal matching based filtering to the confidence-based thresholding, since the former can simultaneously implement the OOD sample filtering and promote the discrimination between features of different categories.

Efficacy of Warming Up Stage The warm-up stage plays a crucial role in fully leveraging all ID and OOD samples for representation learning, providing a decent initialization for the K-way category prediction and the cross-modal matching tasks. Without the warm-up stage (M-7 in Table 5), the classification accuracy is reduced by 2.07% and 1.38% on CIFAR-10 and CIFAR-ID-50, respectively.

4.3.3 Justification of Cross-modal Matching

The performance of the cross-modal matching in identifying OOD samples and misclassified ID samples at different training iterations is visualized in Figure 4. The cross-modal matching head can effectively separate correctly classified samples from OOD samples, throughout the training process. In early iterations, the head is capable of precluding most of the misclassified samples. This helps prevent these misclassified samples from misleading the optimization process. As the iteration progresses, the core classifier gradually becomes better and the number of correctly classified samples continues to increase.

5. Conclusions

In this paper, we propose a novel framework for open-set SSL which can harvest the OOD samples for enhanced feature learning while avoiding its adverse impact on the SSL training. Our key insight is that OOD samples, if exploited in a pretext task of rotation recognition, can be “treasures” for learning more discriminative features that can benefit the final classification task. After the value of OOD samples in representation learning has been explored, the cross-modal matching branch is further utilized to filter out OOD samples without causing much interference to the core category prediction task. Our proposed method can be easily integrated into existing SSL algorithms, and achieves state-of-the-art in open-set SSL on extensive public benchmarks.

Acknowledgments This work was supported in part by the Guangdong Basic and Applied Basic Research Foundation under Grant No.2020B1515020048, in part by the National Natural Science Foundation of China under Grant No.61976250, No.U1811463 and No.62003256, and in part by the Guangzhou Science and technology project under Grant No.202102020633.

References

[1] Mikhail Belkin and Partha Niyogi. Semi-supervised learning on riemannian manifolds. *Machine learning*, 56(1):209–239, 2004. 3

[2] 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*, pages 5049–5059, 2019. 3, 6

[3] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100, 1998. 3

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 8

[5] Xing Chen and Gui-Ying Yan. Semi-supervised learning for potential human microrna-disease associations inference. *Scientific reports*, 4:5501, 2014. 2

[6] Yanbei Chen, Xiatian Zhu, Wei Li, and Shaogang Gong. Semi-supervised learning under class distribution mismatch. In *Proceedings of the AAAI Conference on Artificial Intelligence.*, pages 3569–3576, 2020. 3, 6

[7] C. Doersch, A. Gupta, and A. A. Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1422–1430, 2015. 3

[8] WeiWang Dong-Dong Chen and Zhi-Hua Zhou WeiGao. Trinet for semi-supervised deep learning. In *Proceedings of the International Joint Conferences on Artificial Intelligence*, pages 2014–2020, 2018. 3

[9] Alexey Dosovitskiy, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with convolutional neural networks. In *Advances in Neural Information Processing Systems*, 2014. 3

[10] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *Proceedings of the International Conference on Learning Representations*, 2018. 3

[11] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In *Advances in Neural Information Processing Systems*, 2005. 5

[12] 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*, pages 3897–3906. PMLR, 2020. 3, 6

[13] Jingrui He, Jaime Carbonell, and Yan Liu. Graph-based semi-supervised learning as a generative model. In *Proceedings of the International Joint Conferences on Artificial Intelligence*, pages 2492–2497, 01 2007. 3

[14] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 8

[15] Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondrej Chum. Label propagation for deep semi-supervised learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 5070–5079, 2019. 3

[16] Lukas Käll, Jesse D Canterbury, Jason Weston, William Stafford Noble, and Michael J MacCoss. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nature methods*, 4(11):923–925, 2007. 2

[17] Hisashi Kashima, Tsuyoshi Kato, Yoshihiro Yamanishi, Masashi Sugiyama, and Koji Tsuda. Link propagation: A fast semi-supervised learning algorithm for link prediction. In *Proceedings of the SIAM International Conference on Data Mining*, pages 1100–1111. SIAM, 2009. 2

[18] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *Proceedings of the International Conference on Learning Representations*, 2017. 3

[19] Alexander Kolesnikov, Xiaohua Zhai, and Lucas Beyer. Revisiting self-supervised visual representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2019. 3

[20] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. *Master's thesis, Department of Computer Science, University of Toronto*, 2009. 1, 2, 5

[21] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. In *Proceedings of the International Conference on Learning Representations*, 2016. 3

[22] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7, 2015. 2, 5

[23] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in neural information processing systems*, volume 31, 2018. 7

[24] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *Proceedings of the International Conference on Learning Representations*, 2017. 7

[25] Huixiang Luo, Hao Cheng, Yuting Gao, Ke Li, Mengdan Zhang, Fanxu Meng, Xiaowei Guo, Feiyue Huang, and Xing Sun. On the consistency training for open-set semi-supervised learning. *arXiv preprint arXiv:2101.08237*, 2021. 3, 6

[26] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008. 1

[27] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *Proceedings of the European Conference on Computer Vision*, volume 9910, pages 69–84, 10 2016. 3

[28] Avital Oliver, Augustus Odena, Colin Raffel, Ekin D Cubuk, and Ian J Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. In *Advances in Neural Information Processing Systems*, 2018. 2

[29] N Ostu, O Nobuyuki, and N Otsu. A threshold selection method from gray- level histogram ieee transactions on systems. *IEEE Trans.syst.man. & Cybern*, 9(1):62–66, 1979. 5

[30] Siyuan Qiao, Wei Shen, Zhishuai Zhang, Bo Wang, and Alan Yuille. Deep co-training for semi-supervised image recognition. In *Proceedings of the European Conference on Computer Vision*, pages 135–152, 2018. 3

[31] U. Raghavan, R. Albert, and S. Kumara. Near linear time algorithm to detect community structures in large-scale networks. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 76 3 Pt 2:036106, 2007. 3

[32] Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 761–769, 2016. 5

[33] Kihyuk Sohn, David Berthelot, Chun Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *Advances in Neural Information Processing Systems*, 2020. 3, 6

[34] 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. 3

[35] Isaac Triguero, Salvador García, and Francisco Herrera. Self-labeled techniques for semi-supervised learning: taxonomy, software and empirical study. *Knowledge and Information systems*, 42(2):245–284, 2015. 3

[36] Bo Wang, Zhuowen Tu, and John K Tsotsos. Dynamic label propagation for semi-supervised multi-class multi-label classification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 425–432, 2013. 3

[37] Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6256–6268. Curran Associates, Inc., 2020. 3, 4, 6, 7, 8

[38] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015. 6

[39] Qing Yu, Daiki Ikami, Go Irie, and Kiyoharu Aizawa. Multi-task curriculum framework for open-set semi-supervised learning. In *Proceedings of the European Conference on Computer Vision*, 2020. 1, 2, 3, 5, 6, 7

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

[41] Xiaohua Zhai, Avital Oliver, Alexander Kolesnikov, and Lucas Beyer. S4l: Self-supervised semi-supervised learning. In *Proceedings of the IEEE International Conference on Computer Vision*, October 2019. 2, 3

[42] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *Proceedings of the International Conference on Learning Representations*, 2017. 3

[43] Xiaojin Zhu and Zoubin Ghahramani. Learning from labeled and unlabeled data with label propagation. 2002. 3