STEP : OOD Detection in the Presence of Limited In-Distribution Labeled Data

Anonymous Author(s)
Affiliation
Address
email

Abstract

Existing semi-supervised learning (SSL) studies typically assume that unlabeled and test data are drawn from the same distribution as labeled data. However, in many real-world applications, it is more desirable to have SSL algorithms that not only classify the samples drawn from the same distribution of labeled data but also detect out-of-distribution (OOD) samples drawn from an unknown distribution. In this paper, we propose a novel setting called semi-supervised OOD detection. Two main challenges compared with previous OOD detection settings are i) the lack of labeled data and in-distribution data; ii) OOD samples could be unseen during training. Efforts on this direction remain limited. In this paper, we present an approach STEP significantly improving OOD detection performance by introducing a new technique: Structure-Keep Unzipping. We proposed a novel objective formulation that learns a new representation space in which OOD samples could be separated well. An efficient optimization algorithm is derived to solve the objective. Comprehensive experiments across various OOD detection benchmarks clearly show that our STEP approach outperforms other methods by a large margin and achieves remarkable detection performance on several benchmarks. Our code will be open source.

1 Introduction

Deep learning has achieved great success in many application scenarios, such as computer vision [15, 25], speech recognition [1], natural language processing [7]. These successful techniques typically rely on sufficient supervised information. However, collecting large amounts of well-labeled training data is hard in real-world applications due to the expensive cost of the labeling process. Therefore, tremendous efforts have been devoted to SSL [5] which aims to enhance the model performance by exploiting much cheaper unlabeled data over the past decades. These SSL methods have been successfully applied to various real-world applications [45, 32, 28].

Previous SSL studies [35, 37] are typically based on the assumption that unlabeled data and test data are drawn from the same distribution as labeled data. However, it is often the case that such an assumption fails in practical applications. For example, in document classification [10], irrelevant documents readily occur in the testing data leading to high-confidence misclassification. Similar cases commonly appear in other applications, such as medical diagnosis [4] and autonomous driving [9]. In such applications, SSL methods should be not only to classify samples from known distribution accurately but also be equipped with the ability to detect out-of-distribution (OOD) samples from unknown distributions precisely.

OOD detection has been studied for a long history with numerous methods proposed, such as ODIN [29], Mahalanobis [27], DeConf [19], ELOC [39]. These methods perform OOD detection based on the logits of the model or the Mahalanobis distance in the feature space. However, it is hard to adapt

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

these methods to semi-supervised settings because they do not consider the exploitation of unlabeled data. Some methods consider unlabeled data, such as UOOD [43], CSI [36], SSD [34] have been proposed recently. These methods assume that the model can obtain sufficient in-distribution (ID) labeled data or ID unlabeled during the training. Such an assumption also limits their ability to solve practical problems.

Therefore, we introduce a novel setting called semi-supervised OOD detection in our paper. Specifically, only a tiny subset of labeled ID data is observed. The others remain unlabeled and may belong to ID or OOD. This setting is crucial because it is ubiquitous in real-world applications. For example, in web page classification [41], acquiring large numbers of web pages annotated with relevant categories is very expensive, and unlabeled web pages crawled from the Internet according to keywords usually contain irrelevant pages that belong to unseen categories. In medical diagnosis [4], warning users of the model's uncertainty is crucial because any unfaithful diagnosis will bring unimaginable disasters to the patients' health. However, both sufficient labeled data and clean unlabeled data are hard to acquire for training. In ride-sharing liability judgment, detecting abnormal orders is of significant business value. However, collecting training data will meet similar problems stated above. Similar cases ofter occur in other applications such as image recognition and autonomous driving [43, 9, 44]. There are two main challenges for us compared with previous OOD detection settings. First, both our labeled data and ID data are limited, while sufficient unlabeled data is mixed with ID and OOD samples. Second, OOD samples could be unseen during the training, requiring more stringent generalization of the model.

Focusing on semi-supervised OOD detection, we find that the widely-used Mahalanobis distance is no longer suitable as the confidence score for OOD detection. This is because the necessary covariance matrix $\hat{\mathbf{\Sigma}}$ for calculating Mahalanobis distance is complicated to accurately estimate with limited ID samples which will severely affect the performance of OOD detection. To alleviate this issue, we propose a novel approach, called STEP (STructure-keEP). Our high-level idea is to detect OOD samples in a detection-specific space where we maintain the same local topological structures as the original feature space. The relationships between samples need to be confirmed through local topological structures. We formulate this idea into an objective and derive a loss function for optimizing it. Then, we design STEP approach to solve this issue efficiently. We evaluate our approach on a diverse set of ID and OOD data set pairs. The experiments prove that our STEP approach outperforms other methods by a large margin in our setting.

The contributions of our paper are summarized as follows:

- We propose a novel and practical setting for OOD detection, called semi-supervised OOD detection. It is worth mentioning that this setting is widespread in real-world applications.
- We propose a novel approach STEP for semi-supervised OOD detection. We alleviate the problem of Mahalanobis distance that the necessary covariance matrix $\hat{\mathbf{\Sigma}}$ is complex to be estimated accurately with limited ID samples and propose a new distance calculated in a detection-specific space as OOD confidence scores.
- We evaluate our STEP approach with comprehensive experiments across various OOD detection benchmarks. Our STEP approach outperforms other methods by a large margin and achieves remarkable detection performance on several benchmarks.

2 Method

2.1 Notations and Setting

In the semi-supervised OOD detection setting, we assume that a limited label data set $\mathcal{D}_l = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ consisting n samples with labels drawn from ID, and an unlabeled data set $\mathcal{D}_u = \{(\mathbf{x}_i)\}_{i=1}^m$ consisting m unlabeled samples drawn both ID and OOD, are accessible during the training phase. We denote the set of ground-truth classes in the labeled data set $\mathcal{D}_l$ and unlabeled data set $\mathcal{D}_u$ as $\mathcal{C}_l$ and $\mathcal{C}_u$, respectively. The labeled samples can be classified into one of K classes denoted by $\mathcal{C}_l = \{c_1, c_2, \ldots, c_K\}$, and the unlabeled samples can be classified into the seen K classes $\mathcal{C}_l$ and some unseen classes denoted by $\mathcal{C}_n = \mathcal{C}_u \backslash \mathcal{C}_l$. The goal is to distinguish whether a sample in $\mathcal{D}_u$ or an unknown testing sample is drawn from ID or not.

2.2 Inaccurate Mahalanobis Distance

Mahalanobis distance which is widely used in previous studies [27, 34], has been proven to be a powerful metric in OOD detection. $\mathcal{MD}(\mathbf{x}_i, \mathbf{x}_j)$ denotes the function measuring the Mahalanobis distance between sample $\mathbf{x}_i$ and sample $\mathbf{x}_j$ based on estimated covariance matrix $\hat{\mathbf{\Sigma}}$:

$$\mathcal{MD}(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^\top \hat{\mathbf{\Sigma}}^{-1} (\mathbf{x}_i - \mathbf{x}_j)} \tag{1}$$

Previous methods mentioned above calculate the minimum Mahalanobis distance between target sample $\mathbf{x}$ and each class center as the confidence score:

$$\textsc{Score}_{\mathcal{MD}}(\mathbf{x}) = \min_{c \in c_1, c_2, \ldots, c_K} \mathcal{MD}(\mathbf{x}, \mu_c) \tag{2}$$

where μ_c denotes the center of samples which belong to class c and $\hat{\mathbf{\Sigma}}$ is the covariance matrix estimated on all ID samples.

However, $\hat{\mathbf{\Sigma}}$ is hard to accurately estimate in a semi-supervised OOD detection setting because the available ID labeled data set $\mathcal{D}_l$ is insufficient. Inaccurate estimation of $\hat{\mathbf{\Sigma}}$ will affect the calculation of Mahalanobis distance. This makes it difficult for the algorithm to distinguish OOD samples and ID samples near the cluster boundary.

Instead of using inaccurate Mahalanobis distance, we decide to learn a $\mathbf{P}$ to project samples into space where a large margin separates ID samples and OOD samples. Inspired by the topological technology [40] used in noisy label problems and cluster assumption [5] used in SSL, we hope that the projected samples can maintain the same local topological structure as the original space while increasing the distance between samples not directly topologically connected. Because of the inaccurate estimation of $\hat{\mathbf{\Sigma}}$, we consider that relationship between samples that are not topologically adjacent is uncertain. Their relationships need to be confirmed through each local topological structure. We formulate our goal into the objective:

$$\begin{aligned} \max_{\mathbf{P}} \quad & \|\mathbf{P}\mathbf{x}_i - \mathbf{P}\mathbf{x}_j\|_2, \quad \forall \mathbf{x}_i, \mathbf{x}_j \in \mathcal{D}_l \cup \mathcal{D}_u \\ \text{s.t.} \quad & \|\mathbf{P}\mathbf{x}_i - \mathbf{P}\mathbf{x}_n\|_2 = \mathcal{MD}(\mathbf{x}_i, \mathbf{x}_n), \\ \text{if} \quad & \mathbf{x}_n \in \mathcal{B}_k(\mathbf{x}_i) \end{aligned} \tag{3}$$

where, $\mathcal{M}(\mathbf{x}_i, \mathbf{x}_j)$ is the Mahalanobis distance between $\mathbf{x}_i$ and $\mathbf{x}_j$ in the feature space and $\mathcal{B}_k(\mathbf{x}_i)$ is the set of k nearest neighbours of $\mathbf{x}_i$.

Finally, our detection-specific metric can directly calculate as L2 distance in the projected space:

$$\mathcal{N}(\mathbf{x}_i, \mathbf{x}_j) = \|\mathbf{P}\mathbf{x}_i - \mathbf{P}\mathbf{x}_j\|_2 \tag{4}$$

2.3 Our Approach: STEP

2.3.1 Backbone Pretraining

Our semi-supervised OOD detection task considers OOD detection as a clustering problem based on the feature space. Therefore, reliable feature representations are essential. Benefiting from recent progress on self-supervised learning, we adopt a simple contrastive learning method SimCLR [6] to pre-train our backbone network on the whole dataset $\mathcal{D}_u \cup \mathcal{D}_l$ in an unsupervised fashion. We find that representations obtained by SimCLR have a reasonable ability to distinguish ID and OOD samples. Notably, the learned representations could be not only used for our STEP approach but also used as the initialization of downstream tasks.

2.3.2 Structure-Keep Unzipping

Based on the representations obtained by SimCLR, we further train a $\mathbf{P}$ to project samples into a detection-specific space via our objective formulated in Eq.(3). However, there are two main difficulties: i) Building a KNN graph for extracting topological structure needs $\mathcal{O}(n^2 d^2)$ time complexity to calculate Mahalanobis distance between each pair of samples. This step is very time-consuming because we use an ensemble of representations from each backbone network's layer, and feature dimension d is relatively large. ii) The constraint in Eq.(3) can not be directly optimized.

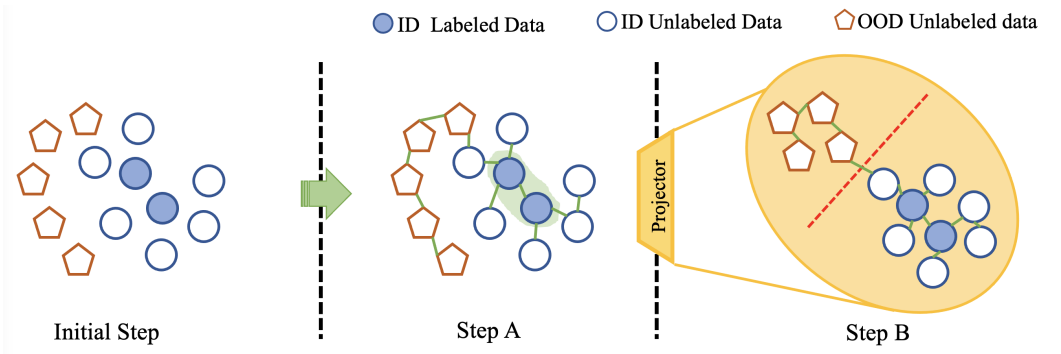


Figure 1: The overall of STEP approach: (a) In initial step, we use contrastive learning to train initial representations. (b) In step A, we estimated statistics information via limited labeled data and extracted topological structure via the KNN algorithm. (c) In step B, we train a $\mathbf{P}$ to project all the samples into a detection-specific space where we can use L2 distances as OOD scores.

First, we transform the process of calculating pairwise Mahalanobis distance into calculating pairwise Euclidean distance in projection space. The time complexity of this step reduces from $\mathcal{O}(n^2d^2)$ to $\mathcal{O}(n^2d)$. Specifically, as shown in Eq.(5), we can perform cholesky decomposition on $\hat{\mathbf{\Sigma}}^{-1}$ to get linear projector the $\mathbf{P}_{\mathcal{MD}}$. Then, we multiply all samples by $\mathbf{P}_{\mathcal{MD}}$ to project them into a new space where Euclidean distance equals to original Mahalanobis distance between each pair of samples. There are n^2 pairwise Euclidean distances to calculate, and each calculation costs $\mathcal{O}(d)$ time complexity. Therefore, the total time complexity of this step is $\mathcal{O}(n^2d)$.

$$\begin{aligned} \mathcal{MD}(\mathbf{x}_i, \mathbf{x}_j) &= \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^\top \hat{\mathbf{\Sigma}}^{-1} (\mathbf{x}_i - \mathbf{x}_j)} = \|\mathbf{P}_{\mathcal{MD}}\mathbf{x}_i - \mathbf{P}_{\mathcal{MD}}\mathbf{x}_j\|_2 \\ \text{s.t.} \quad & \mathbf{P}_{\mathcal{MD}}^\top \mathbf{P}_{\mathcal{MD}} = \hat{\mathbf{\Sigma}}^{-1} \end{aligned} \tag{5}$$

After converting Mahalanobis distance to Euclidean distance, we can further use the advanced KNN toolkit, such as Faiss [21], to speed up the entire process.

Second, we define L_{Keep} and L_{Unzip} that can be directly optimized to approximately achieve our objective shown in Eq.(3). Both L_{Keep} and L_{Unzip} are shown in Eq.(6):

$$\begin{cases} L_{Keep} & = \max(0, \|\mathbf{P}\mathbf{x}_i - \mathbf{P}\mathbf{x}_n\|_2 - \mathcal{MD}(\mathbf{x}_i, \mathbf{x}_n)), \\ L_{Unzip} & = -\|\mathbf{P}\mathbf{x}_i - \mathbf{P}\mathbf{x}_j\|_2. \end{cases} \tag{6}$$

where $\mathbf{x}_i, \mathbf{x}_j$ are randomly sampled from $\mathcal{D}_l \cup \mathcal{D}_u$, and x_n is randomly sampled from $\mathcal{B}_k(\mathbf{x}_i)$. The final loss to optimize $\mathbf{P}$ is $Loss = L_{Keep} + L_{Unzip}$. The overall of our STEP approach is summarized in Fig.(1), and the pseudo-code of our approach is shown in Algo.(1).

In the detection stage, we directly use the minimum L2 distance between the target sample and each class center in the detection-specific space as the confidence score:

$$Score(\mathbf{x}) = \min_{c \in \{c_1, c_2, \ldots, c_K\}} \mathcal{N}(\mathbf{x}, \boldsymbol{\mu}_c) \tag{7}$$

where the $\boldsymbol{\mu}_c$ is the center of class c in the original feature space.

3 Experiments

3.1 Experimental Setup

In-distribution Data Set. We use CIFAR-10 and CIFAR-100 [24] as ID data sets in our experiments. They both contain 50,000 training images and 10,000 testing images. The image size of these two data sets is 32×32. For CIFAR-10, each image belongs to one of 10 classes, and we randomly sample 250 training images as labeled ID data $\mathcal{D}_l$. For CIFAR-100, the size of image classes is 100, and we randomly sample 400 training images as labeled ID data. $\mathcal{D}_l$. We add the remaining training images to the unlabeled data $\mathcal{D}_u$.

Algorithm 1 Training Phase of STEP

Input: $\mathcal{D}_l$: labeled ID data set; $\mathcal{D}_u$: unlabeled mixed data set; K: number of neighbours
Output: pre-trained backbone $f_\theta(\cdot)$; projctor $\mathbf{P}$

```
train backbone f_θ(·) via contrastive learning on D_l ∪ D_u
estimate Σ̂ on D_l with f_θ(·)
calculate P_MD based on Σ̂^{-1}
build KNN on D_l ∪ D_u with P_MD and f_θ(·)
for epoch ∈ {1, 2, ..., epoch_max} do
    randomly sample x_i, x_j from D_l ∪ D_u
    randomly sample x_n from B_k(x_i)
    calculate Loss based on Eq.(6)
    optimize P via SGD according to Loss
end for
return f_θ(·) and P
```

Out-of-distribution Data Set. We use Tiny ImageNet data set [8] and Large-scale Scene Understanding data set [42] as OOD data sets. The Tiny ImageNet data set (TIN) is a subset of ImageNet, which contains 10,000 test images, includes 200 different classes. Following the settings used by previous studies [29, 39, 43], we use two variants of TIN: TinyImageNet-crop (TINc) and TinyImageNet-resize (TINr), by randomly cropping or downsampling each image to 32×32, respectively. The Large-scale Scene Understanding data set (LSUN) contains 10,000 testing images belong to 10 different scene categories. Similarly, we use two variants of LSUN: LSUN-crop (LSUNc) and LSUN-resize (LSUNr). Because some comparison methods in our experiments heavily rely on OOD validation. We randomly draw several images from ID testing images and OOD images as the OOD validation set. The rest of the OOD images are added to unlabeled data $\mathcal{D}_u$ and used as testing data. These OOD data sets are released by ODIN [29] with their code[1].

Comparsion Methods. We compare our STEP approach with representative OOD detection methods, including the state-of-the-art UOOD method. ODIN [29] is a common baseline of OOD detection. It uses maximum softmax score combining temperature scaling and input preprocessing tricks to distinguish ID and OOD samples. MAH [27] uses Mahalanobis distance as the OOD confidence score. For features of each layer in the backbone model, it independently calculates the Mahalanobis distances between the target sample and each known class center. Then it integrates them by weighted averaging via an extra OOD validation set. We denote it as MAH † because it uses a validation set when training. UOOD [43] utilizes a two-head CNN consisting of one common feature extractor and two classifiers which has different decision boundaries to detect OOD samples. This method optimizes a discrepancy loss between two classifiers during the training stage and uses this discrepancy as the OOD score when testing. However, this method relies on extra OOD validation to perform model selection. Therefore, we denoted it as UOOD † in our experiments. For fair comparisons, we also implement a variant of it denoting as UOOD . UOOD which uses discrepancy loss to perform model selection instead of the performance on an extra OOD validation set.

Evaluation Metrics. Follow the settings used by previous studies [43, 39, 29], we evaluate our approach with five common metrics: AUROC, FPR at 95% TPR, Detection Error, AUPR-In, and AUPR-Out. More details about evaluation metrics are presented in the supplementary material.

Implementation Details. In all experiments, we adopt the Densenet-BC [20] as the backbone since it is widely used in previous studies [43, 39, 29]. Our backbone is trained by SOTA contrastive learning method SimCLR [6] for 500 epochs. We set the learning rate to 10^{-3} with a cosine annealing strategy. For fair comparisons, each comparison method can use the pre-trained backbone model. MAH [27] uses the features from different layers extracted from the pre-trained backbone model. A well-trained linear classifier with a pre-trained backbone model is provided for ODIN [29] and UOOD [43]. The hyper-parameter K for STEP is set to 12 for all data set pairs. All experiments are performed on one single NVIDIA 3090 graphics card. More details on implementation are provided in the supplementary material.

[1] https://github.com/ShiyuLiang/odin-pytorch

3.2 Experiment Results

OOD Detection Performance. We evaluate STEP with compared methods on various OOD benchmarks. Analyzed by five common metrics, the results are shown in Tab.(1). From the results, we observe that ODIN suffers from severe performance degradation. Moreover, its performance is close to random guessing in some cases. The limitation of labeled data mainly causes this. We can hardly train a high-quality classification model to provid accurate logits for ODIN . Hence, ODIN can not give the correct judgment based on inaccurate logits. Our STEP approach outperforms methods that do not heavily rely on an OOD validation set by a large margin. Even compared with those methods that heavily rely on the OOD validation set, such as UOOD † and MAH †, our STEP approach is still better than them in most cases. However, a good OOD validation set is expensive and nearly impossible to build in the real world. The number of OOD samples can be infinitely many, and a fixed-size validation set cannot capture the complete OOD information. Therefore, introducing the validation set during training will reduce the model's generalization in the real environment. We will verify this in detail in subsequent experiments.

Table 1: Performance comparison on various OOD benchmarks evaluated by 5 common metrics. Methods with † use extra OOD validation set. The best results are indicated in bold. Our approach outperforms other methods in most cases, even though they use an extra OOD validation set.

Metrics	ID Dataset	OOD Dataset	ODIN	MAH †	UOOD	UOOD †	STEP
AUROC ↑	Cifar10	TINc	81.00 ± 6.30	87.67 ± 2.47	90.46 ± 9.74	99.07 ± 0.48	$\mathbf{99.99 \pm 0.00}$
		TINr	59.10 ± 2.08	86.88 ± 0.87	84.67 ± 9.41	92.63 ± 3.42	$\mathbf{95.61 \pm 0.36}$
		LSUNc	76.17 ± 5.37	97.68 ± 0.09	96.92 ± 2.04	98.79 ± 0.67	$\mathbf{99.99 \pm 0.00}$
		LSUNr	69.05 ± 3.49	90.41 ± 1.00	80.87 ± 24.45	97.81 ± 0.94	$\mathbf{99.07 \pm 0.20}$
	Cifar100	TINc	61.65 ± 6.71	71.15 ± 2.20	98.34 ± 1.57	98.84 ± 0.83	$\mathbf{99.99 \pm 0.01}$
		TINr	54.46 ± 0.74	73.94 ± 1.79	84.80 ± 8.87	$\mathbf{95.31 \pm 0.93}$	93.51 ± 1.17
		LSUNc	46.99 ± 4.99	93.91 ± 3.41	97.49 ± 1.48	99.31 ± 0.62	$\mathbf{99.99 \pm 0.00}$
		LSUNr	52.06 ± 2.24	78.45 ± 1.11	97.61 ± 0.55	$\mathbf{98.96 \pm 0.40}$	98.20 ± 0.56
FPR at 95%TPR ↓	Cifar10	TINc	53.37 ± 10.55	44.17 ± 6.43	29.35 ± 30.05	2.75 ± 1.65	$\mathbf{0.00 \pm 0.00}$
		TINr	89.76 ± 1.45	58.57 ± 3.09	31.72 ± 11.50	19.61 ± 9.50	$\mathbf{17.63 \pm 1.10}$
		LSUNc	64.06 ± 9.12	7.73 ± 0.46	6.59 ± 3.22	3.56 ± 1.93	$\mathbf{0.00 \pm 0.00}$
		LSUNr	76.89 ± 5.04	45.41 ± 3.87	32.69 ± 31.93	6.49 ± 2.89	$\mathbf{4.48 \pm 1.02}$
	Cifar100	TINc	84.24 ± 8.02	90.15 ± 1.99	5.22 ± 5.59	3.16 ± 2.25	$\mathbf{0.00 \pm 0.01}$
		TINr	90.10 ± 0.46	80.55 ± 1.89	29.09 ± 15.68	$\mathbf{11.10 \pm 4.21}$	23.21 ± 4.14
		LSUNc	93.49 ± 2.42	24.93 ± 21.75	6.24 ± 3.80	1.93 ± 2.43	$\mathbf{0.00 \pm 0.00}$
		LSUNr	89.79 ± 0.79	69.69 ± 2.42	4.92 ± 1.33	$\mathbf{2.39 \pm 0.74}$	8.25 ± 3.14
Detection Error ↓	Cifar10	TINc	25.53 ± 4.67	19.93 ± 2.63	11.59 ± 11.35	2.54 ± 1.27	$\mathbf{0.12 \pm 0.01}$
		TINr	43.04 ± 1.48	20.14 ± 0.82	18.07 ± 5.55	11.71 ± 4.56	$\mathbf{10.77 \pm 0.52}$
		LSUNc	29.57 ± 3.82	6.28 ± 0.25	4.20 ± 2.12	2.58 ± 1.32	$\mathbf{0.11 \pm 0.01}$
		LSUNr	35.52 ± 2.46	16.23 ± 0.95	18.40 ± 15.68	4.99 ± 1.91	$\mathbf{4.66 \pm 0.57}$
	Cifar100	TINc	40.95 ± 5.07	32.58 ± 1.64	3.67 ± 3.62	2.76 ± 1.00	$\mathbf{0.32 \pm 0.06}$
		TINr	46.36 ± 0.56	31.09 ± 1.44	16.53 ± 7.87	$\mathbf{6.88 \pm 2.33}$	13.26 ± 1.61
		LSUNc	48.47 ± 1.61	11.20 ± 3.73	4.24 ± 2.34	2.06 ± 1.54	$\mathbf{0.23 \pm 0.04}$
		LSUNr	46.73 ± 0.66	27.33 ± 1.03	3.11 ± 0.78	$\mathbf{1.90 \pm 0.51}$	6.40 ± 1.32
AUPR-In ↑	Cifar10	TINc	76.80 ± 8.20	85.35 ± 2.86	89.31 ± 10.05	98.59 ± 0.67	$\mathbf{99.99 \pm 0.00}$
		TINr	57.10 ± 2.11	86.79 ± 1.17	79.02 ± 12.17	88.72 ± 4.93	$\mathbf{94.71 \pm 0.51}$
		LSUNc	72.16 ± 6.60	96.70 ± 0.21	94.78 ± 4.07	98.31 ± 0.92	$\mathbf{100.00 \pm 0.00}$
		LSUNr	65.37 ± 3.39	89.93 ± 1.23	79.41 ± 19.89	96.86 ± 1.27	$\mathbf{99.02 \pm 0.20}$
	Cifar100	TINc	58.29 ± 5.01	71.18 ± 2.69	97.55 ± 2.04	98.24 ± 1.50	$\mathbf{99.99 \pm 0.01}$
		TINr	52.96 ± 0.59	70.95 ± 2.20	77.32 ± 9.81	91.67 ± 1.29	$\mathbf{91.91 \pm 1.34}$
		LSUNc	47.41 ± 2.86	92.26 ± 2.17	95.45 ± 2.32	99.09 ± 0.88	$\mathbf{99.99 \pm 0.00}$
		LSUNr	50.47 ± 1.75	74.22 ± 1.14	95.53 ± 0.95	$\mathbf{98.11 \pm 0.78}$	98.07 ± 0.52
AUPR-Out ↑	Cifar10	TINc	83.63 ± 5.11	88.67 ± 2.28	91.34 ± 8.69	99.32 ± 0.35	$\mathbf{99.99 \pm 0.00}$
		TINr	58.83 ± 1.77	84.26 ± 0.95	89.21 ± 6.22	94.60 ± 2.70	$\mathbf{96.31 \pm 0.28}$
		LSUNc	78.43 ± 5.12	98.16 ± 0.12	98.01 ± 1.18	99.14 ± 0.48	$\mathbf{99.99 \pm 0.00}$
		LSUNr	70.51 ± 3.97	88.84 ± 1.20	84.45 ± 21.48	98.41 ± 0.70	$\mathbf{99.14 \pm 0.19}$
	Cifar100	TINc	62.88 ± 7.90	65.14 ± 2.21	98.77 ± 1.23	99.08 ± 0.51	$\mathbf{99.99 \pm 0.01}$
		TINr	55.94 ± 0.71	71.57 ± 1.71	89.44 ± 6.96	$\mathbf{96.84 \pm 0.82}$	94.66 ± 1.07
		LSUNc	49.91 ± 4.42	93.77 ± 5.30	98.33 ± 0.99	99.39 ± 0.48	$\mathbf{99.99 \pm 0.00}$
		LSUNr	55.18 ± 1.56	78.19 ± 1.33	98.49 ± 0.37	$\mathbf{99.32 \pm 0.24}$	98.35 ± 0.56

Generalization of OOD Detection. Our STEP approach and some previous studies, e.g., UOOD , MAH , could utilize some OOD samples during the training phase. For example, our STEP performs contrastive learning on both ID and OOD data. UOOD optimizes the discrepancy loss on ID and OOD unlabeled data and uses an OOD validation set to select the final model. MAH tunes their weighting parameters in the OOD validation set. We denote the OOD samples that the model somehow used during the training phase as known OOD samples, contrasting to unknown OOD samples. We want to explore whether the use of known OOD samples will reduce the OOD detection performance of the model on unknown OOD samples. Therefore, we design a novel experimental method. We train the OOD detection model with ID data set and one OOD data set while testing the model with ID data set and a different OOD data set. As an example, we train the model on the ID data set (Cifar10) and OOD data set (TINr) and replace all OOD samples in the test set with a new OOD data set (TINc) when testing. An OOD detection model with strong generalization should obtain consistent performance, no matter what OOD data set we used to construct the testing set. We tested four OOD detection methods on CIFAR-10 with two different OOD data set pairs. From the results shown in Fig.(2), we found that UOOD † and MAH † have severe performance degradation when detecting unknown OOD samples. This phenomenon is because the OOD validation set used by these methods introduces a severe bias to their models. Furthermore, we also conducted experiments to analyze the relationship between OOD detection performance and loss, and verified the instability in the training process of the SOTA UOOD † method. We put this experiment in the supplementary material. ODIN's performance changes very randomly, which is also in line with expectations because it has only seen ID data during the training process. Our STEP approach gives a very high and relatively close performance on both known and unknown OOD data sets, which proves the effectiveness and strong generalization of our approach. Further, we suggest that the experimental method proposed should be verified in all future OOD detection studies that use OOD samples in the training process.

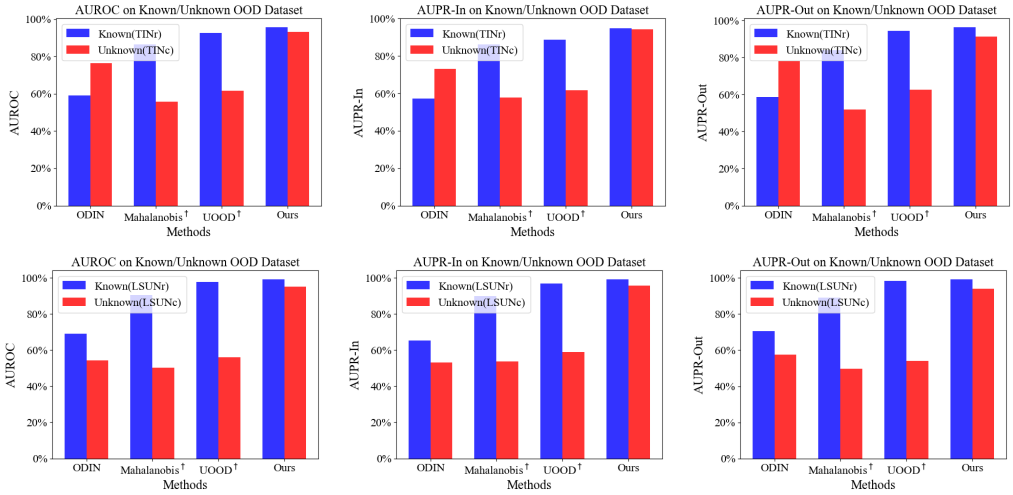


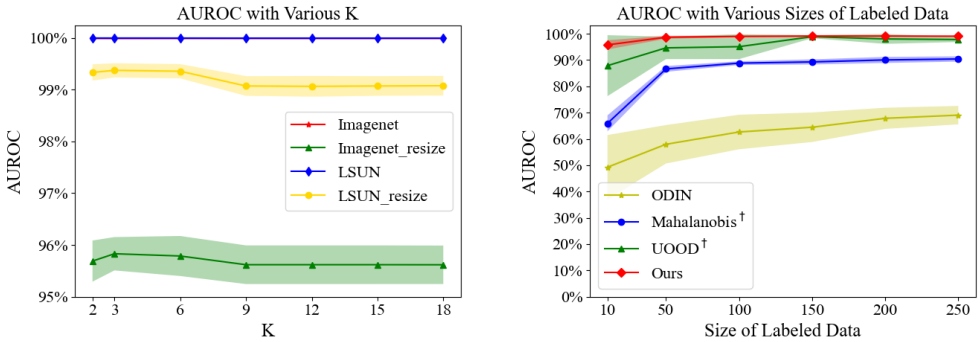
Figure 2: Performance of different methods on Known / Unknown OOD data set evaluated by various metrics. The results shows that our STEP approach not only has very good OOD detection performance, but also can generalize to unknown OOD samples.

Ablation Study. As introduced in Section 2, our STEP approach contains four components in total: MAH, KNN, Unzipping, and Structure-Keep. We conducted comprehensive ablation studies to verify the effectiveness of each component. As shown in Tab.(2), we sequentially add the components of STEP and verify the performance of each model on two OOD benchmarks. The first line in the table shows the results of directly distinguishing the minimum Mahalanobis distance from the target sample to each class center. Since necessary $\hat{\Sigma}$ cannot be accurately estimated, the detection performance is not ideal. The second line proves that the geodetic distance can alleviate the inaccurate estimation problem to a certain extent, thereby improving the detection performance. The third line is the incomplete version of our STEP approach to remove Structure-Keep. The result of this line proves that the Structure-Keep technique is very important. Otherwise, the detection performance will be greatly reduced. The fourth line, our STEP approach, gives the best results. This proves that the four steps proposed in this article can only be integrated together to get the best results.

Table 2: Ablation Study of our STEP approach evaluated by AUROC. This table proves that every part of our approach is indispensable.

Different parts of STEP				Data set pair	
MAH	KNN	Unzipping	Sturture-Keep	Cifar10-TINr	Cifar10-LSUNr
✓				90.96 ± 0.28	93.46 ± 0.51
✓	✓			91.26 ± 1.74	97.35 ± 0.45
✓	✓	✓		79.58 ± 0.69	80.38 ± 0.95
✓	✓	✓	✓	95.62 ± 0.39	99.07 ± 0.20

Robustness. In this paragraph, we verify the robustness of STEP to hyper-parameter K and the number of labeled data $|\mathcal{D}_l|$. We test the performance of STEP on different OOD data sets for different choices of K in a large range from 2 to 18. Fig.(3a) shows that STEP is not sensitive to the hyper-parameter K (the number of neighbors when KNN is built). Furthermore, we find that choosing a smaller K helps improve the detection performance. Then we test how the amount of labeled ID data affects the performance of different methods. From the results shows in Fig.(3b), we find that our STEP is very tolerant of the amount of labeled ID data. Even in the case of extremely insufficient labeled ID data, an acceptable performance still can be achieved by our STEP approach.



(a) AUROC with various K on different OOD benchmarks.

(b) AUROC of different methods with various sizes of labeled data.

Figure 3: The robustness of STEP approach. (a), (b) show that STEP approach is robust on K and size of labeled data, respectively.

4 Related Work

This work is mainly related to self-supervised learning, semi-supervised learning and OOD detection.

Self-supervised learning Self-supervised learning is a powerful framework to learn discriminative feature representations in an unsupervised fashion via artificially designed auxiliary tasks. Recently, contrastive learning [33, 16, 6] shows remarkable progress on it. Benefited from the progress, some studies [38, 14] utilize the learned powerful representations to cluster samples with unseen labels. STEP proposed in this paper takes advantage of the powerful features derived from the use of contrastive learning. Any progress in comparative learning can be used by STEP to further improve OOD detection performance.

Semi-supervised learning. SSL [5] aims to leverage unlabeled data to improve the performance of the model when plenty of labeled data is inconvenient and expensive to access. Our paper is mainly related to deep SSL. The combining of SSL technology and DNNs has significantly improved classification accuracy. Many excellent studies, such as consistency regularization based methods [37, 31], entropy minimization based methods [11] and holistic methods [3, 35], have been proposed in recent years. There are also some studies [13, 44] that focus on improving the safeness of SSL. Specifically, they aim to ensure the performance of SSL when unlabeled data contains OOD samples. However, these studies all consider the classification performance of the model for known categories

under the semi-supervised setting and ignore the problem of overconfidence in the OOD sample when testing. Efforts on this issue remain limited. Therefore, we propose the semi-supervised OOD detection setting and design STEP approach for it.

OOD detection. OOD detection has been studied for a long history. The baseline [17] of this problem attempts to detect OOD samples depending on the predicted softmax class probability. Modified generative adversarial networks [26] are used to generate challenging OOD samples during the training stage, and the algorithm encourages the classifier to assign OOD samples uniform class probabilities. ODIN [29] applies the temperature scaling and input preprocessing to further strengthen the difference between ID samples and OOD samples. ELOC [39] uses the ensemble of K leave-out classifiers to detect OOD samples. There are some other studies that use energy-based models [30, 12], hierarchical relations [22, 23], and so on. The current state-of-the-art method [43] for OOD detection utilizes the discrepancy between two classifiers to separate ID and OOD samples. Nevertheless, these studies either assume that there is an accurate classification model or assume that there is sufficient labeled data, which limits their application in the real world. There are also some unsupervised OOD detection studies [36, 18, 2, 34] utilizing the power of the contrastive learning framework. However, although these studies do not require labels, they still need a large amount of ID data for training. Previous studies [13] have reported that collecting clean unlabeled data is also very difficult in the real world. Hence, in this paper, we study a more general setting that is very common in real-world applications.

5 Conclusions

In this paper, we propose a novel OOD detection setting, called semi-supervised OOD detection. In this setting, we aim to distinguish ID and OOD samples by using limited ID labeled data and large amounts of mixed unlabeled data. Due to the generality of this setting, it commonly occurs in real-world applications. In the case of only having limited ID labeled data, we found that the previous studies have suffered performance degeneration, mainly due to the inaccurate estimation from the limited ID data. Focusing on this setting, we propose a novel STEP approach. Our main idea is to detect OOD samples in a detection-specific space where we maintain the same local topological structures as the original feature space. Our STEP approach outperforms other methods by a large margin in most cases and achieves remarkable detection performance on several benchmarks. Meanwhile, we also conduct comprehensive experiments to verify the robustness and generalization of our STEP approach. The limitation of our work is the lack of solid theoretical results. Broadly speaking, other OOD detection methods also have similar problems. We will put efforts into the theoretically understanding of OOD detection in future work.

Broader Impact

In this work, we study OOD detection, which is a fundamental problem in deep learning. Specifically, we first proposed a novel OOD detection setting. In this setting, only limited labeled ID data and many mixed unlabeled data can be used for OOD detection. This is a novel and practical setting commonly appearing in real-world applications because under this setting, we neither require a large amount of labeled data nor clean unlabeled data. We propose the STEP approach to detect OOD samples in a detection-specific space, greatly improving the performance of OOD detection. Our work will give instructions for those applications having difficulties collecting large quantities of pure ID labeled data while demanding detecting OOD samples to prevent potential dangers in real-world applications. At the same time, there is still much room for exploration in this setting. We hope our work can inspire more discussions about OOD detection in real scenarios and drive more researchers to build practical and robust OOD detection algorithms.

Meanwhile, we are aware that abuse of this technology can pose ethical issues. In particular, we note that people expect that real people rather than algorithms make the judgments behind the system. Despite the risks of such AI research, developing and demonstrating such technologies is essential to understand the technology's practical and potentially troublesome applications. We hope that the responsible use of technology will stimulate discussion about these methods' practices and controls.

References

[1] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al. Deep speech 2 : End-to-end speech recognition in english and mandarin. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 173–182, 2016.

[2] Liron Bergman and Yedid Hoshen. Classification-based anomaly detection for general data. In *Proceedings of the 8th International Conference on Learning Representations*, pages 1–12, 2020.

[3] David Berthelot, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. Remixmatch: Semi-supervised learning with distribution matching and augmentation anchoring. In *Proceedings of the 8th International Conference on Learning Representations*, pages 1–13, 2020.

[4] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1721–1730, 2015.

[5] Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien. *Semi-Supervised Learning*. MIT Press, 2006.

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, pages 1597–1607, 2020.

[7] Gobinda G Chowdhury. Natural language processing. *Annual Review of Information Science and Technology*, 37(1):51–89, 2003.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[9] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1625–1634, 2018.

[10] Geli Fei and Bing Liu. Breaking the closed world assumption in text classification. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 506–514, 2016.

[11] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In *Advances in Neural Information Processing Systems*, pages 529–536, 2004.

[12] Will Grathwohl, Kuan-Chieh Wang, Joern-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model and you should treat it like one. In *Proceedings of the 8th International Conference on Learning Representations*, pages 1–23, 2020.

[13] 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 *Proceedings of the 37th International Conference on Machine Learning*, pages 3897–3906, 2020.

[14] Kai Han, Sylvestre-Alvise Rebuffi, Sebastien Ehrhardt, Andrea Vedaldi, and Andrew Zisserman. Automatically discovering and learning new visual categories with ranking statistics. In *Proceedings of the 8th International Conference on Learning Representations*, pages 1–13, 2020.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[16] 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.

[17] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *Proceedings of the 5th International Conference on Learning Representations*, pages 1–12, 2017.

[18] Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. Using self-supervised learning can improve model robustness and uncertainty. In *Advances in Neural Information Processing Systems*, pages 15637–15648, 2019.

[19] Yen-Chang Hsu, Yilin Shen, Hongxia Jin, and Zsolt Kira. Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10951–10960, 2020.

[20] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4700–4708, 2017.

[21] Jeff Johnson, Matthijs Douze, and Herve Jegou. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, pages 1–1, 2019.

[22] Josef Kittler and Cemre Zor. Delta divergence: A novel decision cognizant measure of classifier incongruence. *IEEE Transactions on Cybernetics*, 49(6):2331–2343, 2018.

[23] Josef Kittler, Cemre Zor, Ioannis Kaloskampis, Yulia Hicks, and Wenwu Wang. Error sensitivity analysis of delta divergence-a novel measure for classifier incongruence detection. *Pattern Recognition*, 77:30–44, 2018.

[24] Alex Krizhevsky and Hinton Geoffrey. Learning multiple layers of features from tiny images. *Tech Report*, 2009.

[25] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

[26] Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. In *Proceedings of the 6th International Conference on Learning Representations*, pages 1–16, 2018.

[27] 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*, pages 7167–7177, 2018.

[28] Yu-Feng Li and De-Ming Liang. Safe semi-supervised learning: a brief introduction. *Frontiers of Computer Science*, 13(4):669–676, 2019.

[29] Shiyu Liang, Yixuan Li, and R Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *Proceedings of the 6th International Conference on Learning Representations*, pages 1–15, 2018.

[30] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *Advances in Neural Information Processing Systems*, pages 21464–21475, 2020.

[31] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):1979–1993, 2018.

[32] Avital Oliver, Augustus Odena, Colin A Raffel, Ekin Dogus Cubuk, and Ian Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. In *Advances in Neural Information Processing Systems*, pages 3235–3246, 2018.

[33] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[34] Vikash Sehwag, Mung Chiang, and Prateek Mittal. Ssd: A unified framework for self-supervised outlier detection. In *Proceedings of the 9th International Conference on Learning Representations*, pages 1–17, 2021.

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

[36] Jihoon Tack, Sangwoo Mo, Jongheon Jeong, and Jinwoo Shin. Csi: Novelty detection via contrastive learning on distributionally shifted instances. In *Advances in Neural Information Processing Systems*, pages 11839–11852, 2020.

[37] 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*, pages 1195–1204, 2017.

[38] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. Scan: Learning to classify images without labels. In *Proceedings of the European Conference on Computer Vision*, pages 268–285, 2020.

[39] Apoorv Vyas, Nataraj Jammalamadaka, Xia Zhu, Dipankar Das, Bharat Kaul, and Theodore L Willke. Out-of-distribution detection using an ensemble of self supervised leave-out classifiers. In *Proceedings of the European Conference on Computer Vision*, pages 550–564, 2018.

[40] Pengxiang Wu, Songzhu Zheng, Mayank Goswami, Dimitris Metaxas, and Chao Chen. A topological filter for learning with label noise. In *Advances in Neural Information Processing Systems*, pages 21382–21393, 2020.

[41] Haiqin Yang, Shenghuo Zhu, Irwin King, and Michael R Lyu. Can irrelevant data help semi-supervised learning, why and how? In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, pages 937–946, 2011.

[42] 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.

[43] Qing Yu and Kiyoharu Aizawa. Unsupervised out-of-distribution detection by maximum classifier discrepancy. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9517–9525, 2019.

[44] 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*, pages 438–454, 2020.

[45] Zhi-Hua Zhou. A brief introduction to weakly supervised learning. *National Science Review*, 5(1):44–53, 2018.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] We have discussed the limitations in Section 5.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] We have discussed it in Section Broader Impact.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] The code will be open source once the paper is accepted. The data and implement details have been discussed in Section 3.1 and supplemental material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] They have been discussed in Section 3.1 and supplemental material.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] They have been discussed in Section 3.1.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]