Exploiting Mixed Unlabeled Data for Detecting Samples of Seen and Unseen Out-of-Distribution Classes

Yi-Xuan Sun, Wei Wang*

National Key Laboratory for Novel Software Technology
Nanjing University, Nanjing 210023, China
{sunyixuan, wangw}@lamda.nju.edu.cn

Abstract

Out-of-Distribution (OOD) detection is essential in real-world applications, which has attracted increasing attention in recent years. However, most existing OOD detection methods require many labeled In-Distribution (ID) data, causing a heavy labeling cost. In this paper, we focus on the more realistic scenario, where limited labeled data and abundant unlabeled data are available, and these unlabeled data are mixed with ID and OOD samples. We propose the Adaptive In-Out-aware Learning (AIOL) method, in which we employ the appropriate temperature to adaptively select potential ID and OOD samples from the mixed unlabeled data and consider the entropy over them for OOD detection. Moreover, since the test data in realistic applications may contain OOD samples whose classes are not in the mixed unlabeled data (we call them unseen OOD classes), data augmentation techniques are brought into the method to further improve the performance. The experiments are conducted on various benchmark datasets, which demonstrate the superiority of our method.

Introduction

Deep neural networks (DNNs) have achieved great success in various applications, but the success heavily relies on the assumption that the training and test data are drawn from the same distribution. In realistic scenarios, however, some Out-of-Distribution (OOD) samples may lead DNNs to make completely incorrect predictions, which is essentially harmful in many real-world applications, e.g., autonomous driving or medical diagnosis. Therefore, it is demanded that the trained model can at least correctly detect these OOD samples during the inference process, and then human intervention can be involved to deal with them.

Recently, Hendrycks and Gimpel (2017) considered the OOD detection problem and proposed a baseline method with the output confidence, i.e., the maximum softmax probability. The method is based on the observation that In-Distribution (ID) samples tend to have higher output confidence than OOD samples. Some other methods (Liang, Li, and Srikant 2018; Lee et al. 2018; Sastry and Oore 2020; Hsu et al. 2020) made further improvements with some post-hoc techniques. However, these methods need many labeled ID data for training, causing a heavy labeling cost.

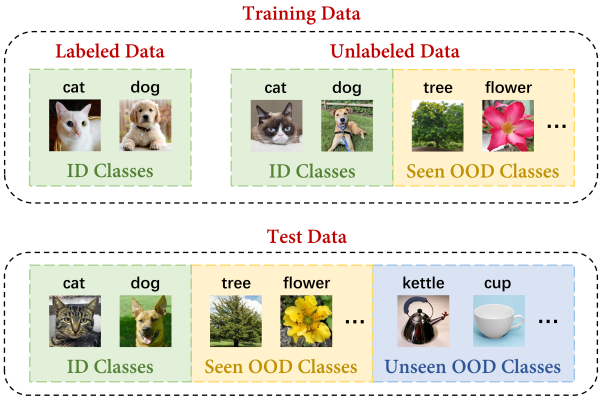


Figure 1: OOD detection with labeled and mixed unlabeled data. Unlabeled data are mixed with samples of ID classes (*cat* and *dog*) and seen OOD classes (*tree* and *flower*). Test data contain samples of ID classes (*cat* and *dog*), seen OOD classes (*tree* and *flower*) and unseen OOD classes (*kettle* and *cup* which are not in the mixed unlabeled data).

To mitigate the labeling overload, the advanced works tried to utilize abundant unlabeled data. Some methods (Hendrycks et al. 2019; Tack et al. 2020; Sehwag, Chiang, and Mittal 2021) employed self-supervised learning on the pure unlabeled ID data, while some other methods (Hendrycks, Mazeika, and Dietterich 2019; Liu et al. 2020) were proposed to exploit the pure unlabeled OOD data. However, these methods require that the unlabeled data must be pure ID or OOD, which is hardly met in realistic applications. Recently, a few works (Chen et al. 2020b; Yu et al. 2020; Guo et al. 2020) attempted to utilize unlabeled data that consist of ID and OOD samples. But these works were not developed for OOD detection, and there was no OOD sample in the test data during the inference process.

For the OOD detection problem, the unlabeled data can be mixed with ID and OOD samples. During the inference process, the test data may contain OOD samples whose classes are in the mixed unlabeled data (we call them seen OOD classes), as well as OOD samples whose classes are not in the mixed unlabeled data (we call them unseen OOD classes). The goal is to train a model which can classify ID samples and detect all OOD samples. This OOD detection

*Corresponding author.

problem for the image recognition task is summarized in Figure 1. A similar task exists in the medical diagnosis of lung diseases, where a model is trained with *Computerized Tomography* (CT) images of the lung. Due to the heavy labeling cost, only limited labeled images (three ID classes include *normal*, *pneumonia*, and *asthma*) and abundant unlabeled ones are available for the diagnosis of lung diseases. The images of other diseases, e.g., *lung cancer*, can also be collected as unlabeled data. During the diagnosis process, the trained model should not only detect samples of seen OOD classes, e.g., *lung cancer*, but also detect samples of unseen OOD classes, e.g., emerging disease *COVID-19*. It is desirable to develop new methods for this realistic OOD detection problem.

In this paper, we focus on the problem discussed above and propose the Adaptive In-Out-aware Learning (AIOL) method, which can utilize limited labeled data and abundant mixed unlabeled data for OOD detection. Specifically, we employ the appropriate temperature and learn the mixture probabilistic model to adaptively select the ID and OOD samples from the mixed unlabeled data. Then, the entropy over the selected samples is considered to make them more distinguishable. Moreover, since the test data may contain samples of unseen OOD classes, data augmentation techniques are brought into the method to enhance the model's generalization capability. We verify the effectiveness of our method on various benchmark datasets, and the results show that our method outperforms the compared methods.

Related Works

Out-of-Distribution (OOD) detection aims to detect OOD samples during the inference process. A baseline method (Hendrycks and Gimpel 2017) was proposed for detecting OOD samples with the output confidence. Later, some methods (Liang, Li, and Srikant 2018; Lee et al. 2018; Sastry and Oore 2020; Hsu et al. 2020) built the advanced detectors in a post-hoc manner. For instance, Liang, Li, and Srikant (2018) combined temperature scaling and input preprocessing to achieve better detection performance. Instead of using the output confidence, Lee et al. (2018) utilized the Mahalanobis distance between the test samples' feature representations and the train samples'. However, these methods require many labeled ID samples for training.

There were some methods that focused on how to exploit unlabeled data for OOD detection. Hendrycks, Mazeika, and Dietterich (2019) enforced the model to produce the low-confidence output on the pure unlabeled OOD data. Some other works (Golan and El-Yaniv 2018; Hendrycks et al. 2019; Winkens et al. 2020; Tack et al. 2020; Sehwag, Chiang, and Mittal 2021) found that self-supervised learning on the pure unlabeled ID data could improve the detection performance. For instance, Sehwag, Chiang, and Mittal (2021) combined contrastive learning (Chen et al. 2020a) and Mahalanobis distance for OOD detection. There was also a line of works (Nalisnick et al. 2019; Huang et al. 2019; Serrà et al. 2019) which employed deep generative models on the pure unlabeled ID data. However, all these methods require that the unlabeled data must be pure ID or OOD, which is hardly met in realistic applications.

Recently, some methods (Chen et al. 2020b; Yu et al. 2020; Guo et al. 2020) considered the class distribution mismatch between labeled and unlabeled data. The mismatched samples in the unlabeled data can be regarded as OOD samples. Chen et al. (2020b) filtered out OOD samples in the unlabeled data with a confidence threshold and trained the model on the remaining data only. Yu et al. (2020) proposed a joint optimization framework to classify ID samples and filter out OOD samples concurrently. Guo et al. (2020) employed the bi-level optimization to weaken the weights of OOD samples. But these methods were developed for ID classification and there was no OOD sample during the inference process. Another work (Yu and Aizawa 2019) tried to utilize mixed unlabeled data for OOD detection. It encouraged two classifiers to maximally disagree on the mixed unlabeled data. However, each unlabeled sample was treated equally, hence the model still needed many labeled samples to distinguish between ID and OOD samples.

Semi-Supervised Learning (SSL) methods (Grandvalet and Bengio 2004; Lee 2013; Berthelot et al. 2019; Xie et al. 2019; Sohn et al. 2020) were also developed for utilizing limited labeled data and abundant unlabeled data. These methods focused on the classification performance and usually ignored the existence of OOD samples.

Method

In this paper, we consider OOD detection with labeled and unlabeled data. Let $L = \{(\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_n, y_n)\}$ be labeled data, and $U = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_m\}$ be unlabeled data. Here, $\boldsymbol{x} \in \mathcal{X} \subseteq \mathbb{R}^d$, $y \in \mathcal{Y} = \{1, \cdots, K\}$, d is the number of the input dimension, K is the number of classes in the labeled data, and $m \gg n$. For the OOD detection problem, the true class of the sample in U may not belong to $\mathcal{Y} = \{1, \cdots, K\}$. This kind of sample is called OOD sample and is generally denoted as $\boldsymbol{x}^{out}$, while the sample whose true class belongs to $\mathcal{Y} = \{1, \cdots, K\}$ is denoted as $\boldsymbol{x}^{in}$. In this way, the labeled data can be rewritten as $L = \{(\boldsymbol{x}_1^{in}, y_1), \cdots, (\boldsymbol{x}_n^{in}, y_n)\}$. Let the ID part of U be $U^{in} = \{\boldsymbol{x}_1^{in}, \cdots, \boldsymbol{x}_{m_1}^{in}\}$, and the OOD part of U be $U^{out} = \{\boldsymbol{x}_1^{out}, \cdots, \boldsymbol{x}_{m_2}^{out}\}$, hence $U = U^{in} \cup U^{out}$. The goal of OOD detection is to learn a model f^* with L and U, i.e., $f^*(\boldsymbol{x}^{in}) = y$ ($y \in \mathcal{Y}$ is the true class of $\boldsymbol{x}^{in}$) for $\boldsymbol{x}^{in}$ in the test data, while $f^*(\boldsymbol{x}^{out}) = \perp$ ($\perp$ means the true class of $\boldsymbol{x}^{out}$ does not belong to $\mathcal{Y}$) for $\boldsymbol{x}^{out}$ in the test data. In realistic applications, the true class of $\boldsymbol{x}^{out}$ in the test data may not be included in the mixed unlabeled data U, which is referred to as the unseen OOD class shown in Figure 1.

We try to learn a deep neural network f_θ with parameters θ for OOD detection. Intuitively, we could utilize the labeled data L and the mixed unlabeled data U to obtain a basic model. For the labeled data L, we use the supervised cross-entropy loss (denoted as CE):

$$\mathcal{L}_S = \frac{1}{|L|} \sum_{(\boldsymbol{x},y)\in L} \mathrm{CE}\big(\boldsymbol{y} \parallel q_\theta(\boldsymbol{x})\big), \tag{1}$$

where $\boldsymbol{y}$ is the K-dimensional one-hot label constructed with y, and $q_\theta(\boldsymbol{x}) \in [0,1]^K$ is the output probability distribution after the softmax layer for $\boldsymbol{x}$. As for the unlabeled

data, consistency regularization (Xie et al. 2019; Sohn et al. 2020) over $U = U^{in} \cup U^{out}$ can be formulated as:

$$\mathcal{L}_{CR} = \frac{1}{|U|} \sum_{\boldsymbol{x} \in U} \mathrm{CE}\Big(q_\theta\big(\mathcal{A}(\boldsymbol{x}), T\big) \parallel q_\theta\big(\mathcal{A}'(\boldsymbol{x})\big)\Big), \quad (2)$$

where $\mathcal{A}(\cdot)$ and $\mathcal{A}'(\cdot)$ are different data augmentations. The soft target in Eq (2) is usually scaled with temperature T:

$$q_\theta^{(i)}(\boldsymbol{x}, T) = \frac{\exp\left(z_i/T\right)}{\sum_{j=1}^{K} \exp\left(z_j/T\right)}, \quad (3)$$

where z_i is the output logit of class i for $\boldsymbol{x}$. For the OOD samples in U^{out}, $T < 1$ will sharpen the output probability distribution, leading to a high probability of identifying OOD samples as ID samples, while $T > 1$ will soften the output probability distribution, leading to a low probability of identifying OOD samples as ID samples. Therefore, temperature T should be larger than 1 in order to exploit the OOD samples in U^{out}. As for the ID samples in U^{in}, although $T < 1$ was used in previous works (Xie et al. 2019; Sohn et al. 2020) to encourage the high-confidence output, we still set $T > 1$ in Eq (2) since exploiting the OOD samples in U plays an important role in OOD detection. Actually, we can provide the following Theorem 1 to show that $T > 1$ will push the output confidence of ID and OOD samples further apart from each other.

Theorem 1. *Let $C_\theta(\boldsymbol{x}, T) = q_\theta^{(\hat{y})}(\boldsymbol{x}, T)$ be the output confidence of the sample $\boldsymbol{x}$ with temperature T, where $\hat{y}$ is the predicted class on $\boldsymbol{x}$. Under $K = 2$, for the samples $\boldsymbol{x}^{in}$ and $\boldsymbol{x}^{out}$ that satisfy $C_\theta(\boldsymbol{x}^{in}, 1) > C_\theta(\boldsymbol{x}^{out}, 1)$:*
if $c > T_1 > T_2 \geq 1$, we have

$$C_\theta(\boldsymbol{x}^{in}, T_1) - C_\theta(\boldsymbol{x}^{out}, T_1) > C_\theta(\boldsymbol{x}^{in}, T_2) - C_\theta(\boldsymbol{x}^{out}, T_2); \quad (4)$$

if $0 < T_1 < T_2 \leq 1$, we have

$$C_\theta(\boldsymbol{x}^{in}, T_1) - C_\theta(\boldsymbol{x}^{out}, T_1) < C_\theta(\boldsymbol{x}^{in}, T_2) - C_\theta(\boldsymbol{x}^{out}, T_2). \quad (5)$$

Here, the constant c depends on $C_\theta(\boldsymbol{x}^{in}, 1)$ and $C_\theta(\boldsymbol{x}^{out}, 1)$.

Remark. Previous works (Hendrycks and Gimpel 2017; Liang, Li, and Srikant 2018; Hsu et al. 2020) have shown that ID samples tend to have higher output confidence than OOD samples, i.e., $C_\theta(\boldsymbol{x}^{in}, 1) > C_\theta(\boldsymbol{x}^{out}, 1)$ holds for most ID and OOD samples. For these ID and OOD samples, Theorem 1 tells that the output confidence gap between them relates to temperature T, and $T > 1$ will make the gap larger. Due to space constraints, we postpone the omitted proof of Theorem 1 to Appendix A in the supplementary material.

For $K \geq 3$, it is difficult to analyze the case since $C_\theta(\boldsymbol{x}, T)$ depends on the exponential functions. Intuitively, similar results to Theorem 1 can be got for $K \geq 3$. For a sample $\boldsymbol{x}$, the output probability distribution of the trained neural network is denoted as $\boldsymbol{p} = (p_1, \cdots, p_K)$, where $p_i = \frac{\exp(z_i)}{\sum_{j=1}^{K} \exp(z_j)}$. In deep learning, neural networks are trained on the data with one-hot labels. Because of the expressive power of neural networks, the output $\boldsymbol{p}$ is usually close to a one-hot vector. Let $\tilde{\boldsymbol{p}} = (\tilde{p}_1, \cdots, \tilde{p}_K)$ be the decreasingly sorted version of $\boldsymbol{p}$, where $\tilde{p}_1 > \cdots > \tilde{p}_K$. Since $\boldsymbol{p}$ is close to a one-hot vector, it is reasonable to assume that $\tilde{p}_1 \gg \tilde{p}_i$ for $2 \leq i \leq K$. Here, we consider a weaker assumption that $\tilde{p}_1 \gg \tilde{p}_j$ for $3 \leq j \leq K$, which can be met in realistic applications (we will verify this assumption with experiments in Appendix C in the supplementary material). For the trained neural network f_θ, we have

$$\begin{aligned} C_\theta(\boldsymbol{x}, T) &= 1/\big(1 + \sum_{i \neq \hat{y}} \exp\left(z_i/T - z_{\hat{y}}/T\right)\big) \\ &= 1/\big(1 + \sum_{j=2}^{K} (\tilde{p}_j/\tilde{p}_1)^{1/T}\big). \end{aligned} \quad (6)$$

For $3 \leq j \leq K$, $\tilde{p}_j/\tilde{p}_1 \approx 0$ since $\tilde{p}_1 \gg \tilde{p}_j$, so we have $C_\theta(\boldsymbol{x}, T) \approx 1/\big(1 + (\tilde{p}_2/\tilde{p}_1)^{1/T}\big)$. With similar proof of Theorem 1, we can obtain Eq (4) and Eq (5) under $K \geq 3$.

However, $T > 1$ will soften the output probability distribution of the ID samples in U^{in}. In order to guarantee the performance of ID classification, we employ the calibration technique (Guo et al. 2017) to set the value of T adaptively. Specifically, T is optimized with respect to the negative log likelihood loss on the ID validation set V in epoch t:

$$T_t = \arg\min_{T} \; - \sum_{(\boldsymbol{x}, y) \in V} \log\Big(q_\theta^{(y)}(\boldsymbol{x}, T)\Big). \quad (7)$$

The temperature T_t will push the output confidence of the ID and OOD samples in U further apart from each other, which motivates us to select these samples from U with the output confidence. A straightforward way is to employ the constant confidence thresholds τ^{in} and τ^{out}, i.e., the sample $\boldsymbol{x}$ is determined as ID if $C_\theta(\boldsymbol{x}, T) > \tau^{in}$ or OOD if $C_\theta(\boldsymbol{x}, T) < \tau^{out}$. However, since the output confidence distribution of the samples in U varies during the training process, we should choose the thresholds τ^{in} and τ^{out} dynamically. We fit a two-component Gaussian Mixture Model (GMM) on the output confidence of the samples in U with the Expectation-Maximization algorithm in each epoch and calculate the average confidence of the samples in each component as the confidence threshold. This process is summarized in Procedure 1, and the selected ID and OOD samples

Procedure 1 Obtaining the thresholds τ_t^{in} and τ_t^{out}

Input: Unlabeled data U, epoch t, neural network f_θ.
Parameter: Components g_1 (for ID) and g_2 (for OOD) of Gaussian Mixture Model (GMM).

1: Fit g_1 and g_2 on $\big\{C_\theta(\boldsymbol{x}, T_t) \mid \boldsymbol{x} \in U\big\}$ with EM;
2: Separate U with g_1 and g_2 by the posterior probability:
$U_t^{g_1} \leftarrow \big\{\boldsymbol{x} \mid p\big(g_1|C_\theta(\boldsymbol{x}, T_t)\big) > p\big(g_2|C_\theta(\boldsymbol{x}, T_t)\big) \wedge \boldsymbol{x} \in U\big\}$;
$U_t^{g_2} \leftarrow \big\{\boldsymbol{x} \mid p\big(g_1|C_\theta(\boldsymbol{x}, T_t)\big) \leq p\big(g_2|C_\theta(\boldsymbol{x}, T_t)\big) \wedge \boldsymbol{x} \in U\big\}$;
3: Calculate the thresholds:
$\tau_t^{in} \leftarrow \frac{1}{|U_t^{g_1}|} \sum_{\boldsymbol{x} \in U_t^{g_1}} C_\theta(\boldsymbol{x}, T_t)$;
$\tau_t^{out} \leftarrow \frac{1}{|U_t^{g_2}|} \sum_{\boldsymbol{x} \in U_t^{g_2}} C_\theta(\boldsymbol{x}, T_t)$.

Output: τ_t^{in} and τ_t^{out}.

in epoch t can be written as:

$$U_t^{in} = \{\boldsymbol{x} \mid \boldsymbol{x} \in U \wedge C_\theta(\boldsymbol{x}, T_t) > \tau_t^{in}\}, \quad (8)$$

$$U_t^{out} = \{\boldsymbol{x} \mid \boldsymbol{x} \in U \wedge C_\theta(\boldsymbol{x}, T_t) < \tau_t^{out}\}. \quad (9)$$

There may be samples of unseen OOD classes in the test data during the inference process (shown in Figure 1), so it is important to enhance the model's generalization capability to deal with them. One reasonable way is to increase the diversity of data with augmentation techniques. Here we employ RandAugment (Cubuk et al. 2020) and mixup (Zhang et al. 2018). RandAugment can produce heavily distorted versions of a given image, and mixup aims to generate new samples and targets by linear combination. We modify the vanilla mixup since there are OOD samples in U. Specifically, we first choose λ from $Beta(\alpha, \alpha)$, where α is a hyperparameter of the Beta distribution. Then, for a pair $(\boldsymbol{x}, \boldsymbol{x}')$, we obtain the new sample $\hat{\boldsymbol{x}} = \lambda'\boldsymbol{x} + (1-\lambda')\boldsymbol{x}'$, where $\boldsymbol{x}, \boldsymbol{x}' \in U$. $\lambda' = \max(\lambda, 1-\lambda)$ leads $\hat{\boldsymbol{x}}$ closer to $\boldsymbol{x}$ to preserve the semantic information in the original image $\boldsymbol{x}$. We combine these two augmentation techniques to further improve the richness of data, which can be formulated as:

$$\begin{aligned}\tilde{\boldsymbol{x}} &= \mathcal{M}\big(\mathcal{R}(\boldsymbol{x}), \mathcal{R}(\boldsymbol{x}')\big) \\ &= \lambda'\mathcal{R}(\boldsymbol{x}) + (1-\lambda')\mathcal{R}(\boldsymbol{x}'),\end{aligned} \quad (10)$$

where $\mathcal{R}(\cdot)$ denotes RandAugment for a given sample.

In order to make the model distinguish between ID and OOD samples better with the output confidence, we employ entropy minimization and maximization on the augmented samples in U_t^{in} and U_t^{out} respectively. The learning objectives can be written as:

$$\mathcal{L}_{Emin} = \frac{1}{|U|} \sum_{\boldsymbol{x} \in U_t^{in}} \mathrm{CE}\big(\hat{q}_\theta(\boldsymbol{x}) \parallel q_\theta(\tilde{\boldsymbol{x}})\big), \quad (11)$$

$$\mathcal{L}_{Emax} = -\frac{1}{|U|} \sum_{\boldsymbol{x} \in U_t^{out}} \mathrm{H}\big(q_\theta(\tilde{\boldsymbol{x}})\big). \quad (12)$$

In Eq (11), $\hat{q}_\theta(\boldsymbol{x})$ is the K-dimensional one-hot pseudo label for $\boldsymbol{x}$, in which the ith element $\hat{q}_\theta^{(i)}(\boldsymbol{x}) = 1$ if and only if $i = \hat{y}$ ($\hat{y}$ is the predicted class on $\boldsymbol{x}$). It enforces the model to produce the low-entropy (high-confidence) output on ID samples. In Eq (12), $\mathrm{H}(\cdot)$ calculates the entropy of a given distribution and enforces the model to produce the high-entropy (low-confidence) output on OOD samples. Similar to Sohn et al. (2020), we normalize these losses with $|U|$ to take the capacity of the selected sets into consideration.

In this way, the overall loss can be formulated as:

$$\mathcal{L} = \mathcal{L}_S + \omega\mathcal{L}_{CR} + \beta\mathcal{L}_{Emin} + \gamma\mathcal{L}_{Emax}, \quad (13)$$

where ω, β, and γ are hyperparameters to balance each loss. We first obtain a basic model with L and U, and use $\omega = 1$ and $\beta = \gamma = 0$ at the first stage of the training process. After getting the basic model, we select the ID and OOD samples in U with the output confidence and use $\omega = 0$ and $\beta = \gamma = 1$. This intuition is similar to curriculum learning (Bengio et al. 2009), which starts with an easier learning objective and then faces a more difficult one. The overall training process of Adaptive In-Out-aware Learning (AIOL) is shown in Algorithm 1, and the output confidence is used as the detection score during the inference process.

Algorithm 1 Adaptive In-Out-aware Learning (AIOL)

Input: Labeled data L, Unlabeled data U.
Parameter: Neural network f_θ.

for $t = 1$ **to** *max_epoch* **do**
  Obtain the temperature T_t with Eq (7);
  Obtain the thresholds τ_t^{in} and τ_t^{out} with Procedure 1;
  for $k = 1$ **to** *max_iteration* **do**
    Draw a batch of labeled data B_L from L, and draw a batch of unlabeled data B_U from U;
    Compute $\mathcal{L}_S$ with Eq (1) on B_L;
    Compute $\mathcal{L}_{CR}$, $\mathcal{L}_{Emin}$, and $\mathcal{L}_{Emax}$ with Eq (2), Eq (11), and Eq (12) respectively on B_U;
    Update neural network parameters θ with Eq (13).
  end for
end for

Output: Trained neural network f_θ.

Experiments

Datasets

The training set of CIFAR10 (CIFAR100) is used as ID training data, and we split them into labeled and unlabeled data. For the labeled data, the number of each class is set as 100, which results in 1000 (10000) labeled data for CIFAR10 (CIFAR100). The remaining data is used as the unlabeled ID data U^{in}. As for U^{out}, the unlabeled OOD data are drawn from the following datasets:

- **ImageNet.** The test set of ImageNet, i.e., 50000 images with 1000 classes, are drawn to construct U^{out}.
- **SVHN.** The test set of SVHN, i.e., 26032 images with 10 classes, are drawn to construct U^{out}.
- **CIFAR10** & **CIFAR100.** If CIFAR10 is considered as the ID dataset, the test set of CIFAR100, i.e., 10000 images with 100 classes, will be drawn to construct U^{out}; if CIFAR100 is considered as the ID dataset, the test set of CIFAR10, i.e., 10000 images with 10 classes, will be drawn to construct U^{out}.
- **Split.** Except for the mixtures of two different datasets, we also consider the splits of one single dataset. Specifically, we split CIFAR10 into *animal* group as ID and *non-animal* group as OOD, i.e., 6 classes and 4 classes. Similarly, we split CIFAR100 into *living* group as ID and *non-living* group as OOD, i.e., 65 classes and 35 classes. Details of the splits are given in Appendix B in the supplementary material. We also keep 100 labeled data per class for the ID group and use the remaining data of the training set as the mixed unlabeled data U because they naturally include the ID and OOD groups.

The test set of CIFAR10 (CIFAR100) is used to evaluate the performance. Following that in Yu and Aizawa (2019) and Sastry and Oore (2020), we split 10% of the test set as the ID validation data and use the rest as the ID test data. As for the OOD test data, samples of seen and unseen OOD classes are drawn as follows:

- **Seen.** If U^{in} and U^{out} are from two different datasets, samples of seen OOD classes will be drawn from the

U^{in}	CIFAR10				CIFAR100			
U^{out}	ImageNet	SVHN	CIFAR100	Split	ImageNet	SVHN	CIFAR10	Split
OOD type	Seen / Unseen				Seen / Unseen			
Baseline	66.2 / 65.1	54.2 / 66.4	63.9 / 65.3	66.4 / 61.4	70.3 / 64.3	70.2 / 64.2	68.1 / 64.7	75.5 / 74.3
OE	69.6 / 70.7	82.9 / 62.3	63.9 / 59.3	65.3 / 59.2	86.5 / 82.5	98.7 / 66.8	70.3 / 67.9	76.1 / 74.6
MCD	98.4 / 83.7	97.3 / 57.7	59.0 / 51.6	73.5 / 57.8	96.1 / 84.4	99.3 / 69.3	79.8 / 70.0	92.2 / 82.3
SSD	73.7 / 86.9	33.3 / **94.8**	81.6 / 93.3	39.1 / 94.5	49.3 / 67.8	29.9 / 71.5	47.4 / 79.7	44.9 / 75.4
FixMatch	44.6 / 85.4	40.2 / 93.2	76.8 / 90.7	39.4 / 91.4	47.3 / 72.6	25.0 / 77.1	60.4 / 78.0	39.1 / 73.2
UASD	86.0 / 85.7	88.4 / 84.7	80.0 / 86.0	80.4 / 82.1	74.6 / 72.9	77.1 / 71.3	71.3 / 72.2	74.6 / 73.6
Ours	**99.9 / 96.9**	**100.** / 93.4	**94.7 / 95.0**	**93.8 / 94.6**	**99.7 / 90.0**	**100. / 77.4**	**80.0 / 80.9**	**97.7 / 88.9**

Table 1: OOD detection results with percentage of AUROC.

training set of the dataset for U^{out}. For example, when U^{in} is from CIFAR10 and U^{out} is from CIFAR100, samples of seen OOD classes are drawn from the training set of CIFAR100; if U^{in} and U^{out} are from the splits of one single dataset, samples of seen OOD classes will be drawn from the OOD group in the test set. For example, when U^{in} and U^{out} are from the splits of CIFAR10, samples of seen OOD classes are drawn from the *non-animal* group in the test set of CIFAR10.

- **Unseen.** Samples of unseen OOD classes are drawn from various benchmark datasets: CIFAR10, CIFAR100, SVHN, ImageNet, Blobs, Texture, iSUN, LSUN, and Places365. The dataset will not be used to construct the unseen OOD data if it is considered for training. Following that in Hendrycks and Gimpel (2017), each benchmark dataset is used to evaluate the detection performance respectively, and the average result is reported.

More details of the used datasets are given in Appendix B in the supplementary material.

Setup

Following that in Sohn et al. (2020), the augmentation $\mathcal{A}(\cdot)$ is implemented with the standard data augmentations (random flip and crop), and the augmentation $\mathcal{A}'(\cdot)$ is implemented with RandAugment. In the experiments, we use the standard Wide ResNet (Zagoruyko and Komodakis 2016), i.e., WRN-28-2, as the base network and use SGD optimizer for training. The experiments are run for 256 epochs with 512 iterations per epoch. We set $\omega = 1$ and $\beta = \gamma = 0$ at the beginning, and set $\omega = 0$ and $\beta = \gamma = 1$ after 80% of the training epochs. We set $\alpha = 0.2$ for mixup. We employ EMA model (Tarvainen and Valpola 2017), and limit $\tau_t^{in} \leq 0.95$ and $\tau_t^{out} \geq 1/K + 0.05$ for stability. Other hyperparameters are the same as that of Sohn et al. (2020) for a fair comparison. We evaluate the detection performance with three metrics: AUROC, AUPR, and FPR95. Due to space constraints, we present results with AUROC (the area under the ROC curve) in this section. More results with other metrics are in Appendix C in the supplementary material.

Compared Methods

We compare our method with various related methods, including: Baseline (Hendrycks and Gimpel 2017); OE (Hendrycks, Mazeika, and Dietterich 2019); MCD (Yu and Aizawa 2019); SSD (Sehwag, Chiang, and Mittal 2021); FixMatch (Sohn et al. 2020); UASD (Chen et al. 2020b). The Baseline method only uses labeled data for training, but the others consider unlabeled data. Note that the first four methods are developed for OOD detection, while the last two methods are barely SSL methods. The OOD detector based on the output confidence is built for the last two methods, which is the same as ours. For a fair comparison, the experiments are also conducted for 256 epochs with 512 iterations per epoch on WRN28-2 for all the compared methods except for SSD. For SSD, it is a self-supervised learning method and needs more training resources, so we follow the original paper and run it on ResNet-18 (He et al. 2016) for 500 epochs. The hyperparameters are set according to the original paper for all the compared methods.

Results

Since the test data contain samples of seen and unseen OOD classes, we evaluate the detection performance on them respectively, and the results are shown in Table 1.

Performance of seen OOD detection. The results for detecting samples of seen OOD classes are summarized on the left of the slash in Table 1, which indicates that our method outperforms the compared methods on seen OOD detection. The MCD method performs worse than ours since it treats each unlabeled sample equally. The SSD method and the FixMatch method are developed for the pure unlabeled ID data, while the OE method is developed for the pure unlabeled OOD data. These three methods all produce poor and unstable results since they are confused with the mixed ID and OOD samples in U. The UASD method produces more stable results since it eliminates the effect of OOD samples in U, but it still performs worse than ours since we can learn from OOD samples in U rather than ignoring them.

Performance of unseen OOD detection. The results for detecting samples of unseen OOD classes are summarized on the right of the slash in Table 1. Except when U^{in} is from CIFAR10 and U^{out} is from SVHN, our method outperforms the compared methods because the augmentation techniques enhance the model's generalization capability on unseen OOD detection. When U^{in} is from CIFAR10 and U^{out} is from SVHN, our method performs slightly worse than the

$\|L\|$	250	1000	4000
OOD type	Seen / Unseen		
Baseline	53.3 / 52.5	66.2 / 65.1	76.7 / 76.8
FixMatch	42.1 / 84.6	44.6 / 85.4	68.2 / 88.5
UASD	70.4 / 70.0	86.0 / 85.7	90.5 / 91.1
Ours	**99.8 / 96.1**	**99.9 / 96.9**	**99.3 / 97.0**

Table 2: Results with different numbers of labeled samples.

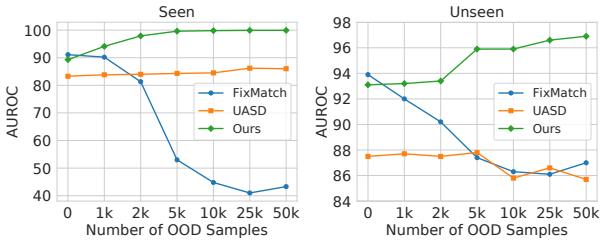


Figure 2: Results with various numbers of OOD samples.

SSD method. The reason is that the samples in SVHN are images of house numbers, and it is hard to get more diverse information from these plain images. Furthermore, the SSD method requires more computational resources, which is run for 500 epochs on ResNet-18.

Since our method is developed for OOD detection with the labeled data L and the mixed unlabeled data U, number of labeled samples and number of OOD samples in U are parameters of the experiments. We set that U^{in} is from CIFAR10 and U^{out} is from ImageNet, and provide further results to investigate the effectiveness of the two parameters:

Number of labeled samples. We conduct the experiments with different numbers of labeled samples (250, 1000, and 4000), and the results are shown in Table 2. It can be found that our method performs better than the other methods with different numbers of labeled samples and is not sensitive to this parameter. Moreover, our method can achieve superior performance even with very few labeled samples, i.e., 250. The reason lies in that our method can make full use of the mixed unlabeled data.

Number of OOD samples in U. We conduct the experiments with various numbers of OOD samples in U ($|U^{out}|$ varies from 0 to 50000), and the results are shown in Figure 2. We set $\gamma = 0$ when there is no OOD sample in U. Figure 2 shows that our method outperforms the other methods with various numbers of OOD samples in U except for $|U^{out}| = 0$. When $|U^{out}| = 0$, our method is slightly worse than the FixMatch method, which is developed for the pure unlabeled ID data ($|U^{out}| = 0$). But our method is appropriate for the mixed unlabeled data ($|U^{out}| > 0$) and can produce better results with more OOD samples in U.

Ablation Study

We set that U^{in} and U^{out} are from the splits of CIFAR100 and run the experiments to study the details of our method and provide additional insight into what makes it successful.

Ablation	Seen / Unseen
Supervised loss (Eq(1))	75.5 / 74.3
+ Consistency regularization (Eq (2)) with $T = 1$	59.2 / 80.4
+ Adaptive temperature T_t	89.7 / 84.3
+ Emin and Emax (Eq (11) and Eq (12)) without augmentation	96.6 / 83.6
+ RandAugment	94.0 / 85.6
+ Modified mixup	**97.7 / 88.9**

Table 3: Ablation study on the used modules.

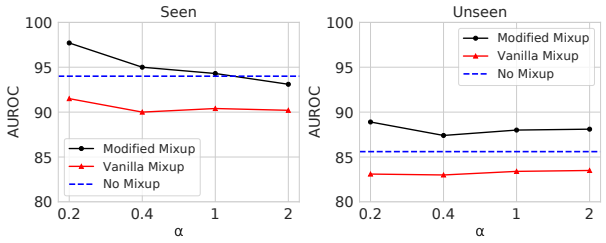
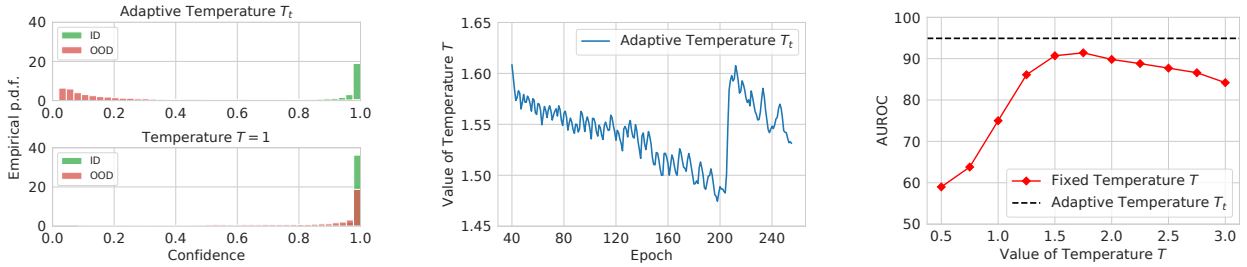


Figure 3: Ablation study on the modified mixup.

The effectiveness of the used modules. There are several modules in our method: supervised loss (Eq (1)); consistency regularization (Eq (2)) with $T = 1$; adaptive temperature T_t; Emin and Emax (Eq (11) and Eq (12)) without augmentation; two augmentation techniques. The ablation study on these modules is summarized in Table 3. We can see that the method with supervised loss and consistency regularization ($T = 1$) performs badly, but after using the adaptive temperature T_t, the method performs much better. Combined with Eq (11) and Eq (12) over U_t^{in} and U_t^{out}, which are selected according to the adaptive temperature T_t and the dynamic thresholds τ_t^{in} and τ_t^{out}, the method achieves significantly better results on seen OOD detection. Further combined with the two augmentation techniques, i.e., RandAugment and the modified mixup, the model's generalization capability on unseen OOD detection is improved.

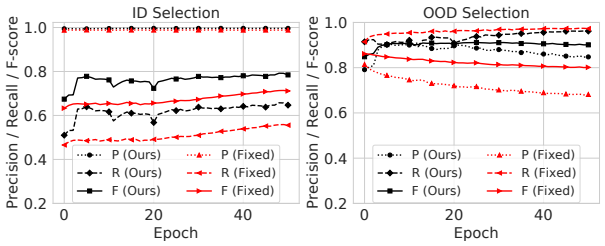
The effectiveness of the modified mixup. The results for the mixup with different parameters are shown in Figure 3. On the one hand, the method with the vanilla mixup (without the maximum operation $\lambda' = \max(\lambda, 1-\lambda)$) performs worse than that with no mixup since the vanilla mixup directly confuses the ID and OOD samples. On the other hand, our method with the modified mixup performs better since λ' is close to 1 and can produce diverse data points around the original one, which can protect the semantic information. As for the hyperparameter α, the smaller one can achieve better results since smaller α leads λ closer to 1 and bigger α leads λ closer to 0.5.

The effectiveness of the adaptive temperature T_t. After 80% of the training epochs, we depict the output confidence of the samples in U with the adaptive temperature T_t in the upper part of Figure 4(a). As a comparison, we use $T = 1$ and depict the output confidence of the samples in U in the lower part of Figure 4(a). It can be found that the adaptive

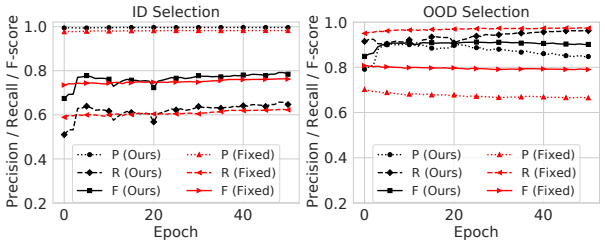


(a) (b) (c)

Figure 4: Ablation study on the adaptive temperature T_t. (a): Empirical p.d.f. of the output confidence of the samples in U with T_t (upper part) and $T = 1$ (lower part). (b): The values of T_t during training. (c) Detection results on U with different T.



(a)



(b)

Figure 5: Precision (P), Recall (R) and F-score (F) of the selection for ID samples (left part) and OOD samples (right part) with our dynamic thresholds (τ_t^{in} and τ_t^{out}) and the fixed thresholds (τ^{in} and τ^{out}). (a): $\tau^{in} = 0.9$ and $\tau^{out} = 0.3$. (b): $\tau^{in} = 0.7$ and $\tau^{out} = 0.5$.

temperature T_t pushes the output confidence of the ID and OOD samples further apart from each other. The values of the adaptive temperature T_t during the training process are shown in Figure 4(b) (we use $T = 1$ before epoch 40 for warming up), which indicates that T_t is always larger than 1. It rapidly increases at 80% of the training epochs, i.e., epoch 205, because the large temperature is required to calibrate the output confidence of ID samples for Eq (11). For comparison, we also run the experiments with the fixed temperature T, and the detection performance on U is shown in Figure 4(c). It demonstrates that the adaptive temperature T_t can consistently outperform the fixed ones.

The effectiveness of the thresholds τ_t^{in} and τ_t^{out}. For obtaining U_t^{in} in Eq (8) and U_t^{out} in Eq (9) from U, we choose the dynamic thresholds τ_t^{in} and τ_t^{out} based on the two GMM components g_1 and g_2. For comparison, we also run the experiments with the fixed thresholds ($\tau^{in} = 0.9$ and $\tau^{out} = 0.3$, or $\tau^{in} = 0.7$ and $\tau^{out} = 0.5$), and report the Precision, Recall, and F-score of the selection on U in Figure 5(a) and Figure 5(b). F-score, i.e., the harmonic mean of Precision and Recall, is a metric that can generally evaluate the selection quality. From Figure 5(a) and Figure 5(b), we can see that the F-score of our dynamic thresholds is always higher than that of the fixed thresholds. Detailed results for more fixed thresholds are given in Appendix C in the supplementary material.

U^{out}	ImageNet	SVHN	CIFAR	Split
U^{in}	CIFAR10 / CIFAR100			
Baseline	58.1 / 57.1	58.1 / 57.1	58.1 / 57.1	51.6 / 53.3
FixMatch	93.3 / 67.9	94.4 / 69.3	93.9 / 70.4	93.6 / 67.1
UASD	84.7 / 64.9	85.1 / 65.3	85.8 / 65.8	82.5 / 62.7
Ours	92.6 / 68.4	93.3 / 67.7	93.0 / 67.8	93.8 / 66.0

Table 4: Classification results with percentage of accuracy.

Performance of ID classification. The results of ID classification are reported in Table 4. Our method performs slightly worse than the FixMatch method, perhaps because we use temperature $T > 1$ for the samples of U^{in} in Eq (2). But our method is developed for detecting OOD samples and performs much better than the FixMatch method in OOD detection, which is shown in Table 1.

Conclusion

In this paper, we focus on the more realistic OOD detection scenario, where limited labeled data and abundant mixed unlabeled data are available for training. During the inference process, the trained model should not only detect samples of seen OOD classes but also detect samples of unseen OOD classes. We propose the Adaptive In-Out-aware Learning (AIOL) method, in which we adaptively select potential ID and OOD samples from the mixed unlabeled data and optimize the entropy over them. Moreover, data augmentation techniques are brought into the method to further improve the performance of unseen OOD detection. The experimental results show that our method outperforms the compared methods on various benchmark datasets.

Acknowledgments

This work is supported by the National Key Research and Development Program of China (2018AAA0101100), the National Science Foundation of China (61921006, 61673202), and the Collaborative Innovation Center of Novel Software Technology and Industrialization.

References

Bengio, Y.; Louradour, J.; Collobert, R.; and Weston, J. 2009. Curriculum learning. In *Proceedings of the 26th International Conference on Machine Learning (ICML'09)*, 41–48.

Berthelot, D.; Carlini, N.; Goodfellow, I.; Papernot, N.; Oliver, A.; and Raffel, C. 2019. Mixmatch: A holistic approach to semi-supervised learning. *arXiv:1905.02249*.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. E. 2020a. A Simple Framework for Contrastive Learning of Visual Representations. In *Proceedings of the 37th International Conference on Machine Learning (ICML'20)*, 1597–1607.

Chen, Y.; Zhu, X.; Li, W.; and Gong, S. 2020b. Semi-Supervised Learning under Class Distribution Mismatch. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI'20)*, 3569–3576.

Cubuk, E. D.; Zoph, B.; Shlens, J.; and Le, Q. 2020. RandAugment: Practical Automated Data Augmentation with a Reduced Search Space. In *Advances in Neural Information Processing Systems 33 (NeurIPS'20)*.

Golan, I.; and El-Yaniv, R. 2018. Deep Anomaly Detection Using Geometric Transformations. In *Advances in Neural Information Processing Systems 31 (NeurIPS'18)*, 9781–9791.

Grandvalet, Y.; and Bengio, Y. 2004. Semi-supervised learning by entropy minimization. In *Advances in Neural Information Processing Systems 17 (NIPS'04)*.

Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. Q. 2017. On Calibration of Modern Neural Networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML'17)*, 1321–1330.

Guo, L.; Zhang, Z.; Jiang, Y.; Li, Y.; and Zhou, Z. 2020. Safe Deep Semi-Supervised Learning for Unseen-Class Unlabeled Data. In *Proceedings of the 37th International Conference on Machine Learning (ICML'20)*, 3897–3906.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR'16)*, 770–778.

Hendrycks, D.; and Gimpel, K. 2017. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. In *Proceedings of the 5th International Conference on Learning Representations (ICLR'17)*.

Hendrycks, D.; Mazeika, M.; and Dietterich, T. G. 2019. Deep Anomaly Detection with Outlier Exposure. In *Proceedings of the 7th International Conference on Learning Representations (ICLR'19)*.

Hendrycks, D.; Mazeika, M.; Kadavath, S.; and Song, D. 2019. Using Self-Supervised Learning Can Improve Model Robustness and Uncertainty. In *Advances in Neural Information Processing Systems 32 (NeurIPS'19)*, 15637–15648.

Hsu, Y.; Shen, Y.; Jin, H.; and Kira, Z. 2020. Generalized ODIN: Detecting Out-of-Distribution Image Without Learning From Out-of-Distribution Data. In *Proceedings of the European Conference on Computer Vision (ECCV'20)*, 10948–10957.

Huang, Y.; Dai, S.; Nguyen, T.; Baraniuk, R. G.; and Anandkumar, A. 2019. Out-of-Distribution Detection Using Neural Rendering Generative Models. *arXiv:1907.04572*.

Lee, D.-H. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *ICML Workshop on Challenges in Representation Learning*.

Lee, K.; Lee, K.; Lee, H.; and Shin, J. 2018. A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks. In *Advances in Neural Information Processing Systems 31 (NeurIPS'18)*, 7167–7177.

Liang, S.; Li, Y.; and Srikant, R. 2018. Enhancing The Reliability of Out-of-distribution Image Detection in Neural Networks. In *Proceedings of the 6th International Conference on Learning Representations (ICLR'18)*.

Liu, W.; Wang, X.; Owens, J. D.; and Li, Y. 2020. Energy-based Out-of-distribution Detection. In *Advances in Neural Information Processing Systems 33 (NeurIPS'20)*.

Nalisnick, E.; Matsukawa, A.; Teh, Y. W.; and Lakshminarayanan, B. 2019. Detecting Out-of-Distribution Inputs to Deep Generative Models Using Typicality. *arXiv:1906.02994*.

Sastry, C. S.; and Oore, S. 2020. Detecting Out-of-Distribution Examples with Gram Matrices. In *Proceedings of the 37th International Conference on Machine Learning (ICML'20)*, 8491–8501.

Sehwag, V.; Chiang, M.; and Mittal, P. 2021. SSD: A Unified Framework for Self-Supervised Outlier Detection. In *Proceedings of the 9th International Conference on Learning Representations (ICLR'21)*.

Serrà, J.; Álvarez, D.; Gómez, V.; Slizovskaia, O.; Núñez, J. F.; and Luque, J. 2019. Input complexity and out-of-distribution detection with likelihood-based generative models. *arXiv:1909.11480*.

Sohn, K.; Berthelot, D.; Carlini, N.; Zhang, Z.; Zhang, H.; Raffel, C.; Cubuk, E. D.; Kurakin, A.; and Li, C. 2020. FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence. In *Advances in Neural Information Processing Systems 33 (NeurIPS'20)*.

Tack, J.; Mo, S.; Jeong, J.; and Shin, J. 2020. CSI: Novelty Detection via Contrastive Learning on Distributionally Shifted Instances. In *Advances in Neural Information Processing Systems 33 (NeurIPS'20)*.

Tarvainen, A.; and Valpola, H. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems 30 (NIPS'17)*.

Winkens, J.; Bunel, R.; Roy, A. G.; Stanforth, R.; Natarajan, V.; Ledsam, J. R.; MacWilliams, P.; Kohli, P.; Karthikesalingam, A.; Kohl, S.; Cemgil, T.; Eslami, S. M. A.; and Ronneberger, O. 2020. Contrastive Training for Improved Out-of-Distribution Detection. *arXiv:2007.05566*.

Xie, Q.; Dai, Z.; Hovy, E.; Luong, M.-T.; and Le, Q. V. 2019. Unsupervised Data Augmentation for Consistency Training. *arXiv:1904.12848*.

Yu, Q.; and Aizawa, K. 2019. Unsupervised Out-of-Distribution Detection by Maximum Classifier Discrepancy. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV'19)*, 9517–9525.

Yu, Q.; Ikami, D.; Irie, G.; and Aizawa, K. 2020. Multi-task Curriculum Framework for Open-Set Semi-supervised Learning. In *Proceedings of the European Conference on Computer Vision (ECCV'20)*, 438–454.

Zagoruyko, S.; and Komodakis, N. 2016. Wide Residual Networks. In *Proceedings of the British Machine Vision Conference (BMVC'16)*.

Zhang, H.; Cissé, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2018. mixup: Beyond Empirical Risk Minimization. In *Proceedings of the 6th International Conference on Learning Representations (ICLR'18)*.