Out-of-Distributed Semantic Pruning for Robust Semi-Supervised Learning

Yu Wang[1,4*] Pengchong Qiao[1,2,4*] Chang Liu[3] Guoli Song[2,4] Xiawu Zheng[2,4†] Jie Chen[1,2,4†]
[1] School of Electronic and Computer Engineering, Peking University, Shenzhen, China
[2] Peng Cheng Laboratory, Shenzhen, China
[3] Department of Automation and BNRist, Tsinghua University, Beijing, China
[4] AI for Science(AI4S)-Preferred Program,Peking University Shenzhen Graduate School, China
{rain_wang,pcqiao}@stu.pku.edu.cn, {songgl,zhengxw01,chenj}@pcl.ac.cn,
{liuchang2022}@tsinghua.edu.cn

Abstract

*Recent advances in robust semi-supervised learning (SSL) typically filter out-of-distribution (OOD) information at the sample level. We argue that an overlooked problem of robust SSL is its corrupted information on **semantic** level, practically limiting the development of the field. In this paper, we take an initial step to explore and propose a unified framework termed OOD Semantic Pruning (OSP), which aims at pruning OOD semantics out from in-distribution (ID) features. Specifically, (i) we propose an aliasing OOD matching module to pair each ID sample with an OOD sample with semantic overlap. (ii) We design a soft orthogonality regularization, which first transforms each ID feature by suppressing its semantic component that is collinear with paired OOD sample. It then forces the predictions before and after soft orthogonality decomposition to be consistent. Being practically simple, our method shows a strong performance in OOD detection and ID classification on challenging benchmarks. In particular, OSP surpasses the previous state-of-the-art by **13.7%** on accuracy for ID classification and **5.9%** on AUROC for OOD detection on TinyImageNet dataset. The source codes are publicly available at https://github.com/rain305f/OSP.*

1. Introduction

Deep neural networks have obtained impressive performance on various tasks [30, 44, 46]. Their success is partially dependent on a large amount of labeled training data, of which the acquisition is expensive and time-consuming [20, 25, 40, 45]. A prevailing way to reduce the dependency on human annotation is semi-supervised learning (SSL). It learns informative semantics using annotation-free and acquisition-easy unlabeled data to extend the label information from limited labeled data and has achieved promising results in various tasks [38, 50, 51, 54].

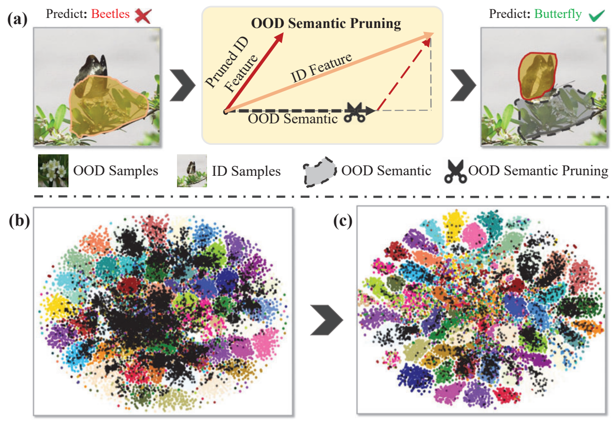

Figure 1. (a) Intuitive diagram of OOD Semantic Pruning (OSP), that pruning OOD semantics out from ID features. (b) t-SNE visualization [52] from the baseline [23]. (c) t-SNE visualization from our OSP model. The colorful dots donate ID features, while the black dots mark OOD features. The dots with the same color represent the features of the same class. Here, our OSP and the baseline are trained on CIFAR100 with 100 labeled data per class and 60% OOD in unlabeled data.

Unfortunately, classical SSL relies on a basic assumption that the labeled and unlabeled data are collected from the same distribution, which is difficult to hold in real-world applications. In most practical cases, unlabeled data usually contains classes that are not seen in the labeled data. Existing works [12, 17, 21, 40] have shown that training the SSL model with these OOD samples in unlabeled data leads to a large degradation in performance. To solve this problem, robust semi-supervised learning (Robust SSL) has been investigated to train a classification model that performs sta-

*Equal contribution.
†Corresponding author.

bly when the unlabeled set is corrupted by OOD samples. Typical methods focus on discarding OOD information at the sample level, that is, detecting and filtering OOD samples to purify the unlabeled set [12, 17, 21, 57]. However, these methods ignore the semantic-level pollution caused by the classification-useless semantics from OOD samples, which improperly disturbs the feature distribution learned from ID samples, eventually resulting in weak ID and OOD discrimination and low classification performance. We provide an example to explain such a problem in Fig. 1. As we can see, due to the semantics of *Orchid* in OOD examples, the model pays too much attention to the background and misclassifies the *Butterfly* as *Beetle*.

In this paper, we propose Out-of-distributed Semantic Pruning (OSP) method to solve the problem mentioned above and achieve effective robust SSL.

Concretely, our OSP approach consists of two main modules. We first develop an aliasing OOD matching module to pair each ID sample with an OOD sample with which it has feature aliasing. Secondly, we propose a soft orthogonality regularization, which constrains the predictions of ID samples to keep consistent before and after soft-orthogonal decomposition according to their matching OOD samples.

We evaluate the effectiveness of our OSP in extensive robust semi-supervised image recognition benchmarks including MNIST [53], CIFAR10 [29], CIFAR100 [29] and TinyImageNet [15]. We show that our OSP obtains significant improvements compared to state-of-the-art alternatives (e.g., **13.7%** and **15.0%** on TinyImagetNet with an OOD ratio of 0.3 and 0.6 respectively). Besides, we also empirically demonstrate that OSP indeed increases the feature discrimination between ID and OOD samples. To summarize, the contributions of this work are as follows:

- To the best of our knowledge, we are the first to exploit the OOD effects at the semantic level by regularization ID features to be orthogonal to OOD features.
- We develop an aliasing OOD matching module that adaptively pairs each ID sample with an OOD sample. In addition, we propose a soft orthogonality regularization to restrict ID and OOD features to be orthogonal.
- We conduct extensive experiments on four datasets, i.e., MNIST, CIFAR10, CIFAR100, and TIN200, and achieve new SOTA performance. Moreover, we analyze that the superiority of OSP lies in the enhanced discrimination between ID and OOD features.

2. Related work

2.1. Semi-Supervised Learning

Semi-supervised learning aims to learn informative semantics from unlabeled data to reduce the dependence on human annotations. Recently, many efforts have been made in SSL classification [2,4,5,8,10,11,14,19,24,26,35,43,48]. Powerful methods based on entropy minimization enforce their networks to make low-entropy predictions on unlabeled data [3, 27, 32, 32, 34, 42]. Another spectrum of popular approaches is consistency regularization, whose core idea is to obtain consistent prediction under various perturbations [31, 38, 50, 51, 54]. VAT [38] enforces prediction invariance under adversarial noises on unlabeled images. UDA [54] and FixMatch [50] employ weak and strong augmentation to compute the consistency loss.

The effectiveness of these SSL methods relies on an assumption that the labeled and unlabeled data are drawn from the same distribution. However, in practice, such an assumption is difficult to satisfy, resulting in severe performance degeneration of close-set SSL [18, 40, 57]. Thus, there is an urgent need to develop SSL algorithms that could work robustly with an unlabeled dataset that contains OOD samples.

2.2. Robust Semi-Supervised Learning

Robust SSL aims to train a classification model that performs stably when the unlabeled set is corrupted by OOD samples [1, 6, 18, 22, 24, 28, 33, 47]. This paper considers a common case: unlabeled data contains classes not seen in the labeled data [55]. Current typical approaches focus on removing the effects of OOD information at the sample-level [12, 17, 21, 57]. UASD [12] utilizes self-distillation to detect OOD samples and filter them out later from unlabeled data. MTC [55] proposes a multi-task curriculum learning framework, which detects the OOD samples in unlabeled data and simultaneously estimates the probability of the sample belonging to OOD. DS^{3}L [17] trains a soft weighting function to assign small weights to OOD unlabeled samples and large weights to ID unlabeled samples. More recently, some works have proposed utilizing OOD samples to improve the feature representation capacity of their models [23, 37]. Simultaneously, they also inherited the idea of previous work to filter out OOD samples in classification supervision. [37] extracts style features of ID samples and transfers OOD samples to ID style. T2T [23] employs an agent self-supervised task on both ID and OOD samples to enhance representation learning. Different from existing methods, we propose to prune the harmful OOD semantics out from ID features by regularizing ID and OOD features to be orthogonal, resulting in accurate ID classification and OOD detection.

3. Method

3.1. Preliminaries

Give a small set of labeled data $\mathcal{D}^l = \{(x_i^l, y_i^l)\}_{i=1}^{N_l}$ and a large set of unlabeled data $\mathcal{D}^u = \{(x_i^u)\}_{i=1}^{N_u}$ ($N_l \ll N_u$), where x_i^l, y_i^l and x_i^u are the image and label of the i-th la-

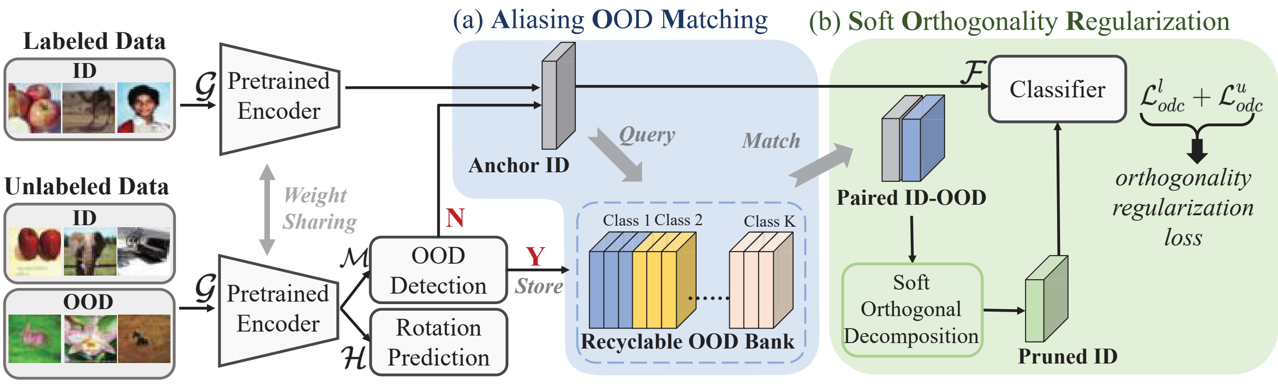

Figure 2. The overall architecture of our proposed OSP for robust semi-supervised classification. The core modules are aliasing OOD matching (AOM) and soft orthogonality regularization (SOR). The training process of our OSP consists of two-stage. At the pre-training stage, we pre-train the model with rotation prediction and K-ways predictor (Sec. 3.5). At the fine-tuning stage, we utilize the pre-trained OOD detection module to detect OOD samples in unlabeled data and store them in a class-wise memory bank, named recyclable OOD bank. To prune harmful OOD semantics out from ID features, the AOM selects an OOD sample with semantic overlap for each ID feature and composes ID-OOD pairs. Then the SOR applies a Soft Orthogonal Transform on ID-OOD pairs and generates pruned ID features. Finally, our proposed Orthogonality Regularization Loss constrains the predictions of ID features and corresponding pruned ID features to be consistent. During inference, the encoder and classifier are applied to K-ways ID classification. The details are shown in Sec. 3.

beled data and the image of the i-th unlabeled data. The label space of labeled data contains K labels, that is, $y_i^l \in C^l = \{1, ..., K\}$. The difference from classic SSL is that there exist OOD samples of unseen classes in the unlabeled training set. Formally, $C^l \subset C^u$ and $C^{\text{OOD}} = C^u \setminus C^l$. Robust SSL aims to train a classification model that performs stably when the unlabeled set is corrupted by OOD samples.

3.2. Overview

The architecture of our OSP is summarized in Fig. 2. The previous state-of-the-art robust SSL method T2T [23] is selected as our baseline. Following T2T, OSP has a shared encoder $\mathcal{G}(\cdot)$, a K-ways classifier $\mathcal{F}(\cdot)$, a rotation prediction head $\mathcal{H}(\cdot)$ and an OOD detection module $\mathcal{M}(\cdot)$. Different from T2T [23], we design two novel modules, named aliasing OOD matching (AOM) and soft orthogonality regularization (SOR) respectively, to prune out-of-distributed semantic and obtain a robust classifier simultaneously. The AOM module and SOR module are elaborated in Sec. 3.3 and Sec. 3.4, respectively. Inheriting the training paradigm of current robust SSL methods [21, 23, 55], our OSP contains two training stages: the pre-training stage and fine-tuning stage, where the detailed descriptions are as follows.

Pre-training stage. The purpose of this stage is to obtain a pre-trained model that could detect OOD samples reasonably. Following T2T [23], we carry out a K-way classification on $\mathcal{D}^l$ and a self-supervised task [39] [9] (i.e., rotation recognition [16]) on $\mathcal{D}^u$ to pre-train the encoder $\mathcal{G}(\cdot)$, the classifier $\mathcal{F}(\cdot)$, and the rotation predictor $\mathcal{H}(\cdot)$. Given a labeled input $x_i^l \in D^l$ and an unlabeled input $x_j^u \in D^u$, we denote their representations as $z_i^l = \mathcal{G}(x_i^l)$ and $z_j^u = \mathcal{G}(x_j^u)$. The training of model parameters is optimized by minimizing a supervised cross-entropy loss $\mathcal{L}_{ce}$ and a rotation loss $\mathcal{L}_{rot}$. Details are described in Sec. 3.5.

Meanwhile, we pre-train the OOD detection module $\mathcal{M}(\cdot)$ on $\mathcal{D}^l$ to calculate OOD scores $S(x^u)$ for unlabeled samples, which is used to distinguish ID samples and OOD samples in unlabeled data. Formally, we define the classifier as follows:

$$g(x^u) = \begin{cases} \text{ID}, & if\ S(x^u) \geq \gamma, \\ \text{OOD}, & if\ S(x^u) < \gamma, \end{cases} \tag{1}$$

where γ is calculated by the Ostu algorithm [41] in our experiments [23]. Additionally, we enforce our model to predict consistent predictions before and after adding Gaussian noises on feature maps $\mathcal{G}(\cdot)$, which helps to obtain more robust features.

Fine-tuning stage. The fine-tuning stage aims to refine the pre-trained model to obtain an accurate and robust classifier, which is achieved by the proposed AOM and SOR.

As illustrated in Fig. 2, we first utilize the OOD detection module $\mathcal{M}(\cdot)$ to periodically split unlabeled data into subsets: ID unlabeled set and OOD unlabeled set, referring to [23]. The ID unlabeled set is then used to learn semantics from unlabeled data. Due to OOD samples having conflicting targets with the classification, the compared baseline T2T [23] drops the OOD unlabeled set. In contrast, we argue that the dropped set still contains useful information, which needs to be pruned in optimization. To this end, we propose the AOM and SOR to achieve such a purpose. Specifically, the AOM pairs each ID sample with an OOD sample with which it has feature aliasing. And then, the

SOR constrains the predictions of ID samples to keep consistent before and after soft-orthogonal decomposition according to their matching OOD samples.

3.3. Aliasing OOD Matching

In this section, we introduce our aliasing OOD matching (AOM) Module and discuss how to select anchor ID samples and pair them with OOD samples with which they have feature aliasing.

Anchor ID features. During training, we sample anchor ID images (queries) for each target category that appears in the current mini-batch. We denote the feature set of labeled candidate anchor images for category c as $\mathcal{A}_c^l$, which contains features of labeled images with high confidence. Formally,

$$\mathcal{A}_c^l = \{z_i^l | z_i^l = \mathcal{G}(x_i^l), y_i^l = c, p_i^l[c] > \delta\}, \tag{2}$$

where y_i^l, z_i^l and p_i^l are the ground-truth label, feature representation, and class probability for the labeled image x_i^l, respectively. Here, δ denotes the positive threshold and is set to 0.8 following [23], and $p_i^l[c]$ is the predicted probability of class c. For unlabeled data, counterpart $\mathcal{A}_c^u$ is computed as:

$$\mathcal{A}_c^u = \{z_i^u | z_i^u = \mathcal{G}(x_i^u), \hat{y}_i^u = c, max_c(p_i^u[c]) > \delta\}, \tag{3}$$

where $y_i^u = \arg\max_c(p_i^u[c])$ is the pseudo label of the image x_i^u. This $\mathcal{A}_c^u$ is similar to $\mathcal{A}_c^l$, the only difference is that it uses the pseudo-label for class determination. Based on $\mathcal{A}_c^l$ and $\mathcal{A}_c^u$, we obtain the set of all qualified ID anchors $\mathcal{A}_c$:

$$\mathcal{A}_c = \mathcal{A}_c^l \cup \mathcal{A}_c^u. \tag{4}$$

Recyclable OOD samples. We define a binary variable $n_i(c)$ to identify whether an unlabeled image $x_i^u \in \mathcal{D}^u$ is qualified to be a recyclable OOD sample of category c. For a target category c, a qualified recyclable OOD sample should highly probably belong to OOD samples and share class-agnostic features with ID samples belonging to the category c. Therefore, $n_i(c)$ is formalized as follows:

$$n_i(c) = \mathbb{1}[\hat{y}_i^u = c] \cdot \mathbb{1}[g(x_i^u) = \text{OOD}] \cdot \mathbb{1}[p_i^u[c] < \gamma_{\text{OOD}}], \tag{5}$$

where γ_{ood} is a threshold set as 0.2, which prevents us from selecting some ID samples that are wrongly classified as OOD. Considering that each minibatch contains ID samples and not necessarily OOD samples, we store the recyclable OOD samples of each category in a category-wise first-in-first-out memory queue $\mathcal{B}(\cdot)$.

Aliasing OOD Matching. In training iterations, we first collect the $\mathcal{A}_c$ of the current minibatch and then match each ID feature in it with a random OOD feature in $\mathcal{B}(c)$ as ID-OOD pairs $\{t_i\}$:

$$t_i = (z_i; o_i), z_i \in \mathcal{A}_c, o_i \in \mathcal{B}(c). \tag{6}$$

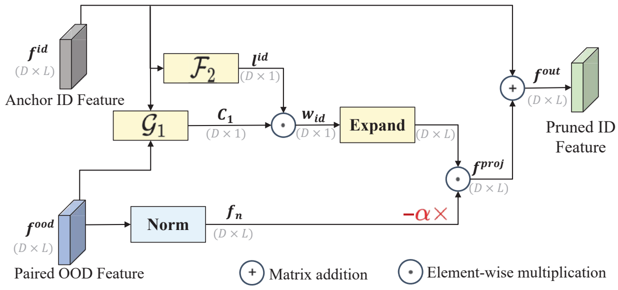

Figure 3. The pipeline of the soft orthogonal decomposition (SOD). The input of SOD is ID-OOD pairs, and its output is pruned ID features. Here, the function $\mathcal{G}_1$ calculates the cosine of the angle between two vectors, while the function $\mathcal{F}(\cdot)$ calculates the l_2-norm of a vector. First, we obtain the cosine of the angle of f^{ID} and f^{OOD}, named as C_1 by $\mathcal{G}_1$. Then we get the l_2-norm l^{ID} of anchor ID feature f^{ID} by $\mathcal{F}(\cdot)$. We normalize anchor ID feature f^{ID} and obtain f_n. Then we get the projector of f^{ID} to f^{OOD}, named as f^{proj}. Finally, we get the pruned ID feature $f^{\text{out}} = f^{\text{ID}} - \alpha f^{\text{proj}}$.

At the end of each iteration, we update each $\mathcal{B}(c)$ by determining whether there are qualified OOD samples (i.e., $n_i(c) = 1$) in this minibatch.

3.4. Soft Orthogonality Regularization

In this section, we introduce our proposed SOR in detail, which includes two parts, as follows:

- We perform a soft orthogonal decomposition (SOD) on ID-OOD pairs to generate pruned ID features.
- We design two losses $\mathcal{L}_{odc}^u$ and $\mathcal{L}_{odc}^l$, which regularize prediction invariance on original ID features and pruned ID features generated by soft orthogonal decomposition.

Proposition 1 *Feature Orthogonal Decomposition (FOD). Any vector $\mathcal{V}$ in the high-dimensional space can be transformed into two mutually orthogonal vectors $\mathcal{V}_a$ and $\mathcal{V}_b$ along a certain basis vector $\mathcal{U}$ direction, formally:*

$$\begin{aligned}
&\vec{\mathcal{V}} = \vec{\mathcal{V}}_a + \vec{\mathcal{V}}_b, \\
&\vec{\mathcal{V}}_a = \vec{\varepsilon} * ||\vec{\mathcal{V}} * \sin < \vec{\mathcal{U}}, \vec{\mathcal{V}} > ||, \\
&\mathcal{V}_b = \vec{\sigma} * ||\vec{\mathcal{V}} * \cos < \vec{\mathcal{U}}, \vec{\mathcal{V}} > ||, \\
&s.t. \quad \vec{\varepsilon} \perp \vec{\mathcal{U}}, \vec{\sigma} \parallel \vec{\mathcal{U}}, \quad ||\vec{\varepsilon}|| = ||\vec{\sigma}|| = 1,
\end{aligned} \tag{7}$$

where ϵ and σ both are unit vectors, and $< \cdot, \cdot >$ denotes the angle between two vectors, $*$ denotes scalar multiplication of vectors.

Soft Orthogonal Decomposition. As shown in Fig. 3, given ID-OOD pairs $t_i = (z_i, o_i)$, SOD applies soft feature orthogonal decomposition on each ID feature z_i^c along with its matching OOD feature o_i. Then we obtain the pruned ID feature $z^c_{i,r}$, which has less similarity with paired OOD features since the OOD semantic component is pruned out of the original ID feature. According to proposition 1, the process is formulated as follows:

$$\begin{aligned} &\vec{z}_i = \vec{z}_{i,a} + \vec{z}_{i,b}, \\ &\vec{z}_{i,r} = \vec{z}_i - \alpha\vec{z}_{i,b}, \\ &s.t.\ \vec{z}_{i,a} \perp \vec{o}_i, \vec{z}^c_{i,b} \parallel \vec{o}_i, \end{aligned} \tag{8}$$

where α (we set α = 0.8) is a hyperparameter to slow down the drastic changes in the feature space caused by FOT, which named soft orthogonal decomposition (SOD). With the pruned ID feature $\vec{z}_{i,r}$ for the anchor ID image $\vec{z}_i$, we obtain its corresponding probability vector $p_{i,r}$ as follows:

$$p_{i,r} = \mathcal{F}(\vec{z}_{i,r}). \tag{9}$$

Orthogonality Regularization Loss. Moreover, we design orthogonality regularization loss $\mathcal{L}^l_{odc}$ and $\mathcal{L}^u_{odc}$ to encourage the predictions of our model to be consistent before and after SOD as:

$$\begin{aligned} \mathcal{L}^l_{odc} &= \frac{1}{\sum_{c=0}^{K}|\mathcal{A}^l_c|}\sum_{c=0}^{K}\sum_{z^l_i\in\mathcal{A}^l_c} KL(p^l_i, p^l_{i,r}) \\ &\quad - \frac{1}{|\mathcal{A}^l_c|}\sum_{z^l_i\in\mathcal{A}^l_c}\ln(p^l_{i,r}[c]), \\ \mathcal{L}^u_{odc} &= \frac{1}{M}\sum_{c=0}^{K}\sum_{z^u_i\in\mathcal{A}^u_c} KL(p^u_i, p^u_{i,r}), \end{aligned} \tag{10}$$

where $\mathcal{L}^l_{odc}$ and $\mathcal{L}^u_{odc}$ are orthogonality regularization losses for labeled and unlabeled data, respectively. For unlabeled data, the $\mathcal{L}^u_{odc}$ is formulated as the KL divergence between p^u_i and $p^u_{i,r}$, while for labeled data, we additionally minimize the cross-entropy between $p^l_{i,r}$ and y^l_i to utilize the label information.

3.5. Total Loss

In this section, we describe the training processing and loss functions in detail. As mention above, we use T2T [23] as our plain baseline.

At pre-training stage, our OSP follows baseline, which learns a K-ways predictor with labeled data and a rotation recognizer [16] with all unlabeled data to enhance the representation capacity. For the K-ways prediction branch, $\mathcal{F}$ calculates a K-dimensional class probability vector $p^l_i = \mathcal{F}\circ\mathcal{G}(x^l_i)$. During training, cross entropy is used to regularize the class probability vectors of labeled images:

$$\mathcal{L}_{ce} = -\frac{1}{||\mathcal{D}^l||}\sum_{(x^l_i,y^l_i)\in\mathcal{D}^l}\log p^l_i[y^l_i], \tag{11}$$

For rotation recognition, we denote four counterparts images x^u_j generated via rotating by $(k-1)*90^\circ$ as $x^u_{j,k}$, then the rotation prediction head $\mathcal{H}(\cdot)$ is responsible for predicting x^u_j with rotation label k with cross entropy loss,

$$\mathcal{L}_{rot} = -\frac{1}{4*||\mathcal{D}^u||}\sum_{(x^u_i)\in\mathcal{D}^u}\sum_{k=1}^{4}\log q^l_{i,k}[k], \tag{12}$$

To sum up, the total loss of OSP at the pre-training stage is described as follows:

$$\mathcal{L}_{pre} = \mathcal{L}_{ce} + \mathcal{L}_{rot} + \mathcal{L}^l_{ood}, \tag{13}$$

where $\mathcal{L}^l_{ood}$ is used to train the OOD detection module $\mathcal{M}(\cdot)$, referring to [23].

At the fine-tuning stage, we apply our proposed orthogonality regularization losses on the baseline, which aims to prune OOD semantic from ID features. Referring to [23], the fine-tuning loss of baseline is described as follows:

$$\mathcal{L}_{t2t} = \underbrace{\mathcal{L}_{ce} + \mathcal{L}_u}_{\text{Classic SSL Loss}} + \underbrace{\mathcal{L}^l_{ood} + \mathcal{L}^u_{ood}}_{\text{OOD Detection Loss}} + \mathcal{L}_{rot}. \tag{14}$$

With our proposed orthogonality regularization losses $\mathcal{L}^l_{odc}$ and $\mathcal{L}^u_{odc}$, the total loss of OSP at the fine-tuning stage is described as follows:

$$\mathcal{L}_{ft} = \mathcal{L}_{t2t} + \underbrace{\mathcal{L}^l_{odc} + \mathcal{L}^u_{odc}}_{\text{Our OSR Loss}} \tag{15}$$

where $\mathcal{L}^l_{ood}$ and $\mathcal{L}^u_{ood}$ are used to train the OOD detection model $\mathcal{D}(\cdot)$ [23].

4. Experiments

4.1. Experimental Setup

Datasets. Referring to [21] [23] [17], we evaluate the effectiveness of our OSP on four widely used datasets: MNIST [53], CIFAR10 [29], CIFAR100 [29] and TinyImageNet [15].

OOD setting. In this paper, we use inter-dataset and intra-dataset OOD settings to verify the superiority of OSP.

(a) Intra-dataset OOD Setting: Following [21] [23] [17], we select some categories as ID categories and the rest as OOD categories in MNIST [53], CIFAR10 [29], CIFAR100 [29] and TinyImageNet (a subset of ImageNet [15]). During training, we random sample labeled and unabeled images for ID categories as ID samples and unlabeled images from OOD categories as OOD samples. For MNIST and CIFAR10, we select first six classes as ID categories. For CIFAR100 and TinyImageNet, we select first 50 classes and 100 classes as ID catrgories, respectively. Moreover, we use the mismatch ratio $\gamma \in [0, 1]$ to adjust the ratio of OOD samples in the unlabeled data, which modulates class distribution mismatch. For example, when the mismatch ratio γ is 0.3, 30% unlabeled samples come from unseen classes. The details are shown in Tab. 3. More details about datasets and settings refers to Appendix.

(b) Inter-dataset OOD setting: Following [23], we random sample ID samples from CIFAR-10 and use other

Method	MNIST		CIFAR10		CIFAR100		TinyImagetNet	
	γ=0.3	γ=0.6	γ=0.3	γ=0.6	γ=0.3	γ=0.6	γ=0.3	γ=0.6
Supervised	93.2	93.2	76.3	76.3	58.6	58.6	36.5	36.5
Classic SSL Methods								
UDA† [54]	-	-	90.7	88.3	67.1	64.5	-	-
Pi-Model [48]	92.4	86.6	75.7	74.5	59.4	57.9	36.9	36.4
PL [32]	90.0	86.0	75.8	74.6	60.2	57.5	36.6	35.8
VAT [38]	94.5	90.4	76.9	75.0	61.8	59.6	36.7	36.3
Fixmatch [50]	-	-	81.5	80.9	65.9	65.2	-	-
Robust SSL Methods								
DS3L [17]	96.8	94.5	78.1	76.9	-	-	-	-
UASD [12]	96.2	94.3	77.6	76.0	61.8	58.4	37.1	36.9
CL [7]	96.9	95.6	83.2	82.1	63.6	61.5	37.3	36.7
Safe-Students [21]	98.3	96.5	85.7	83.8	68.4	68.2	37.7	37.1
MTC [55]	93.7	88.5	85.5	81.7	63.1	61.1	37.0	36.6
T2T [23]	99.1	98.7	**91.6**	**89.3**	70.0	68.2	39.0	35.0
Ours	**99.3**(+0.2)	**99.4**(+0.7)	90.5(-1.1)	88.2(-1.1)	**72.4**(+2.4)	**70.9**(+2.7)	**52.7**(+13.7)	**52.1**(+15.0)

Table 1. **Intra-dataset:** ID categories classification accuracy (%) of different methods on the four datasets. In this paper, the **bold** numbers denote the best results across all approaches. The (+number) denotes the absolute improvements.

Method	TIN		LSUN		Gaussian		Uniform	
	N_l=250	N_l=1000	N_l=250	N_l=1000	N_l=250	N_l=1000	N_l=250	N_l=1000
Classic SSL Methods								
UDA [54]	88.8	91.8	88.5	91.1	88.9	89.2	88.7	89.7
MixM [4]	82.4	88.0	76.3	87.0	75.8	85.7	72.9	84.5
Robust SSL Methods								
DS3L [17]	-	70.1	-	69.7	-	62.9	-	62.9
UASD [12]	83.6	-	-	80.9	-	-	-	-
MTC [55]	86.4	89.9	86.7	90.2	87.3	89.8	85.6	89.9
OTCT [36]	-	91.1	-	91.3	-	92.3	-	91.8
T2T [23]	91.5	93.3	91.1	94.4	90.8	93.6	90.0	94.1
Ours	**92.4**(+0.9)	**93.7**(+0.5)	**91.9**(+0.8)	**94.8**(+0.4)	**91.0**(+0.2)	**93.7**(+0.1)	**90.8**(+0.8)	**94.2**(+0.1)

Table 2. **Inter-dataset:** ID categories classification accuracy (%) of different methods on CIFAR10 and other four datasets as OOD.

Dataset	ID classes	OOD classes	labeled samples N_l	OOD samples
MNIST	6	4	6×10	30,000×γ
CIFAR10	6	4	6×400	20,000×γ
CIFAR100	50	50	50×100	20,000×γ
TinyImageNet	100	100	100×100	40,000×γ

Table 3. Intra-dataset OOD setting details.

dataset to synthesize OOD samples. Specifically, 10,000 unlabeled images are sampled from each of the TIN dataset, the Large-scale Scene Understanding (LSUN) dataset, Guassian noise dataset, and uniform noise dataset, forming into 4 inter-dataset OOD setting.

Metrics. Following [17] [23] [21], we choose the mean accuracy (Acc.) to evaluate the classification performance. For OOD detection, we use the area under the receiver operating characteristic (AUROC) as metrics [23].

Implementation Details. Existing methods including UDA [54], FixMatch [50], VAT [38], PL [32], Pi-Model [48], MTC [55], DS3L [17], UASD [12], CL [7], T2T [23] and Safe-Student [21] are used for comparison. For our method, SGD is used to optimize network weights. The learning rate is initially set to 0.03 at the pre-training stage and 0.001 at the fine-tuning stage, which is adjusted via the cosine decay strategy [50, 54]. The momentum is set to 0.9. In each training batch, the batch size of labeled data and unlabeled data are 64 and 320. And the pre-training stage costs 50,000 iterations, and the fine-tuning stage takes 200,000

$\mathcal{L}^{u}_{odc}$	$\mathcal{L}^{l}_{odc}$	AOM	**TinyImageNet**	**CIFAR100**
			35.0	68.2
	✓	✓	49.5	69.9
✓		✓	48.4	70.4
✓	✓		46.5	70.5
✓	✓	✓	**52.1**	**69.1**

Table 4. Abaltion results on CIFAR100 ($\gamma = 0.6$) and TinyImageNet ($\gamma = 0.6$)

Method	MNIST	CIFAR10	CIFAR100	TinyImageNet
T2T [23]	92.6	67.4	64.8	40.5
Ours	**99.8**	**88.3**	**71.8**	**54.4**

Table 5. The OOD detection performance comparison accross different datasets (AUROC(%)).

iterations. We set the size of recyclable OOD Bank $\mathcal{B}(\cdot)$ is 5000. For UDA [54] and FixMatch [50], models are trained with 250,000 iterations for a fair comparison.For far comparison, when training MTC [55] and T2T [23], we follow their original settings in [55] and [23], respectively. In MNIST, we adopt a simple two-layer CNN model as a backbone network [21] [17], while in CIFAR10, CIFAR100 and TinyImageNet, we use the Wide-ResNet28-2 [56] as the backbone model.

4.2. Main Results.

OOD proportion of datasets. Here, we report the proportion of OOD samples in different datasets to help understand the performances of OSP. As Tab. 3 shows, hard datasets like TinyImageNet contain more OOD classes and samples, for which obtaining a clear ID/OOD discrimination is very hard. In other words, the 'feature aliasing' problem corrupts learning more heavily on hard datasets (e.g., TinyImageNet) than on easy ones (e.g., CIFAR10).

Performance on intra-dataset setting. As shown in Tab. 1, our OSP achieves the best performance on MNIST, CIFAR100, and TinyImageNet with various class mismatch ratios γ. Prominently, on TinyImageNet, most existing methods have low accuracy but our OSP improves the best baseline by **13.7%** and **15.0%** when the class mismatch ratio $\gamma = 0.3$ and 0.6, respectively. This is because our OSP is designed to tackle the "feature aliasing" problem, and this problem matters heavily in hard datasets like TinyImageNet as mentioned above. While for easy datasets, our OSP also obtains competitive performances to SOTA alternatives. These comparisons highlight the superiority of OSP in addressing the corruption from OOD data.

Performance on inter-dataset setting. As shown in Tab. 2, OSP outperforms previous methods on CIFAR10 with various OOD datasets (e.g. TIN, LSUN, Gaussian, and Uniform). This indicates the good versatility of OSP for different OOD sources, reflecting its potential in real complex dataset settings.

Results on various class mismatch ratio. To verify the robustness of our OSP to corruption of unlabeled data, we illustrate the performance of our model under various mismatch ratios in CIFAR100 with 100 labeled data per class. The results are shown in Fig. 4(a). We see that our OSP achieves SOTA in all settings. Moreover, most baselines display significant performance degradation as γ increases, whereas OSP remains competitive. These observations clearly validate the superiority of OSP.

Results on different labeled data amount. Moreover, we further verify the effectiveness of our OSP under different labeled data amounts. Here, we carry out all experiments on CIFAR100 with $\gamma = 0.6$. As shown in Fig. 4(b), our OSP obtains the best performances on all labeled data amount settings, reflecting the broad applicability of our approach. A notable point is that the advantages of previous robust SSL methods [23] [55] gradually fade away with the increase in the amount of labeled data.

4.3. Ablation Studies

Effect of Soft Orthogonality Regularization. To verify the effectiveness of our SOR, we compare four variants: (1) Row 1: the baseline without our proposed AOM and SOR and use Eq.14 as finetuning loss function. (2) Row 2: only applies SOR on labeled ID anchor features $\mathcal{A}^{l}_{c}$. (3) Row 3: only applies SOR on unlabeled ID anchor features $\mathcal{A}^{u}_{c}$. (4) Row 5: our OSP which applies SOR on all ID anchor features $\mathcal{A}^{u}_{c} \cup \mathcal{A}^{l}_{c}$. As shown in Tab. 4, our SOR module outperforms baseline obviously and our proposed regularization loss $\mathcal{A}^{l}_{c}$ and $\mathcal{A}^{u}_{c}$ both contribute to performance improvements.

Effect of Aliasing OOD Matching. To quantify the impact of AOM, we compare two variants: (1) Row 4: random selects OOD features to pair ID features (2) Row 5: our OSP which matches each ID sample with an OOD sample that has a large semantic overlap with it, as described in Sec. 3.3. From Tab. 4, the results indicate that our ID-OOD pairs procedure (AOM) is beneficial to pruning OOD semantic and further improves performance.

4.4. Further Analysis

OOD detection. In Tab. 5, we compare our method against T2T [23] under combinations of ID and OOD datasets, to validate the efficacy of our OSP . The AUROC is used as the metric here. We see that our OSP outperforms T2T [23] under all settings with a large margin. reflecting the superority of OSP in ID/OOD discrimination.

Visualization of class activation map. We use Grad-CAM [49] to visualize the class activation map. As shown in Fig. 5, we notice that the baseline (row 2) is distracted and even focuses on non-foreground object regions, thus

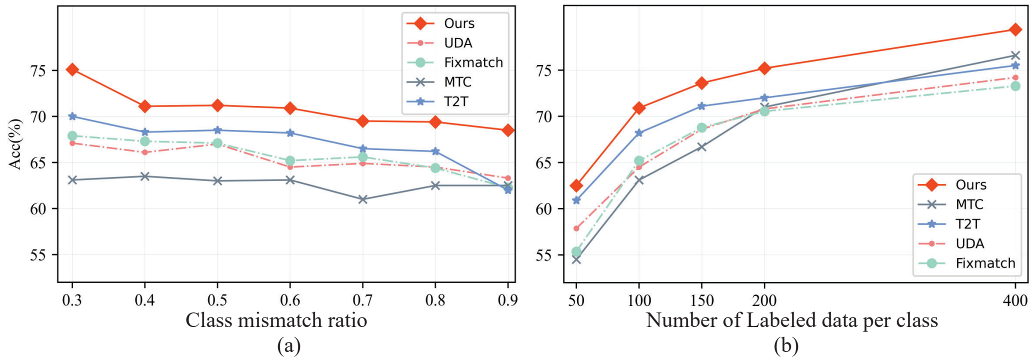

Figure 4. (a) Effect of the class mismatch ratio. (b) Effect of the labeled data amount. All these results are obtained on the CIFAR100 dataset with 100 labeled data per class.

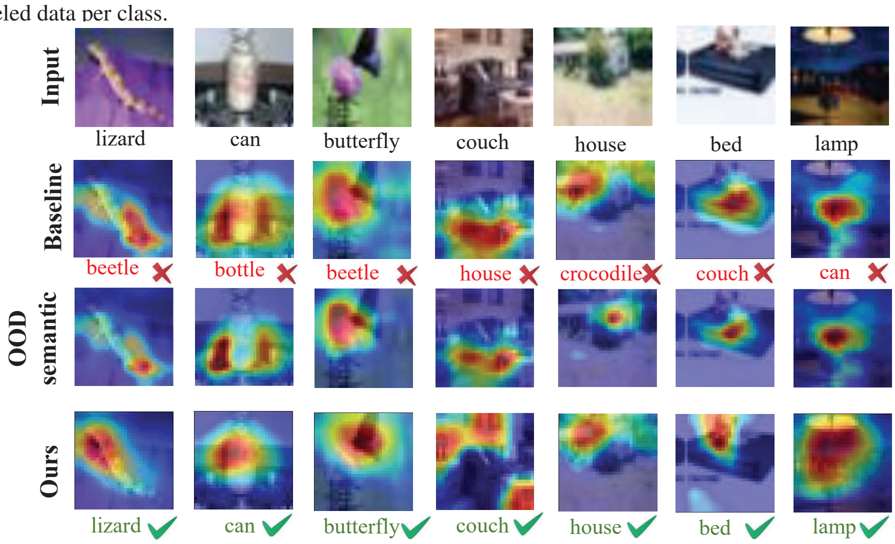

Figure 5. Activation maps of baseline [23] and OSP using Grad-CAM [49]. The red (blue) color represents more (less) attention from the model. Rows 1-4 represent input images, CAMs from baseline [23], paired OOD features in OSP, and CAMs from OSP, respectively.

has wrong predictions. In contrast, OSP focuses on the object regions more accurately and comprehensively (row4), indicating the superiority of OSP in learning semantic structure. This is because OSP encourages our model to only reserve classification-related ID semantics by pruning classification-useless OOD semantics, which is mostly activated in the background region (row 3).

More results on real-world dataset. STL-10 [13] is a dataset for real-world image recognition, while each class has fewer labeled training examples (ID samples) and a very large set of unlabeled OOD examples. The unlabeled OOD samples comes from a similar but different distribution from the labeled data. The primary challenge is to make use of the unlabeled data to improve recognition for the ID samples. Here, we resize the images as 32×32. We applied our OSP on STL-10 with 20,000 OOD samples, 100 labeled and 200 unlabeled ID samples per class. Our OPS improves T2T [23] by 3.1% (Acc.78.0% v.s. 74.9%).

5. Conclusion

In this paper, we introduce a novel method named OSP for robust semi-supervised learning [18, 57], which first exploits the value of OOD at the semantic level. Our OSP mitigates the corruption from OOD samples by pruning OOD semantics out from ID features at the semantics level. Specifically, we propose an aliasing OOD matching module to pair each ID sample with an OOD sample with which it has semantic overlap. We then develop a soft orthogonality regularization to regularize the ID and OOD features to be orthogonal. Further, we will extend our OSP to more challenging open-set scenarios [22, 28, 33].

Acknowledgements. This work was supported in part by the National Key R&D Program of China (No.2022ZD0118201), Natural Science Foundation of China (No.61972217, 32071459, 62176249, 62006133, 62271465), and the Natural Science Foundation of Guangdong Province in China (No.2019B1515120049).

References

[1] Maximilian Augustin and Matthias Hein. Out-distribution aware self-training in an open world setting. *arXiv: Learning*, 2020.

[2] David Berthelot, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. *arXiv: Learning*, 2019.

[3] David Berthelot, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. *In ICLR*, 2019.

[4] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. Mixmatch: A holistic approach to semi-supervised learning. *In NeurIPS*, 2019.

[5] Zhaowei Cai, Avinash Ravichandran, Paolo Favaro, Manchen Wang, Davide Modolo, Rahul Bhotika, Zhuowen Tu, and Stefano Soatto. Semi-supervised vision transformers at scale. *In NeurIPS*, 2022.

[6] Kaidi Cao, Maria Brbic, and Jure Leskovec. Open-world semi-supervised learning. *In ICLR*, 2021.

[7] Paola Cascante-Bonilla, Fuwen Tan, Yanjun Qi, and Vicente Ordonez. Curriculum labeling: Revisiting pseudo-labeling for semi-supervised learning. *In AAAI*, pages 6912–6920, 2021.

[8] Dong-Dong Chen, Wei Wang, Wei Gao, and Zhi-Hua Zhou. Tri-net for semi-supervised deep learning. *In IJCAI*, pages 2014–2020, 2018.

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. *In ICML*, pages 1597–1697, 2020.

[10] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E. Hinton. Big self-supervised models are strong semi-supervised learners. *In NeurIPS*, 2020.

[11] Yanbei Chen, Massimiliano Mancini, Xiatian Zhu, and Zeynep Akata. Semi-supervised and unsupervised deep visual learning: A survey. *In TPAMI*, 2022.

[12] Yanbei Chen, Xiatian Zhu, Wei Li, and Shaogang Gong. Semi-supervised learning under class distribution mismatch. *In AAAI*, pages 3569–3576, 2020.

[13] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. *In AISTATS*, pages 215–223, 2011.

[14] Ekin D. Cubuk, Barret Zoph, Dandelion Mane, Vijay K. Vasudevan, and Quoc V. Le. Autoaugment: Learning augmentation strategies from data. *In CVPR*, pages 113–123, 2019.

[15] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. *In CVPR*, pages 248–255, 2009.

[16] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *In ICLR*, 2018.

[17] 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 ICML*, pages 3897–3906, 2020.

[18] Lan-Zhe Guo, Zhi Zhou, and Yu-Feng Li. Robust deep semi-supervised learning: A brief introduction. *arXiv: Learning*, 2022.

[19] Zhongyi Han, Xian-Jin Gui, Chaoran Cui, and Yilong Yin. Towards accurate and robust domain adaptation under noisy environments. *In IJCAI*, pages 2269–2276, 2020.

[20] Zhongyi Han, Benzheng Wei, Xiaoming Xi, Bo Chen, Yilong Yin, and Shuo Li. Unifying neural learning and symbolic reasoning for spinal medical report generation. *Medical Image Analysis*, 67:101872, 2021.

[21] Rundong He, Zhongyi Han, Xiankai Lu, and Yilong Yin. Safe-student for safe deep semi-supervised learning with unseen-class unlabeled data. *In CVPR*, pages 14585–14594, 2022.

[22] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *In ICLR*, 2016.

[23] Junkai Huang, Chaowei Fang, Weikai Chen, Zhenhua Chai, Xiaolin Wei, Pengxu Wei, Liang Lin, and Guanbin Li. Trash to treasure: Harvesting ood data with cross-modal matching for open-set semi-supervised learning. *In CVPR*, pages 8310–8319, 2021.

[24] Zhuo Huang, Chao Xue, Bo Han, Jian Yang, and Chen Gong. Universal semi-supervised learning. *In NeurIPS*, pages 26714–26725, 2021.

[25] Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondrej Chum. Label propagation for deep semi-supervised learning. *In CVPR*, pages 5070–5079, 2019.

[26] Juho Kannala, Alex Lamb, Kenji Kawaguchi, Vikas Verma, Yoshua Bengio, David Lopez-Paz, Vikas Verma, Kenji Kawaguchi, Alex Lamb, Juho Kannala, Yoshua Bengio, and David Lopez-Paz. Interpolation consistency training for semi-supervised learning. *arXiv: Machine Learning*, 2019.

[27] Rihuan Ke, Angelica I. Aviles-Rivero, Saurabh Pandey, Saikumar Reddy, and Carola-Bibiane Schönlieb. A three-stage self-training framework for semi-supervised semantic segmentation. *In TIP*, 31:1805–1815, 2022.

[28] Jaehyung Kim, Youngbum Hur, Sejun Park, Eunho Yang, Sung Ju Hwang, and Jinwoo Shin. Distribution aligning refinery of pseudo-label for imbalanced semi-supervised learning. *In NeurIPS*, 2020.

[29] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[30] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

[31] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *In ICLR*, 2016.

[32] Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. *In ICML Workshops*, pages 1163–1171, 2022.

[33] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *In NeurIPS*, 2018.

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

[35] Yu-Feng Li and Zhi-Hua Zhou. Towards making unlabeled data never hurt. *In TPAMI*, pages 175–188, 2015.

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

[37] Huixiang Luo, Hao Cheng, Fanxu Meng, Yuting Gao, Ke Li, Mengdan Zhang, and Xing Sun. An empirical study and analysis on open-set semi-supervised learning. *arXiv preprint*, 2021.

[38] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *In TPAMI*, 41(8):1979–1993, 2018.

[39] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. *In ECCV*, pages 69–84, 2016.

[40] Avital Oliver, Augustus Odena, Colin A Raffel, Ekin Dogus Cubuk, and Ian Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. *In NeurIPS*, 31:3239–3250, 2018.

[41] Nobuyuki Otsu. A threshold selection method from gray level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 1979.

[42] Hieu Pham, Zihang Dai, Qizhe Xie, and Quoc V Le. Meta pseudo labels. pages 11557–11568, 2021.

[43] Siyuan Qiao, Wei Shen, Zhishuai Zhang, Bo Wang, and Alan L. Yuille. Deep co-training for semi-supervised image recognition. *In ECCV*, pages 142–159, 2018.

[44] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *In ICML*, pages 8748–8763, 2021.

[45] Zhongzheng Ren, Raymond A. Yeh, and Alexander G. Schwing. Not all unlabeled data are equal: Learning to weight data in semi-supervised learning. *In NeurIPS*, 33:21786–21797, 2020.

[46] Mamshad Nayeem Rizve, Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. *In ICLR*, 2021.

[47] Kuniaki Saito, Donghyun Kim, and Kate Saenko. Openmatch: Open-set consistency regularization for semi-supervised learning with outliers. *arXiv: Computer Vision and Pattern Recognition*, 2021.

[48] Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. *In NeurIPS*, pages 1163–1171, 2016.

[49] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *In IJCV*, 2016.

[50] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *arXiv preprint arXiv:2001.07685*, 2020.

[51] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *In NeurIPS*, pages 1195–1204, 2017.

[52] Laurens Van Der Maaten. Learning a parametric embedding by preserving local structure. *In PMLR*, pages 384–391, 2009.

[53] Hayden Walles, Anthony Robins, Alistair Knott, Hayden Walles, Anthony Robins, and Alistair Knott. the mnist handwritten digit database. *In IEEE Signal Process. Mag.*, 29(6):141–142, 2012.

[54] Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. Unsupervised data augmentation for consistency training. *In NeruIPS*, page 6256–6268, 2019.

[55] Qing Yu, Daiki Ikami, Go Irie, and Kiyoharu Aizawa. Multi-task curriculum framework for open-set semi-supervised learning. *In ECCV*, pages 438–454, 2020.

[56] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *In BMVC*, 2016.

[57] Xujiang Zhao, Killamsetty Krishnateja, Rishabh Iyer, and Feng Chen. Robust semi-supervised learning without of distribution data. *arXiv preprint arXiv:2010.03658*, 2020.