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

Yu Wang\textsuperscript{1,4*} Pengchong Qiao\textsuperscript{1,2,4*} Chang Liu\textsuperscript{3} Guoli Song\textsuperscript{2,4} Xiawu Zheng\textsuperscript{2,4†} Jie Chen\textsuperscript{1,2,4†}

\textsuperscript{1} School of Electronic and Computer Engineering, Peking University, Shenzhen, China
\textsuperscript{2} Peng Cheng Laboratory, Shenzhen, China
\textsuperscript{3} Department of Automation and BNRist, Tsinghua University, Beijing, China
\textsuperscript{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].

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-
by 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-unease 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 TinyImageNet 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 \[3, 27, 32, 34, 42\]. 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\[1\] 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 \(\mathcal{D}^u = \{(x_i^u)\}_{i=1}^{N_u} (N_u \ll N_l)\), where \(x_i^l, y_i^l\) and \(x_i^u\) are the image and label of the \(i\)-th la-
beled data and the image of the $i$-th unlabeled data. The label space of labeled data contains $K$ labels, that is, $y^l_i \in C^l = \{1, \ldots, K\}$. 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 $C^{OOD} = C^u \setminus C^l$. 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 $G(\cdot)$, a K-ways classifier $F(\cdot)$, a rotation prediction head $H(\cdot)$ and an OOD detection module $M(\cdot)$. Different from T2T [23], we design two novel modules, namely 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 $D^l$ and a self-supervised task [39] [9] (i.e., rotation recognition [16]) on $D^u$ to pre-train the encoder $G(\cdot)$, the classifier $F(\cdot)$, and the rotation predictor $H(\cdot)$. Given a labeled input $x^l_i \in D^l$ and an unlabeled input $x^u_j \in D^u$, we denote their representations as $z^l_i = G(x^l_i)$ and $z^u_j = G(x^u_j)$.

The training of model parameters is optimized by minimizing a supervised cross-entropy loss $L_{ce}$ and a rotation loss $L_{rot}$. Details are described in Sec. 3.5.

Meanwhile, we pre-train the OOD detection module $M(\cdot)$ on $D^l$ 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:

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

where $\gamma$ 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 $G(\cdot)$, 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 $M(\cdot)$ 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 $\mathcal{A}^l_c$, which contains features of labeled images with high confidence. Formally,

$$\mathcal{A}^l_c = \{ z_i | z_i = G(x^l_i), y_i^l = c, p_i^l[c] > \delta \},$$  \hspace{1cm} (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 $x^l_i$, respectively. Here, $\delta$ 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 $\mathcal{A}^u_c$ is computed as:

$$\mathcal{A}^u_c = \{ z_i^u | z_i^u = G(x^u_i), \hat{y}_i^u = c, max_v(p_i^u[v]) > \delta \},$$ \hspace{1cm} (3)

where $\hat{y}_i^u = \arg \max_v(p_i^u[v])$ is the pseudo label of the image $x^u_i$. This $\mathcal{A}^u_c$ is similar to $\mathcal{A}^l_c$, the only difference is that it uses the pseudo-label for class determination. Based on $\mathcal{A}^l_c$ and $\mathcal{A}^u_c$, we obtain the set of all qualified ID anchors $\mathcal{A}_c$:

$$\mathcal{A}_c = \mathcal{A}^l_c \cup \mathcal{A}^u_c. \hspace{1cm} (4)$$

**Recollectable OOD samples.** We define a binary variable $n_i(c)$ to identify whether an unlabeled image $x^u_i \in \mathcal{D}^u$ is qualified to be a collectable OOD sample of category $c$. For a target category $c$, a qualified collectable 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:

$$n_i(c) = 1[\hat{y}_i^u = c] \cdot 1[g(x^u_i) = \text{OOD}] \cdot 1[p_i^u[c] < \gamma_{ood}],$$ \hspace{1cm} (5)

where $\gamma_{ood}$ 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 collectable OOD samples of each category in a category-wise first-in-first-out memory queue $\mathcal{B}(c)$.

**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 $\{t_i\}$:

$$t_i = (z_i; o_i), z_i \in \mathcal{A}_c, o_i \in \mathcal{B}(c). \hspace{1cm} (6)$$

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 $\mathcal{L}^n_{gcd}$ and $\mathcal{L}^t_{gcd}$, 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 $V$ in the high-dimensional space can be transformed into two mutually orthogonal vectors $V_a$ and $V_b$ along a certain basis vector $U$ direction, formally:

$$\tilde{V} = \tilde{V}_a + \tilde{V}_b,$$

$$\tilde{V}_a = \tilde{\epsilon} \cdot ||\tilde{V} \cdot \sin \phi < \tilde{U}, \tilde{V} > ||,$$

$$\tilde{V}_b = \tilde{\sigma} \cdot ||\tilde{V} \cdot \cos \phi < \tilde{U}, \tilde{V} > ||,$$

s.t. $\tilde{\epsilon} \perp \tilde{U}, \tilde{\sigma} \parallel \tilde{U}, \ ||\tilde{\epsilon}|| \cdot ||\tilde{\sigma}|| = 1,$$ \hspace{1cm} (7)

where $\epsilon$ and $\sigma$ both are unit vectors, and $< \cdot, \cdot >$ denotes the angle between two vectors, $\cdot$ 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 $z^c_i$ along with
its matching OOD feature \(a_i\). Then we obtain the pruned ID feature \(\tilde{z}_{i,a}\), 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:

\[
\begin{align*}
\tilde{z}_i &= \tilde{z}_{i,a} + \tilde{z}_{i,b}, \\
\tilde{z}_{i,r} &= \tilde{z}_i - \alpha \tilde{z}_{i,b},
\end{align*}
\]

\(\alpha\) (we set \(\alpha = 0.8\)) 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 \(\tilde{z}_{i,r}\) for the anchor ID image \(\tilde{z}_i\), we obtain its corresponding probability vector \(p_i, r\) as follows:

\[
p_{i, r} = F(\tilde{z}_{i, r}).
\]

**Orthogonality Regularization Loss.** Moreover, we design orthogonality regularization loss \(L^l_{\text{ode}}\) and \(L^l_{\text{ode}}\) to encourage the predictions of our model to be consistent before and after SOD as:

\[
L^l_{\text{ode}} = \frac{1}{\sum_{c=0}^{K} |A|} \sum_{c=0}^{K} \sum_{x_i \in A} KL(p_i^l, p_{i, r}^l)
\]

\[
= \frac{1}{|A|} \sum_{x_i \in A} \ln(p_i^l[c]),
\]

\[
L^u_{\text{ode}} = \frac{1}{\sum_{c=0}^{K} |A|} \sum_{c=0}^{K} \sum_{x_i \in A} KL(p_i^u, p_{i, r}^u),
\]

where \(L^l_{\text{ode}}\) and \(L^u_{\text{ode}}\) are orthogonality regularization losses for labeled and unlabeled data, respectively. For unlabeled data, the \(L^u_{\text{ode}}\) is formulated as the KL divergence between \(p_i^u\) and \(p_{i, r}^u\), while for labeled data, we additionally minimize the cross-entropy between \(p_i^l, r\) and \(y_i\) to utilize the label information.

### 3.5. Total Loss

In this section, we describe the training processing and loss functions in detail. As mentioned above, we use T2T [23] as our 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, \(F\) calculates a K-dimensional class probability vector \(p_i^l = F \circ G(x_i)\). During training, cross entropy is used to regularize the class probability vectors of labeled images:

\[
L_{ce} = -\frac{1}{|D^l|} \sum_{(x_i^l, y_i^l) \in D^l} \log p_i^l[y_i^l],
\]

For rotation recognition, we denote four counterparts images \(x_{i,k}^u\) generated via rotating by \((k-1) \cdot 90^\circ\) as \(x_{i,k}^u\), then the rotation prediction head \(H(\cdot)\) is responsible for predicting \(x_{i,k}^u\) with rotation label \(k\) with cross entropy loss:

\[
L_{rot} = -\frac{1}{|D^u|} \sum_{(x_i^u, k) \in D^u} \log q_i^u[k],
\]

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

\[
L_{pre} = L_{ce} + L_{rot} + L_{ood}^l + L_{ood}^u.
\]

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:

\[
L_{ft} = L_{ce} + L_{ood}^l + L_{ood}^u + L_{rot}.
\]

With our proposed orthogonality regularization losses \(L_{ood}^l\) and \(L_{ood}^u\), the total loss of OSP at the fine-tuning stage is described as follows:

\[
L_{ft} = L_{pre} + L_{ood}^l + L_{ood}^u.
\]

where \(L_{ood}^l\) and \(L_{ood}^u\) are used to train the OOD detection model \(D(\cdot)\) [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 unlabeled 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 categories, 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 \(\gamma = 0.3\), 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 randomly sample ID samples from CIFAR-10 and use other
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.

<table>
<thead>
<tr>
<th>Method</th>
<th>MNIST</th>
<th>CIFAR10</th>
<th>CIFAR100</th>
<th>TinyImagNet</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>γ=0.3</td>
<td>γ=0.6</td>
<td>γ=0.3</td>
<td>γ=0.6</td>
</tr>
<tr>
<td>Supervised</td>
<td>93.2</td>
<td>93.2</td>
<td>76.3</td>
<td>76.3</td>
</tr>
</tbody>
</table>

**Classic SSL Methods**

- UDA† [54]: 90.7 / 88.3 / 67.1 / 64.5
- Pi-Model [48]: 92.4 / 86.6
- PL [32]: 90.0 / 86.0
- VAT [38]: 94.5 / 90.4
- Fixmatch [50]: 81.5 / 80.9

**Robust SSL Methods**

- DS3L [17]: 96.8 / 94.5 / 87.1 / 76.9
- UASD [12]: 96.2 / 94.3 / 77.6 / 76.0
- CL [7]: 96.9 / 95.6 / 83.2 / 82.1
- Pi-Model [48]: 93.7 / 88.5 / 85.5 / 81.7
- MTC [55]: 99.1 / 98.7
- T2T [23]: 96.8 / 94.5

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

<table>
<thead>
<tr>
<th>Dataset</th>
<th>ID classes</th>
<th>OOD classes</th>
<th>labeled samples N_l</th>
<th>OOD samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>MNIST</td>
<td>6</td>
<td>4</td>
<td>6×10</td>
<td>30,000×γ</td>
</tr>
<tr>
<td>CIFAR10</td>
<td>6</td>
<td>4</td>
<td>6×400</td>
<td>20,000×γ</td>
</tr>
<tr>
<td>CIFAR100</td>
<td>50</td>
<td>50</td>
<td>50×100</td>
<td>20,000×γ</td>
</tr>
<tr>
<td>TinyImageNet</td>
<td>100</td>
<td>100</td>
<td>100×100</td>
<td>40,000×γ</td>
</tr>
</tbody>
</table>

**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 iterations.
Table 4. Abalation results on CIFAR100 (γ = 0.6) and TinyImageNet (γ = 0.6)

<table>
<thead>
<tr>
<th>Method</th>
<th>MNIST</th>
<th>CIFAR10</th>
<th>CIFAR100</th>
<th>TinyImageNet</th>
</tr>
</thead>
<tbody>
<tr>
<td>T2T [23]</td>
<td>92.6</td>
<td>67.4</td>
<td>64.8</td>
<td>40.5</td>
</tr>
<tr>
<td>Ours</td>
<td>99.8</td>
<td>88.3</td>
<td>71.8</td>
<td>54.4</td>
</tr>
</tbody>
</table>

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

iterations. We set the size of recyclable OOD Bank $B_h(\cdot)$ 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 $\gamma$. 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 $\gamma$ 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 $A^l_c$. (3) Row 3: only applies SOR on unlabeled ID anchor features $A^u_c$. (4) Row 5: our OSP which applies SOR on all ID anchor features $A^l_c \cup A^u_c$. As shown in Tab. 4, our SOR module outperforms baseline obviously and our proposed regularization loss $A^l_c$ and $A^u_c$ 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 superiority of OSP in ID/OOD discrimination.

Visualization of class activation map. We use GradCAM [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


