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

A. The performance of OOD detection

To further measure the potential to identify OOD classes, we compare our OSP with T2T [4] in CIFAR100 with 100 labeled data per class. We utilize AUROC [3, 4] to evaluate performance for OOD detection. The results are shown in Table. 1. Our OSP surpasses the baseline [4] under all setting. This indicates that our OSP promotes OOD detection by keeping ID features and OOD semantic orthogonal.

	Class Mismatch Ratio						
Method	0.3	0.4	0.5	0.6	0.7	0.8	0.9
T2T [4]	65.5	61.0	59.0	60.9	56.3	56.7	59.4
Ours	68.3	73.2	69.4	69.1	66.7	64.3	63.0

Table 1. AUROC(%) for OOD detection on CIFAR100 with 100 labeled data per class.

B. Ablation Study

Effect of soft weight α. The parameter α is a hyper-parameters to adjust the drastic changes in the feature space caused by the orthogonal operation (in Eq. **??**). We conduct ablation experiments on different soft weights α to explore the effect of it on OSP. As shown in Fig. 1(a), small α weakens the effect of our OSP, while large α leads to dramatic changes in feature space. Given our observation of the trade-off, we adopt $\alpha = 0.8$ in all our experiments.

C. Further Analysis

Analysis of the angle of ID and OOD features. As shown in Fig. 1 (b), we see the baseline T2T has an angle around $50°$, the cosine similarity is 64% (i.e., $\cos(50°)$), which means there is an amount of meaning aliasing between ID and OOD features. In contrast, the feature angles after our OSP are around $80°$, which remarkably suppresses their similarity to about only 17% (i.e., $\cos(80°)$). This indicates that our model effectively prunes OOD semantics out from ID features, enhancing the discrimination of ID and OOD samples.

Analysis on the inter-class variance. As shown in Fig. 1 (c), our OSP obtains a significantly larger inter-class variance than the baseline [4], reflecting OSP obtains inter-class discrimination with higher generalizability [9]. Another interesting property is that OSP encourages the inter-class variance to increase within training, whereas the baseline [4] does not. This suggests that OSP progressively acquires discriminative ID class semantics during training.

D. Datasets

D.1. intra-dataset setting

For intra-dataset setting, we follow [3] [2] to evaluate OSP on image classification datasets: MNIST [6], CIFAR10 [5], CIFAR100 [5] and TinyImageNet(a subset of ImageNet [1]), with different class mismatch ratio γ.

MNIST includes 60,000 training images and 10,000 testing images of size 28×28, which contains 10 categories from digit *0* to digit *9*. In this paper, we consider the first six categories (from digit *0* to digit *5*) as $\mathcal{C}^l$ and the remaining categories as OOD categories, $\mathcal{C}^{ood}$. Moreover, we respectively select ten images from $\mathcal{C}^l$ to construct the labeled data set $\mathcal{D}^l$, i.e., a total of 60 labeled data, and select 30,000 images in total from digit *0* to digit *9* as unlabeled data $\mathcal{D}^u$ .Moreover, we use the mismatch ratio γ to adjust the ratio of OOD samples in the unlabeled data to modulate class distribution mismatch. For example, when the extent of labeled/unlabeled class mismatch ratio is 0%, all unlabeled data come from digit *0* to digit *5*.

CIFAR10 includes 60,000 training images and 10,000 testing images of size 32×32 which contains ten categories: *airline, automobile, bird, cat, deer, dog, frog, horse, ship* and *trunk*. Our experiment carries out six-categories classification tasks. We consider animal categories (*birds, cats, deer, dogs, frogs and horses*) as ID categories and the rest as OOD categories. We select 400 images from each ID category to construct the labeled data set $\mathcal{D}^l$, i.e., 2400 labeled instances. Meanwhile, 20,000 images in total are randomly selected as the unlabeled data set $\mathcal{D}^u$ from all the ten categories. We adjust the ratio of OOD images in the unlabeled data to modulate class distribution mismatch γ.

CIFAR100 includes 50,000 training images and 10,000 testing images of size 32×32 which contains 100 cate-

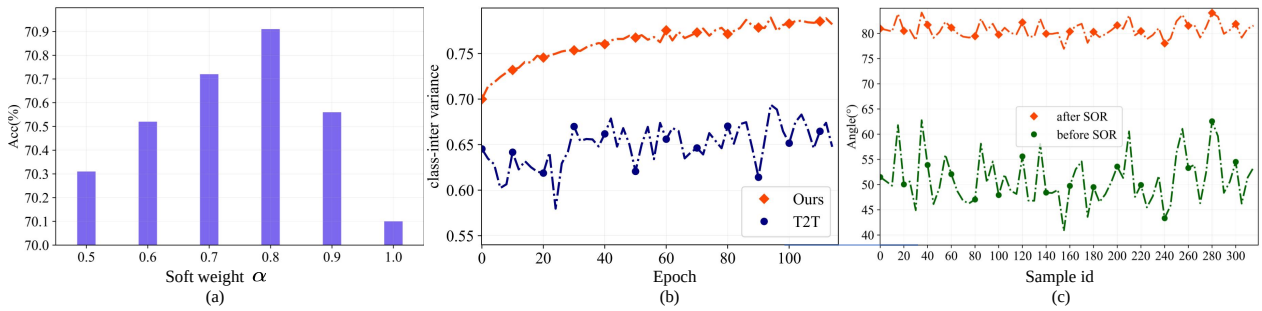

Figure 1. (a) The ablation study about soft weight α. (b) The class-inter variance between ID features. (c) The angle between ID and OOD features. All these results are obtained on CIFAR100 with 100 labeled data per class and γ=0.6.

gories. We use the first half categories (1-50) as ID categories, and the remaining classes as OOD categories. We select 100 images from each ID category to construct the labeled data set $\mathcal{D}^l$, i.e., 5000 labeled instances. Meanwhile, 20,000 images in total are randomly selected as the unlabeled data set $\mathcal{D}^u$ from all the 100 categories with different ratios of OOD classes.

TinyImagetNet contains 200 categories which includes 500 training images and 50 testing images in each category. We resize all images to 32 × 32. We use the first 100 categories as ID classes, and the remaining classes as OOD categories. We select 100 images from each ID category to construct the labeled data set $\mathcal{D}^L$, i.e., 10000 labeled instances. Meanwhile, 40,000 images in total are randomly selected as the unlabeled data set $\mathcal{D}^u$ from all the 200 categories with different ratios of OOD classes.

D.2. inter-dataset setting

For inter-dataset setting, we follow [4] to evaluate OSP on CIFAR10 [5] with different amounts of labeled data. Here, CIFAR10 [5] is used as ID samples, and we samples 10,000 images from other dataset as OOD samples, e.g. the TIN dataset, the Large-scale Scene Understanding (LSUN) dataset [7], Gaussian noise dataset, and uniform noise dataset, forming into our inter-dataset settings. For CIFAR10 [5], following the original split, 10,000 images are used for testing and the same splits in [4] [8] are adopted for training and validating.

E. Algorithm

We provide our training algorithm in Alg. 1. The training processing consists of two stages: pre-training stage and fine-tuning stage.

References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. *computer vision and pattern recognition*, 2009. 1

[2] Lan-Zhe Guo, Zhen-Yu Zhang, Yuan Jiang, Yu-Feng Li, and Zhi-Hua Zhou. Safe deep semi-supervised learning for unseen-class unlabeled data. In *International Conference on Machine Learning*, pages 3897–3906. PMLR, 2020. 1

[3] Rundong He, Zhongyi Han, Xiankai Lu, and Yilong Yin. Safe-student for safe deep semi-supervised learning with unseen-class unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14585–14594, 2022. 1

[4] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8310–8319, 2021. 1, 2

[5] Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009. 1, 2

[6] Hayden Walles, Anthony Robins, Alistair Knott, Hayden Walles, Anthony Robins, and Alistair Knott. the mnist handwritten digit database. 2011. 1

[7] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015. 2

[8] Qing Yu, Daiki Ikami, Go Irie, and Kiyoharu Aizawa. Multitask curriculum framework for open-set semi-supervised learning. In *European Conference on Computer Vision*, pages 438–454. Springer, 2020. 2

[9] Shaofeng Zhang, Lyn Qiu, Feng Zhu, Junchi Yan, Hengrui Zhang, Rui Zhao, Hongyang Li, and Xiaokang Yang. Align representations with base: A new approach to self-supervised learning. 2022. 1

Algorithm 1 Training algorithm

Input: Labeled data, $\mathbb{D}^l$, a set of unlabeled data $\mathbb{D}^u$, an encoder $\mathcal{G}(\cdot)$, a K-ways classifier $\mathcal{F}(\cdot)$, a rotation prediction head $\mathcal{H}(\cdot)$, an OOD detection module $\mathcal{M}(\cdot)$, pre-training epochs E_1, Fine-tuning epochs E_2, max iteration per epochs I, temperature T.

Output: Trained encoder $\mathcal{G}(\cdot)$ and Trained K-ways classifier $\mathcal{F}(\cdot)$.

/********************** Pre-training Stage ************************/**
for e = 1...E_1 **do**
  for i = 1...I **do**
    compute $\mathcal{L}_{pre} = \mathcal{L}_{ce} + \mathcal{L}_{rot} + \mathcal{L}^l_{ood}$ ▷ Eq. 13
    update $\mathcal{G}(\cdot)$,$\mathcal{F}(\cdot)$, $\mathcal{H}(\cdot)$,$\mathcal{M}(\cdot)$ $\longleftarrow$ SGD with $\mathcal{L}_{pre}$.
/********************** Fine-tuning Stage ************************/**
Initialize OOD samples set $\mathbb{U}^{ood} = \emptyset$.
for e = 1...E_2 **do**
  if $e\%10 = 0$ **then**
    update $\mathbb{U}^u$, $\mathbb{U}^{ood}$ $\longleftarrow$ **split old unlabeled data** $\mathbb{D}^u$ **with** $\mathcal{M}(\cdot)$. ▷ Eq. 6
  if $\mathbb{U}^{ood} \neq \emptyset$ **then**
    select recyclable OOD samples for $\mathbb{U}^{ood}$ **and update recyclable OOD Bank** $\mathcal{B}(c)$ $(c = 1...K)$. ▷ Eq. 5

  for i = 1...I **do**
    $(\mathbf{B}_l = \{x^l, y^l\}, \mathbf{B}_u = \{x^u\}) \leftarrow \mathrm{SampleBatch}(\mathbb{D}^l, \mathbb{D}^u)$.
    $z^l_i = \mathcal{G}(x^l_i)$ and $z^u_j = \mathcal{G}(x^u_j)$,where, $x^l_i \in \mathbf{B}_l, x^u_j \in \mathbf{B}_u$.
    for c = 1...K **do** // compute anchor ID samples set
      $\mathcal{A}^l_c = \{z^l_i | z^l_i = \mathcal{G}(x^l_i), y^l_i = c, p^l_i[c] > \delta\}$, ▷ Eq. 2
      $\mathcal{A}^u_c = \{z^u_i | z^u_i = \mathcal{G}(x^u_j), \hat{y}^u_j = c, p(z^u_j) > \delta\}$, ▷ Eq. 3
      $\mathcal{A}_c = \mathcal{A}^l_c \cup \mathcal{A}^u_c$. ▷ Eq. 4
    obtain all anchor ID samples set $\mathcal{A} = \{\mathcal{A}_c\}^K_{c=1}$.
    compute ID-OOD pairs t_i for each z_i in $\mathcal{A}$ // our AOM module ▷ Eq. 6
    compute pruned ID features $z_{i,r}$ for each t_i in $\mathcal{A}$. // our SOT module ▷ Eq. 8,9
    compute $\mathcal{L}_{ft} = \underbrace{\mathcal{L}_{ce} + \mathcal{L}_u}_{\text{Classic SSL Loss}} + \underbrace{\mathcal{L}^l_{ood} + \mathcal{L}^u_{ood}}_{\text{OOD Detection Loss}} + \underbrace{\mathcal{L}^l_{odc} + \mathcal{L}^u_{odc}}_{\text{Our OSR Loss}} + \mathcal{L}_{rot}$ ▷ Eq. 15

    update $\mathcal{G}(\cdot)$,$\mathcal{F}(\cdot)$, $\mathcal{H}(\cdot)$,$\mathcal{M}(\cdot)$ $\longleftarrow$ SGD with $\mathcal{L}_{ft}$.
return encoder $\mathcal{G}(\cdot)$ and K-ways classifier $\mathcal{F}(\cdot)$