Rethinking Safe Semi-supervised Learning: Transferring the Open-set Problem to A Close-set One
-Supplementary Material-

1. Detailed Datasets

In this section, we will give a more detailed introduction to the four public benchmark datasets used in the experiments.

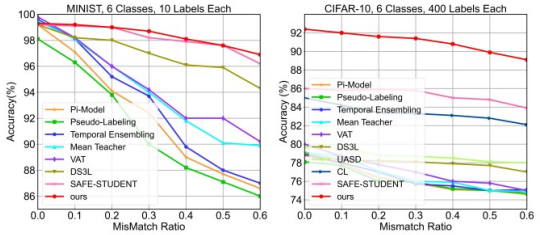
MNIST contains 10 classes from digit "0" to digit "9", including 60,000 training images and 10,000 testing images of size 28×28 [5]. We select 10 images from each class of "0"-"5" as the labeled set, and 30,000 images from all classes as unlabeled data.

CIFAR-10 has 10 classes, each containing 6,000 natural images of size 32×32 [4]. We set the animal classes (bird, cat, deer, dog, frog, horse) as seen classes and the rest (airline, automobile, ship, trunk) as unseen classes. 2,400 labeled images (400 from each seen class) and 20,000 unlabeled images (randomly sampled from all classes) are selected for training dataset.

CIFAR-100 is an extension of CIFAR-10 which has 100 classes [4]. We set the first 50 classes as seen classes and the rest as unseen classes. For training, 5,000 labeled images are selected (100 from each seen class) to construct the labeled set, and the rest data settings remain the same as CIFAR-10.

TinyImageNet is a subset of ImageNet [1], consisting 120,000 images of 200 classes. We resize all images to 32×32. The first 100 classes are set as seen classes and the rest as unseen classes. We select 100 images from each seen class to build the labeled set. Meanwhile, 40,000 images are randomly sampled from all classes to construct the unlabeled set.

Note that, to investigate the effectiveness of our method under different amounts of OOD data, we select OOD data from the unseen classes of the unlabeled set according to different mismatch ratios. For example, the 0% mismatch ratio represents that there is no OOD data and all the unlabeled data is from the seen classes. The 50% mismatch ratio indicates half of the unlabeled data comes from the unseen classes and the other half from the seen classes. For reference, when all the images are labeled, our method achieves 99.6%, 54.0%, 94.7% and 76.0% seen-class classification accuracy on MNIST, TinyImageNet, CIFAR-10 and CIFAR-100, respectively.



(a) (b)

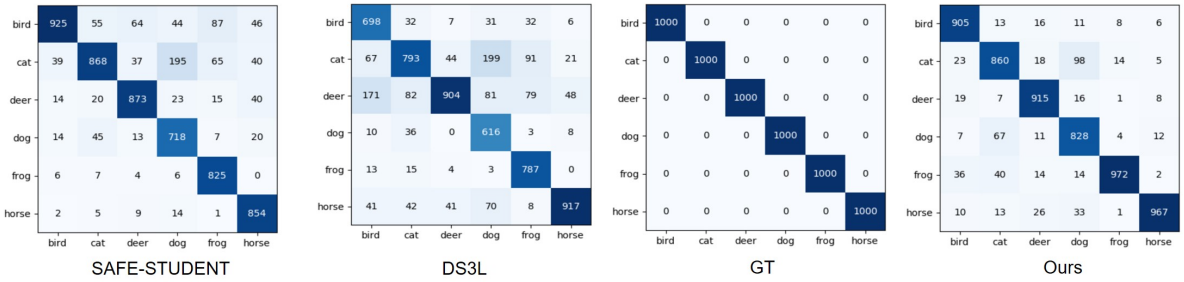
Figure 1. Seen-class classification accuracy (%) of our method and compared methods on MNIST and CIFAR-10 with different mismatch ratios.

2. Qualitative Comparison with SOTA

Figure 1 shows the average accuracy of different SSL methods over five runs with different mismatch ratios on MNIST and CIFAR-10. From Figure 1(a), we can notice that on the MNIST dataset, all the SSL methods achieve satisfying performance when the mismatch ratio is 0. As the ratio increases, the performance of those conventional SSL methods shows a rapid downward trend. Although the SOTA safe SSL methods (e.g., DS3L [2], SAFE-STUDENT [3]) is not that sensitive to the increasing ratio, our method can perform better than them and reach 96.9% averaged accuracy even when the ratio is 0.6. Figure 1(b) displays the results on a more challenging dataset CIFAR-10. Even at a mismatch ratio of 0, our method can surpass the conventional SSL methods by 10% accuracy on average. When the ratio increases from 0 to 0.6, our method maintains consistently higher performance than the SOTA safe SSL methods by more than 5%. These results verify the superiority of our method.

In Figure 2, we further visualize the confusion matrices of ground-truth (GT), our method, SAFE-STUDENT, and DS3L on CIFAR-10. Compared with DS3L, it's easy to find that our method reaches higher accuracy on almost all the given classes. Even in comparison with the competitive SAFE-STUDENT, our method can correctly clas-

Figure 2. Confusion matrices of the different methods over classes of bird, cat, deer, dog, frog, and horse on CIFAR-10.



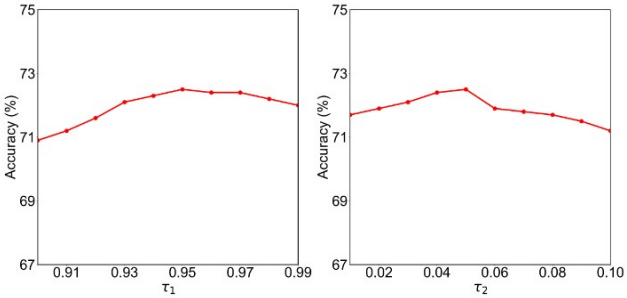


Figure 3. Seen-class classification accuracy (%) of hyperparameters (a) τ_1 and (b) τ_2 under different values on CIFAR-100 with mismatch ratio of 0.4.

sify much more instances on classes of bird, cat, frog, and horse, and exhibit comparable performance on only two other classes. These results intuitively prove the excellent recognition ability of our method on most classes.

3. Analysis of Hyperparameters

In this experiment, we measure the two important hyperparameters on CIFAR-100 with a mismatch ratio of 0.4, including τ_1 as the threshold to determine whether the pseudo labels will engage in positive learning or negative learning and τ_2 as the threshold in negative learning to decide if need to update the complementary label. Figure 3 (a) and (b) show the results under the wider values of τ_1 and τ_2. As observed, the best accuracy of 72.5% is achieved when τ_1 equals 0.95 and τ_2 is set to 0.05.

4. Ablation Study on Self-supervised Learning

We conduct an ablation study on the rotation prediction task in stage one. Compared with using cross entropy loss only, the rotation prediction task can boost the seen-class classification accuracy by 3.3% and 2.7% on CIFAR-10 and CIFAR-100 datasets, respectively. Moreover, we also compare the rotation prediction task with DINO and MoCo. We find the rotation prediction task can still outperform them by 1.7% and 0.7% averagely.

5. Insight of Feature Maps for ID&OOD

We provide extra results of the analysis of feature maps from both ID and OOD data in this section. Notably, in the CIFAR-10, we select the cat, bird, horse, and dog as ID classes, and the truck, airplane, automobile, ship as OOD classes.

Analysis of the Feature Vectors. To discover the more intuitive behavior of feature vectors in the feature space, we record the five channels with the highest response corresponding to different categories of ID and OOD, as Table 1 shows. The results show that the different OOD classes have similar high response channels, while the ID classes have diverse ones. This further shows that the OOD features are gathering in the feature space.

Visualization of Class Activation Maps. Inspired by the Class Activation Maps (CAMs) [6], we can find out exactly what areas the model is focusing on for ID and OOD data. The original approach to generate class activation maps is computing a weighted sum of the feature maps of the last convolutional layer. The weight indicates the importance of the corresponding feature maps in the final class activation map. Intuitively, we can use the weights in the final linear classifier, as they also indicate their importance in image classification. However, in the safe SSL settings, there are no corresponding weights for the OOD classes in the final linear classifier. Thus, instead of using the weighted sum of the feature maps, we visualize the feature maps separately. In Figure 4, we visualize the highest value channel of each ID class, respectively. The results show that the model can successfully extract the high-level features of each class. For the OOD data, we visualize the 89-*th* and 116-*th* feature maps in Figure 5 and 6. The figures have shown that the model is mainly focusing on the class-agnostic low-level features, which explains why the OOD classes show a gathering tendency in the feature space.

Table 1. Highest response channels of ID and OOD classes.

	Class	Top-5 highest value channel
ID	Dog	88,6,121,46,34
	Horse	25,118,62,49,14
	Deer	23,67,96,51,76
	Cat	60,124,99,5,2
	Frog	35,72,50,8,127
	Bird	22,11,74,120,61
OOD	Airplane	89,116,76,109,49
	Truck	116,89,109,49,76
	Automobile	89,116,22,11,76
	Ship	116,89,109,49,76

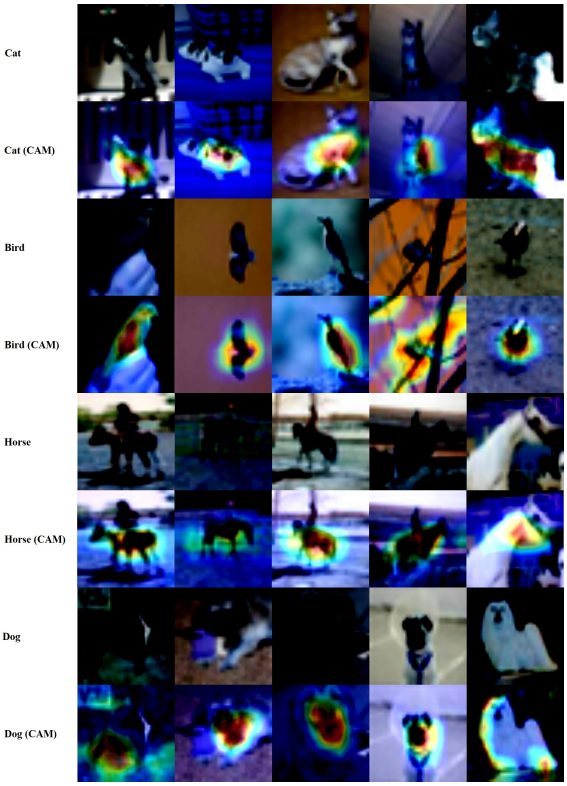


Figure 4. CAMs for ID data.

6. Limitations

We further explore the limitations of our method regarding the classification accuracy on various categories. As the Figure 3 shows, we found that our method obtains only 82.8% and 86.0% accuracy on dog and cat categories, respectively, about 11.2% and 8.0% lower than the accuracy on other categories averagely. After further analysis, we attribute the failure cases to the similar appearance between dog and cat in some cases, or called inter-class similarity. Regretfully, our current method did not consider how to better discriminate classes with similar characteristics. We will take this into consideration in future studies.

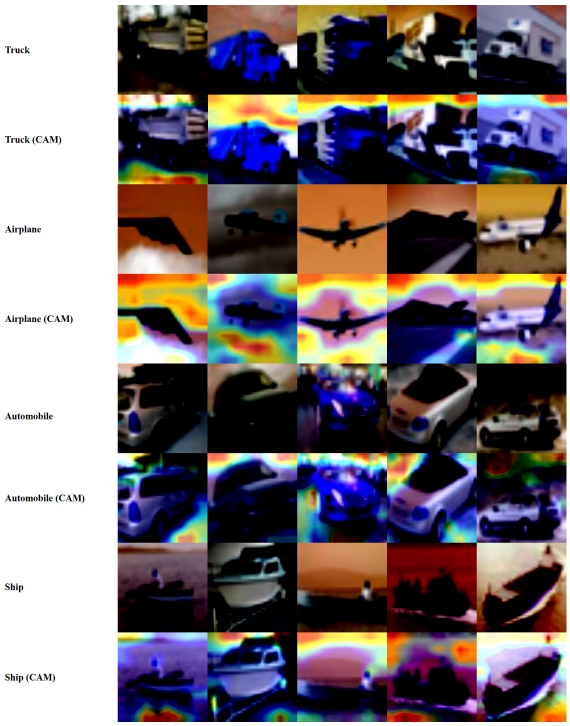


Figure 5. 89-*th* feature maps for OOD data.

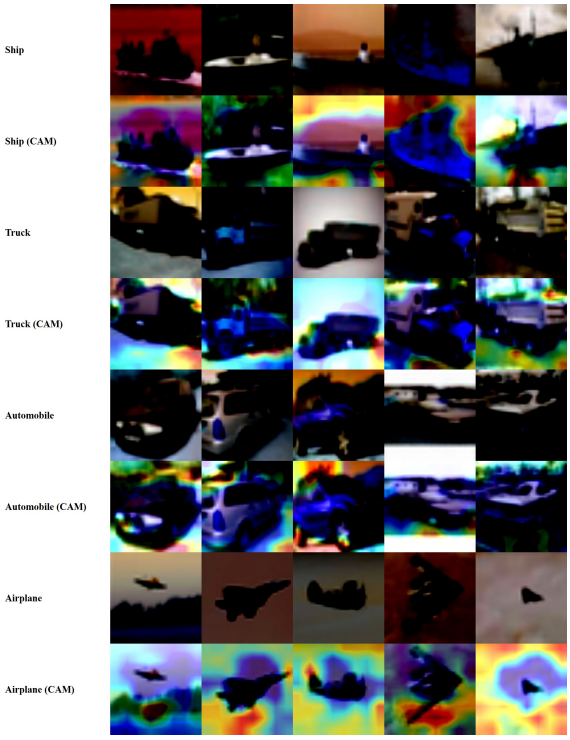


Figure 6. 116-*th* feature maps for OOD data.

References

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

In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 1

[2] Lan-Zhe Guo, Zhenyu 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, 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 conference on Computer Vision and Pattern Recognition*, pages 14565–14574, 2022. 1

[4] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 1

[5] Yann LeCun. The mnist database of handwritten digits. 1998. 1

[6] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2016. 2