SSB: Simple but Strong Baseline for Boosting Performance of Open-Set Semi-Supervised Learning

Yue Fan Anna Kukleva Dengxin Dai Bernt Schiele
{yfan, akukleva, ddai, schiele}@mpi-inf.mpg.de
Max Planck Institute for Informatics, Saarbrücken, Germany
Saarland Informatics Campus

In this appendix, we first show the tabulated breakdown of our main result in Section A. Then we compare our method with more recent approaches in Section B and show additional ablation studies in Section C. Finally, we present the pseudo-code and more visualizations of pseudo-inliers in Section D and Section E, respectively.

A. Main Results Breakdown

Here we present tabulated breakdown results of Fig. 3 and 4 from the main paper. We summarize the inlier classification accuracy and outlier detection in AUROC for different settings in Table 1, 2, 3, 4, 5, 6, and 7, respectively. To provide a more comprehensive analysis, we further provide the separate outlier detection performance results for seen outliers and unseen outliers in Table 8 and Table 9, respectively. SSB achieves competitive results in all settings. In particular, for CIFAR-10 and CIFAR-100 with 25 labels, SSB outperforms other methods by a large margin.

Test Acc. / AUROC		CIFAR-10
inlier / outlier classes		6 / 4
labels per class		25
SSL	FixMatch [13]	$\mathbf{91.94}_{\pm 0.16}$ / $62.58_{\pm 0.53}$
	FlexMatch [15]	$82.91_{\pm 0.92}$ / $69.60_{\pm 4.11}$
	SimMatch [16]	$89.22_{\pm 2.24}$ / $63.85_{\pm 0.70}$
OSSL	MTC [14]	$71.91_{\pm 10.82}$ / $85.57_{\pm 6.63}$
	OpenMatch [12]	$54.88_{\pm 2.33}$ / $53.32_{\pm 4.62}$
	T2T [9]	$83.21_{\pm 0.98}$ / $44.79_{\pm 17.26}$
	SSB (FixMatch)	$\underline{91.74}_{\pm 0.24}$ / $\underline{95.86}_{\pm 1.37}$
	SSB (FlexMatch)	$84.72_{\pm 0.49}$ / $90.32_{\pm 0.75}$
	SSB (SimMatch)	$88.51_{\pm 2.86}$ / $\mathbf{97.54}_{\pm 0.08}$

Table 1. **CIFAR-10 with 25 labels and 6 inlier classes.** We report test accuracy (%) / AUROC (%) for inliers classification and outlier detection, respectively. The numbers are averaged over 3 different random seeds. The best number is in **bold**, and the second best is in underline.

Test Acc. / AUROC		CIFAR-10
inlier / outlier classes		6 / 4
labels per class		50
SSL	FixMatch [13]	$\underline{91.33}_{\pm 0.18}$ / $63.77_{\pm 0.14}$
	FlexMatch [15]	$83.98_{\pm 0.31}$ / $64.47_{\pm 0.10}$
	SimMatch [16]	$91.10_{\pm 0.52}$ / $65.34_{\pm 0.09}$
OSSL	MTC [14]	$81.03_{\pm 5.21}$ / $92.01_{\pm 2.62}$
	OpenMatch [12]	$91.31_{\pm 1.18}$ / $\underline{95.88}_{\pm 0.60}$
	T2T [9]	$90.56_{\pm 0.07}$ / $39.73_{\pm 8.94}$
	SSB (FixMatch)	$\mathbf{92.18}_{\pm 0.33}$ / $\mathbf{97.65}_{\pm 0.19}$
	SSB (FlexMatch)	$84.26_{\pm 1.36}$ / $93.16_{\pm 3.63}$
	SSB (SimMatch)	$90.82_{\pm 0.47}$ / $94.07_{\pm 0.40}$

Table 2. **CIFAR-10 with 50 labels and 6 inlier classes.** We report test accuracy (%) / AUROC (%) for inliers classification and outlier detection, respectively. The numbers are averaged over 3 different random seeds. The best number is in **bold**, and the second best is in underline.

Test Acc. / AUROC		CIFAR-100
inlier / outlier classes		55 / 45
labels per class		25
SSL	FixMatch [13]	$69.89_{\pm 0.00}$ / $63.81_{\pm 0.26}$
	FlexMatch [15]	$67.87_{\pm 0.58}$ / $67.65_{\pm 1.12}$
	SimMatch [16]	$70.25_{\pm 0.96}$ / $65.21_{\pm 0.68}$
OSSL	MTC [14]	$58.13_{\pm 2.11}$ / $71.62_{\pm 1.36}$
	OpenMatch [12]	$67.09_{\pm 1.44}$ / $80.18_{\pm 0.09}$
	T2T [9]	$65.71_{\pm 0.93}$ / $60.11_{\pm 6.25}$
	SSB (FixMatch)	$\underline{70.64}_{\pm 0.36}$ / $82.91_{\pm 0.30}$
	SSB (FlexMatch)	$68.28_{\pm 0.74}$ / $\underline{83.62}_{\pm 0.43}$
	SSB (SimMatch)	$\mathbf{70.77}_{\pm 0.54}$ / $\mathbf{84.77}_{\pm 0.52}$

Table 3. **CIFAR-100 with 25 labels and 55 inlier classes.** We report test accuracy (%) / AUROC (%) for inliers classification and outlier detection, respectively. The numbers are averaged over 3 different random seeds. The best number is in **bold**, and the second best is in underline.

	Test Acc. / AUROC	CIFAR-100
	inlier / outlier classes	55 / 45
	labels per class	50
SSL	FixMatch [13]	$73.28_{\pm0.59}$ / $66.03_{\pm0.41}$
	FlexMatch [15]	$71.74_{\pm0.01}$ / $70.36_{\pm0.60}$
	SimMatch [16]	$\underline{74.15}_{\pm0.57}$ / $67.34_{\pm0.19}$
OSSL	MTC [14]	$65.97_{\pm0.77}$ / $68.96_{\pm1.08}$
	OpenMatch [12]	$71.90_{\pm1.05}$ / $82.82_{\pm0.47}$
	T2T [9]	$70.69_{\pm0.11}$ / $61.97_{\pm0.50}$
	SSB (FixMatch)	$73.70_{\pm0.75}$ / $\underline{85.89}_{\pm0.07}$
	SSB (FlexMatch)	$72.65_{\pm0.25}$ / $\mathbf{85.97}_{\pm0.46}$
	SSB (SimMatch)	$\mathbf{75.15}_{\pm0.34}$ / $84.60_{\pm0.18}$

Table 4. **CIFAR-100 with 50 labels and 55 inlier classes.** We report test accuracy (%) / AUROC (%) for inliers classification and outlier detection, respectively. The numbers are averaged over 3 different random seeds. The best number is in **bold**, and the second best is in underline.

	Test Acc. / AUROC	CIFAR-100
	inlier / outlier classes	80 / 20
	labels per class	25
SSL	FixMatch [13]	$63.58_{\pm0.36}$ / $56.40_{\pm0.21}$
	FlexMatch [15]	$59.83_{\pm1.78}$ / $62.73_{\pm0.62}$
	SimMatch [16]	$\underline{65.92}_{\pm0.81}$ / $60.61_{\pm0.60}$
OSSL	MTC [14]	$52.32_{\pm0.13}$ / $67.43_{\pm0.38}$
	OpenMatch [12]	$52.13_{\pm4.81}$ / $68.32_{\pm4.68}$
	T2T [9]	$47.58_{\pm10.38}$ / $51.95_{\pm4.44}$
	SSB (FixMatch)	$64.20_{\pm0.41}$ / $\underline{81.71}_{\pm0.86}$
	SSB (FlexMatch)	$60.39_{\pm1.89}$ / $79.85_{\pm0.94}$
	SSB (SimMatch)	$\mathbf{66.48}_{\pm0.77}$ / $\mathbf{82.39}_{\pm2.97}$

Table 5. **CIFAR-100 with 25 labels and 80 inlier classes.** We report test accuracy (%) / AUROC (%) for inliers classification and outlier detection, respectively. The numbers are averaged over 3 different random seeds. The best number is in **bold**, and the second best is in underline.

	Test Acc. / AUROC	CIFAR-100
	inlier / outlier classes	80 / 20
	labels per class	50
SSL	FixMatch [13]	$67.06_{\pm0.10}$ / $58.05_{\pm0.49}$
	FlexMatch [15]	$65.22_{\pm0.18}$ / $65.00_{\pm0.07}$
	SimMatch [16]	$\underline{69.35}_{\pm0.26}$ / $61.44_{\pm0.16}$
OSSL	MTC [14]	$59.17_{\pm0.01}$ / $69.34_{\pm1.81}$
	OpenMatch [12]	$66.90_{\pm0.19}$ / $79.95_{\pm0.26}$
	T2T [9]	$64.18_{\pm0.64}$ / $65.26_{\pm13.73}$
	SSB (FixMatch)	$67.97_{\pm0.20}$ / $80.81_{\pm1.02}$
	SSB (FlexMatch)	$65.79_{\pm0.06}$ / $\mathbf{83.32}_{\pm0.36}$
	SSB (SimMatch)	$\mathbf{70.27}_{\pm0.19}$ / $\underline{81.16}_{\pm2.10}$

Table 6. **CIFAR-100 with 50 labels and 80 inlier classes.** We report test accuracy (%) / AUROC (%) for inliers classification and outlier detection, respectively. The numbers are averaged over 3 different random seeds. The best number is in **bold**, and the second best is in underline.

	Test Acc. / AUROC	ImageNet-30
	inlier / outlier classes	20 / 10
	labels per class	5%
SSL	FixMatch [13]	$90.33_{\pm0.66}$ / $75.60_{\pm1.28}$
	FlexMatch [15]	$86.33_{\pm0.92}$ / $72.24_{\pm0.45}$
	SimMatch [16]	$\underline{91.32}_{\pm0.73}$ / $71.72_{\pm0.14}$
OSSL	MTC [14]	$81.05_{\pm0.35}$ / $80.66_{\pm2.23}$
	OpenMatch [12]	$78.75_{\pm0.35}$ / $\mathbf{84.21}_{\pm0.03}$
	T2T [9]	$88.75_{\pm0.90}$ / $73.11_{\pm1.11}$
	SSB (FixMatch)	$\mathbf{91.80}_{\pm0.05}$ / $\underline{82.80}_{\pm1.18}$
	SSB (FlexMatch)	$86.90_{\pm0.70}$ / $75.42_{\pm0.24}$
	SSB (SimMatch)	$91.30_{\pm0.65}$ / $75.54_{\pm0.10}$

Table 7. **ImageNet-30 with 5% labels and 20 inliers classes.** We report test accuracy (%) / AUROC (%) for inliers classification and outlier detection, respectively. The numbers are averaged over 3 different random seeds. The best number is in **bold**, and the second best is in underline.

B. Results on More Benchmarks

In this section, we compare SSB with more recent methods on their benchmarks.

Comparison with methods of class-mismatched SSL. Here, we compare with Safe-Student [7] and SPL [8] on CIFAR-10 and CIFAR-100 with different levels of class distribution mismatch between the labeled and unlabeled data. Following [8, 7], on CIFAR-10, we consider the six animal classes as inlier classes and use 400 labels per class. The unlabeled set contains 20,000 images coming from all ten classes with different class mismatch ratios. For example, when the ratio is 0.3, 70% of the samples are from the six inlier classes and the rest samples are from the remaining four classes. Similarly, for CIFAR-100, the first 50 classes are used as inlier classes, and the unlabeled set has 20,000 samples with different class mismatch ratios. We compare the inlier accuracy of different methods in Table 10. SSB outperforms other methods by significant margins across all settings, which indicates the effectiveness of our method for class-mismatched SSL.

Seen AUORC		CIFAR-10		CIFAR-100		CIFAR-100		ImageNet-30
Inlier / outlier classes		6 / 4		55 / 45		80 / 20		20 / 10
Labels per class		25	50	25	50	25	50	5%
SSL	FixMatch [13]	37.37$_{\pm 0.84}$	39.41$_{\pm 0.15}$	54.48$_{\pm 1.05}$	55.77$_{\pm 0.95}$	41.43$_{\pm 0.10}$	44.33$_{\pm 0.79}$	65.09$_{\pm 2.09}$
	FlexMatch [15]	51.32$_{\pm 8.24}$	41.01$_{\pm 0.20}$	60.82$_{\pm 1.07}$	63.72$_{\pm 1.15}$	53.68$_{\pm 0.48}$	57.45$_{\pm 0.70}$	61.73$_{\pm 0.31}$
	SimMatch [16]	38.39$_{\pm 0.87}$	41.15$_{\pm 0.26}$	55.28$_{\pm 1.35}$	57.69$_{\pm 0.16}$	49.19$_{\pm 0.43}$	49.53$_{\pm 0.26}$	56.41$_{\pm 0.55}$
OSSL	MTC [14]	92.00$_{\pm 3.49}$	94.47$_{\pm 2.05}$	76.93$_{\pm 1.48}$	72.53$_{\pm 0.18}$	69.15$_{\pm 0.86}$	72.38$_{\pm 1.86}$	82.70$_{\pm 2.60}$
	OpenMatch [12]	62.46$_{\pm 4.19}$	$\underline{99.41}_{\pm 0.18}$	84.93$_{\pm 0.08}$	86.99$_{\pm 0.23}$	74.87$_{\pm 3.78}$	$\underline{86.19}_{\pm 0.48}$	**91.79**$_{\pm 0.49}$
	T2T [9]	34.90$_{\pm 27.50}$	23.85$_{\pm 8.45}$	52.95$_{\pm 6.15}$	59.50$_{\pm 1.50}$	50.45$_{\pm 9.15}$	61.40$_{\pm 21.10}$	62.35$_{\pm 2.05}$
	SSB (FixMatch)	$\underline{99.35}_{\pm 0.38}$	**99.63**$_{\pm 0.15}$	89.39$_{\pm 0.44}$	$\underline{90.62}_{\pm 0.46}$	**90.25**$_{\pm 1.34}$	85.29$_{\pm 2.18}$	$\underline{83.84}_{\pm 2.31}$
	SSB (FlexMatch)	96.81$_{\pm 0.25}$	93.41$_{\pm 6.22}$	**89.84**$_{\pm 0.05}$	**91.21**$_{\pm 0.26}$	88.32$_{\pm 1.36}$	**90.68**$_{\pm 0.27}$	67.58$_{\pm 0.55}$
	SSB (SimMatch)	**99.61**$_{\pm 0.10}$	93.07$_{\pm 0.70}$	$\underline{89.75}_{\pm 0.90}$	87.38$_{\pm 0.13}$	$\underline{88.65}_{\pm 3.86}$	83.05$_{\pm 3.49}$	62.63$_{\pm 0.09}$

Table 8. **AUROC (%) for seen outliers.** The best number is in **bold**, and the second best is in underline. Note that a random OOD detector gives an AUROC of 50%.

Unseen AUROC		CIFAR-10		CIFAR-100		CIFAR-100		ImageNet-30
Inlier / outlier classes		6 / 4		55 / 45		80 / 20		20 / 10
Labels per class		25	50	25	50	25	50	5%
SSL	FixMatch [13]	87.80$_{\pm 0.23}$	88.12$_{\pm 0.42}$	73.15$_{\pm 0.53}$	76.28$_{\pm 0.13}$	71.37$_{\pm 0.32}$	71.77$_{\pm 0.20}$	86.11$_{\pm 0.54}$
	FlexMatch [15]	87.88$_{\pm 0.02}$	87.92$_{\pm 0.00}$	74.48$_{\pm 1.18}$	77.01$_{\pm 0.06}$	71.79$_{\pm 0.76}$	72.54$_{\pm 0.84}$	82.76$_{\pm 0.59}$
	SimMatch [16]	89.32$_{\pm 0.54}$	89.54$_{\pm 0.45}$	75.14$_{\pm 0.01}$	76.98$_{\pm 0.21}$	72.03$_{\pm 0.78}$	73.36$_{\pm 0.06}$	$\underline{87.04}_{\pm 0.27}$
OSSL	MTC [14]	79.14$_{\pm 9.78}$	89.55$_{\pm 3.20}$	66.32$_{\pm 1.25}$	65.40$_{\pm 1.99}$	65.71$_{\pm 0.09}$	66.30$_{\pm 5.49}$	78.61$_{\pm 1.86}$
	OpenMatch [12]	44.18$_{\pm 5.05}$	92.35$_{\pm 1.01}$	75.43$_{\pm 0.11}$	78.66$_{\pm 0.70}$	61.78$_{\pm 5.58}$	73.70$_{\pm 0.04}$	76.63$_{\pm 0.54}$
	T2T [9]	54.68$_{\pm 7.01}$	55.61$_{\pm 9.43}$	67.26$_{\pm 6.35}$	64.44$_{\pm 0.50}$	53.45$_{\pm 0.27}$	69.11$_{\pm 6.36}$	83.88$_{\pm 0.17}$
	SSB (FixMatch)	$\underline{92.37}_{\pm 2.36}$	**95.67**$_{\pm 0.23}$	76.44$_{\pm 0.15}$	$\underline{81.16}_{\pm 0.61}$	$\underline{73.18}_{\pm 0.38}$	$\underline{76.34}_{\pm 0.14}$	81.77$_{\pm 0.04}$
	SSB (FlexMatch)	83.83$_{\pm 1.25}$	92.91$_{\pm 1.03}$	$\underline{77.40}_{\pm 0.82}$	80.72$_{\pm 0.65}$	71.38$_{\pm 0.53}$	75.96$_{\pm 0.99}$	83.26$_{\pm 1.02}$
	SSB (SimMatch)	**95.47**$_{\pm 0.26}$	$\underline{95.07}_{\pm 0.10}$	**79.80**$_{\pm 0.13}$	**81.82**$_{\pm 0.23}$	**76.14**$_{\pm 2.08}$	**79.27**$_{\pm 0.72}$	**88.46**$_{\pm 0.10}$

Table 9. **AUROC (%) for unseen outliers.** The best number is in **bold**, and the second best is in underline. Note that a random OOD detector gives an AUROC of 50%.

Comparison in cross-dataset scenarios. Now, we compare our method in cross-dataset settings, where the labeled set and the unlabeled set are constructed using different datasets. Following [10], we use CIFAR-100 for labeled set and ImageNet for unlabeled set. Specifically, we take 60 classes of CIFAR-100 as inlier classes, which are also contained in ImageNet. Then, 20,000 images are sampled from 100 classes of ImageNet to form the unlabeled set, where 60 classes are the same as the inlier classes of CIFAR-100 and the rest 40 are randomly chosen from the remaining 940 classes. Please refer to [10] for details of the 60 inlier classes. In Table 11, we compare the results of our method with others under different numbers of labeled data. We can see that our method improves the SOTA in all settings. The large performance gap over TOOR [10] shows that the simple confidence-based pseudo-labeling used in SSB is a more effective way for recycling OOD data to improve the classification performance.

C. Ablation Study

In this section, we provide more analysis of the hyperparameters used in SSB. Specifically, we study the effect of depth and width of the projection head, deferred outlier detector training, the effect of the threshold θ for selecting the pseudo-outliers, the effect of the threshold τ for pseudo-labeling, the loss weight λ_{det}^{u} for unlabeled detection loss, and different data augmentation schemes used in pseudo-negative mining.

Effect of depth and width of the projection head. Table 12 and 13 study the effect of the depth and the width of the projection head, respectively. We choose the 2-layer MLP with a hidden dimension of 1024 as it shows the best inlier

Method	CIFAR-10		CIFAR-100	
	ratio=0.3	ratio=0.6	ratio=0.3	ratio=0.6
DS^{3}L [6]	$78.1_{\pm 0.4}$	$76.9_{\pm 0.5}$	-	-
UASD [2]	$77.6_{\pm 0.4}$	$76.0_{\pm 0.4}$	$61.8_{\pm 0.4}$	$58.4_{\pm 0.5}$
MTC [14]	$85.5_{\pm 0.6}$	$81.7_{\pm 0.5}$	$63.1_{\pm 0.6}$	$61.1_{\pm 0.3}$
CL [1]	$83.2_{\pm 0.4}$	$82.1_{\pm 0.4}$	$63.6_{\pm 0.4}$	$61.5_{\pm 0.5}$
Safe-Student [7]	$85.7_{\pm 0.3}$	$83.8_{\pm 0.1}$	$68.4_{\pm 0.2}$	$68.2_{\pm 0.1}$
CL+SPL [8]	$87.8_{\pm 0.3}$	$84.1_{\pm 0.5}$	$65.9_{\pm 0.3}$	$65.5_{\pm 0.4}$
SSB (Ours)	$\mathbf{92.5}_{\pm 0.1}$	$\mathbf{90.6}_{\pm 0.3}$	$\mathbf{74.7}_{\pm 0.6}$	$\mathbf{73.2}_{\pm 0.3}$

Table 10. **Test accuracy (%) with different class mismatch ratios on CIFAR-10 and CIFAR-100.** This benchmark is adopted by [7, 8]. The number of unlabeled data is 20,000. The best number is in **bold**.

Method	CIFAR100+ImageNet with different numbers of labeled data			
	4800 (80 for each class)	6000 (100 for each class)	7200 (120 for each class)	8400 (140 for each class)
UASD [2]	$42.07_{\pm 0.58}$	$44.90_{\pm 0.47}$	$46.38_{\pm 0.79}$	$48.20_{\pm 0.40}$
DS^{3}L [6]	$43.99_{\pm 0.54}$	$45.10_{\pm 1.25}$	$47.11_{\pm 0.73}$	$48.96_{\pm 0.63}$
MTC [14]	$45.69_{\pm 0.74}$	$46.34_{\pm 0.81}$	$48.92_{\pm 0.58}$	$50.05_{\pm 0.77}$
TOOR [10]	$47.19_{\pm 0.83}$	$49.15_{\pm 0.76}$	$51.34_{\pm 0.65}$	$52.98_{\pm 0.79}$
SSB (Ours)	$\mathbf{63.91}_{\pm 0.71}$	$\mathbf{65.95}_{\pm 0.20}$	$\mathbf{67.97}_{\pm 0.12}$	$\mathbf{69.42}_{\pm 0.28}$

Table 11. **Test accuracy (%) with different numbers of labeled samples.** This benchmark is adopted by [10]. The number of unlabeled data is 20,000. The best number is in **bold**.

accuracy and OOD detection performance. Note that the projection head shows good robustness over a large range of design choices. Even using a single fully-connected layer with a ReLU activation function already gives better performance.

Projection head	Inlier Cls. (Acc.)	Outlier Det. (seen AUROC)	Outlier Det. (unseen AUROC)
None	90.28	44.11	82.81
1-layer MLP	91.48	98.89	89.96
2-layer MLP	**91.65**	**99.16**	**90.35**
3-layer MLP	91.48	98.49	88.08
4-layer MLP	91.28	98.29	86.44
2-layer MLP + BN	90.00	90.53	78.33

Table 12. **Effect of different projection heads.** *None* denotes not using projection head; *MLP* denotes multilayer perceptron; *BN* denotes using BatchNorm [11] between different layers. All models are trained with confidence-based pseudo-labeling and pseudo-negative mining on the same data split on CIFAR-10 with 25 labeled data.

Hidden dim.	Inlier Cls. (Acc.)	Outlier Det. (seen AUROC)	Outlier Det. (unseen AUROC)
128	91.25	97.30	82.98
256	90.83	97.47	85.43
512	90.63	98.91	88.36
1024	**91.65**	**99.16**	**90.35**
2048	91.07	97.74	86.59

Table 13. **Effect of the hidden dimension of the projection head.** We use a 2-layer MLP as the projection and train all models with confidence-based pseudo-labeling and pseudo-negative mining on the same data split on CIFAR-10 with 25 labeled data.

Deferring the outlier detector training. Here we study the effect of different training lengths on the outlier detector. In Table 14, we can see that while the performance are similar under different training epochs, the training cost can be largely reduced by deferring the detector training. With 37 training epochs, our method can reach the best performance while reducing the training costs.

Effect of threshold τ. We follow FixMatch [13] and set the threshold τ for pseudo-labeling as 0.95 for our main results. Here we provide an ablation study of this hyper-parameter in Table 15, with SSB exhibiting robustness within a wide range of τ.

Different loss weights. In the main paper, we use $\lambda_{det}^{u} = 1$ for the unlabeled data detection loss with pseudo-negative mining. Here, we provide more results of different loss weights in Table 16. We can see that, except for very small loss weights (0.1 or 0), the OOD detection performance is quite robust to various λ_{det}^{u}.

Total epochs	Starting epochs	GPU hours reduced	Inlier Cls. (Acc.)	Outlier Det. (avg. AUROC)
512	0	0	90.20	87.92
512	100	5.9	90.58	89.11
512	200	10.7	91.01	90.39
512	300	16.0	91.55	91.71
512	400	19.6	91.62	93.13
512	475	**23.2**	**91.65**	**94.76**

Table 14. **Effect of different starting epochs for outlier detector training.** We defer the detector training by enabling the detection losses at a later stage. All models are trained on a single NVIDIA V100 and we compute the reduced GPU hours with respect to the model using detection losses throughout the entire training. The setting is CIFAR-10 with 25 labeled data.

Threshold τ	0.97	0.95	0.85	0.75	0.5
Inlier Cls.	90.28	91.65	92.28	90.52	**92.35**
Outlier Det.	93.67	94.76	**97.41**	97.17	92.64

Table 15. **Effect of different pseudo-labeling thresholds τ.** SSB shows good robustness within a wide range of τ. The experimental setting is CIAR-10 with 25 labels.

Loss weight λ^u_{det}	Inlier Cls. (Acc.)	Outlier Det. (seen AUROC)	Outlier Det. (unseen AUROC)
10	**92.45**	99.18	89.40
5	91.98	98.61	88.20
2	91.98	98.82	88.96
1	91.65	99.16	90.35
0.5	92.08	**99.22**	90.18
0.1	91.77	88.98	**90.74**
0	91.52	89.78	90.27

Table 16. **Effect of the loss weight for pseudo-negative mining.** Our method is robust to a wide range of loss weights. The setting is CIFAR-10 with 25 labeled data.

Effect of data augmentation scheme. In pseudo-negative mining, there are two types of data augmentations used for loss computing. Given an unlabeled image, the OOD score is computed from a weak augmentation, which consists of random crop and horizontal flipping. Then, if the confidence is low enough, a strong augmentation will be used to compute the binary cross-entropy loss. Following [13], the strong augmentation is RandAugment [3] with CutOut [4]. In Table 17, we study the effect of different types of augmentation for computing the loss. Using the strong augmentation gives the best OOD detection performance while having similar inlier classification performance to other strategies.

Augmentation	Inlier Cls. (Acc.)	Outlier Det. (seen AUROC)	Outlier Det. (unseen AUROC)
weak+strong	**91.88**	92.36	77.55
weak	91.73	92.10	77.80
strong	91.65	**99.16**	**90.35**

Table 17. **Effect of data augmentation in pseudo-negative mining.** It is important to use a different type of data augmentation for loss computing from the one used to generate pseudo-outliers. The setting is CIFAR-10 with 25 labeled data.

D. Pseudo-Code

We present the complete pseudo-code of SSB with deferred outlier detector training in Algorithm 1.

E. Visualization of Pseudo-Inliers

We visualize the OOD samples selected for different classes in Fig. 1. The model is trained on CIFAR-100 with 55 inlier classes and 25 labels. We can see that the selected pseudo-inliers contain semantic information of the corresponding classes, which indicates that some OOD data are natural data augmentations and can be used to improve the generalization performance if used properly.

References

[1] Paola Cascante-Bonilla, Fuwen Tan, Yanjun Qi, and Vicente Ordonez. Curriculum labeling: Revisiting pseudo-labeling for semi-supervised learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021. 4

[2] Yanbei Chen, Xiatian Zhu, Wei Li, and Shaogang Gong. Semi-supervised learning under class distribution mismatch. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020. 4

[3] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE conference on computer vision and pattern recognition Workshops*, 2020. 5

[4] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. 5

[5] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In *Advances in neural information processing systems*, 2005. 6

[6] 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*, 2020. 4

[7] 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*, 2022. 2, 4

Algorithm 1

```
Input: Labeled set $\mathcal{D}_{\text{labeled}} = \{(\mathbf{x}_i^l, y_i)\}_{i=1}^N$, unlabeled set $\mathcal{D}_{\text{unlabeled}} = \{(\mathbf{x}_i^u)\}_{i=1}^M$, feature encoder $f$, inlier classifier $g_c$, outlier detector $g_d$, two MLP projection heads $h_c$ and $h_d$, thresholds $\tau$ and $\theta$, batch size $B_l$ and $B_u$, loss weights $\lambda_{det}^u$, $\lambda_{OC}^u$, and $\lambda_{em}^u$, warm-up iterations $T_0$, total number of training iterations $T$
Initialize the parameters of $f$, $g_c$, $g_d$, $h_c$, and $h_d$ randomly
for $t = 1$ to $T$ do
    // Sample labeled and unlabeled data
    $\{\mathbf{x}_i^l, y_i\}_{i=1}^{B_l} \sim$ Random sampler($\mathcal{D}_{\text{labeled}}$)
    $\{\mathbf{x}_i^u\}_{i=1}^{B_u} \sim$ Random sampler($\mathcal{D}_{\text{unlabeled}}$)
    // Compute classification losses
    $\hat{p}_i^u = \text{softmax}(g_c(h_c(f(\mathbf{x}_i^u)))), i = 1, ..., B_u$ // Compute pseudo-label distributions
    $\hat{y}_i^u = \text{argmax}\hat{p}_i^u, i = 1, ..., B_u$ // Compute pseudo-labels
    $L_{cls}^l = \frac{1}{B_l}\sum_{i=1}^{B_l} H(g_c(h_c(f(\mathbf{x}_i^l)), y_i)$ // Labeled data loss
    $L_{cls}^u = \frac{1}{B_u}\sum_{i=1}^{B_u} \mathbb{1}(\max\hat{p}_i^u \geq \tau)H(\hat{p}_i^u, \hat{y}_i^u)$ // Unlabeled data loss as in Equation (2)
    $L_{cls} = L_{cls}^l + L_{cls}^u$
    // Compute detection losses
    $L_{det}^l = -\frac{1}{B_l}\sum_{i=1}^{B_l} log(p_{y_i}(\mathbf{x}_i^l)) + \frac{1}{|\mathcal{C}|-1}\sum_{j\neq y_i} log(1 - p_j(\mathbf{x}_i^l))$ // Detection loss for labeled data as in Equation (4)
    $L_{det}^u = -\frac{1}{B_u}\sum_{i=1}^{B_u} \frac{1}{\sum_c \mathbb{1}(p_c>\theta)}\sum_{c=1}^{|\mathcal{C}|} \mathbb{1}(p_c > \theta)log(1 - p_c(\mathbf{x}_i^u))$ // Pseudo-negative mining as in Equation (5)
    $L_{em}^u = \frac{1}{B_u}\sum_{i=1}^{B_u} entropy(g_d(h_d(f(\mathbf{x}_i^u))))$ // Entropy minimization loss as in [5]
    $L_{OC}^u = \frac{1}{B_u}\sum_{i=1}^{B_u} ||g_d(h_d(f(\mathcal{T}_1(\mathbf{x}_i^u)))) - g_d(h_d(f(\mathcal{T}_2(\mathbf{x}_i^u))))||^2$ // Open-set consistency loss as in [12]
    $L_{det} = L_{det}^l + \lambda_{det}^u L_{det}^u + \lambda_{OC}^u L_{OC}^u + \lambda_{em}^u L_{em}^u$
    // Total loss
    $L_{total} = L_{cls} + \mathbb{1}(t > T_0)L_{det}$
    Update parameters in $f$, $g_c$, $g_d$, $h_c$, and $h_d$ with SGD
end for
return $f$, $g_c$, $g_d$, $h_c$, and $h_d$
```

[8] Rundong He, Zhongyi Han, Yang Yang, and Yilong Yin. Not all parameters should be treated equally: Deep safe semi-supervised learning under class distribution mismatch. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022. 2, 4

[9] 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*, 2021. 1, 2, 3

[10] Zhuo Huang, Jian Yang, and Chen Gong. They are not completely useless: Towards recycling transferable unlabeled data for class-mismatched semi-supervised learning. *IEEE Transactions on Multimedia*, 2022. 3, 4

[11] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015. 4

[12] Kuniaki Saito, Donghyun Kim, and Kate Saenko. Openmatch: Open-set semi-supervised learning with open-set consistency regularization. *Advances in Neural Information Processing Systems*, 2021. 1, 2, 3, 6

[13] 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. In *Advances in Neural Information Processing Systems*, 2020. 1, 2, 3, 4, 5

[14] Qing Yu, Daiki Ikami, Go Irie, and Kiyoharu Aizawa. Multi-task curriculum framework for open-set semi-supervised learning. In *European Conference on Computer Vision*, 2020. 1, 2, 3, 4

[15] Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. *Advances in Neural Information Processing Systems*, 2021. 1, 2, 3

[16] Mingkai Zheng, Shan You, Lang Huang, Fei Wang, Chen Qian, and Chang Xu. Simmatch: Semi-supervised learning with similarity matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 1, 2, 3

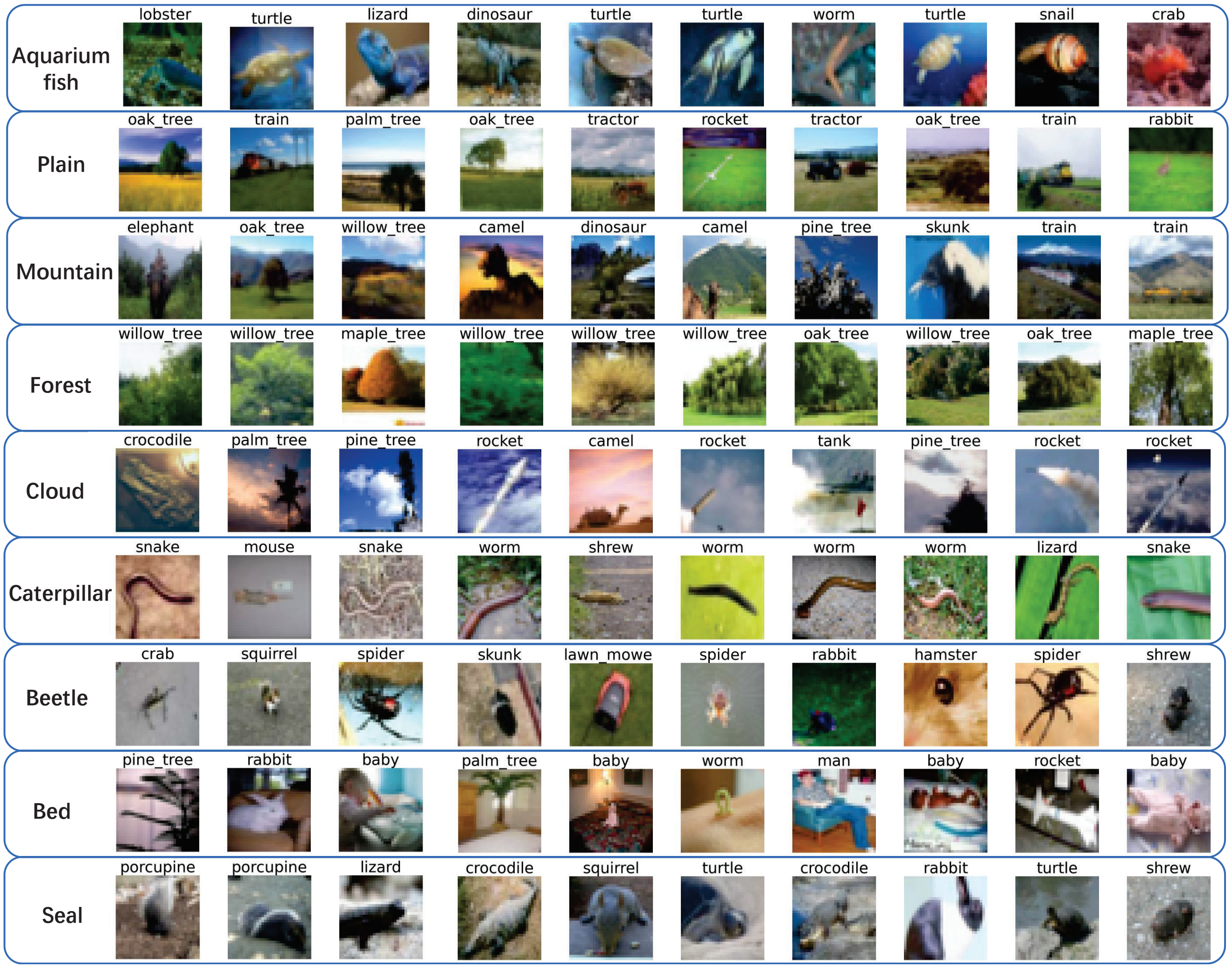

Figure 1. Selected pseudo-inliers for different classes. Each row lists 10 most confident images with the pseudo-label on the left of the row. The ground-truth class of the OOD sample is shown on the top of each image.