

PSEUDO-LABELS ARE ALL YOU NEED FOR OUT-OF-DISTRIBUTION DETECTION (SUPPLEMENTAL MATERIAL)

Anonymous authors

Paper under double-blind review

1 COMPARATIVE STUDY WITH SUPERVISED METHODS

We compared our approach with supervised methods across different backbones. The averaged results are presented in the main paper (Section 4.2), while the detailed statistics are shown in Tables 1 and 2.

WRN-28-10														
Method	SVHN		Textures		LSUN-C		LSUN-R		iSUN		Places365		Average	
	FPR↓	AUR↑	FPR↓	AUR↑	FPR↓	AUR↑	FPR↓	AUR↑	FPR↓	AUR↑	FPR↓	AUR↑	FPR↓	AUR↑
Hendrycks & Gimpel (2016)	42.10	91.85	53.30	87.45	24.85	96.37	37.81	93.71	40.11	93.73	50.73	88.58	41.49	91.84
Liang et al. (2017)	37.08	88.36	47.58	82.85	6.14	98.65	20.51	95.04	22.95	94.22	41.03	86.57	29.22	90.95
Liu et al. (2020)	33.11	90.54	46.06	85.09	5.86	98.76	22.68	94.90	25.12	94.17	39.08	88.50	28.65	91.99
Sun et al. (2021)	39.94	98.31	60.80	91.85	57.11	96.76	77.63	80.15	79.48	79.48	73.29	77.98	65.78	86.22
Sun & Li (2022)	37.84	86.99	50.77	79.70	2.54	99.43	26.30	92.89	28.30	92.89	43.46	84.65	31.53	89.30
(Zhang & Xiang, 2023)	16.43	95.62	20.31	94.81	2.5	99.03	24.59	96.14	20.74	96.28	40.26	83.77	19.8	94.27
Yu et al. (2023)	3.83	99.18	<u>14.23</u>	<u>97.06</u>	0.32	99.81	8.13	98.32	5.98	98.71	48.69	90.91	<u>13.53</u>	<u>97.33</u>
Ours	<u>4.36</u>	97.22	13.85	99.16	<u>1.76</u>	<u>99.60</u>	<u>18.67</u>	<u>96.64</u>	<u>12.06</u>	<u>97.63</u>	22.27	95.07	12.16	97.55

VGG11														
Method	SVHN		Textures		LSUN-C		LSUN-R		iSUN		Places365		Average	
	FPR↓	AUR↑	FPR↓	AUR↑	FPR↓	AUR↑	FPR↓	AUR↑	FPR↓	AUR↑	FPR↓	AUR↑	FPR↓	AUR↑
Hendrycks & Gimpel (2016)	68.07	90.02	63.86	89.37	46.63	93.73	70.19	86.29	71.81	85.71	68.08	87.25	64.77	88.73
Liang et al. (2017)	53.86	92.23	48.09	91.94	19.95	97.01	54.29	89.47	56.61	88.57	52.34	<u>89.86</u>	47.52	91.56
Liu et al. (2020)	53.11	92.26	47.04	<u>92.08</u>	18.51	97.20	<u>53.02</u>	<u>89.58</u>	55.39	<u>88.97</u>	<u>51.67</u>	89.95	46.46	91.67
Sun et al. (2021)	58.16	83.28	51.73	87.47	23.40	94.77	47.19	89.68	<u>51.30</u>	87.39	50.47	87.39	47.15	88.44
Sun & Li (2022)	47.81	93.27	50.95	91.77	16.73	97.06	64.26	87.83	65.83	87.83	59.23	88.53	50.80	90.98
(Zhang & Xiang, 2023)	68.62	91.30	33.3	92.5	19.46	95.49	79.27	82.70	59.34	88.14	66.59	86.88	54.43	89.17
Yu et al. (2023)	8.84	98.24	24.62	95.11	<u>3.38</u>	<u>99.36</u>	71.17	83.12	62.80	86.05	65.25	85.20	<u>39.34</u>	<u>91.18</u>
Ours	<u>26.92</u>	<u>93.63</u>	<u>27.48</u>	91.97	1.34	99.65	73.70	84.97	16.31	96.77	61.67	85.66	36.41	91.83

Table 1: We compared our method with baseline methods using the supervised benchmark setting across different network architectures. The ID dataset was CIFAR-10, and the OOD datasets are listed in the first row of each table. For clarity, the best results are highlighted in bold, and the second-best results are underlined.

Method	SVHN		Textures		LSUN-C		LSUN-R		iSUN		Places365		Average	
	FPR↓	AUR↑	FPR↓	AUR↑	FPR↓	AUR↑	FPR↓	AUR↑	FPR↓	AUR↑	FPR↓	AUR↑	FPR↓	AUR↑
Hendrycks & Gimpel (2016)	79.88	79.12	63.87	86.42	78.21	80.84	68.19	84.36	66.97	84.91	80.98	78.92	73.02	82.43
Liang et al. (2017)	72.22	81.63	74.80	84.45	72.70	85.75	42.38	92.23	40.07	92.54	81.05	79.06	63.87	85.94
Liu et al. (2020)	78.94	80.68	<u>59.09</u>	<u>89.84</u>	78.04	84.63	63.95	<u>87.56</u>	63.27	88.05	83.58	79.02	71.45	84.96
Sun et al. (2021)	76.01	80.17	60.45	89.78	67.11	86.97	70.25	86.00	67.60	87.04	83.76	78.76	70.95	84.79
Sun & Li (2022)	78.35	81.75	58.06	90.16	77.73	84.89	<u>63.89</u>	87.23	63.22	<u>87.85</u>	83.47	79.17	70.78	85.17
(Zhang & Xiang, 2023)	80.26	83.01	85.07	93.63	88.66	79.63	65.26	55.13	80.09	77.61	<u>77.04</u>	76.74	75.79	81.23
Yu et al. (2023)	23.34	92.75	68.62	71.47	<u>19.46</u>	<u>93.88</u>	79.27	81.08	89.34	82.94	81.59	<u>82.42</u>	<u>60.27</u>	84.0
Ours	<u>53.64</u>	<u>89.64</u>	63.81	82.53	7.4	98.51	77.65	77.81	<u>59.03</u>	86.46	67.57	82.87	54.92	86.21

Table 2: We compared our method with supervised baseline methods trained on ResNet18. The ID dataset was CIFAR-100, and the OOD datasets are listed in the first row of each table. For clarity, the best results are highlighted in bold, and the second-best results are underlined.

2 NUMBER OF PSEUDO-CATEGORIES

We evaluated the impact of the number of pseudo-categories using the supervised benchmark setting on CIFAR-10 and CIFAR-100 datasets. The average results are presented in the main paper. Here, we provide the detailed statistics in Table 3.

ID dataset: Cifar-10														
Method	SVHN		Textures		LSUN-C		LSUN-R		iSUN		Places365		Average	
	FPR↓	AUR↑	FPR↓	AUR↑	FPR↓	AUR↑	FPR↓	AUR↑	FPR↓	AUR↑	FPR↓	AUR↑	FPR↓	AUR↑
Ours(K5)	23.44	93.62	29.68	92.64	3.36	99.01	<u>34.15</u>	<u>92.80</u>	<u>8.47</u>	<u>98.15</u>	<u>34.01</u>	91.52	<u>22.18</u>	<u>94.62</u>
Ours(K10)	22.01	<u>93.78</u>	38.95	89.11	<u>2.30</u>	<u>99.47</u>	38.81	92.23	10.64	97.96	34.05	91.52	24.46	94.01
Ours(K20)	22.82	93.91	<u>25.18</u>	<u>94.25</u>	1.84	99.61	32.47	93.98	6.71	98.55	32.38	92.63	20.23	95.4
Ours(K50)	56.62	84.29	42.81	92.74	4.71	99.02	47.24	93.61	67.14	78.44	41.15	<u>92.38</u>	43.28	90.08
Ours(K100)	61.72	82.41	7.16	98.72	7.58	98.24	47.14	91.72	82.81	65.18	42.97	91.51	41.56	87.96
Ours(K200)	78.27	69.47	74.82	88.48	18.8	94.02	57.21	56.82	78.21	76.42	55.18	88.90	60.42	79.02

ID dataset: Cifar-100														
Method	SVHN		Textures		LSUN-C		LSUN-R		iSUN		Places365		Average	
	FPR↓	AUR↑	FPR↓	AUR↑	FPR↓	AUR↑	FPR↓	AUR↑	FPR↓	AUR↑	FPR↓	AUR↑	FPR↓	AUR↑
Ours(K10)	63.07	83.10	72.69	83.61	10.42	81.88	87.72	60.82	92.87	62.46	77.76	80.25	68.23	74.50
Ours(K20)	60.10	84.34	78.03	80.03	10.30	95.91	83.48	69.27	82.87	69.17	85.68	69.00	67.15	76.28
Ours(K50)	29.63	92.66	73.47	79.38	<u>6.07</u>	<u>98.65</u>	87.68	67.83	86.09	69.84	86.10	69.54	61.78	79.26
Ours(K100)	55.78	88.45	64.47	82.13	8.63	98.50	79.20	<u>75.40</u>	<u>57.69</u>	84.32	70.43	81.43	56.28	85.71
Ours(K200)	53.64	<u>89.64</u>	63.81	82.53	7.4	98.51	77.65	77.81	59.03	86.46	67.57	82.87	54.92	86.21
Ours(K500)	60.27	79.63	77.93	<u>82.9</u>	4.59	98.97	81.86	70.77	62.37	73.55	76.62	65.63	60.72	78.59

Table 3: We evaluated the OOD detection performance of our method across different numbers of pseudo-categories, using ResNet18 as the backbone and following the supervised benchmark setting. The best-performing model is highlighted in bold, while the second-best results are underlined.

REFERENCES

- Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.
- Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017.
- Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33:21464–21475, 2020.
- Yiyao Sun and Yixuan Li. Dice: Leveraging sparsification for out-of-distribution detection. In *European Conference on Computer Vision*, pp. 691–708. Springer, 2022.
- Yiyao Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. *Advances in Neural Information Processing Systems*, 34:144–157, 2021.
- Yeonguk Yu, Sungho Shin, Seongju Lee, Changhyun Jun, and Kyoobin Lee. Block selection method for using feature norm in out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15701–15711, 2023.
- Zihan Zhang and Xiang Xiang. Decoupling maxlogit for out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3388–3397, 2023.