Leveraging Perturbation Robustness to Enhance Out-of-Distribution Detection

Supplementary Material

The appendix is organized as follows:

- In Sec. A1, we provide analysis following derivations in Sec. 4.2. Similarly, we intend to provide an IND Entropy score bound from adversarial training.
- In Sec. A2, we present additional experimental results to better illustrate the effectiveness of PRO.
- In Sec. A3, we introduce the implementation details including hardware and hyperparameters.

In addition to the appendix, we have attached the code for reference if more details are needed.

A1. Adversarial robustness of entropy score

We aim to show the relationship between adversarial robustness to the lower bound of the perturbed IND entropy score by demonstrating that perturbation has a limited effect on attenuating IND entropy scores.

The entropy score for OOD detection is defined as the negative Shannon entropy, aligning with the conventional setting where higher scores indicate IND inputs:

$$g_{\texttt{ENT}}(\mathbf{x}) = -H(f(\mathbf{x})) = \sum_{i=1}^{C} p_i(\mathbf{x}) \log p_i(\mathbf{x}) \tag{A11}$$

The analysis follows the derivation for the MSP score in Sec. 4.2. We begin by rewriting the negative prediction entropy in terms of the MSP score and the probabilities of the remaining classes:

$$\begin{aligned} & - H(f(\mathbf{x}+\delta)) \\ & = \sum_{i=1}^{C} p_i(f(\mathbf{x}+\delta)) \log p_i(f(\mathbf{x}+\delta)) \\ & = p_{\text{max}} \log p_{\text{max}} + \sum_{j=2}^{C} p_j \log p_j \\ & > p_{\text{max}} \log p_{\text{max}} + (C-1) p_a \log p_a, \end{aligned} \tag{A12}$$

where $p_a = (1-p_{\text{max}})/(C-1)$ denotes the probability evenly distributed among the remaining classes, leading to the maximum prediction entropy given the dominant class probability p_{max}. Next, we continue to rewrite the lower bound of the entropy score:

$$\begin{aligned} \Rightarrow & = p_{\text{max}} \log p_{\text{max}} + (1-p_{\text{max}}) \log \frac{1-p_{\text{max}}}{C-1} \\ & = p_{\text{max}} \log p_{\text{max}} + (1-p_{\text{max}}) \log(1-p_{\text{max}}) \\ & \quad + p_{\text{max}} \log(C-1) - \log(C-1) \end{aligned} \tag{A13}$$

Denote $h(p) = p \log p + (1-p)\log(1-p) + p \log(C-1) - log(C-1)$, this function is convex and non-decreasing when $c \in [1/C, 1]$. Apply Jensen's inequality, and substitute Eq. (10), we have [4]:

$$E[-H(f(\mathbf{x}))] \geq E[h(p_{\text{max}})] \geq h(E[p_{\text{max}}]) \geq h(\exp((-\mathcal{E}))) \tag{A14}$$

A2. Additional results

Perturbation robustness analysis. We extend the analysis of robustness differences using the metric of score shift. Similar to Fig. 3, we evaluate score shifts under one-step perturbations of varying magnitudes. In Fig. A1, the MSP score shifts are shown for a CIFAR-10 model without adversarial training. These results illustrate that OOD scores are generally more susceptible to perturbations compared to IND scores even without adversarial training. Additionally, we analyze score shifts on ImageNet models in Fig. A2 and Fig. A3. While a significant proportion of IND scores remain robust, forming a peak distribution near zero, a notable portion of IND scores still experience significant decreases under perturbation.

Score distribution shift. To provide further intuition on how PRO reshapes the original score distribution, Fig. A4 compares the PRO-enhanced scores with original MSP and ENT scores. As demonstrated in the plots, PRO effectively reduces the score values for OOD inputs, resulting in a distribution shift toward lower values. However, we can observe that the shifts also happened within IND scores. These shifts are particularly notable for the ImageNet model, especially MSP scores, limiting the enhancement from PRO.

OOD detection performance on ImageNet. Detailed OOD detection performance metrics for ImageNet are provided in Tab. A1. We present a default model and three models trained with data augmentation procedures PixMix [18], AugMix [17], and RegMixup [34]. We focus on the comparison with softmax scores and other gradient-based methods. The gradient-based method GradNorm [20] shows significant performance degradation for models trained with PixMix and AugMix, indicating that gradients with respect to weights are highly sensitive to data augmentations. ODIN exhibits reduced far-OOD performance across all models compared to the MSP baseline.

In contrast, our proposed method, PRO, provides consistent improvements over basic scores such as MSP and Entropy, establishing PRO as the most competitive post-hoc method for near-OOD detection among the compared

[4] We thank the anonymous reviewer for their helpful suggestion regarding the derivation in Eq. (A14).

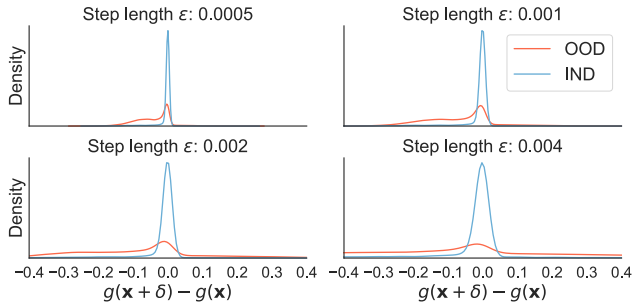


Figure A1. Distribution plots of MSP score shift introduced by a bounded perturbation. It is tested on a robust CIFAR-10 model without adversarial training.

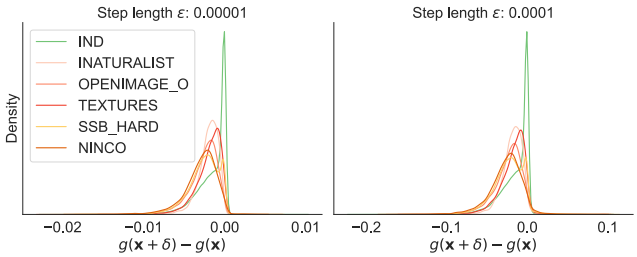


Figure A2. Distribution plots of MSP score shift introduced by one-step perturbation on a default ImageNet model without adversarial training.

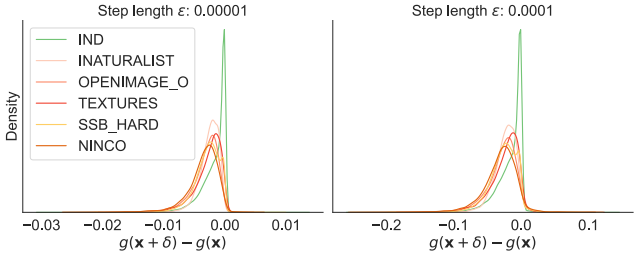


Figure A3. Distribution plots of MSP score shift introduced by one-step perturbation on an ImageNet model trained with PixMix [18] data augmentation

baselines. However, the effect of PRO on Temperature-scaled MSP and GEN is inconsistent across models. We attribute this variability to the additional hyperparameters in these methods, which increase dependence on the evaluation set's comprehensiveness.

Additional metrics on CIFAR-10. Tab. A2 provides the OOD detection performance tested on the other three robust models as an extension to Tab. 1. Both LRR-CARD-Deck and Binary-CARD-Deck have adopted an ensemble of models, making most post-hoc methods perform similarly to MSP baseline. We average the activations between models in an ensemble to implement Scale, Ash, and React. The binary model has an unconventional linear layer thus we have not implemented activation-modification methods on it. PRO has improved most averaged metrics of four softmax scores on Augmix models, achieving leading performance among baselines.

Metrics on different CIFAR-100 models. We present OOD detection metrics on five different CIFAR-100 models in Tab. A3. For this analysis, we focus on original softmax scores and ODIN as baselines to emphasize the enhancements achieved by PRO across different models. For a detailed comparison with other representative state-of-the-art methods, please refer to Tab. 2. PRO provides consistent improvements across different models, particularly for temperature-scaled confidence, entropy, and GEN. As shown in the averaged metrics, PRO-MSP-T, PRO-MSP-ENT, and PRO-GEN demonstrate leading performance across most models.

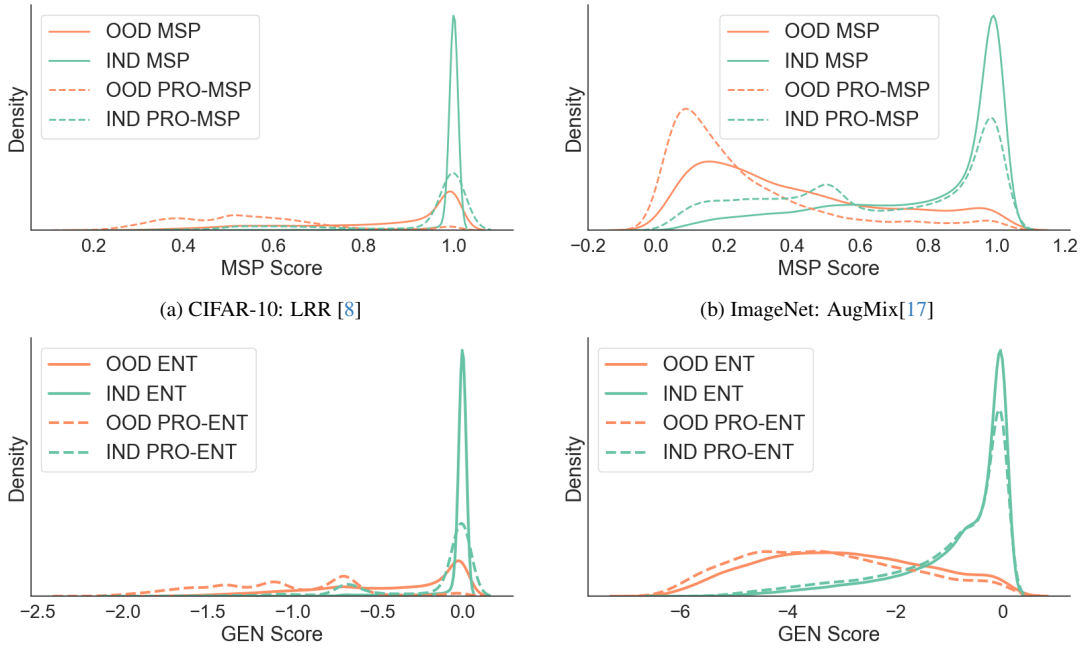
A3. Implementation details & hyperparameter

All experiments presented in this work are conducted on a workstation with four NVIDIA RTX 2080 Ti GPUs and an Intel CPU running at 2.90 GHz. The results can be reproduced by following the experimental platform established by the OpenOOD benchmark [44] [5].

In addition to the overview of PRO provided in Algorithm 1, we highlight a few additional implementation details. PRO has two hyperparameters which are determined by the evaluation set of test benchmarks. We consider step length ϵ to be within the range of 0.00005 to 0.01, with perturbations applied to normalized image tensors. As for update steps K, we limit it to a maximum of 7 to manage computational overhead. Additional hyperparameters introduced by temperature scaling and GEN have reduced search space for higher efficiency. Tab. A4 provides the considered hyperparameter settings for different datasets. It is important to note that the optimal hyperparameters may vary across different pre-trained models.

Hyperparameter sensitivity analysis. Please see Fig. A5 as an ablation study on hyperparameter. The key takeaway here is to use a small perturbation step ϵ which stably improve performance as step number K increases.

[5] https://github.com/Jingkang50/OpenOOD



(a) CIFAR-10: LRR [8]

(b) ImageNet: AugMix[17]

(c) CIFAR-10: LRR [8]

(d) ImageNet: AugMix[17]

Figure A4. PRO method reshapes scores distribution. We select MSP and ENT from a robust CIFAR-10 model and a robust ImageNet model. In above plots, OOD for CIFAR-10 is SVHN [33] and OOD for ImageNet is Texture[4].

	Default Model		**PixMix**		**AugMix**		**RegMixup**	
Method	Near-OOD	Far-OOD	Near-OOD	Far-OOD	Near-OOD	Far-OOD	Near-OOD	Far-OOD
MSP[15]	65.68/76.02	51.45/85.23	65.89/76.86	51.11/85.63	64.45/77.49	46.94/86.67	65.33/77.04	48.91/86.31
TempScaling[13]	**64.5**/77.14	46.64/87.56	64.85/78.02	46.82/87.59	62.61/78.57	42.07/88.75	64.26/77.87	44.6/87.95
Entropy[15]	64.96/77.38	47.86/88.01	64.69/78.38	46.16/88.41	63.16/78.78	41.81/89.41	63.69/78.24	41.9/88.95
GEN[30]	65.32/76.85	**35.61**/89.76	66.77/77.78	38.13/**89.54**	64.0/78.72	32.98/90.99	63.16/77.65	**34.78/89.65**
ODIN[27]	72.5/74.75	43.96/89.47	75.32/74.32	61.36/84.45	67.71/77.69	36.52/**91.1**	74.5/75.18	49.47/88.79
GradNorm[20]	78.89/72.96	47.92/**90.25**	85.37/63.42	79.68/72.27	76.3/72.14	60.35/85.01	81.96/69.22	58.99/85.75
MLS[16]	67.82/76.46	38.22/89.57	67.57/78.28	41.36/89.21	63.36/**79.14**	33.47/90.87	67.99/77.43	38.93/89.25
EBO[29]	68.56/75.89	38.39/89.47	68.75/77.75	41.04/89.3	64.17/78.76	33.45/90.95	69.06/76.48	39.97/88.87
RankFeat[36]	91.83/50.99	87.17/53.93	95.36/42.27	90.32/42.62	93.09/51.18	81.14/60.44	96.92/41.4	94.68/38.39
PRO-MSP	65.0/76.9	52.87/85.54	63.36/77.66	47.2/87.15	63.49/78.21	47.77/87.01	64.59/77.58	50.87/86.2
PRO-MSP-T	67.5/76.54	37.96/89.61	65.21/78.77	40.19/88.92	63.33/79.14	33.48/90.86	67.59/77.5	38.61/89.29
PRO-ENT	64.55/**77.66**	46.57/87.85	**61.71/78.8**	41.78/88.49	62.41/79.01	39.85/89.24	63.52/**78.26**	41.73/88.9
PRO-GEN	65.13/76.62	37.21/89.32	64.05/78.2	**37.57**/89.37	**62.08**/78.56	**32.35**/90.65	**62.96**/77.48	35.82/88.99

Table A1. OOD detection performance on ImageNet. Besides general table format **best metric**, second best metric, and our methods .

IND: CIFAR-10		OOD detection performance: FPR@95 ↓ / AUROC ↑						
	Method	Cifar100	TIN	MNIST	SVHN	Texture	Places365	Average
LRR-CARD-Deck	MSP[15]	**27.76**/91.72	**22.86**/92.99	19.84/93.79	17.90/94.27	16.37/95.28	**23.72**/92.80	**21.41**/93.47
	TempScaling[13]	**27.76**/91.72	**22.86**/92.99	19.84/93.80	17.90/94.27	16.37/95.29	**23.72**/92.80	**21.41**/93.48
	Entropy[15]	**27.76**/91.85	**22.86**/93.15	19.84/94.02	17.90/94.41	16.37/95.49	**23.72**/92.97	**21.41**/93.65
	GEN[30]	**27.76**/91.87	22.89/93.17	19.84/94.05	17.91/94.42	16.37/**95.51**	**23.72**/93.00	21.42/93.67
	ODIN[27]	32.51/91.03	27.02/92.28	**13.40/96.29**	19.68/94.11	**15.96**/95.49	28.24/91.94	22.80/93.52
	MLS[16]	**27.76**/91.73	**22.86**/93.00	19.84/93.81	17.90/94.28	16.37/95.30	**23.72**/92.81	**21.41**/93.49
	EBO[29]	**27.76**/91.87	22.87/93.18	19.84/94.05	17.91/94.43	16.37/**95.51**	**23.72**/93.00	**21.41**/93.67
	ASH[9]	70.56/79.54	68.27/81.34	50.40/88.65	82.06/65.88	61.62/85.02	57.06/85.46	64.99/80.98
	ReAct[37]	74.18/74.78	71.72/78.00	59.22/82.77	82.07/59.04	73.30/75.50	58.23/85.20	69.79/75.88
	Scale[43]	64.16/83.67	60.18/85.64	49.46/88.89	74.01/76.93	56.48/87.22	50.30/88.56	59.10/85.15
	PRO-MSP	29.01/92.09	23.41/93.61	28.76/92.27	**12.62**/95.71	20.29/94.87	24.43/93.53	23.09/93.68
	PRO-MSP-T	29.64/91.83	24.30/93.48	28.86/92.14	14.64/95.48	22.72/94.67	25.84/93.43	24.33/93.50
	PRO-ENT	28.82/**92.45**	23.47/94.08	28.46/93.13	14.43/95.76	19.93/95.43	24.24/94.07	23.23/**94.15**
	PRO-GEN	29.57/92.37	24.08/**94.09**	28.62/93.16	14.06/**95.79**	21.62/95.28	25.50/**94.11**	23.91/94.13
Binary-CARD-Deck	MSP[15]	**31.58**/90.25	27.87/91.24	17.26/95.32	23.33/92.10	**21.39**/93.45	**29.47**/91.04	25.15/92.23
	TempScaling[13]	**31.58**/90.26	27.87/91.24	17.26/95.33	23.33/92.10	**21.39**/93.46	**29.47**/91.05	25.15/92.24
	Entropy[15]	**31.58**/90.50	27.87/91.52	17.18/95.89	23.33/92.25	21.42/93.82	**29.47**/91.36	25.14/92.56
	GEN[30]	**31.58**/90.53	27.87/91.56	17.17/95.97	23.33/92.28	21.46/93.86	29.48/91.40	25.15/92.60
	ODIN[27]	32.37/90.17	29.59/90.84	**6.86/98.53**	25.42/91.50	23.53/93.54	30.83/90.70	**24.77**/92.55
	MLS[16]	**31.58**/90.27	27.87/91.26	17.26/95.36	23.33/92.11	**21.39**/93.48	**29.47**/91.07	25.15/92.26
	EBO[29]	**31.58/90.54**	**27.86**/91.57	17.17/95.97	23.31/92.28	21.43/**93.87**	**29.47**/91.41	25.14/92.61
	PRO-MSP	35.08/89.67	30.52/91.29	31.03/92.45	18.56/93.35	27.98/92.50	31.58/91.38	29.12/91.77
	PRO-MSP-T	34.01/89.98	29.77/91.47	27.72/92.93	**17.37**/93.78	24.96/92.97	30.63/91.50	27.41/92.11
	PRO-ENT	34.84/90.17	30.20/91.94	30.51/93.55	20.17/93.30	27.34/93.21	31.42/92.07	29.08/92.37
	PRO-GEN	33.88/90.43	29.57/**92.03**	27.14/93.98	17.40/**93.91**	24.68/93.64	30.61/**92.11**	27.21/**92.68**
AugMix-ResNeXt	MSP[15]	29.66/91.03	26.22/92.03	13.66/96.09	27.87/90.84	27.79/91.46	25.93/92.12	25.19/92.26
	TempScaling[13]	**29.11**/91.42	25.49/92.48	12.66/96.75	27.91/90.96	**27.56**/91.83	25.31/92.61	24.67/92.67
	Entropy[15]	29.38/91.51	25.90/92.59	13.09/97.02	27.83/91.04	27.80/91.94	25.63/92.72	24.94/92.80
	GEN[30]	29.43/**92.13**	23.51/93.63	6.43/98.60	33.93/89.04	29.30/91.90	22.76/94.03	24.23/93.22
	ODIN[27]	42.48/89.40	39.19/90.18	**0.97/99.74**	77.19/73.85	51.96/87.73	32.21/91.64	40.67/88.76
	MLS[16]	29.92/92.08	23.91/93.59	6.56/98.51	36.03/88.82	29.72/91.80	22.73/94.00	24.81/93.13
	EBO[29]	29.90/92.04	23.97/93.60	6.04/98.67	36.12/88.52	29.98/91.70	22.71/94.04	24.79/93.09
	ASH[9]	35.47/90.95	29.67/92.34	4.18/99.13	54.46/84.75	30.01/92.59	22.84/93.80	29.44/92.26
	Scale[43]	34.53/91.59	28.28/93.07	3.46/99.22	56.68/85.31	28.70/**93.13**	23.24/94.02	29.15/92.72
	PRO-MSP	30.23/90.70	26.89/91.95	19.27/94.53	**17.23/93.31**	30.13/90.95	27.02/92.06	25.13/92.25
	PRO-MSP-T	29.90/92.08	23.73/93.61	6.67/98.47	34.67/89.32	29.94/91.81	22.73/94.01	24.61/93.22
	PRO-ENT	30.96/91.37	27.22/92.99	19.50/96.21	24.76/92.10	32.24/91.37	27.52/93.18	27.03/92.87
	PRO-GEN	29.36/92.12	**23.46/93.66**	6.61/98.55	31.97/89.87	29.53/91.89	**22.60/94.05**	**23.92/93.36**

Table A2. OOD detection performance on three CIFAR-10 robust models.

IND: CIFAR-100		OOD detection performance: FPR@95 ↓ / AUROC ↑						
	Method	Cifar10	TIN	MNIST	SVHN	Texture	Places365	Average
Default Model	MSP[15]	58.91/78.47	50.70/82.07	57.23/76.08	59.07/78.42	61.88/77.32	56.62/79.22	57.40/78.60
	TempScaling[13]	**58.72**/79.02	50.26/82.79	56.05/77.27	57.71/79.79	61.56/78.11	56.46/79.80	56.79/79.46
	Entropy[15]	58.83/79.21	50.33/83.08	56.73/77.46	58.47/80.11	61.68/78.32	56.43/79.99	57.08/79.70
	GEN[30]	58.87/**79.38**	49.98/83.25	53.92/78.29	55.45/81.41	**61.23**/78.74	56.25/**80.28**	55.95/80.23
	ODIN[27]	60.64/78.18	55.19/81.63	**45.94/83.79**	67.41/74.54	62.37/**79.33**	59.71/79.45	58.54/79.49
	PRO-MSP	60.84/78.75	51.36/82.82	62.38/73.31	48.30/84.35	66.45/75.91	57.00/79.47	57.72/79.10
	PRO-MSP-T	60.18/79.05	51.13/83.03	56.13/76.32	**44.29**/85.48	64.43/77.46	57.24/79.59	**55.57**/80.15
	PRO-ENT	60.17/79.09	50.21/83.34	60.69/74.72	46.62/**86.06**	64.77/77.21	56.63/79.79	56.51/80.04
	PRO-GEN	59.83/79.24	**49.62/83.47**	58.07/75.80	46.81/85.51	63.45/77.85	**56.18**/80.07	55.66/**80.32**
Robust Model LRR	MSP[15]	57.19/78.88	50.36/81.49	57.46/74.67	52.73/78.87	62.81/74.53	56.52/78.17	56.18/77.77
	TempScaling[13]	57.48/79.91	49.02/83.07	55.00/77.96	52.17/79.79	62.31/75.62	56.14/79.15	55.35/79.25
	Entropy[15]	**57.03**/79.85	49.97/82.97	56.83/77.01	52.50/79.73	62.83/75.31	56.43/79.07	55.93/78.99
	GEN[30]	58.52/**80.68**	**46.41/84.43**	49.08/80.70	47.88/81.82	**60.02/77.30**	54.01/80.56	52.65/80.92
	ODIN[27]	68.01/76.36	56.21/80.59	**20.62/94.96**	75.73/66.27	70.17/73.40	65.94/74.85	59.45/77.74
	PRO-MSP	59.16/79.14	51.32/82.73	69.49/70.39	51.13/81.35	69.44/74.01	55.88/78.97	59.40/77.77
	PRO-MSP-T	61.94/79.94	49.18/84.01	47.60/83.33	39.06/84.15	64.18/76.02	57.11/79.09	53.18/**81.09**
	PRO-ENT	58.64/80.23	49.98/84.12	66.20/74.01	42.50/84.58	67.42/75.45	55.23/80.10	56.66/79.75
	PRO-GEN	59.57/80.47	46.63/84.41	55.73/76.58	**37.30/85.16**	61.18/76.86	**52.58/80.84**	**52.16**/80.72
Binary	MSP[15]	**62.81**/77.05	53.92/80.85	71.29/67.01	51.81/80.96	69.90/74.64	59.13/78.02	61.48/76.42
	TempScaling[13]	63.20/77.75	53.03/81.90	69.83/69.10	49.90/82.22	67.57/76.06	57.87/79.04	60.23/77.68
	Entropy[15]	63.08/78.41	52.99/82.81	70.48/70.51	49.71/83.29	68.00/77.04	58.02/79.88	60.38/78.66
	GEN[30]	64.22/78.63	50.74/83.40	66.94/74.32	45.08/84.00	64.97/78.93	55.94/80.72	57.98/80.00
	ODIN[27]	73.06/73.28	67.00/77.42	**32.00/91.08**	82.71/64.36	67.63/78.46	66.87/76.59	64.88/76.86
	PRO-MSP	63.39/77.91	53.43/82.02	78.78/62.85	45.18/83.60	73.17/74.83	58.00/79.25	61.99/76.74
	PRO-MSP-T	66.02/78.40	50.63/83.33	61.53/75.71	38.79/86.49	**60.08/79.94**	**54.50/80.96**	**55.26/80.80**
	PRO-ENT	63.48/78.63	51.63/83.22	73.51/68.52	**34.88/89.64**	69.08/77.66	56.98/80.31	58.26/79.66
	PRO-GEN	64.34/**78.64**	**50.36/83.50**	68.34/73.47	37.96/87.06	65.40/79.14	55.58/80.83	57.00/80.44
LRR-CARD-Deck	MSP[15]	**57.77**/79.48	48.12/83.35	59.53/70.34	47.17/84.15	55.59/79.36	54.12/80.54	53.72/79.54
	TempScaling[13]	**57.77**/79.48	48.12/83.35	59.53/70.34	47.17/84.16	55.59/79.37	54.12/80.54	53.72/79.54
	Entropy[15]	**57.77**/79.83	48.12/83.86	59.53/71.10	47.16/84.78	55.59/79.60	54.11/80.91	53.71/80.01
	GEN[30]	57.78/79.85	48.14/83.91	59.54/71.20	47.16/84.85	55.58/79.63	54.11/80.94	53.72/80.06
	ODIN[27]	58.73/77.26	49.40/81.51	**47.80/80.06**	44.83/84.80	**55.17/80.57**	54.31/79.24	**51.71/80.57**
	PRO-MSP	58.54/79.69	48.26/84.10	76.64/64.89	37.41/86.72	65.57/77.44	53.29/81.14	56.62/79.00
	PRO-MSP-T	58.52/79.69	48.17/84.10	76.64/64.90	37.31/86.77	65.56/77.45	53.29/81.15	56.58/79.01
	PRO-ENT	58.46/**80.31**	46.73/84.68	72.91/66.93	**32.80/88.13**	62.20/78.08	**52.21/81.67**	54.22/79.97
	PRO-GEN	58.78/80.26	**45.98/84.74**	74.50/67.05	33.23/87.17	63.48/77.84	52.33/81.61	54.72/79.78
AugMix-ResNeXt	MSP[15]	**55.42**/80.12	51.20/81.69	51.33/80.55	53.92/78.48	67.49/73.44	55.39/79.32	55.79/78.93
	TempScaling[13]	55.52/80.84	50.17/82.65	49.06/82.76	53.20/78.82	66.66/73.99	55.12/80.03	54.96/79.85
	Entropy[15]	55.51/81.08	50.63/82.95	50.24/83.29	53.46/78.84	67.41/73.95	55.17/80.18	55.40/80.05
	GEN[30]	57.81/81.02	51.04/83.17	40.81/86.32	52.54/78.21	**66.01**/74.34	54.60/80.26	53.80/80.55
	ODIN[27]	66.33/77.34	62.21/78.68	**19.21/95.70**	78.33/66.72	74.37/72.57	62.93/76.59	60.56/77.93
	PRO-MSP	58.48/80.08	52.24/82.51	61.60/79.00	53.24/79.91	73.07/72.05	56.01/79.77	59.11/78.89
	PRO-MSP-T	58.33/80.93	51.24/83.02	41.50/85.99	49.43/79.96	67.62/74.15	55.04/80.18	53.86/80.70
	PRO-ENT	56.56/81.17	50.14/**83.48**	52.33/83.06	**40.87/85.80**	68.74/74.30	54.51/**80.75**	53.86/**81.43**
	PRO-GEN	57.58/**81.20**	**50.11**/83.44	44.07/84.84	45.74/82.11	66.44/**74.57**	**53.59**/80.63	**52.92**/81.13

Table A3. OOD detection performance on CIFAR-100 models across one default model and four robust models.

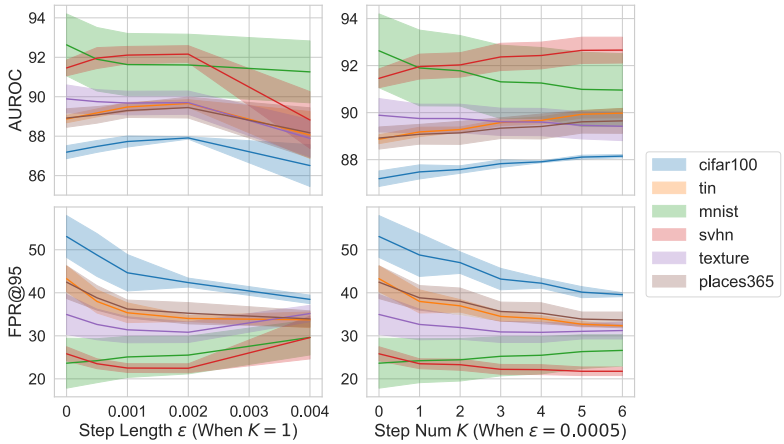


Figure A5. Hyperparameter sensitivity analysis of PRO-MSP: Statistics are evaluated across three default CIFAR-10 models. obtained from independent training runs without applying robust training.

Method	**Dataset**	**Hyperparameters**
PRO-MSP $\{\epsilon, K\}$	Cifar-10	$\{0.0003, 3\}$
	Cifar-100	$\{0.001, 5\}$
	ImageNet	$\{0.0005, 3\}$
PRO-MSP-T $\{\epsilon, K, T\}$	Cifar-10	$\{0.001, 5, 1000\}$
	Cifar-100	$\{0.001, 5, 10\}$
	ImageNet	{1.0e-05, 1, 10}
PRO-ENT $\{\epsilon, K\}$	Cifar-10	$\{0.001, 1\}$
	Cifar-100	$\{0.0005, 7\}$
	ImageNet	{5.0e-05, 7}
PRO-GEN $\{\gamma, M, \epsilon, K\}$	Cifar-10	$\{0.1, 10, 0.001, 5\}$
	Cifar-100	$\{0.01, 100, 0.0008, 5\}$
	ImageNet	$\{0.1, 100, 0.0003, 1\}$

Table A4. Example hyperparameters of PRO