Appendix

A Results under more attacks

In order to verify the effectiveness of the proposed method, in this section, we further evaluate the robustness of our method under a broader range of powerful attacks: 1) AutoAttack [2] (an ensemble of four strong diverse attacks, which is widely considered as the strongest attack for robustness evaluation), 2) CW attack [1] (CW-200), 3) PGD attack with restart [4] (PGD-200), 4) One-pixel attack [5], 5) Spatial Transformation attack [6], as well as 6) Color Channel attack [3]. PGD-200 and CW-200 both restart 5 times with 40 optimization steps each restart.

In Table 1, we report the robust accuracy under these attacks with AdvCL serving as baseline on CIFAR100. The results show that our methods can improve robustness under all different attacks across almost all settings, *e.g.*, 21.43% vs. 19.57% under AutoAttack and 29.56% vs. 27.13% under PGD-200 attack, with loss function $\mathcal{L}^{IP+HN}$, under Linear Probing.

Table 1. Robustness evaluation under diverse attacks on CIFAR100 with AdvCL as baseline.

Training Methods		PGD-200	CW-200	AA	One-pix.	Spatial-Tr.	Color-Ch.
Linear Probing	AdvCL	27.13	21.85	19.57	72.10	47.94	25.62
	w/ $\mathcal{L}^{IP}$	27.87	22.10	19.80	69.60	49.31	25.88
	w/ $\mathcal{L}^{HN}$	29.43	23.10	21.23	**73.20**	51.57	28.01
	w/ $\mathcal{L}^{IP+HN}$	**29.56**	**23.60**	**21.43**	73.00	**52.62**	**28.94**
Adversarial Linear Finetuning	AdvCL	27.29	22.01	20.09	**72.80**	47.31	24.98
	w/ $\mathcal{L}^{IP}$	27.84	22.37	20.06	71.60	46.22	24.23
	w/ $\mathcal{L}^{HN}$	**29.79**	**23.79**	21.52	70.80	**51.04**	**27.84**
	w/ $\mathcal{L}^{IP+HN}$	29.58	23.64	**21.66**	71.70	49.87	27.14
Adversarial Full Finetuning	AdvCL	29.48	25.73	24.46	**72.20**	57.86	25.12
	w/ $\mathcal{L}^{IP}$	30.10	26.05	24.73	71.00	58.95	25.55
	w/ $\mathcal{L}^{HN}$	**30.46**	**26.60**	**25.22**	69.30	59.04	**26.02**
	w/ $\mathcal{L}^{IP+HN}$	**30.46**	26.54	25.06	69.00	**59.33**	25.70

Table 2 provides results on CIFAR10 under canonical optimization-based attack methods: PGD-200, CW-200 and AutoAttack. Our methods also yield robustness gain in almost all settings.

Besides, we also report results compared with RoCL under PGD-200, CW-200 and AutoAttack in Table 3, which further validate the effectiveness of the proposed methods. For instance, 25.09% vs. 23.51% under CW-200 attack, Adversarial Full Finetuning scheme, on CIFAR100.

Table 2. Robustness evaluation under optimization-based attacks on CIFAR10, with AdvCL as baseline.

Training Methods		PGD-200	CW-200	AutoAttack
Linear Probing	AdvCL	51.05	45.65	43.48
	w/ $\mathcal{L}^{IP}$	51.99	46.02	43.57
	w/ $\mathcal{L}^{HN}$	**52.36**	**46.09**	**43.68**
	w/ $\mathcal{L}^{IP+HN}$	52.01	45.35	42.92
Adversarial Linear Finetuning	AdvCL	52.30	46.04	43.93
	w/ $\mathcal{L}^{IP}$	52.77	46.60	**44.22**
	w/ $\mathcal{L}^{HN}$	**53.22**	**46.44**	44.15
	w/ $\mathcal{L}^{IP+HN}$	52.77	45.55	43.01
Adversarial Full Finetuning	AdvCL	52.90	50.92	49.58
	w/ $\mathcal{L}^{IP}$	**53.61**	51.25	49.90
	w/ $\mathcal{L}^{HN}$	53.25	51.11	49.93
	w/ $\mathcal{L}^{IP+HN}$	53.51	**51.46**	**50.28**

Table 3. Robustness evaluation under optimization-based attacks, with RoCL as baseline, on CIFAR-10 and CIFAR-100.

Dataset	Training Methods		PGD-200	CW-200	AutoAttack
CIFAR10	Linear Probing	RoCL	32.47	33.33	24.11
		w/ $\mathcal{L}^{IP+HN}$	**34.13**	**34.59**	**24.58**
	Adversarial Linear Finetuning	RoCL	42.58	40.21	**31.81**
		w/ $\mathcal{L}^{IP+HN}$	**43.54**	**41.26**	30.37
	Adversarial Full Finetuning	RoCL	50.33	47.57	46.69
		w/ $\mathcal{L}^{IP+HN}$	**51.47**	**48.26**	**47.05**
CIFAR100	Linear Probing	RoCL	14.93	14.75	7.58
		w/ $\mathcal{L}^{IP+HN}$	**17.95**	**16.57**	**8.58**
	Adversarial Linear Finetuning	RoCL	22.59	18.99	**11.93**
		w/ $\mathcal{L}^{IP+HN}$	**24.46**	**20.69**	11.69
	Adversarial Full Finetuning	RoCL	27.95	23.51	22.70
		w/ $\mathcal{L}^{IP+HN}$	**29.37**	**25.09**	**24.01**

References

1. Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In: 2017 ieee symposium on security and privacy (sp). pp. 39–57. IEEE (2017)
2. Croce, F., Hein, M.: Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In: International conference on machine learning. pp. 2206–2216. PMLR (2020)
3. Kantipudi, J., Dubey, S.R., Chakraborty, S.: Color channel perturbation attacks for fooling convolutional neural networks and a defense against such attacks. IEEE Transactions on Artificial Intelligence **1**(2), 181–191 (2020)
4. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. In: International Conference on Learning Representations (2018)
5. Su, J., Vargas, D.V., Sakurai, K.: One pixel attack for fooling deep neural networks. IEEE Transactions on Evolutionary Computation **23**(5), 828–841 (2019)
6. Xiao, C., Zhu, J.Y., Li, B., He, W., Liu, M., Song, D.: Spatially transformed adversarial examples. arXiv preprint arXiv:1801.02612 (2018)