Failure Cases Are Better Learned But Boundary Says Sorry: Facilitating Smooth Perception Change for Accuracy-Robustness Trade-Off in Adversarial Training

Yanyun Wang Li Liu*
The Hong Kong University of Science and Technology (Guangzhou)

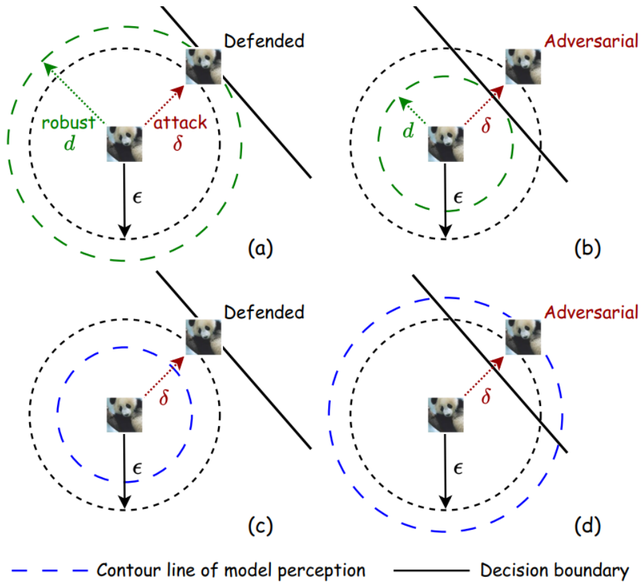
ICCV OCT 19-23, 2025 HONOLULU HAWAII

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

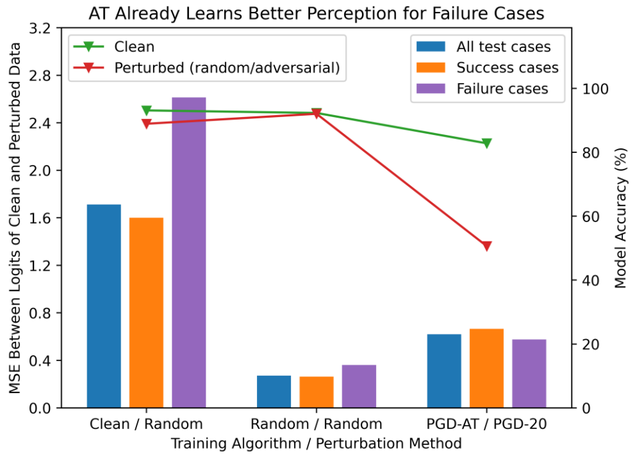
Introduction

Adversarial Training (AT) faces a ***trade-off*** between accuracy and robustness. In this work, we reveal that it is ***not insufficient but oversufficient*** learning of ***failure cases*** (those can still attack the robust model after AT) that contributes to the more complicated decision boundary and finally results in the trade-off in AT. To deal with this, we define a new AT objective named ***Robust Perception***, encouraging the model perception to ***change smoothly*** with input perturbations.

Concept (prediction vs. perception)



Motivation



A surprising phenomenon: Different from the clean training (with or without random noise), where failure cases are ***worse*** learned as we intuitively expected, AT already learns a ***better*** model perception for ***failure cases***.

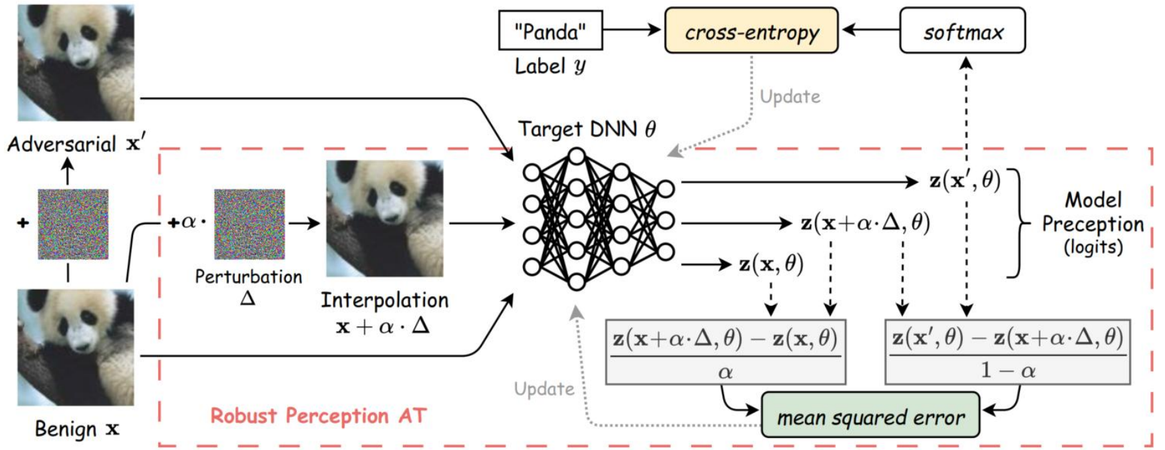
Core Idea: Robust Perception

Definition 1 (***Robust Perception***). *Provided an AT task with target model* $\boldsymbol{\theta}$*, let* $\mathbf{x}$ *and* $\mathbf{x}'$ *be any pairs of corresponding benign and adversarial samples with* $\Delta = \mathbf{x}' - \mathbf{x}$*, for any hidden representation* $h_{\boldsymbol{\theta}}(\mathbf{x})$ *denoting the model perception, our additional AT objective can be formulated as:*

$$\forall\, \alpha \in [0,1],\ \|h_{\boldsymbol{\theta}}(\mathbf{x}+\alpha\cdot\Delta) - h_{\boldsymbol{\theta}}(\mathbf{x})\| = \alpha\cdot\|h_{\boldsymbol{\theta}}(\mathbf{x}') - h_{\boldsymbol{\theta}}(\mathbf{x})\|.$$

A new regularization to encourage **smoother perception change** towards the decision boundary along with the input perturbation.

Method (Robust Perception Adversarial Training, RPAT)



$$\mathcal{L}^{RPAT}(\boldsymbol{\theta},\mathcal{D},\lambda,\alpha) := \frac{1}{n}\sum_{i=1}^{n}\left(\mathcal{L}^{CE}(\mathbf{p}(\hat{\mathbf{x}}_i',\boldsymbol{\theta}),y_i) + \lambda\cdot\mathcal{L}^{MSE}\left(\frac{\mathbf{z}(\tilde{\mathbf{x}}_i,\boldsymbol{\theta})-\mathbf{z}(\mathbf{x}_i,\boldsymbol{\theta})}{\alpha}\middle\|\frac{\mathbf{z}(\hat{\mathbf{x}}_i',\boldsymbol{\theta})-\mathbf{z}(\tilde{\mathbf{x}}_i,\boldsymbol{\theta})}{1-\alpha}\right)\right)$$

Evaluation (3 models, 3 datasets, 4 baselines + 12 SOTAs involved)

PreActResNet-18

Norm	Method		CIFAR-10 Clean	CIFAR-10 AA	CIFAR-10 Mean	CIFAR-10 NRR	CIFAR-100 Clean	CIFAR-100 AA	CIFAR-100 Mean	CIFAR-100 NRR
ℓ_∞	WA	UAI'18	83.50	49.89	66.695	62.461	57.26	25.83	41.545	35.601
	MMA	ICLR'20	**85.50**	37.20	61.350	51.844	**60.60**	18.40	39.500	28.229
	AWP	NeurIPS'20	81.11	50.09	65.600	61.933	54.10	25.16	39.630	34.347
	GAIRAT	ICLR'21	78.70	37.70	58.200	50.979	52.00	19.80	35.900	28.680
	KD+SWA	ICLR'21	84.06	49.82	**66.940**	62.562	57.17	25.66	41.415	35.422
	EWAT	ICML'21	82.80	48.20	65.500	60.931	54.20	23.52	38.860	32.805
	MAIL	NeurIPS'21	79.50	39.60	59.550	52.867	46.50	16.70	31.600	24.574
	TE	ICLR'22	82.04	50.12	66.080	62.225	56.41	25.84	41.125	35.444
	SOVR	ICML'23	81.90	49.40	65.650	61.628	52.10	24.30	38.200	33.142
	ReBAT	NeurIPS'23	82.09	50.72	66.405	62.700	56.13	27.60	41.865	37.004
	RPAT++	**Ours**	82.63	**51.00**	66.815	**63.072**	56.84	**27.68**	**42.260**	**37.230**
ℓ_2	ReBAT	NeurIPS'23	88.79	71.00	79.895	78.905	65.58	42.67	54.125	51.701
	RPAT++	**Ours**	**89.06**	**71.26**	**80.160**	**79.172**	**65.63**	**42.85**	**54.240**	**51.848**

WideResNet-34-10 (CIFAR10, Linf)

Method		Clean	AA	Mean	NRR
WA	UAI'18	87.66	52.65	70.155	65.787
MMA	ICLR'20	**87.80**	43.10	65.450	57.818
AWP	NeurIPS'20	85.63	53.32	69.475	65.718
GAIRAT	ICLR'21	83.00	41.80	62.400	55.599
KD+SWA	ICLR'21	87.45	53.59	70.520	66.456
EWAT	ICML'21	86.00	51.60	68.800	64.500
MAIL	NeurIPS'21	82.20	43.30	62.750	56.721
TE	ICLR'22	85.97	52.88	69.425	65.482
SOVR	ICML'23	85.00	53.10	69.050	65.366
ReBAT	NeurIPS'23	85.25	54.78	70.015	66.700
ADR	ICLR'24	84.67	53.25	68.960	65.381
CURE	ICLR'24	87.05	52.10	69.575	65.186
RPAT++	**Ours**	86.76	**54.97**	**70.865**	**67.300**
ReBAT*	NeurIPS'23	86.66	55.64	71.150	67.769
RPAT++*	**Ours**	**87.57**	**55.79**	**71.680**	**68.158**

[Acknowledgements] This work was supported by the National Natural Science Foundation of China (No. 62471420), GuangDong Basic and Applied Basic Research Foundation (2025A1515012296), and CCF-Tencent Rhino-Bird Open Research Fund.

[Code] https://github.com/FlaAI/RPAT

[Email] ywang856@connect.hkust-gz.edu.cn; avrillliu@hkust-gz.edu.cn