Appendix

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

A. Experimental Setup

We employ common experimental settings aligned with previous AT works. For the outer minimization, we adopt SGD optimizer with momentum 0.9, batch size 128, weight decay 5×10^{-4} and initial learning rate 0.1. In Section 5.2, following Pang et al. [31], which explores various settings for AT, we train 110 epochs with learning rate decay by a factor of 0.1 at 100 and 105 epochs for better experimental efficiency. While in Section 5.3, we adopt 200 training epochs with learning rate decay at 100 and 150 epochs to ensure fairness in comparison with the current SOTAs that also use this default setting. For the inner maximization, under the ℓ_∞ threat model with perturbation budget $\epsilon = 8/255$, we employ PGD-10 adversary with step size $\alpha = 2/255$ and maximum optimization step 10, except for TRADES which crafts adversarial samples through maximizing its *KL* regularization term [58]. While under the ℓ_2 threat model with maximal perturbation budget $\epsilon = 128/255$, we have step size $\alpha = 32/255$. In the main experiments, following their original papers, we set the regularization parameter $\lambda = 6$ for TRADES and MART while $\lambda = 1$ for Consistency-AT and ours, and fix our specific hyper-parameter $\alpha = 0.5$. Other values of these parameters are further considered by our ablation studies in Section 5.4. Then, for the data pre-processing, we normalize benign images into [0, 1], and employ standard data augmentations, including random crop with 4-pixel zero padding and random horizontal flip with 50% of probability.

The experiments are conducted on Ubuntu 22.04 OS with Intel Xeon Gold 6226R @ 2.90GHz CPU, 512GB RAM and 8 $\times$ NVIDIA GeForce RTX 3090 GPUs, and are implemented with Python 3.8.19 and PyTorch 1.11.0+cu113.

B. Proof of Theorems

In this section, we supplement the proofs of Theorem 1 and Theorem 2 in Section 4.1, which respectively demonstrate the effectiveness of the proposed RPAT from the perspectives of local linearity [12] and *Lipschitz* regularization [1, 11, 21, 33].

B.1. Proof of Theorem 1

Theorem 1 (Section 4.1) **.** *Let H be the Hessian Matrix such that $H_{h_{\boldsymbol{\theta}}}(\mathbf{x}) = \nabla^2_{\mathbf{x}} h_{\boldsymbol{\theta}}(\mathbf{x})$, then with the new optimization objective of Robust Perception, we have:*

$$\forall\, \Delta,\ \ \Delta^\top \cdot H_{h_{\boldsymbol{\theta}}}(\mathbf{x}) \cdot \Delta \to 0. \tag{4}$$

Proof.

Based on Definition 1 in Section 4.1, for any robust model $\boldsymbol{\theta}$ satisfies the proposed *Robust Perception*, we have:

$$\forall\, \alpha \in [0,1],\ \ h_{\boldsymbol{\theta}}(\mathbf{x} + \alpha \cdot \Delta) - h_{\boldsymbol{\theta}}(\mathbf{x}) = \alpha \cdot (h_{\boldsymbol{\theta}}(\mathbf{x} + \Delta) - h_{\boldsymbol{\theta}}(\mathbf{x})). \tag{5}$$

Expand the $h_{\boldsymbol{\theta}}(\mathbf{x} + \alpha \cdot \Delta)$ term of Equation (5) as a *Taylor series*:

$$h_{\boldsymbol{\theta}}(\mathbf{x} + \alpha \cdot \Delta) = h_{\boldsymbol{\theta}}(\mathbf{x}) + \alpha \cdot J_{h_{\boldsymbol{\theta}}}(\mathbf{x}) \cdot \Delta + \frac{\alpha^2}{2} \cdot \Delta^\top \cdot H_{h_{\boldsymbol{\theta}}}(\mathbf{x}) \cdot \Delta + \mathcal{O}(\alpha^3), \tag{6}$$

and also expand the $h_{\boldsymbol{\theta}}(\mathbf{x} + \Delta)$ term of Equation (5) as a *Taylor series*:

$$h_{\boldsymbol{\theta}}(\mathbf{x} + \Delta) = h_{\boldsymbol{\theta}}(\mathbf{x}) + J_{h_{\boldsymbol{\theta}}}(\mathbf{x}) \cdot \Delta + \frac{1}{2} \cdot \Delta^\top \cdot H_{h_{\boldsymbol{\theta}}}(\mathbf{x}) \cdot \Delta + \mathcal{O}(\|\Delta\|^3). \tag{7}$$

Substitute Equation (6) and Equation (7) back into Equation (5), we have:

$$\forall\, \alpha \in [0,1],\ \ \alpha \cdot J_{h_{\boldsymbol{\theta}}}(\mathbf{x}) \cdot \Delta + \frac{\alpha^2}{2} \cdot \Delta^\top \cdot H_{h_{\boldsymbol{\theta}}}(\mathbf{x}) \cdot \Delta + \mathcal{O}(\alpha^3) = \alpha \cdot (J_{h_{\boldsymbol{\theta}}}(\mathbf{x}) \cdot \Delta + \frac{1}{2} \cdot \Delta^\top \cdot H_{h_{\boldsymbol{\theta}}}(\mathbf{x}) \cdot \Delta + \mathcal{O}(\|\Delta\|^3)), \tag{8}$$

which can be then simplified as:

$$\forall\, \alpha \in [0,1],\ \ (\frac{\alpha^2}{2} - \frac{\alpha}{2}) \cdot \Delta^\top \cdot H_{h_{\boldsymbol{\theta}}}(\mathbf{x}) \cdot \Delta + \mathcal{O}(\alpha^3) - \alpha \cdot \mathcal{O}(\|\Delta\|^3)) = 0. \tag{9}$$

Equation (9) directly means:

$$\Delta^\top \cdot H_{h_{\boldsymbol{\theta}}}(\mathbf{x}) \cdot \Delta = 0, \;\; \mathcal{O}(\alpha^3) = 0, \;\text{ and }\; \mathcal{O}(\|\Delta\|^3)) = 0, \tag{10}$$

otherwise we can easily find a specific value of α except 0 and 1 that violates Equation (9).

Therefore, we have Theorem 1 proven, meaning that *Robust Perception* limits the second-order and higher-order nonlinear effects within the adversarial perturbation to the model perception, just as we suggested in Section 4.1.

B.2. Proof of Theorem 2

Theorem 2 (Section 4.1) **.** *Let J be the Jacobian Matrix such that $J_{h_{\boldsymbol{\theta}}}(\mathbf{x}) = \nabla_{\mathbf{x}} h_{\boldsymbol{\theta}}(\mathbf{x})$, then with the new optimization objective of Robust Perception, we have:*

$$\forall\, \alpha \in [0, 1], \;\; J_{h_{\boldsymbol{\theta}}}(\mathbf{x} + \alpha \cdot \Delta) \to J_{h_{\boldsymbol{\theta}}}(\mathbf{x}), \tag{11}$$

then with $\|\cdot\|_{spec}$ denoting the Spectral Norm and γ being any micro value, given $\|J_{h_{\boldsymbol{\theta}}}(\mathbf{x} + \alpha \cdot \Delta) - J_{h_{\boldsymbol{\theta}}}(\mathbf{x})\|_{\text{spec}} \leq \gamma$, the function $h_{\boldsymbol{\theta}}(\mathbf{x})$ can be referred to as K-Lipschitz with the Lipschitz constant K upper-bounded by:

$$K \;\leq\; \sup_{\mathbf{x}} \|J_{h_{\boldsymbol{\theta}}}(\mathbf{x})\|_{\text{spec}} + \gamma. \tag{12}$$

Proof.

Provided the result in Equation (10), we can simplify the *Taylor series* at $\mathbf{x}$ in Equation (6) as:

$$h_{\boldsymbol{\theta}}(\mathbf{x} + \alpha \cdot \Delta) \approx h_{\boldsymbol{\theta}}(\mathbf{x}) + \alpha \cdot J_{h_{\boldsymbol{\theta}}}(\mathbf{x}) \cdot \Delta. \tag{13}$$

For $h_{\boldsymbol{\theta}}(\mathbf{x} + \Delta)$, this time we expand it at $\mathbf{x} + \alpha \cdot \Delta$ (*i.e.*, this is the point where the derivatives are considered), such that the variable here becomes $\Delta - \alpha \cdot \Delta = (1 - \alpha) \cdot \Delta$, with which we have:

$$h_{\boldsymbol{\theta}}(\mathbf{x} + \Delta) \approx h_{\boldsymbol{\theta}}(\mathbf{x} + \alpha \cdot \Delta) + (1 - \alpha) J_{h_{\boldsymbol{\theta}}}(\mathbf{x} + \alpha \cdot \Delta) \cdot \Delta, \tag{14}$$

in which the higher-order terms are also ignored based on Equation (10).

Then substitute Equation (13) into Equation (14), we have:

$$h_{\boldsymbol{\theta}}(\mathbf{x} + \Delta) \approx h_{\boldsymbol{\theta}}(\mathbf{x}) + \alpha \cdot J_{h_{\boldsymbol{\theta}}}(\mathbf{x}) \cdot \Delta + (1 - \alpha) J_{h_{\boldsymbol{\theta}}}(\mathbf{x} + \alpha \cdot \Delta) \cdot \Delta. \tag{15}$$

Recall the definition of *Robust Perception* in Equation (5), substitute Equation (13) and Equation (15) into it, we have:

$$\forall\, \alpha \in [0, 1], \;\; \alpha \cdot J_{h_{\boldsymbol{\theta}}}(\mathbf{x}) \cdot \Delta \approx \alpha \cdot (\alpha \cdot J_{h_{\boldsymbol{\theta}}}(\mathbf{x}) \cdot \Delta + (1 - \alpha) J_{h_{\boldsymbol{\theta}}}(\mathbf{x} + \alpha \cdot \Delta) \cdot \Delta), \tag{16}$$

which can be directly simplified as:

$$\forall\, \alpha \in [0, 1], \;\; J_{h_{\boldsymbol{\theta}}}(\mathbf{x}) \approx J_{h_{\boldsymbol{\theta}}}(\mathbf{x} + \alpha \cdot \Delta). \tag{17}$$

This result tells us the proposed *Robust Perception* encourages the stability in *Jacobian* along with the adversarial perturbation, as given in Equation (11) of Theorem 2.

Based on Equation (17), let us assume that the change in *Jacobian* along with the perturbation satisfies:

$$\|J_{h_{\boldsymbol{\theta}}}(\mathbf{x} + \alpha \cdot \Delta) - J_{h_{\boldsymbol{\theta}}}(\mathbf{x})\|_{\text{spec}} \;\leq\; \gamma, \tag{18}$$

where γ is a micro value and *Spectral Norm* $\|\cdot\|_{\text{spec}}$ indicates the maximum singular value of the matrix.

Then, according to the triangular inequality of *Spectral Norm*, we have:

$$\|J_{h_{\boldsymbol{\theta}}}(\mathbf{x} + \alpha \cdot \Delta)\|_{\text{spec}} - \|J_{h_{\boldsymbol{\theta}}}(\mathbf{x})\|_{\text{spec}} \;\leq\; \|J_{h_{\boldsymbol{\theta}}}(\mathbf{x} + \alpha \cdot \Delta) - J_{h_{\boldsymbol{\theta}}}(\mathbf{x})\|_{\text{spec}} \;\leq\; \gamma. \tag{19}$$

Since the local *Lipschitz* constant $K_{\text{local}}(\cdot)$ can be denoted with *Jacobian* as $\|J_{h_{\boldsymbol{\theta}}}(\cdot)\|_{\text{spec}}$ [30], we further have:

$$K_{\text{local}}(\mathbf{x} + \alpha \cdot \Delta) - \|J_{h_{\boldsymbol{\theta}}}(\mathbf{x})\|_{\text{spec}} \;\leq\; \gamma, \tag{20}$$

with which we can finally represent the upper bound of the global *Lipschitz* constant K under the adversarial perturbation as:

$$K = \sup_{\mathbf{x}} K_{\text{local}}(\mathbf{x} + \alpha \cdot \Delta) \leq \sup_{\mathbf{x}} \|J_{h_{\boldsymbol{\theta}}}(\mathbf{x})\|_{\text{spec}} + \gamma, \tag{21}$$

just as Equation (12) of Theorem 2.

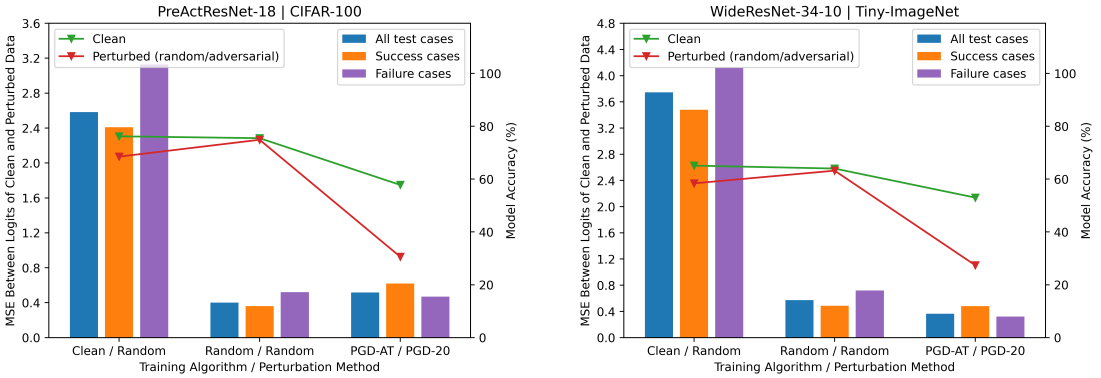
With Theorem 2, we can refer to the function $h_{\boldsymbol{\theta}}(\mathbf{x})$ learned under *Robust Perception* as K-*Lipschitz* with an upper-bounded global *Lipschitz* constant, which is expected to limit the complexity of the decision boundary as suggested in Section 4.1.

C. Additional Results

This section supplements more experimental results to further support our ideas and statements in this work, including more empirical evidence for our motivation in Appendix C.1, different options of the proxy for model perception in Appendix C.2, and further comparison with the current SOTA in Appendix C.3.

C.1. More Empirical Evidence for Motivation

Corresponding to the proof-of-concept experiment illustrated in Figure 2, which is conducted on CIFAR-10 with ResNet-18, we provide more empirical evidence respectively on CIFAR-100 with PreActResNet-18 and Tiny-ImageNet with WideResNet-34-10, as illustrated in Figure 5, both of which show similar patterns to Figure 2.



(a) Results on CIFAR-100 with PreActResNet-18.

(b) Results on Tiny-ImageNet with WideResNet-34-10.

Figure 5. More empirical evidence for our motivation, aligned with the proof-of-concept results in Figure 2.

C.2. Various Proxies for Model Perception

Regarding the proxy of model perception in our RPAT, embeddings from other layers than *logits* could also be alternatives for comparing perception consistency. In this section, we further provide the results with two additional proxies, respectively the embeddings from the second-to-last layer and the third-to-last layer, as demonstrated in Table 5. It can be found that there is no significant difference in the final performance, which also reflects the universality and stability of our new AT objective. Thus, for simplicity, we uniformly utilize *logits* as our proxy in the main experiments.

Table 5. Comparison of different proxies for model perception in calculating the perception consistency. The results are acquired on CIFAR-10 with ResNet-18 and ℓ_∞ norm.

Method / Proxy		Clean	AA	Mean	NRR
PGD-AT (*i.e.*, the baseline as in Table 1)		82.92	46.74	64.83	59.78
+ RPAT (Ours)	*logits* (*i.e.*, the current proxy as in Table 1)	**83.20**	48.00	**65.60**	**60.88**
	Embedding from the second-to-last layer	83.19	47.89	65.54	60.79
	Embedding from the third-to-last layer	83.14	**48.02**	65.58	**60.88**

C.3. Further Comparison with Current SOTA

Except from the ones considered in Section 5.3, ReBAT also suggests another additional training strategy, which is to utilize a stronger training-time adversary with larger perturbation budgets (*e.g.*, $\epsilon = 10/255$) after the first learning rate decay. Although this seems not completely aligned with the default fairness setting of AT, we still supplement comparison with it, as ReBAT is the current SOTA method on the accuracy-robustness trade-off problem. The additional consideration of such a stronger adversary is marked by "$" in Table 6.

Table 6. Comparison of the proposed RPAT^{++} with the current SOTA, ReBAT, under stronger training adversary on CIFAR-10 with PreActResNet-18 and ℓ_∞ norm.

Method	Clean		PGD-20		AA		Mean	NRR
	best	final	best	final	best	final		
ReBAT	82.09	82.05	55.77	56.03	50.72	50.70	66.405	62.700
RPAT^{++}	**82.63**	**82.76**	56.27	56.02	51.00	50.71	**66.815**	**63.072**
ReBAT$	77.57	78.82	56.79	56.46	50.91	**51.09**	64.240	61.474
RPAT^{++}$	79.25	79.47	**57.04**	**56.52**	**51.25**	**51.09**	65.250	62.246

The results demonstrate that, although the stronger training adversary strategy helps further improve the robustness, its destruction on the clean accuracy is more significant. As a consequence, for the Mean and NRR scores measuring the trade-off, adopting such a strategy rather leads to worse results. Therefore, we would not suggest using this strategy for the proposed RPAT method by default.