Adversarial Training for Probabilistic Robustness

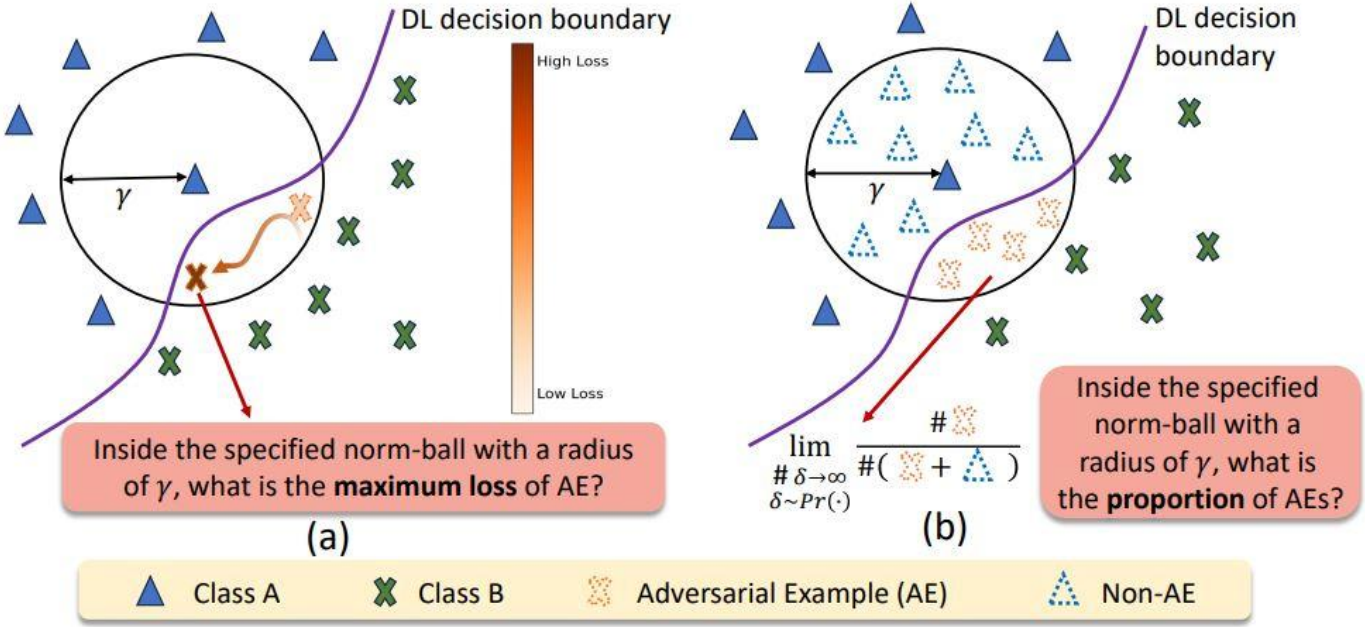
Yi Zhang
PhD Student within Verification & Validation
WMG, University of Warwick, UK

WMG
THE UNIVERSITY OF WARWICK

ICCV
OCT 19-23, 2025
HONOLULU
HAWAII

2025-08-30

Worst-case vs. Probabilistic Robustness

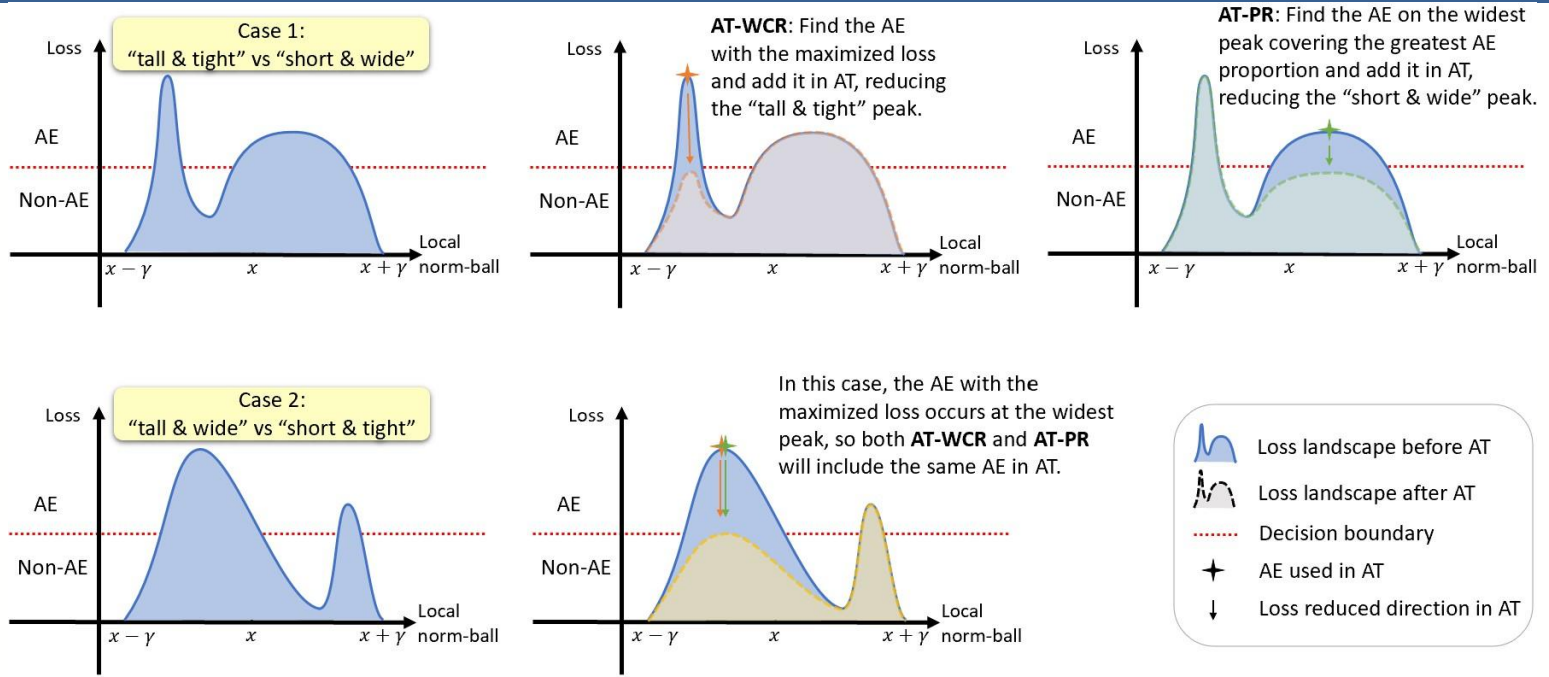


Definition 1 (Probabilistic Robustness) *For a DL classifier f_θ that takes input x and returns a prediction label, the PR of an input x in a norm ball of radius γ is:*

$$PR(x,\gamma) = \mathbb{E}_{\substack{\delta\sim Pr(\cdot|x)\\ \|\delta\|\le\gamma}}[I_{\{f_\theta(x+\delta)=y\}}(x+\delta)],$$

where $I_S(x)$ is an indicator function—it is equal to 1 when S is true and 0 otherwise; $Pr(\cdot)$ is the local distribution of inputs representing how perturbations δ are generated, which is precisely the "input model" used by [59, 72].

Loss landscape



- Two cases of the local loss landscape and their AEs identified by AT-WCR and AT-PR for AT.

Algorithms of AT-PR

Definition 2 (AT for PR) *For a DL model $f_{\boldsymbol{\theta}}$, trained on a dataset $\mathcal{D}$ consisting of pairs $(\boldsymbol{x}, y)$, the objective of AT-PR can be formulated as:*

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}} \left[\max_{\|\boldsymbol{\delta}\|\leq\gamma, PR(\boldsymbol{x}+\boldsymbol{\delta},k)=0} k \right],$$

The intuition behind the maximization is to find an optimal $\boldsymbol{\delta}^*$ that produces an AE $\boldsymbol{x}' = \boldsymbol{x} + \boldsymbol{\delta}^*$ and to maximize the radius k of a smaller norm-ball centered on $\boldsymbol{x}'$ such that the PR within this k-sized norm-ball is 0 (i.e., all inputs within this smaller norm-ball are AEs).

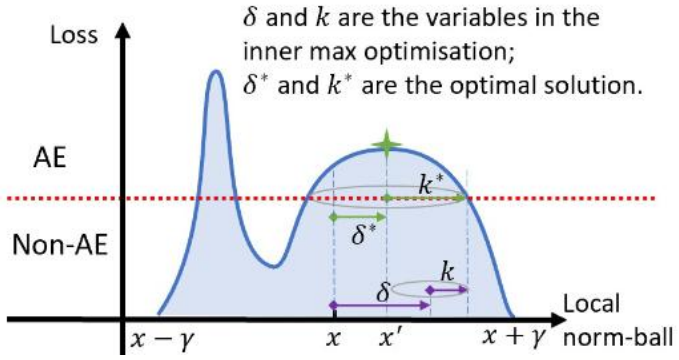


Illustration of the new inner-max optimization in Def 2.

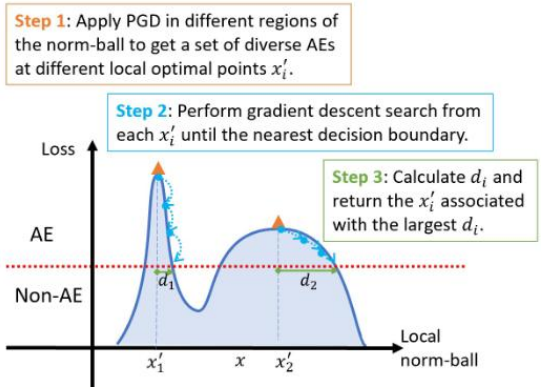


Illustration of the 3 main steps of Algorithm. 1.

Algorithm 1 PGD and gradient-based search for $\boldsymbol{x}'_{pr}$

Require: Inputs $[X, Y]$, N, $\alpha_{min}, \alpha_{max}, step_{min}, step_{max}$
Initialize $AE_s \leftarrow [\,]$ ▷ Initialize AE candidate set
Apply PGD to get different AE candidates
for idx $= 0$ **to** N **do**
  $\boldsymbol{x}_{init} \leftarrow \boldsymbol{x} + \text{Uniform}(-\gamma, \gamma)$
  $\alpha \leftarrow \text{Uniform}(\alpha_{min}, \alpha_{max})$
  $steps \leftarrow \text{Uniform}(step_{min}, step_{max})$
  $\boldsymbol{x}'_{idx} \leftarrow pgd(\boldsymbol{x}_{init}, y, \gamma, \alpha, steps)$
  Append $\boldsymbol{x}'_{idx}$ to AE_s
end for
Require: $AE_s, \boldsymbol{x}, y, C$
Initialize $Max\_d \leftarrow 0$; $\boldsymbol{x}'_{pr} \leftarrow$ None
for each $\boldsymbol{x}' \in AE_s$ **do**
  $\tilde{\boldsymbol{x}} = \boldsymbol{x}'$
  while $iter < C$ **do**
    $y' = f(\tilde{x})$
    if $y' = y$ **then**
      break ▷ Exit if $\tilde{\boldsymbol{x}}$ is classified correctly
    end if
    $g \leftarrow \nabla_{\tilde{x}} L(\tilde{\boldsymbol{x}}, y)$ ▷ Compute gradient
    $\tilde{\boldsymbol{x}} = \tilde{\boldsymbol{x}} - \alpha \cdot g$ ▷ Update $\tilde{\boldsymbol{x}}$
  end while
  $d \leftarrow D(\tilde{\boldsymbol{x}}, \boldsymbol{x}')$ ▷ Distance between $\tilde{\boldsymbol{x}}$ & $\boldsymbol{x}'$
  if $d > Max\_d$ **then**
    $\boldsymbol{x}'_{pr} = \boldsymbol{x}'$
    $Max\_d = d$
  end if
end for
Output $\boldsymbol{x}'_{pr}$ ▷ Farthest AE from decision boundary.

Experiment of AT-PR

Table 1. Evaluation results of DL models trained with different AT schemes, showing standard accuracy, adversarial accuracy (indicating WCR: PGD-10, C&W-10), and PR estimated (mean ± std) for four norm-ball sizes γ. All results are calculated over the test datasets.

Dataset	Model	AT Scheme	Accu. %	PGD Ac. %	CW Ac. %	PR(·,0.08) %	PR(·,0.1)%	PR(·,0.12) %	PR(·,0.15) %
CIFAR-10	ResNet-18	W/O AT	**94.96**	0.02	0.51	78.38 ± 33.14	66.3 ± 38.06	53.41 ± 40.76	37.01 ± 39.77
		AT-WCR(FGSM)	88.6	0.09	0.09	93.99 ± 17.5	88.02 ± 25.03	80.42 ± 31.81	67.77 ± 38.6
		AT-WCR(PGD)	81	**45.01**	**44.56**	97.28 ± 13.29	95.77 ± 17.07	94.1 ± 20.2	89.78 ± 26.53
		AT-PR	80.14	39.77	39.03	**98.22±10.38**	**97.28±13.16**	**96.05±16.18**	**93.21±21.47**
	WRN-50-2	W/O AT	**94.14**	0.23	0.14	76.96±32.76	65.65±37.41	54.43±39.88	40.19±39.49
		AT-WCR(FGSM)	92.04	0.08	0.04	95.68±14.53	91.9±20.25	86.67±25.98	76.79±33.1
		AT-WCR(PGD)	82.62	**49.51**	**48.13**	97.86±12.61	97.08±14.54	95.84±±17.14	92.62±22.87
		AT-PR	82.71	43.27	43.07	**98.09±10.81**	**97.33±13.03**	**96.22±15.58**	**93.56±20.89**
CIFAR-100	ResNet-18	W/O AT	**75.5**	0.02	0.0	44.58 ± 42.64	32.1 ± 40.1	23.32 ± 36.41	15.62 ± 31.57
		AT-WCR(FGSM)	67.80	1.41	0.66	83.90 ± 31.11	72.89 ± 37.94	61.06 ± 42.23	44.71 ± 43.04
		AT-WCR(PGD)	52.85	**19.93**	**19.51**	95.4 ± 18.15	92.7 ± 22.72	88.98 ± 28.19	80.47 ± 36.4
		AT-PR	53.21	19.38	18.39	**97.59±12.96**	**95.96±16.62**	**93.15±21.77**	**87.38±29.63**
	WRN-50-2	W/O AT	**75.46**	0.49	0.45	51.11±40.54	37.96±39.23	28.33±36.72	19.11±31.97
		AT-WCR(FGSM)	73.17	0.29	0.1	85.68±27.45	76.62±33.89	66.37/38.66	52.04±41.42
		AT-WCR(PGD)	57.46	**23.17**	**22.5**	95.31±17.38	92.0±23.79	88.14±29.15	80.67±36.21
		AT-PR	59.46	22.06	21.59	**96.05±16.17**	**93.72±20.47**	**89.48±26.87**	**81.6±34.86**
SVHN	ResNet-18	W/O AT	**94.96**	0.08	0.12	97.11 ± 9.66	95.63 ± 11.98	93.74 ± 14.48	90.12 ± 18.40
		AT-WCR(FGSM)	93.49	4.68	5.32	98.79 ± 7.06	98.42 ± 7.70	97.86 ± 8.67	96.57 ±10.78
		AT-WCR(PGD)	89.21	44.57	**43.70**	98.96 ± 5.85	98.50 ± 7.40	97.72 ± 9.61	95.41 ± 14.41
		AT-PR	90.29	**45.81**	42.21	**99.12 ± 5.15**	**98.87 ± 5.93**	**98.54 ± 6.84**	**97.84 ± 8.60**
	WRN-50-2	W/O AT	**95.91**	1.74	1.29	96.60 ± 10.36	94.77 ± 13.01	92.40 ± 15.87	87.97 ± 20.04
		AT-WCR(FGSM)	94.31	8.13	8.54	98.34 ± 7.49	97.71 ± 8.39	96.80 ± 9.69	95.12 ± 12.06
		AT-WCR(PGD)	88.91	**54.99**	50.58	98.72 ± 6.44	98.28 ± 7.66	**97.76 ± 8.95**	**96.77 ± 11.18**
		AT-PR	91.97	52.22	**51.58**	**98.88 ± 6.12**	**98.38 ± 7.53**	97.72 ± 9.19	96.26 ± 12.35
Tiny-ImageNet	ResNet-18	W/O AT	**59.26**	0.0	0.0	80.78 ± 33.1	72.95 ± 37.69	64.17 ± 40.91	51.14 ± 42.91
		AT-WCR(FGSM)	54.81	0.05	0.01	89.9 ± 24.97	85.33 ± 29.95	80.18 ± 33.74	69.03 ± 39.39
		AT-WCR(PGD)	47.43	**21.3**	19.31	96.32 ± 17.01	93.79 ± 22.1	90.16 ± 27.76	82.46 ± 35.97
		AT-PR	51.61	21.1	**19.89**	**98.23 ± 11.06**	**96.1 ± 17.84**	**93.12 ± 23.32**	**86.65 ± 32.26**
	ResNet-34	W/O AT	**61.66**	0.0	0.0	82.39 ± 31.3	75.2 ± 36.0	66.95 ± 39.42	53.7 ± 42.17
		AT-WCR(FGSM)	57	0.13	0.02	90.67 ± 23.92	85.81 ± 29.5	80.57 ± 34.13	70.22 ± 39.36
		AT-WCR(PGD)	49.33	20.97	19.58	95.45 ± 18.55	93.08 ± 23.48	89.85 ± 28.3	83.53 ± 34.99
		AT-PR	53.98	**21.72**	**20.95**	**96.71 ± 15.74**	**95.02 ± 19.89**	**92.5 ± 23.87**	**86.23 ± 32.29**

Table 2. ViT models trained on CIFAR-10 with TRADES and AT-PR, showing standard accuracy, adversarial accuracy (indicating WCR: PGD, C&W, AA), and PR estimated (mean ± std) for four norm-ball sizes γ.

Dataset	Model	AT Scheme	Accu. %	PGD-20. %	CW-20. %	AA. %	PR(·,0.08) %	PR(·,0.1)%	PR(·,0.12) %	PR(·,0.15) %
CIFAR-10	Deit-Ti	TRADES	79.44	**49.78**	46.45	**45.96**	96.61 ± 15.23	94.73 ± 19.36	92.69 ± 23.01	89.18 ± 27.9
		+AT-PR	**81.07**	49.08	**46.92**	45.95	**97.26 ± 14.0**	**95.94 ± 17.12**	**94.16 ± 20.5**	**90.65 ± 25.88**
	Vit-S	TRADES	81.03	**50.96**	48.14	**47.40**	96.93 ± 14.99	95.48 ± 18.4	93.58 ± 21.62	90.06 ± 26.81
		+AT-PR	**83.29**	50.44	**48.20**	47.34	**97.76 ± 11.89**	**96.03 ± 16.46**	**93.98 ± 20.69**	**90.48 ± 26.19**
	Vit-B	TRADES	83.55	**53.14**	50.94	**49.96**	97.34 ± 13.57	95.87 ± 16.99	93.78 ± 21.17	90.77 ± 25.88
		+AT-PR	**86.03**	52.94	**51.24**	49.34	**97.9 ± 11.8**	**96.49 ± 15.2**	**94.44 ± 19.62**	**91.02 ± 25.27**

Table 3. Comparison of PR(·, γ) and training time (sec/epoch) for more recent AT methods (TRADES, MART, ALP, CLP) and Fast-AT-PR

Data	Model	Metric	AT Scheme						
			PGD	TRADES	MART	ALP	CLP	Fast-AT-PR	AT-PR
CIFAR-10	ResNet-18	PR(·,0.08)%	97.28 ± 13.29	97.74 ± 12.42	97.4 ± 13.7	96.73 ± 15.9	97.61 ± 13.06	98.12 ± 11.52	**98.22 ± 10.38**
		PR(·,0.10)%	95.77 ± 17.07	96.33 ± 16.39	95.92 ± 17.43	95.16 ± 19.53	96.29 ± 16.66	97.1 ± 14.34	**97.28 ± 13.16**
		PR(·,0.12)%	94.10 ± 20.20	94.55 ± 20.31	93.93 ± 21.38	93.24 ± 23.08	94.54 ± 20.28	95.4 ± 17.90	**96.05 ± 16.18**
		PR(·,0.15)%	89.78 ± 26.53	91.22 ± 25.70	90.35 ± 26.91	89.76 ± 28.14	90.86 ± 26.23	92.71 ± 23.54	**93.21 ± 21.47**
		Time (s/ep.)	54	61	44	64	64	97	184
	WRN-50-2	PR(·,0.08)%	97.86 ± 12.61	97.55 ± 13.14	97.40 ± 13.61	97.75 ± 12.43	97.40 ± 13.91	97.96 ± 10.82	**98.09 ± 10.81**
		PR(·,0.10)%	97.08 ± 14.54	96.14 ± 16.70	96.15 ± 16.78	96.42 ± 16.11	96.20 ± 17.58	97.22 ± 12.96	**97.33 ± 13.03**
		PR(·,0.12)%	95.84 ± 17.14	94.33 ± 20.41	94.41 ± 20.37	94.64 ± 20.01	94.95 ± 19.88	**96.22 ± 15.62**	**96.22 ± 15.58**
		PR(·,0.15)%	92.62 ± 22.87	91.01 ± 25.79	91.11 ± 25.65	91.22 ± 25.70	91.48 ± 25.69	**94.74 ± 19.88**	93.56 ± 20.89
		Time (s/ep.)	131	105	71	155	157	229	305
CIFAR-100	ResNet-18	PR(·,0.08)%	95.40 ± 18.15	95.02 ± 19.27	95.45 ± 18.11	96.20 ± 16.07	96.16 ± 16.36	97.09 ± 13.48	**97.59 ± 12.96**
		PR(·,0.10)%	92.70 ± 22.72	92.09 ± 24.22	92.73 ± 23.39	94.40 ± 20.08	93.81 ± 21.42	94.67 ± 18.91	**95.96 ± 16.62**
		PR(·,0.12)%	88.98 ± 28.19	87.80 ± 29.80	89.35 ± 28.32	92.05 ± 24.12	91.05 ± 25.86	92.63 ± 23.91	**93.15 ± 21.77**
		PR(·,0.15)%	80.47 ± 36.40	79.20 ± 37.73	81.93 ± 35.72	85.89 ± 31.65	84.97 ± 33.00	85.87 ± 31.99	**87.38 ± 29.63**
		Time (s/ep.)	55	61	43	64	65	97	180
	WRN-50-2	PR(·,0.08)%	95.31 ± 17.38	95.24 ± 18.61	95.30 ± 16.87	95.34 ± 17.79	95.29 ± 17.70	**96.52 ± 15.11**	96.05 ± 16.17
		PR(·,0.10)%	92.00 ± 23.79	92.30 ± 23.61	92.59 ± 22.81	92.36 ± 23.19	91.79 ± 23.87	93.49 ± 20.93	**93.72 ± 20.47**
		PR(·,0.12)%	88.14 ± 29.15	88.65 ± 28.46	88.38 ± 28.65	88.51 ± 28.22	87.54 ± 29.92	88.82 ± 28.38	**89.48 ± 26.87**
		PR(·,0.15)%	80.67 ± 36.21	80.35 ± 36.54	80.79 ± 36.01	80.37 ± 36.86	79.71 ± 36.76	81.02 ± 37.09	**81.60 ± 34.86**
		Time (s/ep.)	131	102	70	155	156	225	312
SVHN	ResNet-18	PR(·,0.08)%	98.96 ± 5.85	99.03 ± 5.76	98.99 ± 6.15	99.05 ± 5.73	99.12 ± 5.33	98.90 ± 5.97	**99.12 ± 5.15**
		PR(·,0.10)%	98.50 ± 7.40	98.46 ± 8.27	98.31 ± 8.20	98.63 ± 7.15	98.60 ± 7.21	98.53 ± 7.24	**98.87 ± 5.93**
		PR(·,0.12)%	97.72 ± 9.61	97.30 ± 11.08	97.10 ± 11.26	97.85 ± 9.13	97.66 ± 9.78	98.11 ± 8.19	**98.54 ± 6.84**
		PR(·,0.15)%	95.41 ± 14.41	94.15 ± 16.57	94.12 ± 16.65	95.54 ± 14.10	94.93 ± 15.00	97.10 ± 10.05	**97.84 ± 8.60**
		Time (s/ep.)	78	86	61	91	93	141	202
	WRN-50-2	PR(·,0.08)%	98.72 ± 6.44	98.69 ± 6.90	98.61 ± 7.66	98.54 ± 7.52	98.59 ± 7.46	98.43 ± 6.69	**98.88 ± 6.12**
		PR(·,0.10)%	98.28 ± 7.66	98.11 ± 8.66	98.07 ± 9.27	98.12 ± 8.74	98.00 ± 8.89	98.15 ± 7.29	**98.38 ± 7.53**
		PR(·,0.12)%	97.76 ± 8.95	97.26 ± 10.89	97.34 ± 10.90	97.65 ± 8.85	96.98 ± 11.13	**97.79 ± 8.12**	97.72 ± 9.19
		PR(·,0.15)%	96.77 ± 11.18	95.20 ± 14.84	95.45 ± 14.72	95.67 ± 14.32	94.74 ± 15.37	**96.94 ± 9.80**	96.26 ± 12.35
		Time (s/ep.)	190	146	102	223	225	321	420

Key Conclusions:

- **Table 1.** AT-PR achieves larger improvements in PR compared to AT-WCR, while incurring a smaller trade-off in model generalization.
- **Table 2.** Combining AT-PR with TRADES improves both clean accuracy and PR on ViT.
- **Table 3.** AT-PR is compared with four advanced AT methods (TRADES, MART, ALP, and CLP), as well as with Fast-AT-PR (where PGD is replaced with Fast-AT.

WMG
THE UNIVERSITY OF WARWICK

Thank you…

Yi Zhang

Yi.Zhang.16@warwick.ac.uk

WMG
THE UNIVERSITY OF WARWICK