DAT: Improving Adversarial Robustness via Generative Amplitude Mix-up in Frequency Domain

Fengpeng Li[1], Kemou Li[1], Haiwei Wu[2], Jinyu Tian[3], and Jiantao Zhou[1⋆]

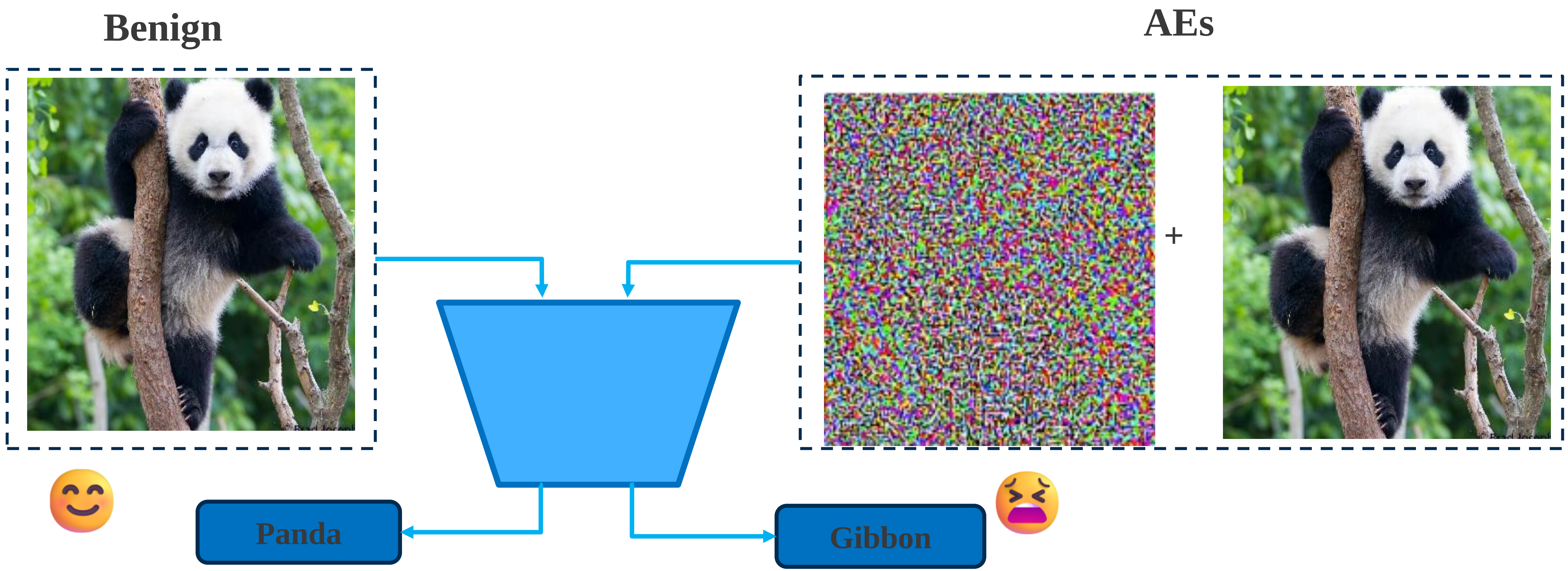
- [1] State Key Laboratory of Internet of Things for Smart City, University of Macau
- [2] Department of Computer Science, City University of Hong Kong
- [3] Faculty of Innovation Engineering, Macau University of Science and Technology

Outline

- Background
- Motivation
- Dual Adversarial Training
- Experiments

Adversarial Attacks

- Adversarial Attacks generates Adversarial Examples (AEs) by adding subtle yet deceptive adversarial perturbations to benign samples.



Motivation

The adversarial perturbation severely damages phase patterns (especially in red rectangular) and the frequency spectrum, while amplitude patterns are rarely impacted.

DFT

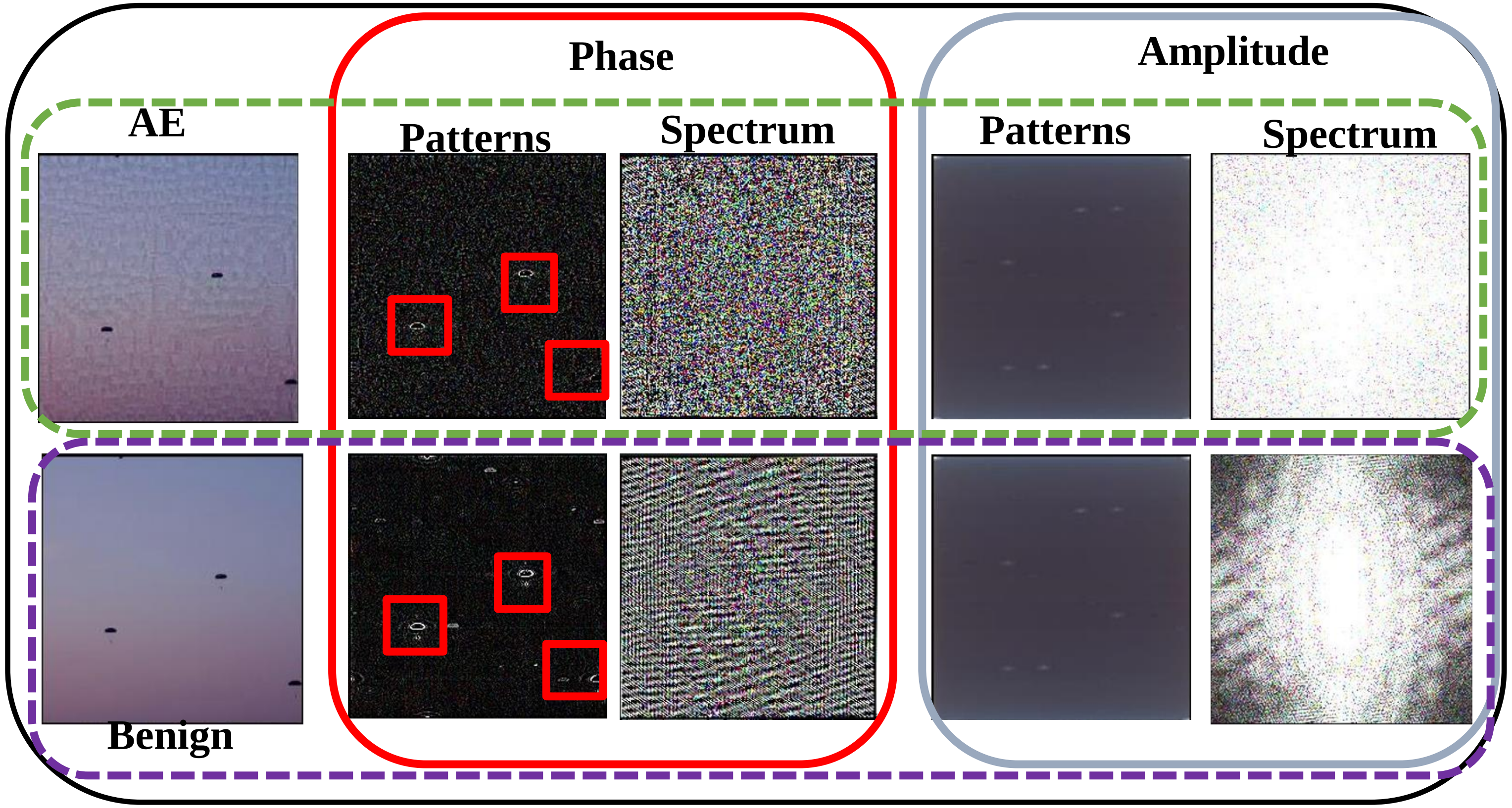
$$\mathcal{F}(\mathbf{x})(u,v)=\sum_{h=1}^{H}\sum_{w=1}^{W}\mathbf{x}(h,w)\,\mathrm{e}^{-\mathrm{i}2\pi\left(u\frac{h}{H}+v\frac{w}{W}\right)}$$

Amplitude

$$\mathcal{A}(\mathbf{x})=\left(\mathrm{Re}^2(\mathcal{F}(\mathbf{x}))+\mathrm{Im}^2(\mathcal{F}(\mathbf{x}))\right)^{\frac{1}{2}},$$

Phase

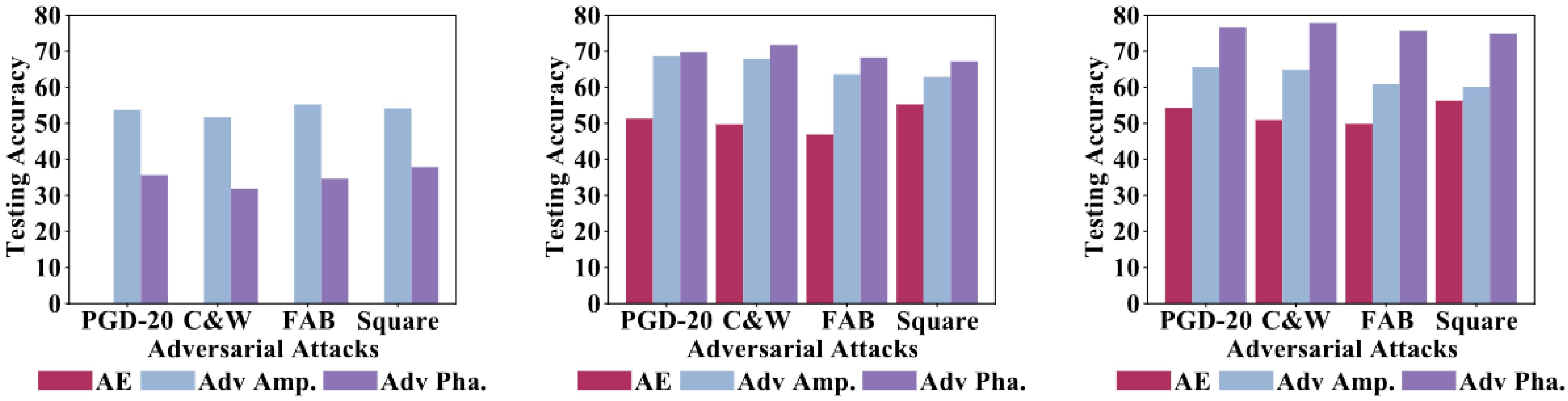
$$\mathcal{P}(\mathbf{x})=\arctan\left(\frac{\mathrm{Im}(\mathcal{F}(\mathbf{x}))}{\mathrm{Re}(\mathcal{F}(\mathbf{x}))}\right)$$



Dual Adversarial Training(DAT)

Motivation:

$$\mathbf{x}'_{amp} = \mathcal{F}^{-1}(\mathcal{A}(\mathbf{x}'), \mathcal{P}(\mathbf{x})), \quad \mathbf{x}'_{pha} = \mathcal{F}^{-1}(\mathcal{A}(\mathbf{x}), \mathcal{P}(\mathbf{x}'))$$



(a) Standard model (b) Robust model (c) Perturbed model

The robust and perturbed models are trained by PGD-AT-10.

Conclusion:

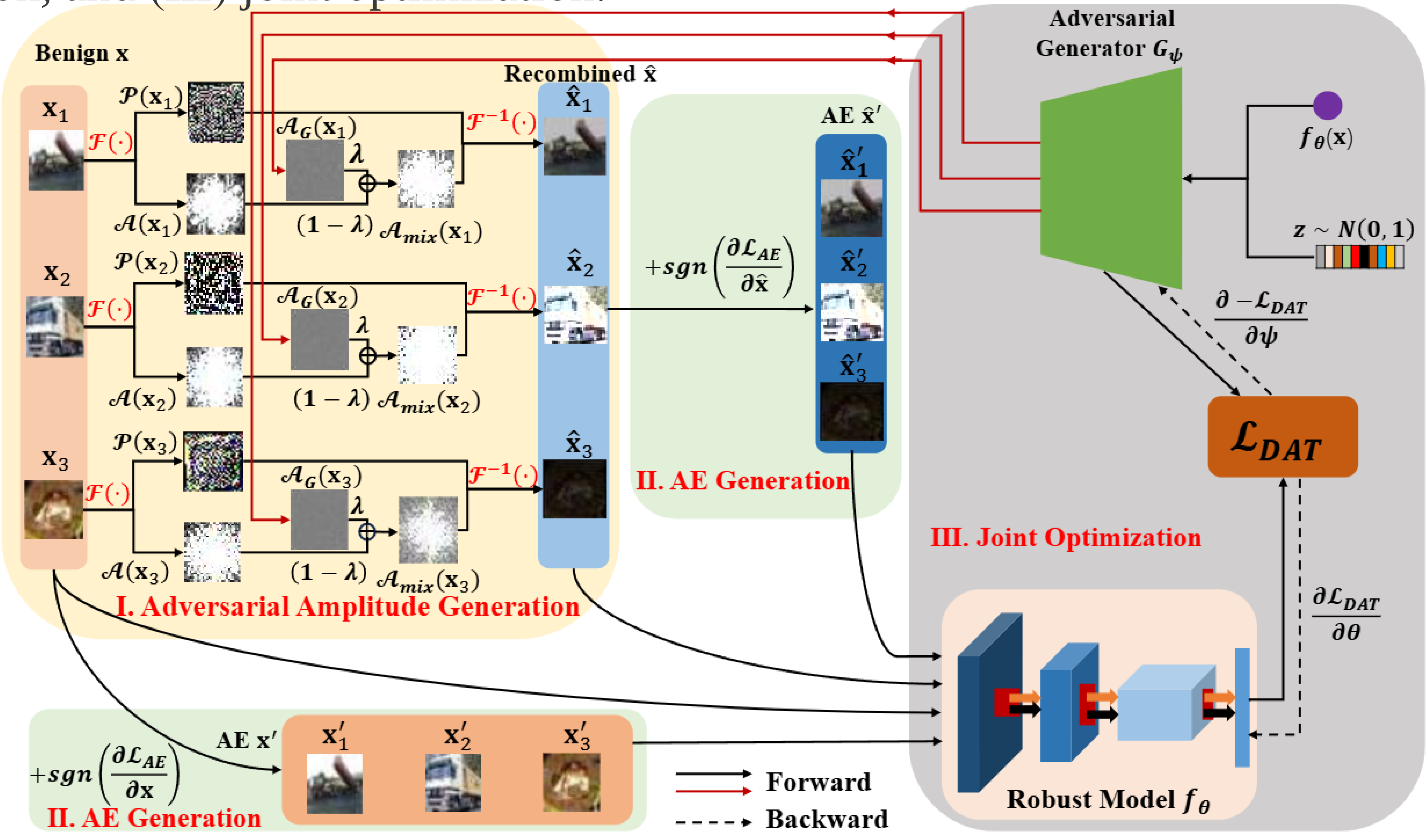
1.fig(a) shows phase patterns are severely damaged.

2.fig(b)Some phase patterns are still unaffected by adversarial perturbations.

3.fig(c)Perturbing the amplitude can force the model to focus on phase patterns.

Dual Adversarial Training(DAT)

The overview of DAT, which consists of three stages:(I) adversarial amplitude generation, (II) AE generation, and (III) joint optimization.



Dual Adversarial Training(DAT)

Adversarial Amplitude Generator

- **C1.** $|h_p(\mathbf{x}) - h_p(\hat{\mathbf{x}})| < \epsilon_1$: Ensuring $\hat{\mathbf{x}}$ retains the same semantics in the phase spectrum as $\mathbf{x}$.
- **C2.** $F_{\boldsymbol{\theta}}(\mathbf{x}) = F_{\boldsymbol{\theta}}(\hat{\mathbf{x}})$: Ensuring $\hat{\mathbf{x}}$ remains distinguishable with the same label as $\mathbf{x}$ by $f_{\boldsymbol{\theta}}$.
- **C3.** $|h_a(\mathbf{x}) - h_a(\hat{\mathbf{x}})| > \epsilon_2$: Making $\hat{\mathbf{x}}$ maximize the $\mathcal{L}_{\mathsf{DAT}}$, causing the model's difficulty fitting the amplitude of images, and forcing the model to focus on phase patterns.

$$\mathcal{A}_G(\mathbf{x}) = G_{\boldsymbol{\psi}}(\mathbf{z}, f_{\boldsymbol{\theta}}(\mathbf{x})), \quad \text{where } \mathbf{z} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{I}).$$

Ensuring that a portion of the original amplitude information is preserved following:

$$\mathcal{A}_{mix}(\mathbf{x}) = \lambda \cdot \mathcal{A}_G(\mathbf{x}) + (1 - \lambda) \cdot \mathcal{A}(\mathbf{x}), \quad \text{where } \lambda \sim \text{Uniform}(0, 1).$$

The recombined $\hat{\mathbf{x}}$ is obtained by IDFT:

$$\hat{\mathbf{x}} = \mathcal{F}^{-1}(\mathcal{A}_{mix}(\mathbf{x}), \mathcal{P}(\mathbf{x})).$$

Dual Adversarial Training(DAT)

Efficient AE Generation

Issues: reducing t difficulty of AEs' reaching the actual maximum in the $\ell_\infty - ball$. Generally, t=10, doubling the training time with vanilla-AT.

$$\min_{\boldsymbol{\theta}} \max_{\mathbf{x}' \in \mathcal{B}_\epsilon[\mathbf{x}]} \mathcal{L}_{\mathrm{CE}}(f(\mathbf{x}'), y) \quad \mathbf{x}'^{(t+1)} = \prod_{\mathcal{B}_\epsilon[\mathbf{x}]} (\mathbf{x}'^{(t)} + \alpha \cdot sign(\nabla_{\mathbf{x}'^{(t)}} \mathcal{L}(f(\mathbf{x}'^{(t)}), y)))$$

Solution: increase adversarial perturbation length in each iteration without change α.

$$\mathcal{L}_{\mathsf{AE}}(f_{\boldsymbol{\theta}}(\mathbf{x}), f_{\boldsymbol{\theta}}(\mathbf{x}'), y) = \mathcal{L}_{\mathsf{CE}}(f_{\boldsymbol{\theta}}(\mathbf{x}'), y) + \beta \cdot \mathcal{D}_{\mathsf{KL}}(f_{\boldsymbol{\theta}}(\mathbf{x}'), f_{\boldsymbol{\theta}}(\mathbf{x})),$$

Dual Adversarial Training(DAT)

Joint Optimization

Optimization objective for $f_{\boldsymbol{\theta}}$ and $G_{\boldsymbol{\psi}}$

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}} \left[\max_{\boldsymbol{\psi}} \mathbb{E}_{\hat{\mathbf{x}}\sim p(\hat{\mathbf{x}}|\mathbf{x},\boldsymbol{\psi})} \left[\mathcal{L}_{\mathsf{DAT}}(f_{\boldsymbol{\theta}}(\mathbf{x}), f_{\boldsymbol{\theta}}(\hat{\mathbf{x}}), y) \right] \right],$$

$\hat{\mathbf{x}}$ follows a sample-dependent conditional distribution $p(\hat{\mathbf{x}}|\mathbf{x},\boldsymbol{\psi})$

Total loss $\mathcal{L}_{\text{DAT}}$

$$\mathcal{L}_{\mathsf{DAT}}(f_{\boldsymbol{\theta}}(\mathbf{x}), f_{\boldsymbol{\theta}}(\hat{\mathbf{x}}), y) = \frac{1}{2}(\mathcal{L}_{\mathsf{AT}}(f_{\boldsymbol{\theta}}(\mathbf{x}), y) + \mathcal{L}_{\mathsf{AT}}(f_{\boldsymbol{\theta}}(\hat{\mathbf{x}}), y)) + \omega \cdot \mathcal{D}_{\mathsf{JS}}(f_{\boldsymbol{\theta}}(\mathbf{x}), f_{\boldsymbol{\theta}}(\hat{\mathbf{x}})),$$

Adversarial Training Loss $\mathcal{L}_{\text{AT}}$

Consistency Regularization Loss $\mathcal{D}_{\text{JS}}$

Experiments

Settings Training: $\boldsymbol{\epsilon = \frac{8}{255}, \alpha = \frac{2}{255}, t = 5}$ Testing: $\boldsymbol{\epsilon = \frac{8}{255}}$

Baselines

Common methods: PGD-AT, TRADES, MART, ST, SCARL,LAS-AT

Complex methods: OA-AT, DAJAT, IDBH

Backbones: ResNet-18, WideResNet-34-10, WideResNet-28-10

Average natural and robust accuracy (%) of ResNet-18 on CIFAR-10

DATASET	METHOD	Natural	FGSM	PGD-20	PGD-100	C&W$_\infty$	AA
CIFAR-10	PGD-AT [40]	82.78±0.12	56.94±0.17	51.30±0.16	50.88±0.26	49.72±0.24	47.63±0.08
	TRADES [61]	82.41±0.12	58.47±0.19	52.76±0.08	52.47±0.13	50.43±0.17	49.37±0.08
	MART [54]	80.70±0.17	58.91±0.24	54.02±0.29	53.38±0.30	49.35±0.27	47.49±0.23
	ST [37]	83.10±0.10	59.42±0.32	54.53±0.14	54.31±0.17	51.35±0.21	50.51±0.17
	SCARL [33]	80.67±0.31	58.32±0.13	54.24±0.17	54.10±0.13	51.93±0.15	50.45±0.11
	DAT (Ours)	**84.17±0.21**	**62.06±0.19**	**57.55±0.15**	**57.47±0.17**	**52.59±0.13**	**51.36±0.14**
	TRADES+AWP	81.16±0.12	57.86±0.14	54.56±0.06	54.45±0.14	50.95±0.12	50.31±0.10
	SCARL+AWP	81.46±0.15	59.26±0.16	55.38±0.14	55.27±0.13	52.15±0.15	51.08±0.11
	DAT+AWP (Ours)	**82.63±0.15**	**62.78±0.21**	**58.87±0.12**	**58.78±0.15**	**52.88±0.21**	**52.54±0.12**

Experiments

Average natural and robust accuracy (%) of ResNet-18 on CIFAR-100 and Tiny-ImageNet

DATASET	METHOD	Natural	FGSM	PGD-20	PGD-100	C&W$_\infty$	AA
CIFAR-100	PGD-AT [40]	57.27±0.21	31.81±0.11	28.66±0.11	28.49±0.16	26.89±0.08	24.60±0.04
	TRADES [61]	57.94±0.15	32.37±0.18	29.25±0.18	29.10±0.20	25.88±0.16	24.71±0.04
	MART [54]	55.03±0.10	33.12±0.26	30.32±0.18	30.20±0.17	26.60±0.11	25.13±0.15
	ST [37]	58.44±0.12	33.35±0.23	30.53±0.13	30.39±0.17	26.70±0.20	25.61±0.07
	SCARL [33]	57.63±0.11	33.14±0.19	30.83±0.24	30.77±0.21	26.86±0.16	25.82±0.19
	DAT (Ours)	**62.57±0.17**	**36.63±0.12**	**33.37±0.15**	**33.15±0.12**	**28.34±0.14**	**27.11±0.15**
	TRADES+AWP	58.76±0.07	33.82±0.15	31.53±0.14	31.42±0.12	27.03±0.16	26.06±0.12
	SCARL+AWP	58.36±0.12	34.25±0.14	32.32±0.14	32.26±0.13	27.92±0.11	26.83±0.15
	DAT+AWP (Ours)	**63.28±0.11**	**38.22±0.14**	**35.29±0.13**	**35.18±0.12**	**29.43±0.17**	**28.09±0.12**
Tiny ImageNet	PGD-AT [40]	46.36±0.22	23.49±0.39	20.41±0.29	20.35±0.37	17.86±0.28	14.46±0.31
	TRADES [61]	43.65±0.35	21.37±0.48	18.62±0.48	18.56±0.33	15.38±0.35	13.32±0.41
	LAS-AT [29]	45.27±0.35	24.64±0.24	21.82±0.27	21.72±0.23	18.07±0.25	16.25±0.22
	SCARL [33]	49.75±0.17	25.52±0.16	22.64±0.11	22.58±0.18	18.77±0.27	16.31±0.14
	DAT (Ours)	**52.45±0.21**	**28.45±0.15**	**25.47±0.12**	**25.36±0.14**	**20.39±0.17**	**17.51±0.19**
	TRADES+AWP	46.64±0.35	26.58±0.19	22.31±0.20	22.28±0.12	17.84±0.11	15.34±0.12
	LAS-AT+AWP	46.85±0.13	25.76±0.12	23.30±0.11	23.05±0.15	19.68±0.11	17.98±0.15
	DAT+AWP (Ours)	**53.29±0.25**	**30.91±0.11**	**27.25±0.13**	**27.18±0.16**	**22.12±0.12**	**19.29±0.13**

Experiments

Average natural and robust accuracy (%) of WideResNet34-10 on CIFAR-10 and CIFAR-100

Method	CIFAR-10				CIFAR-100			
	Natural	PGD-100	C&W$_\infty$	AA	Natural	PGD-100	C&W$_\infty$	AA
PGD-AT [40]	85.37±0.74	54.61±0.68	53.42±0.82	52.03±0.68	60.63±1.17	30.83±0.51	30.21±0.83	27.93±0.57
TRADES [61]	85.54±0.59	56.04±0.45	53.91±0.46	53.37±0.51	61.26±0.39	33.11±0.42	30.24±0.58	28.32±0.62
MART [54]	85.13±0.52	58.72±0.66	53.02±0.37	51.61±0.48	60.52±0.62	32.34±0.62	29.07±0.43	25.91±0.36
LAS-AT [29]	86.07±0.31	55.97±0.47	55.49±0.54	53.34±0.42	61.87±0.57	32.21±0.45	30.47±0.34	28.91±0.39
SCARL [33]	84.41±0.23	57.81±0.65	56.21±0.47	54.37±0.29	62.41±0.36	34.19±0.46	30.53±0.31	29.52±0.33
DAT (Ours)	**86.78±0.42**	**61.32±0.24**	**57.62±0.34**	**56.46±0.33**	**64.53±0.25**	**36.75±0.43**	**32.21±0.27**	**30.79±0.17**

Average natural and robust accuracy (%) of Complex Methods on CIFAR-10 and CIFAR-100

Method	ResNet-18				WRN-34-10			
	CIFAR-10		CIFAR-100		CIFAR-10		CIFAR-100	
	PGD-20	AA	PGD-20	AA	PGD-20	AA	PGD-20	AA
TRADES+AWP	54.56±0.06	50.31±0.10	31.53±0.14	26.06±0.12	59.26±0.24	55.28±0.21	34.48±0.26	29.74±0.21
TRADES+AWP+SWA	55.21±0.24	51.14±0.13	31.72±0.23	26.21±0.15	60.25±0.26	55.37±0.15	35.16±0.23	29.92±0.16
OA-AT (SWA+variable ϵ and α) [2]	56.47±0.37	50.83±0.24	32.63±0.25	26.84±0.36	60.49±0.31	57.91±0.18	36.18±0.27	30.35±0.23
DAJAT (AWP+SWA+variable ϵ&α) [1]	56.52±0.47	51.85±0.26	32.96±0.32	27.83±0.29	62.34±0.35	56.62±0.23	37.05±0.14	31.51±0.17
IDBH (AWP+SWA+variable ϵ) [35]	57.48±0.34	52.31±0.26	33.67±0.27	27.86±0.32	62.47±0.23	57.64±0.26	36.46±0.23	31.34±0.22
DAT+AWP (Ours)	**58.57±0.14**	**52.54±0.12**	**35.29±0.13**	**28.09±0.12**	**63.34±0.18**	**57.96±0.16**	**38.41±0.17**	**31.62±0.12**
DAT+AWP+SWA (Ours)	**58.84±0.16**	**52.76±0.14**	**35.47±0.11**	**28.31±0.13**	**63.65±0.19**	**58.12±0.18**	**38.59±0.16**	**31.81±0.12**

Experiments

The average experimental results for different augmentations on CIFAR-10 and CIFAR-100 with ResNet-18

METHOD	CIFAR-10		CIFAR-100	
	PGD-20	AA	PGD-20	AA
Baseline	53.13±0.51	49.64±0.62	30.09±0.58	25.43±0.39
CutOut [18]	55.85±0.51	50.28±0.14	31.35±0.44	26.26±0.14
CutMix [60]	55.76±0.42	50.13±0.54	31.26±0.62	26.17±0.19
AutoAugment [14]	56.24±0.45	50.42±0.15	31.69±0.52	26.44±0.17
DAT (Ours)	**57.55±0.15**	**51.36±0.14**	**33.37±0.15**	**27.11±0.15**

Time consumption (s) of each training epoch for different AT methods on ResNet-18

METHOD	CIFAR-10	CIFAR-100
PGD-AT [40]	187	188
TRADES [61]	187	192
ST [37]	320	326
SCARL [33]	221	228
DAT (Ours)	218	221

Thanks