PDE+: Enhancing Generalization via PDE with Adaptive Distributional Diffusion

Paper Code

Yige Yuan, Bingbing Xu, Bo Lin, Liang Hou, Fei Sun, Huawei Shen, Xueqi Cheng

CAS Key Laboratory of AI Safety and Security, Institute of Computing Technology, Chinese Academy of Sciences

Department of Mathematics, National University of Singapore

Homepage WeChat

Email: yuanyige20z@ict.ac.cn

INTRODUCTION

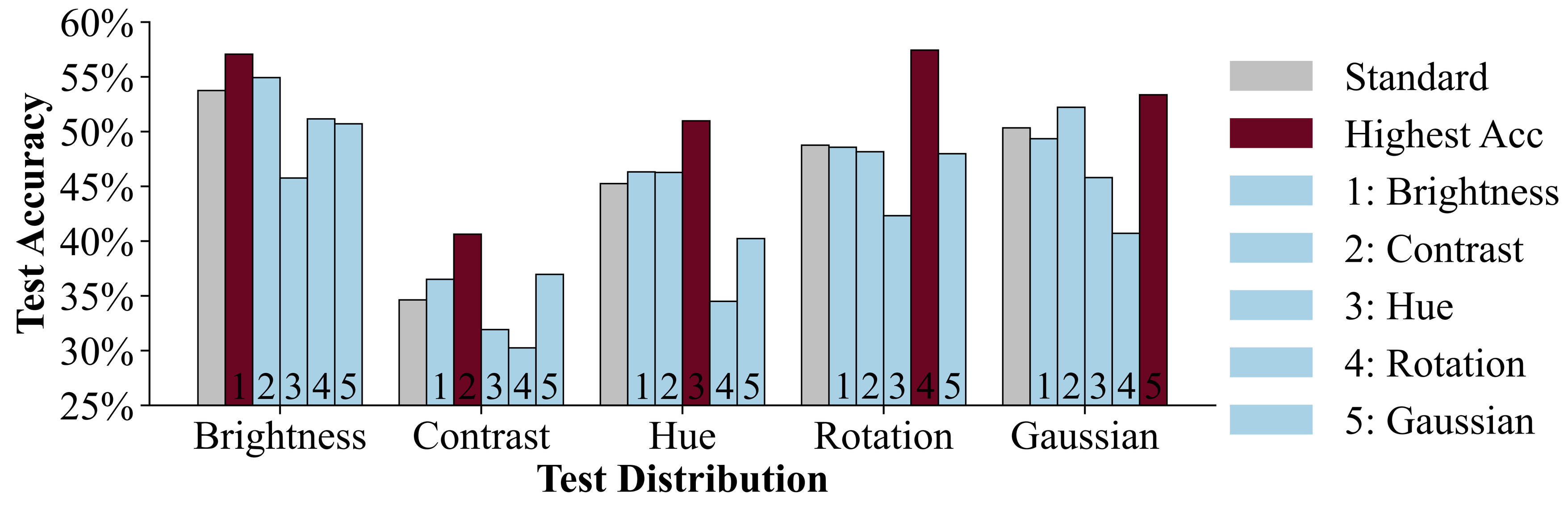
Objective: Enhancing the **generalization** of neural networks especially under distributions that differ from the training distribution.

Weakness of existing methods: Current methods, mainly based on the data-driven paradigm such as data augmentation, adversarial training, and noise injection, may encounter limited generalization due to model **non-smoothness**.

Motivation: Investigating generalization from a **PDE perspective**, aiming to enhance it directly through the **underlying function** of neural networks.

Models can only achieve satisfactory generalization performance when the training data is subjected to augmentation similar to that of the testing data.

Observation



METHOD

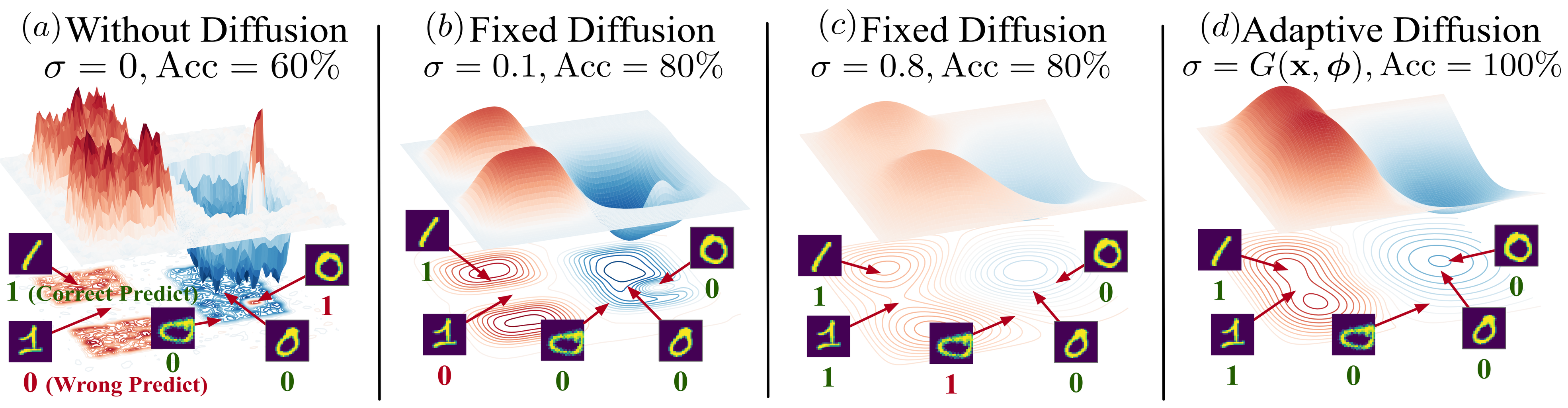
From PDE to Neural Network

① Neural Network as the Solution of PDE

$$\frac{\partial u}{\partial t}(\mathbf{x},t)+F(\mathbf{x},\boldsymbol{\theta}(t))\cdot\nabla u(\mathbf{x},t)=0 \quad\Rightarrow\quad \begin{aligned}&\mathbf{h}_{l+1}=f(\mathbf{h}_l,\boldsymbol{\theta}_l)+\mathbf{h}_l\\ &u(\hat{\mathbf{x}},0)=o\Big(\hat{\mathbf{x}}+\sum_{l=1}^{L}f(\mathbf{h}_l,\boldsymbol{\theta}_l)\Big)\end{aligned}$$

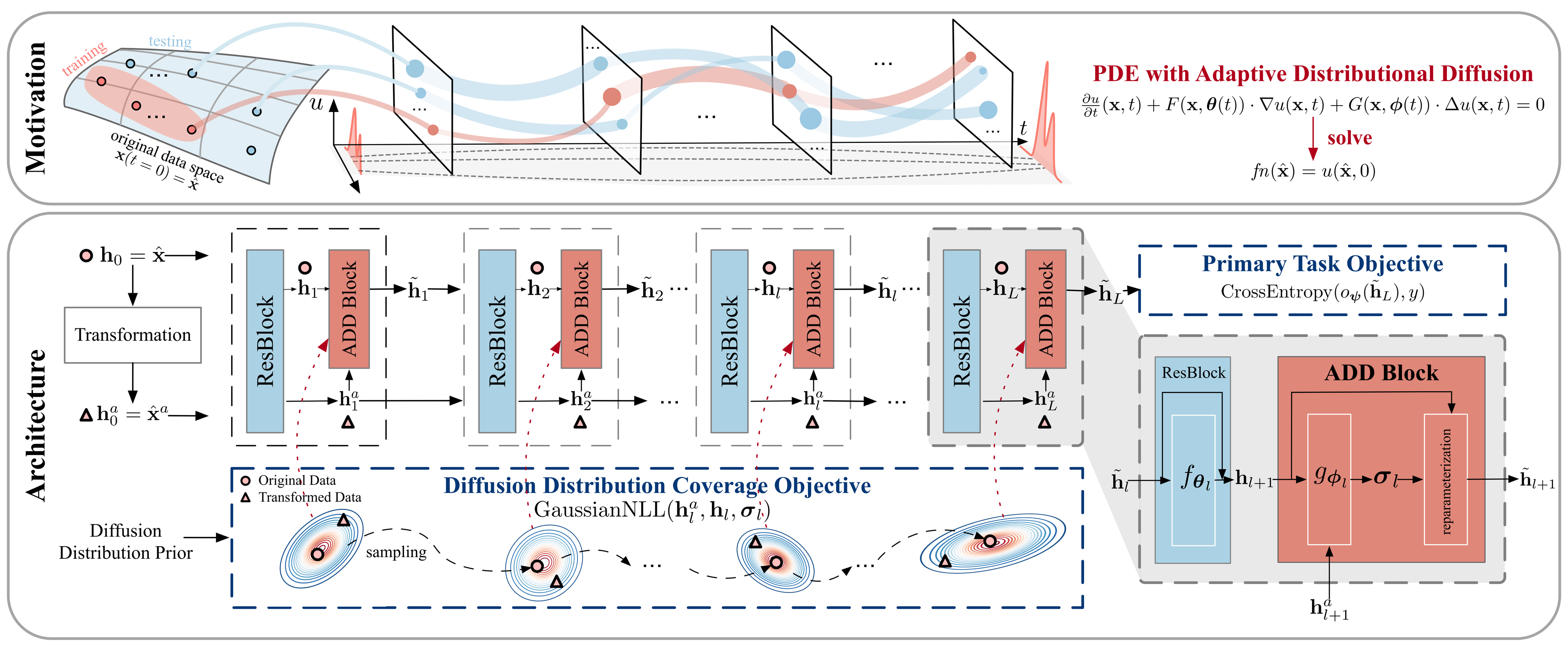
② Adaptive Distributional Diffusion for Generalization

$$\frac{\partial u}{\partial t}(\mathbf{x},t)+F(\mathbf{x},\boldsymbol{\theta}(t))\cdot\nabla u(\mathbf{x},t)+\frac{1}{2}G(\mathbf{x},\boldsymbol{\phi}(t))^2\cdot\Delta u(\mathbf{x},t)=0$$



③ Deriving Neural Network from PDE with ADD

$$\begin{aligned}&u(\hat{\mathbf{x}},0)=\mathbb{E}\left[o(\mathbf{x}(1))\mid\mathbf{x}(0)=\hat{\mathbf{x}}\right]\\ &\mathrm{d}\mathbf{x}(t)=F(\mathbf{x}(t),\boldsymbol{\theta}(t))\,\mathrm{d}t+G(\mathbf{x}(t),\boldsymbol{\phi}(t))\cdot\mathrm{d}B_t\end{aligned} \quad\Rightarrow\quad \begin{aligned}&u(\hat{\mathbf{x}},0)=\mathbb{E}\left[o(\mathbf{h}_L)\mid\mathbf{h}_0=\hat{\mathbf{x}}\right]\\ &\mathbf{h}_{l+1}=\mathbf{h}_l+f(\mathbf{h}_l,\boldsymbol{\theta}_l)+g(\mathbf{h}_l,\boldsymbol{\phi}_l)\cdot\mathcal{N}(\mathbf{0},\mathbf{I})\end{aligned}$$



Network Instantiation

Architecture and Parameterization

$$\mathbf{h}_{l+1}=\mathbf{h}_l+f(\mathbf{h}_l,\boldsymbol{\theta}_l)\qquad \boldsymbol{\sigma}_{l+1}=g_{\boldsymbol{\phi}_{l+1}}(\mathbf{h}_{l+1})\qquad \tilde{\mathbf{h}}_{l+1}=\mathbf{h}_{l+1}+\boldsymbol{\sigma}_{l+1}\cdot\mathcal{N}(\mathbf{0},\mathbf{I})$$

$$\text{PDE+}_{\boldsymbol{\theta},\boldsymbol{\phi}}:(g_{\boldsymbol{\phi}_l}\circ(f_{\boldsymbol{\theta}_{l-1}}+I)\circ\cdots\circ g_{\boldsymbol{\phi}_3}\circ(f_{\boldsymbol{\theta}_2}+I)\circ g_{\boldsymbol{\phi}_2}\circ(f_{\boldsymbol{\theta}_1}+I))$$

Learning Objective

① Diffusion Distribution Coverage Objective

$$\min_{\boldsymbol{\phi}}\ \mathop{\mathbb{E}}_{\mathbf{x}\sim S_N}-\sum_{l=1}^{L}\log p_{\boldsymbol{\phi}_l}(\mathbf{h}_l^a\mid\mathbf{h}_l)=-\frac{1}{2N}\sum_{n=1}^{N}\sum_{l=1}^{L}\left[\log g_{\boldsymbol{\phi}_l}(\mathbf{h}_l)+\frac{(\mathbf{h}_{n,l}^a-\mathbf{h}_{n,l})^2}{g_{\boldsymbol{\phi}_l}(\mathbf{h}_l)}\right]$$

② Primary Task Objective

$$\min_{\boldsymbol{\theta},\boldsymbol{\phi},\boldsymbol{\psi}}\ \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim S_N}-\log p_{\boldsymbol{\theta},\boldsymbol{\phi},\boldsymbol{\psi}}(y\mid\mathbf{x})=-\frac{1}{N}\sum_{n=1}^{N}\left[\log\frac{\exp(o_{\boldsymbol{\psi}}(\tilde{\mathbf{h}}_{n,L})_{y_n})}{\sum_{c=1}^{C}\exp(o_{\boldsymbol{\psi}}(\tilde{\mathbf{h}}_{n,L})_c)}\right]_{y_n}$$

EXPERIMENTS

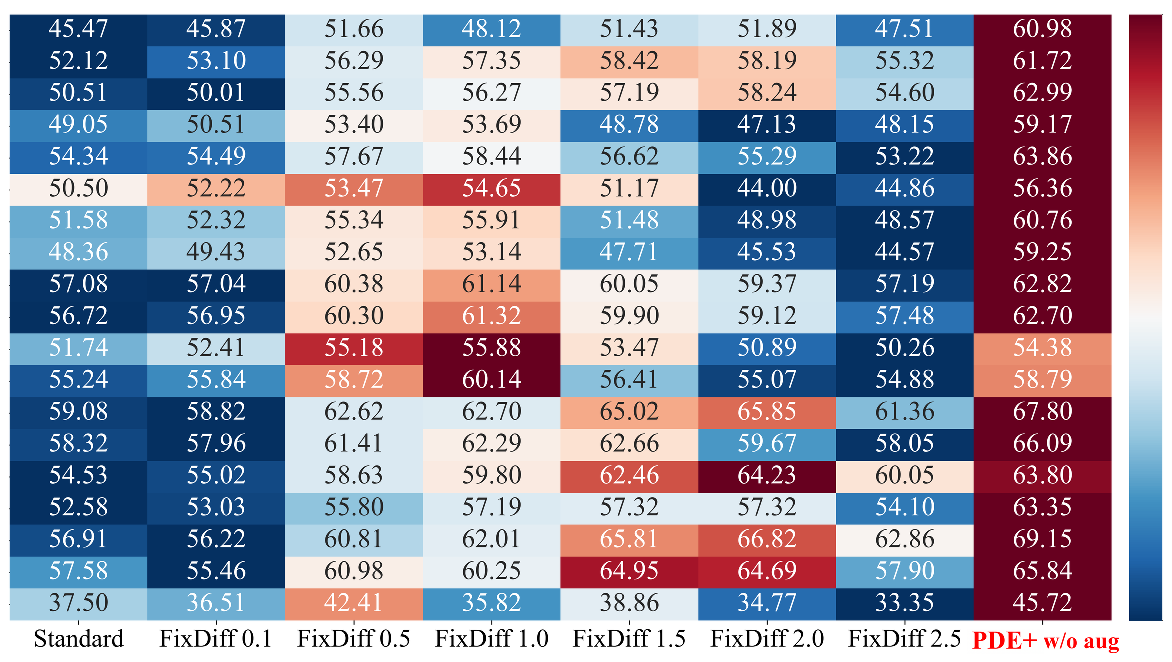
(Q1) Does PDE+ improve generalization compared to SOTA methods on various benchmarks?

(Q2) Does PDE+ learns appropriate diffusion distribution coverage?

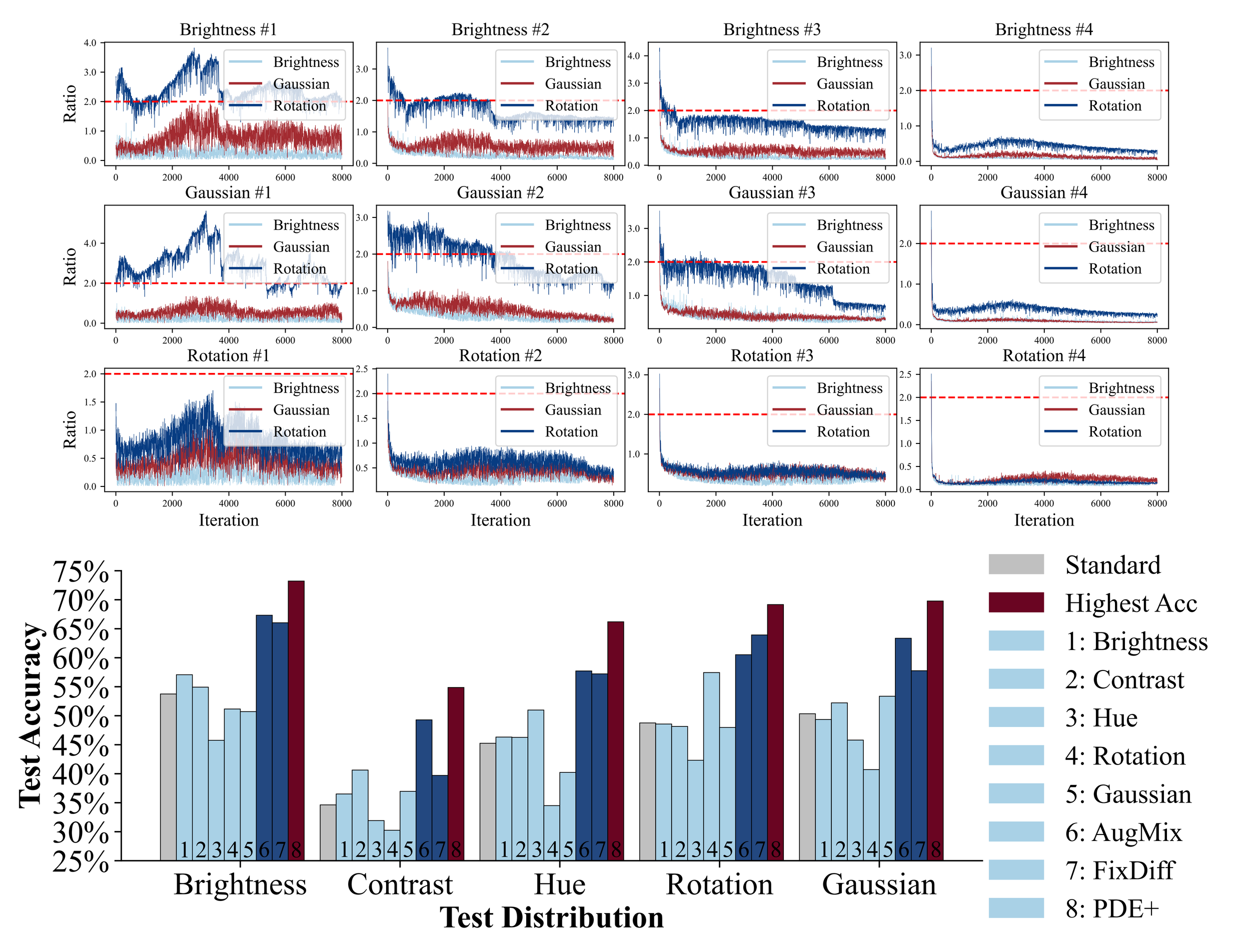
(Q3) Does PDE+ improve generalization beyond observed (training) distributions?

	Method	CIFAR-10(C) Clean					CIFAR-100(C) Clean					Tiny-ImageNet(C) Clean				
		Clean	Corr Severity All		Corr Severity 5		Clean	Corr Severity All		Corr Severity 5		Clean	Corr Severity All		Corr Severity 5	
		Acc (↑)	Acc (↑)	mCE (↓)	Acc (↑)	mCE (↓)	Acc (↑)	Acc (↑)	mCE (↓)	Acc (↑)	mCE (↓)	Acc (↑)	Acc (↑)	mCE (↓)	Acc (↑)	mCE (↓)
Std	ERM	95.35	74.63	100.00	57.19	100.00	77.71	49.27	100.00	33.18	100.00	54.02	25.57	100.00	15.54	100.00
Lip	GradReg	93.64	77.62	96.29	62.33	91.52	73.80	52.16	96.95	37.33	94.49	52.01	29.20	95.13	19.91	94.86
NI	EnResNet	83.33	74.34	137.98	66.87	63.72	67.11	49.28	103.61	40.24	83.56	49.26	25.83	100.18	19.01	96.55
	RSE	95.59	77.86	94.12	63.66	89.08	77.98	53.73	94.10	38.03	92.88	53.74	27.99	96.81	18.92	96.11
	NFM*	95.40	83.30	-	-	-	79.40	59.70	-	-	-	-	-	-	-	-
DA	Gaussian	92.50	80.46	100.03	68.08	87.22	71.87	54.24	98.34	41.77	89.81	48.89	32.92	90.48	24.57	89.56
	Mixup*	**95.80**	80.40	-	-	-	**79.70**	54.20	-	-	-	-	-	-	-	-
	DeepAug*	94.10	85.33	64.63	77.29	60.05	-	-	-	-	-	**54.90**	-	-	-	-
	AutoAug	95.61	85.37	61.74	75.12	62.07	76.34	58.72	83.12	45.38	82.84	52.63	35.14	87.67	25.36	88.54
	AugMix	95.26	86.24	60.44	76.06	59.96	77.11	61.93	77.51	48.99	77.52	52.82	37.74	84.06	28.66	84.69
AT	PGD_{ℓ_∞}	93.52	82.17	86.53	70.10	78.20	71.78	55.03	93.49	42.04	88.17	49.94	32.54	90.65	23.47	90.63
	PGD_{ℓ_2}	93.91	83.07	81.06	70.97	75.17	72.50	56.09	91.65	42.82	87.33	51.08	33.46	89.37	24.00	89.92
	RLAT	93.23	83.67	80.98	72.73	72.59	71.10	56.54	91.98	44.27	86.24	50.24	33.13	89.83	24.46	89.47
	$\text{RLAT}_{\text{Augmix}}$	94.73	88.28	55.60	80.37	51.56	75.06	62.77	77.38	51.60	74.24	51.29	37.92	83.69	29.05	84.17
Ours	PDE+	95.59	**89.11**	**48.07**	**82.81**	**44.97**	78.84	**65.62**	**69.68**	**54.22**	**69.43**	53.72	**39.41**	**81.80**	**30.32**	**82.68**

(AQ1) PDE+ Outperforms SOTA on Corruptions



(AQ2) PDE+ Learns Appropriate Diffusion



(AQ3) PDE+ Generalizes Beyond Observation