Better Diffusion Models Further Improve Adversarial Training

Zekai Wang*[1], Tianyu Pang*[2], Chao Du[2], Min Lin[2], Weiwei Liu[1], Shuicheng Yan[2]

1. School of Computer Science, Wuhan University, China
2. Sea AI Lab, Singapore

Acknowledgment


Zekai Wang


Tianyu Pang

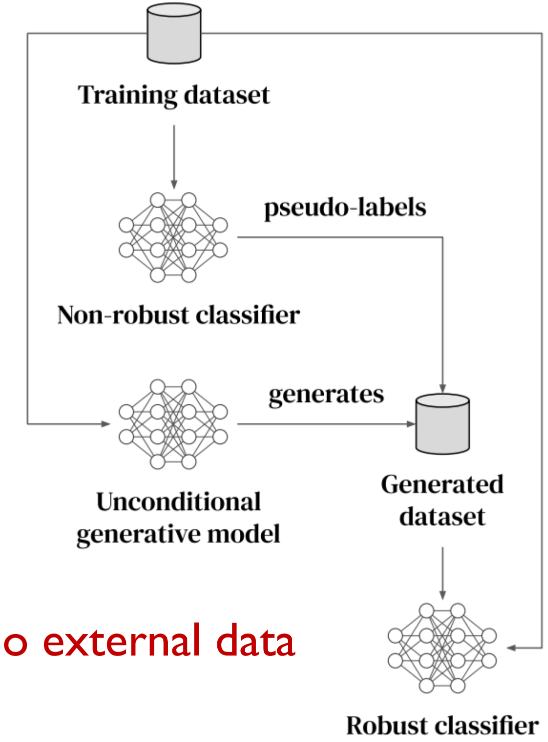

Chao Du


Min Lin

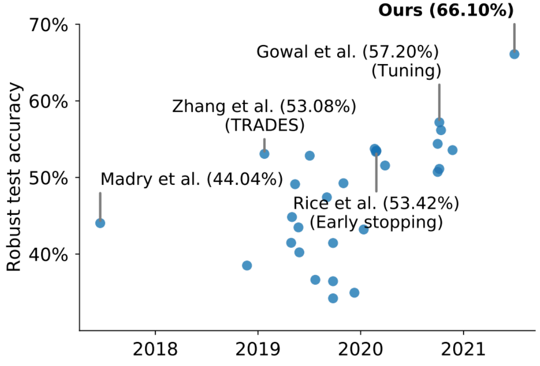

Weiwei Liu

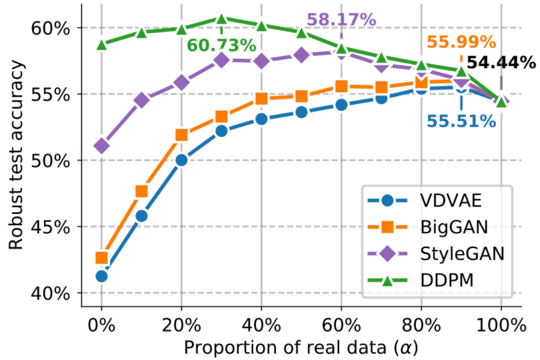

Shuicheng Yan

Diffusion Models for Trustworthy ML



- No external data





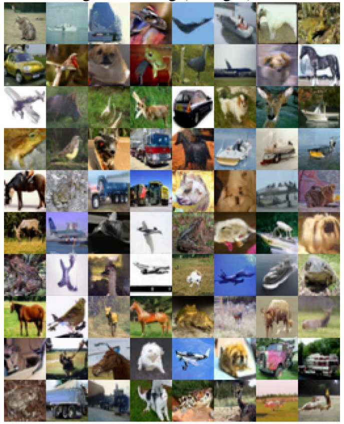
Dominate ROBUSTBENCH A standardized benchmark for adversarial robustness for two years!

[1] Rebuffi et al. Fixing Data Augmentation to Improve Adversarial Robustness. NeurIPS 2021
[2] Gowal et al. Improving Robustness using Generated Data. NeurIPS 2021

Does Lower FID lead to Better Downstream Performance?

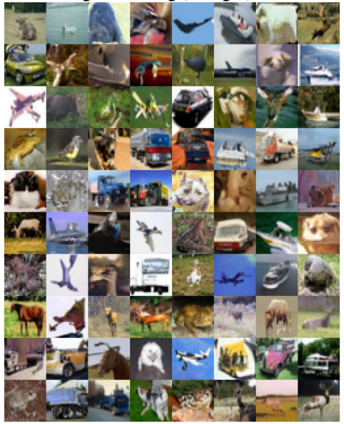
	CIFAR-10 [29] at 32×32				FFHQ [27] 64×64		AFHQv2 [7] 64×64	
	Conditional		Unconditional		Unconditional		Unconditional	
Training configuration	VP	VE	VP	VE	VP	VE	VP	VE
A Baseline [49] (* pre-trained)	2.48	3.11	3.01*	3.77*	3.39	25.95	2.58	18.52
B + Adjust hyperparameters	2.18	2.48	2.51	2.94	3.13	22.53	2.43	23.12
C + Redistribute capacity	2.08	2.52	2.31	2.83	2.78	41.62	2.54	15.04
D + Our preconditioning	2.09	2.64	2.29	3.10	2.94	3.39	2.79	3.81
E + Our loss function	1.88	1.86	2.05	1.99	2.60	2.81	2.29	2.28
F + Non-leaky augmentation	**1.79**	**1.79**	**1.97**	**1.98**	**2.39**	**2.53**	**1.96**	**2.16**
NFE	35	35	35	35	79	79	79	79

Original training (config A), VP



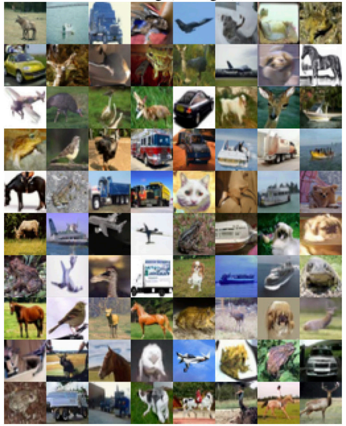
FID 3.01 NFE 35

Original training (config A), VE



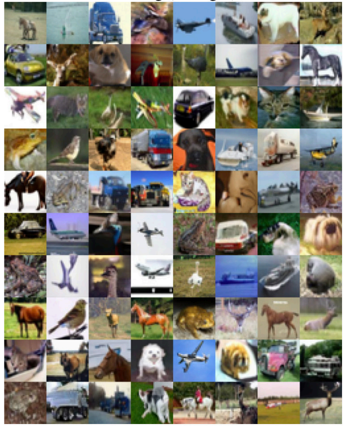
FID 3.77 NFE 35

Our training (config F), VP



FID **1.97** NFE 35

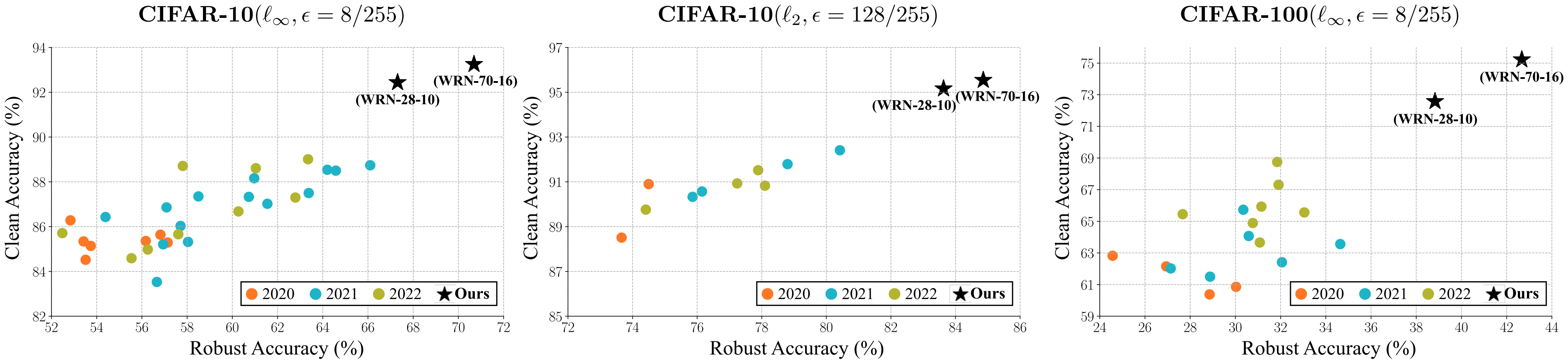
Our training (config F), VE



FID 1.98 NFE 35

[3] Karras et al. Elucidating the Design Space of Diffusion-Based Generative Models. NeurIPS 2022

Yes! Better Diffusion Models are Indeed Better



- New state-of-the-art!

RobustBench

A standardized benchmark for adversarial robustness

Wang et al. Better Diffusion Models Further Improve Adversarial Training. ICML 2023

Yes! Better Diffusion Models are Indeed Better

Table 1. A brief summary comparison of test accuracy (%) between our models and existing Rank #1 models, *with* (✓) and *without* (✗) external datasets, as listed in RobustBench (Croce et al., 2021).

Dataset	Method	External	Clean	AA
CIFAR-10 (ℓ_∞, $\epsilon = 8/255$)	Rank #1	✗	88.74	66.11
		✓	92.23	66.58
	Ours	✗	**93.25**	**70.69**
CIFAR-10 (ℓ_2, $\epsilon = 128/255$)	Rank #1	✗	92.41	80.42
		✓	**95.74**	82.32
	Ours	✗	95.54	**84.86**
CIFAR-100 (ℓ_∞, $\epsilon = 8/255$)	Rank #1	✗	63.56	34.64
		✓	69.15	36.88
	Ours	✗	**75.22**	**42.67**

- Even beat previous SOTA that using external datasets
- No extra training time
(only extra cost for generating data)

Wang et al. Better Diffusion Models Further Improve Adversarial Training. ICML 2023

Yes! Better Diffusion Models are Indeed Better

- Results on SVHN and Tiny-ImageNet

Dataset	Method	Generated	Ratio	Batch	Epoch	Clean	AA
SVHN ($\ell_\infty, \epsilon = 8/255$)	Gowal et al. (2021)	✗	✗	512	400	92.87	56.83
	Gowal et al. (2021)	1M	0.4	1024	800	94.15	60.90
	Rebuffi et al. (2021)	1M	0.4	1024	800	94.39	61.09
	Ours	1M	0.2	1024	800	**95.19**	**61.85**
		50M	0.2	2048	1600	**<u>95.56</u>**	**<u>64.01</u>**
TinyImageNet ($\ell_\infty, \epsilon = 8/255$)	Gowal et al. (2021)	✗	✗	512	400	51.56	21.56
	Ours	1M	0.4	512	400	**53.62**	**23.40**
	Gowal et al. (2021)*	1M	0.3	1024	800	60.95	26.66
	Ours (ImageNet EDM)	1M	0.2	512	400	**<u>65.19</u>**	**<u>31.30</u>**

Wang et al. Better Diffusion Models Further Improve Adversarial Training. ICML 2023

Alleviate Overfitting

- The model performs better with a longer training process

Generated	Epoch	Best epoch	Clean			PGD-40			AA		
			Early	Last	Diff	Early	Last	Diff	Early	Last	Diff
✗	400	86	84.41	82.18	−2.23	55.23	46.21	−9.02	54.57	44.89	−9.68
	800	88	83.60	82.15	−1.45	53.86	45.75	−8.11	53.13	44.58	−8.55
20M	400	370	91.27	91.45	+0.18	64.65	64.80	+0.15	63.69	63.84	+0.15
	800	755	92.08	92.14	+0.06	66.61	66.72	+0.11	65.66	65.63	+0.03
	1200	1154	92.43	92.32	−0.11	67.45	67.64	+0.19	66.31	66.60	+0.29
	1600	1593	92.51	**92.61**	+0.10	68.05	67.98	−0.07	67.14	67.10	−0.04
	2000	1978	92.41	92.55	+0.14	68.32	68.30	−0.02	67.22	67.17	−0.05
	2400	2358	**92.58**	92.54	−0.04	**68.43**	**68.39**	−0.04	**67.31**	**67.30**	−0.01

Wang et al. Better Diffusion Models Further Improve Adversarial Training. ICML 2023

Alleviate Overfitting

- Alleviate overfitting in adversarial training (Amount of Generated Data >500K)

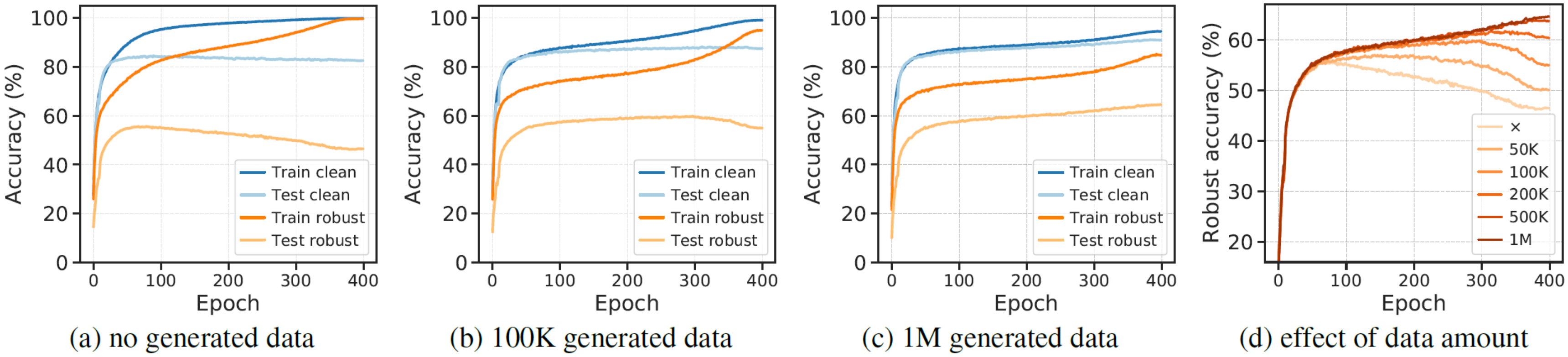


Figure 2. Clean and PGD robust accuracy of AT using (a) no generated data; (b) 100K generated data; (c) 1M generated data. (d) plots the PGD test robust accuracy of AT using different amounts of generated data.

Wang et al. Better Diffusion Models Further Improve Adversarial Training. ICML 2023

THANKS