Better Diffusion Models Further Improve Adversarial Training

WUHAN UNIVERSITY
sea AI Lab

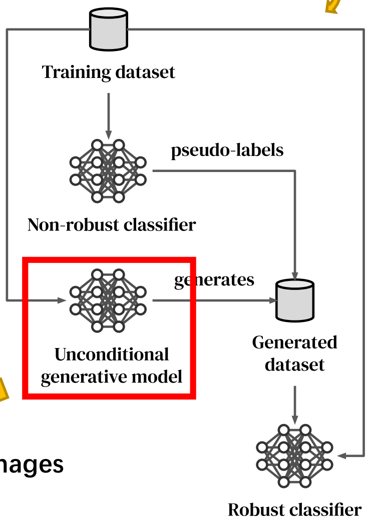
Zekai Wang*[1], Tianyu Pang*[2], Chao Du[2], Min Lin[2], Weiwei Liu[1], Shuicheng Yan[2]

[1]*Wuhan University* [2]*Sea AI Lab*

ICML International Conference On Machine Learning

➢ Previous SOTA in adversarial training (Rebuffi et al.)

- AT requires more data (Schmidt et al.)
- External datasets are not always available
- Use DDPM (FID **3.17** on CIFAR-10)
- Recent FID: **1.97 by EDM**



Can better diffusion models further improve adversarial training?

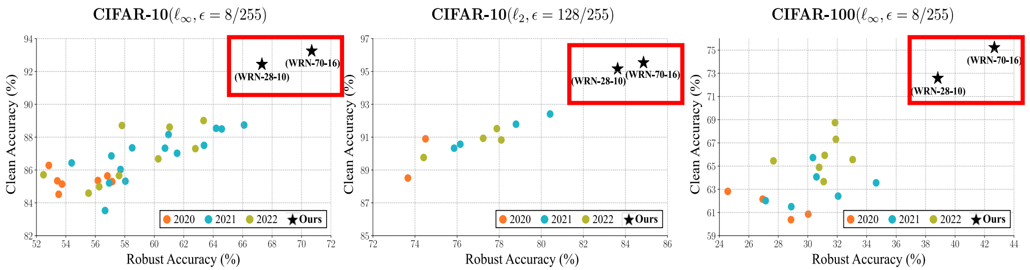
Replace DDPM with EDM (Karras et al.)

class-conditional generation, 50 million generated images

✓ We achieve SOTA results on with a large improvement!

RobustBench
A standardized benchmark for adversarial robustness

	CIFAR-10 l_∞	CIFAR-10 l_2	CIFAR-100 l_∞	SVHN l_∞	TinyImageNet l_∞
Clean	+4.51%	+3.13%	+11.66%	+1.17%	+4.24%
Robust	+4.58%	+4.44%	+8.03%	+2.92%	+4.64%



➢ Even beat the results using external datasets

With the same batch size, the training time per epoch of our method is equivalent to the w/o-generated-data baseline (only extra cost for data generation)

Dataset	Method	External	Clean	AA
CIFAR-10 ($\ell_\infty, \epsilon = 8/255$)	Rank #1	✗	88.74	66.11
		✓	92.23	66.58
	Ours	✗	**93.25**	**70.69**
CIFAR-10 ($\ell_2, \epsilon = 128/255$)	Rank #1	✗	92.41	80.42
		✓	**95.74**	82.32
	Ours	✗	95.54	**84.86**
CIFAR-100 ($\ell_\infty, \epsilon = 8/255$)	Rank #1	✗	63.56	34.64
		✓	69.15	36.88
	Ours	✗	**75.22**	**42.67**

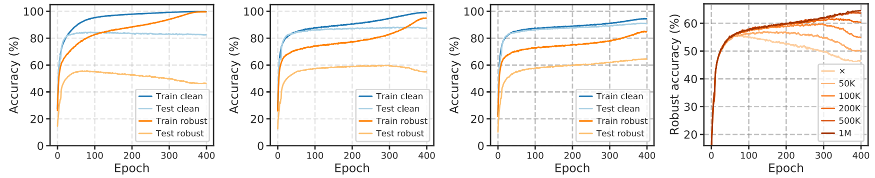
➢ Lower FID is better Conditional > Unconditional

	Step	FID ↓	Clean	PGD-40	AA
Class-cond.	5	35.54	88.92	57.33	57.78
	10	2.477	90.96	66.21	62.81
	15	1.848	91.05	64.56	63.24
	20	**1.824**	**91.12**	**64.61**	**63.35**
	25	1.843	91.07	64.59	63.31
	30	1.861	91.10	64.51	63.25
	35	1.874	91.01	64.55	63.13
	40	1.883	91.03	64.44	63.03
Uncond.	5	37.78	88.00	56.92	57.19
	10	2.637	89.40	62.88	61.92
	15	1.998	89.36	63.47	62.31
	20	**1.963**	**89.76**	**63.66**	**62.45**
	25	1.977	89.61	63.63	62.40
	30	1.992	89.52	63.51	62.33
	35	2.003	89.39	63.56	62.37
	40	2.011	89.44	63.30	62.24

➢ Models perform 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

➢ Alleviate overfitting



(a) no generated data (b) 100K generated data (c) 1M generated data (d) effect of data amount

➢ Sensitivity study on hyper-parameters

Batch Size	Clean	PGD-40	AA
128	91.12	64.77	63.90
256	91.15	65.76	64.72
512	91.81	66.15	65.21
1024	91.90	66.21	65.29
2048	**91.98**	**66.54**	**65.50**

Batch size

LS	Clean	PGD-40	AA
0	90.40	64.32	62.83
0.1	91.12	**64.61**	**63.35**
0.2	**91.23**	64.38	63.27
0.3	91.06	64.35	63.12
0.4	90.82	64.15	62.87

Label smoothing

β	Clean	PGD-40	AA
2	**92.46**	63.66	62.32
3	91.83	64.18	63.03
4	91.30	64.27	63.11
5	91.12	**64.61**	**63.35**
6	90.77	64.42	63.23
7	90.39	64.51	63.29
8	90.25	64.34	63.19

β in TRADES

➢ Data augmentation

Methed	Clean	PGD-40	AA
Common	91.12	**64.61**	**63.35**
Cutout	**91.25**	64.54	63.30
CutMix	91.08	64.34	62.81
AutoAugment	91.23	64.07	62.86
RandAugment	91.14	64.39	63.12
IDBH	91.08	64.41	63.24

Paper Code Twitter

Find more interesting conclusions in our paper!