arXiv:2302.04638v2 [cs.CV] 1 Jun 2023

更好的扩散模型进一步提升对抗训练

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

摘要

人们已经认识到，去噪扩散概率模型（DDPM）生成的数据能够提升对抗训练。扩散模型经过两年的快速发展后，一个自然的问题随之而来：更优的扩散模型能否进一步提升对抗训练？本文采用最新的扩散模型（Karras 等人，2022年）给出了肯定的答案。与 DDPM 相比，该模型效率更高（~ 20采样步数）且图像质量更好（FID 分数更低）。我们的对抗训练模型仅使用生成的数据（不使用外部数据集），便在 RobustBench 上取得当前最先进水平的性能。在采用 ℓ_∞- 范数威胁模型且 $\epsilon = 8/255$ 的条件下，我们的模型在 CIFAR-10 和 CIFAR-100 上分别达到70.69% 和 42.67% 的鲁棒准确率，分别比此前当前最先进的模型提升了 +4.58% 和 +8.03%。在采用 ℓ_2-范数威胁模型且 $\epsilon = 128/255$的条件下，我们的模型在 CIFAR-10 上达到84.86% （+4.44%）。这些结果也优于此前使用外部数据的研究。我们还在 SVHN 和 TinyImageNet 数据集上取得了有说服力的结果。我们的代码见 https://github.com/wzekai99/DM-Improves-AT。

1. 引言

对抗训练（AT）最早由Goodfellow 等人 （2015）提出，并已被证明是抵御对抗攻击最有效的防御方法之一（Madry 等人，2018年；Zhang 等人，2019b；Rice等人，2020年），并在对抗竞赛中成为获胜方案的主流（Kurakin 等，2018年；Brendel 等人，2020年）。人们普遍认为，更多数据的可用性对对抗训练模型的性能至关重要（Schmidt 等人，2018年；Stutz 等人，2019年）。因此，已有多项开创性工作尝试将外部数据集纳入对抗训练（Hendrycks 等人，2019年；Carmon 等人，2019年；Alayrac 等人，2019年；Najafi 等人，2019年；Zhai 等人，2019年；Wang 等人，2020年采用全监督或半监督学习范式。

然而，外部数据集即使没有标签，也并非总是可用。近期研究表明，去噪扩散概率模型（DDPM）生成的数据（Ho 等人，2020年）也能显著提升对抗训练模型在干净样本上的准确率和鲁棒准确率，这被视为一种基于学习的数据增强方法（Rebuffi 等人，2021年；Gowal 等人，2021年；Rade 与 Moosavi- Dezfooli，2021年；Sehwag 等人，2022年；Pang 等人，2022年）。由于效果显著，在 CIFAR-10 和 CIFAR- 100 上（Krizhevsky 与 Hinton，2009年），所有列于 RobustBench（Croce 等人，2021年）的现有顶尖模型（未使用外部数据集）都使用了 DDPM 生成的数据。[1]

经过两年扩散模型的快速发展，采样质量和效率都取得了显著提升

表 *1*。 简要比较了我们的模型与现有排名第一的模型的测试准确率（%），其中 （✓）表示使用，不使用（✗）外部数据集；相关模型列于 RobustBench（Croce 等人，2021年）。所有这些模型均采用 WRN-70-16 架构。

数据集	方法	外部干净样本	AA
CIFAR-10 ($\ell_\infty, \epsilon = 8/255$)	排名第1	✗	88.74 66.11
		✓	92.23 66.58
	Ours	✗	**93.25 70.69**
CIFAR-10 ($\ell_2, \epsilon = 128/255$)	排名第1	✗	92.41 80.42
		✓	**95.74** 82.32
	Ours	✗	95.54 **84.86**
CIFAR-100 ($\ell_\infty, \epsilon = 8/255$)	排名第1	✗	63.56 34.64
		✓	69.15 36.88
	Ours	✗	**75.22 42.67**

*共同贡献。相关工作是在 Zekai Wang 于 Sea AI Lab 实习期间完成的。[1]武汉大学计算机科学学院、国家多媒体软件工程技术研究中心、人工智能研究所及湖北省多媒体与网络通信工程重点实验室。 [2]Sea AI Lab。联系人：Tianyu Pang < tianyupang@sea.com>、Weiwei Liu < liuweiwei863@gmail.com>。

发表于 40th 国际机器学习会议，美国夏威夷州檀香山。PMLR 202，2023年。版权所有 © 2023年作者。

[1]https://robustbench.github.io

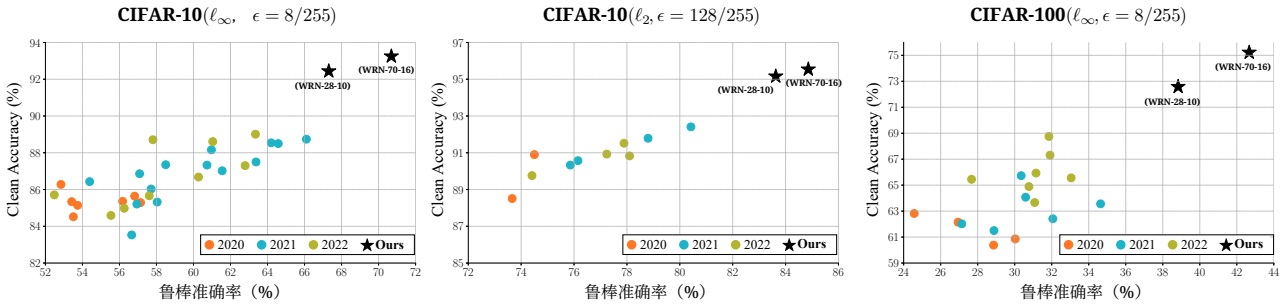


图 *1.* RobustBench 排行榜中排名靠前模型的鲁棒准确率（针对 AutoAttack）和干净准确率（无外部数据集）。排名靠前模型的发表年份以不同颜色标示。我们的模型在每种设置下分别采用 WRN-28-10 和 WRN-70-16 架构，详细准确率数值见表 2 和表 3。

，超越了 DDPM 的早期工作（Song 等人，2021年b）。具体而言，阐释型扩散模型（EDM）（Karras 等人，2022年）取得了新的最先进水平（SOTA）FID 分数（Heusel 等人，2017年），在无条件设置下达到1。97，而 DDPM 的 FID 分数为3。17。尽管 DDPM 和 EDM 生成的图像在视觉上难以区分，但我们想知道，更优的扩散模型（例如，FID 分数更低的模型）能否进一步惠及下游应用（例如，对抗训练任务）。

结果发现，我们的好奇心得到了出乎意料的丰厚回报。我们只需将 DDPM 生成的数据替换为 EDM 生成的数据，并采用与 Rebuffi 等人（2021年）所述几乎相同的训练流程。如表 1所示，无需外部数据集，也无需增加每个训练轮次的训练时间，我们经过对抗训练的 WRN-70-16 模型（Zagoruyko 和 Komodakis，2016年）在 CIFAR-10/CIFAR-100 上取得了 AutoAttack（自动攻击）下新的最先进鲁棒准确率（AA）（Croce 和 Hein，2020年），同时干净准确率也大幅提升。我们的模型甚至超过了依赖外部数据的先前排名第1模型。这些提升十分显著——如图 1 所示，即使是我们规模较小的 WRN-28-10 架构模型，性能也优于此前的基线方法。我们的方法还能显著提升模型在 SVHN 和 TinyImageNet 数据集上的性能。

此外，我们进行了大量消融研究，以更好地揭示扩散模型促进对抗训练（2020年）和Pang 等人（2021）中遵循的类似指导原则，考察了生成数据的数量和质量、提前停止以及数据增强等因素的影响。结果表明，EDM 生成的数据能够消除鲁棒过拟合，并缩小干净样本准确率与鲁棒准确率之间的泛化差距。在对抗训练过程中，我们还对若干重要参数进行了敏感性分析。我们的研究结果拓展了基于学习的数据增强（即使用扩散模型生成的数据）的潜力，并为未来促进对抗训练的研究奠定了坚实基础。

2. 相关工作

扩散模型。近年来，去噪扩散概率模型（Sohl-Dickstein 等人，2015；Ho 等人，2020年）和基于分数的朗之万动力学（Song 与 Ermon，2019年；2020年）在图像生成方面展现出良好效果。Song 等人（2021b）利用随机微分方程（SDE）统一了这两种生成式学习机制，这一统一的模型系列被称为扩散模型。此后，出现了一些新的研究方向，例如加速采样推理（Song 等人，2021a；Lu 等人，2022），优化模型参数化方式和采样计划（Kingma 等人，2021；Karras 等，2022），以及将扩散模型用于文生图（Ramesh 等人，2022；Rombach 等人，2022）。

对抗训练。除了利用外部数据集或生成数据外，研究人员还借鉴其他领域的策略，提出了多种改进对抗训练的方法，包括度量学习（Mao 等人，2019年；Pang 等人，2020年a；b）、自监督学习（Chen 等人，2020年a；b；Naseer 等人，2020年；Wang 和 Liu，2022年）、集成学习（Tram`er 等人，2018年；Pang 等人，2019年）、公平性（Ma 等人，2022年；Li 和 Liu，2023年）以及生成建模（Jiang 等人，2018年；Wang 和 Yu，2019年；Deng 等人，2020年）。Xu 和 Liu（2022年）、Li 等人（2022年）、Zou 和 Liu（2023年）从理论角度研究了对抗鲁棒学习。此外，由于对抗训练的计算成本较高，研究人员尝试通过复用计算（Shafahi 等人，2019年；Zhang 等人，2019a）或单步训练（Wong 等人，2020年；Liu 等人，2020年；Vivek B 和 Venkatesh Babu，2020年）来加速训练阶段。后续的一些研究则关注这些快速对抗训练方法带来的副作用（例如灾难性过拟合）（Andriushchenko 和 Flammarion，2020年；Li 等人，2020年）。

对抗净化。生成模型已被用于净化对抗样本（Song 等人，2018年）或增强可认证防御（Carlini 等人，2022年）。近年来，扩散模型在对抗净化领域广受关注（Yoon 等人，2021年；Nie 等人，2022年；Wang 等人，2022年；

Xiao 等人，2022年），展现出对 AutoAttack 具有良好的鲁棒准确率。另一方面，基于扩散的净化效果取决于随机微分方程求解器的随机性（Ho 等人，2020年；Bao 等人，2022年），这会导致推理计算至少增加数十倍，不利于下游部署。此外，研究表明，随机预处理或测试时防御存在一些共同的局限性（Gao 等人，2022年；Croce 等人，2022年），可能容易受到例如基于迁移的攻击（Kang 等人，2021年）或中间状态攻击（Yang 等人，2022年）。

对抗基准。鉴于提出的防御方法数量众多，开发全面且及时更新的对抗基准，以对现有方法进行排名至关重要。Dong 等人（2020年）开展了大规模实验，生成鲁棒性曲线以评估典型防御方法；Tang 等人（2021年）则全面研究了架构设计和训练技术对鲁棒性的影响。其他基准针对特定场景，包括对抗补丁（Hingun 等人，2022年；Lian 等人，2022年；Pintor 等人，2023年）、语言相关任务（Wang 等人，2021年；Li 等人，2021年）、自动驾驶车辆（Xu 等人，2022年）、多种威胁模型（Hsiung 等人，2022年）以及常见损坏（Mu 和 Gilmer，2019年；Hendrycks 和 Dietterich，2019年；Sun 等人，2021年）。本文采用 RobustBench（Croce 等人，2021年），这是学界广泛使用的基准。RobustBench 基于 AutoAttack 构建，后者已被证明能够可靠地评估对抗训练等确定性防御方法。

3. 实验设置

我们遵循基本设置，并使用 Rebuffi 等人(2021年) 的 PyTorch 实现。[2] 有关实验设置的更多信息，请参阅附录A。

模型架构。 作为骨干网络，我们采用 WideResNet（WRN）（Zagoruyko 和 Komodakis，2016年），并使用 Swish/SiLU 激活函数（Hendrycks 和 Gimpel，2016年）。我们使用 WRN-28-10 和 WRN-70-16，这是 RobustBench 上最常用的两种架构（Croce 等人，2021年）。

生成数据。 为生成新的图像，我们使用阐释型扩散模型（EDM）（Karras 等人，2022年），该模型能够达到最先进水平的 FID 分数。我们采用类别条件 EDM，其训练不依赖外部数据集（TinyImageNet 除外，具体说明见第 4 节）。我们遵循 Carmon 等人 （2019年）的指导，从 500 万张生成图像中筛选出 100 万张 CIFAR-10/CIFAR-100 图像；每张图像均由经过标准预训练的 WRN-28-10 模型进行评分。对于每个类别，我们选取得分最高的20%图像。当生成数据量超过 100 万，或生成数据用于 SVHN/TinyImageNet 时，我们不进行筛选，直接采用所有生成图像。请注意，不同于使用无条件 DDPM 的 Rebuffi 等人 （2021年）和Gowal 等人 （2021年），在我们的实现中，生成图像的伪标签由类别条件直接确定。

训练设置。 我们使用 TRADES（Zhang 等人，2019b）作为对抗训练（AT）的框架，$\beta = 5$用于 CIFAR-10/CIFAR-100，$\beta = 6$用于 SVHN，$\beta = 8$用于 TinyImageNet。我们采用权重平均，衰减率为 $\tau = 0.995$ (Izmailov 等人，2018年）。我们使用带 Nesterov 动量的 SGD 优化器（Nesterov，1983年），其中动量因子和权重衰减分别设为0。9 和 5×10^{-4}。我们使用带余弦退火的循环学习率调度（Smith 和 Topin，2019年），初始学习率设为0。2。

训练时间。 无论使用的生成数据量是多少（例如 1M、20M 或 50M），每个训练轮次的迭代次数都被控制为 $\left\lceil \frac{\text{amount}}{\text{原始数据批量大小的}} \right\rceil$，适用于所有使用生成数据的实验。这样可以确保与不使用生成数据的基线方法（例如表 ✗ 2 中标记为 的基线方法）进行公平比较，因为当训练轮数和批量大小固定时，训练时间保持不变。具体而言，在每个训练批次中，我们按照固定的“原始数据与生成数据比例”从原始数据和生成数据中采样图像。以 *CIFAR-10*（*50K* 张训练图像）和 *0.3* 的原始数据与生成数据比例为例，每个训练轮次包含 *50K* 张图像的模型训练数据：其中 *15K* 张来自原始数据，*35K* 张来自生成数据。对于 *CIFAR-10/CIFAR-100* 实验，当生成数据量为 *1M* 时，原始数据与生成数据比例为0.3；当所需的生成数据量超过 1M 时，该比例为*0*.2。表 4列出了应用于 SVHN 和 TinyImageNet 的原始数据与生成数据比例，而附录B.1 则包含了关于不同比例影响的额外消融研究。

评估指标。我们使用 AutoAttack（Croce 和 Hein，2020年）评估模型鲁棒性。由于对抗训练的计算成本很高，我们无法报告每项实验的标准差。为作说明，我们在 CIFAR-10 上使用 100 万条生成数据训练 WRN-28-10 模型五次，批量大小为512，训练400 轮。干净准确率为 $91.12 \pm 0.15\%$，在$(\ell_\infty, \epsilon = 8/255)$ 威胁模型下的鲁棒准确率为 $63.35 \pm 0.12\%$，表明我们的结果方差较低。

4. 与当前最先进水平的比较

我们将对抗训练模型与 RobustBench 中未使用外部数据集的排名靠前模型进行比较。表2 展示了 CIFAR-10 上（ℓ_∞，$\epsilon = 8/255$）和（ℓ_2，$\epsilon = 128/255$）威胁模型下的结果；表3和4展示了 CIFAR-100、SVHN 和 TinyImageNet 上（ℓ_∞，$\epsilon = 8/255$）威胁模型下的结果。总之，

[2] https://github.com/imrahulr/adversarial_robustness_pytorch

表*2*。CIFAR-10 上干净图像以及遭受 AutoAttack（自动攻击）时的测试准确率（%）。当结果优于采用相同架构的最强基线时，我们将结果以**粗体**标示；当结果在该威胁模型下达到新的最先进水平时，则加下划线。我们没有采用 Pang 等人（Pang 等人 (2022年) †。在批量大小相同的情况下，我们的方法每轮训练时间与不使用生成数据的基线相当（参见第3节“训练时间”段落）。

数据集	架构	方法	生成	批次	训练轮次†	干净样本	AA
CIFAR-10 ($\ell_\infty, \epsilon = 8/255$)	WRN-34-20	Rice等人（2020年）	✗	128	200	85.34	53.42
	WRN-34-10	Zhang 等人（2020年）	✗	128	120	84.52	53.51
	WRN-34-20	Pang 等人（2021年）	✗	128	110	86.43	54.39
	WRN-34-10	Wu等人(2020年)	✗	128	200	85.36	56.17
	WRN-70-16	Gowal 等人(2020年)	✗	512	200	85.29	57.14
	WRN-34-10	Sehwag 等人(2022年)	10M	128	200	87.00	60.60
	WRN-28-10	Rebuffi 等人 (2021年)	1M	1024	800	87.33	60.73
		Pang 等人 (2022年)	1M	512	400	88.10	61.51
		Gowal 等人(2021年)	100M	1024	2000	87.50	63.38
		Ours	1M	512	400	**91.12**	63.35
			1M	1024	800	**91.43**	**63.96**
			50M	2048	1600	**92.27**	**67.17**
			20M	2048	2400	**92.44**	**67.31**
	WRN-70-16	Pang 等人 (2022年)	1M	512	400	88.57	63.74
		Rebuffi 等人 (2021年)	1M	1024	800	88.54	64.20
		Gowal 等人(2021年)	100M	1024	2000	88.74	66.11
		Ours	1M	512	400	**91.98**	65.54
			5M	512	800	**92.58**	**67.92**
			50M	1024	2000	**93.25**	**70.69**
CIFAR-10 ($\ell_2, \epsilon = 128/255$)	WRN-34-10	Wu等人 (2020年)	✗	128	200	88.51	73.66
	WRN-70-16	Gowal 等人(2020年)	✗	512	200	90.9	74.50
	WRN-34-10	Sehwag 等人 (2022年)	10M	128	200	90.80	77.80
	WRN-28-10	Pang 等人 (2022年)	1M	512	400	90.83	78.10
		Rebuffi 等人 (2021年)	1M	1024	800	91.79	78.69
		Ours	1M	512	400	**93.76**	**79.98**
			50M	2048	1600	**95.16**	**83.63**
	WRN-70-16	Rebuffi 等人（2021年）	1M	1024	800	92.41	80.42
		Ours	1M	512	400	**94.47**	**81.16**
			50M	1024	2000	**95.54**	**84.86**

此前排名靠前的模型使用 DDPM 生成的图像，而我们的模型使用 EDM，并显著提升了干净样本准确率和鲁棒准确率。在这些威胁模型下，我们表现最佳的模型击败了 RobustBench 中的所有模型（包括使用外部数据集的模型）。

表 2的说明。 在（ℓ_∞， $\epsilon = 8/255$）威胁模型下的 CIFAR-10 数据集上，即使使用 1M 张生成图像、较小的批量大小 512 以及较短的训练 400 训练轮数，我们的 WRN-28-10 模型仍能达到与使用 100M 张生成图像的 Gowal 等人 (2021年) 相当的鲁棒准确率，同时我们的干净准确率显著提升（+3.62%）。在采用更大的批量大小 2048 并进行更长时间的训练 2400训练轮数后，我们的 WRN-28-10 模型以较大的间隔超过了使用 100M 张生成数据获得的此前最佳结果（干净准确率 +4.94%，鲁棒准确率 +3.93%），甚至超过了此前的最先进水平 WRN-70-16 模型。在 WRN-70-16 上使用 50M 张生成图像并训练 2000 训练轮次后，我们的模型达到 93.25% 的干净准确率和 70.69% 的鲁棒准确率，相较于最先进水平的结果分别提升了 +4.51% 和 +4.58%。这是首个在不使用外部数据集的情况下，干净准确率超过 ~~90% 且鲁棒准确率超过 70% 的对抗训练模型。在 CIFAR-10 数据集的（ℓ_2, $\epsilon = 128/255$）威胁模型下，我们最佳的 WRN-28-10 模型~~ 达到 95.16%（+3.37%）的干净准确率和 83.63%（+4.94%）的鲁棒准确率；我们最佳的 WRN-70-16 模型达到 95.54%（+3.13%）的干净准确率和 84.86%（+4.44%）的鲁棒准确率，明显优于此前的最先进水平模型。

表 3的备注。 在（$\ell_\infty, \epsilon = 8/255$）威胁模型下，使用 EDM 在 CIFAR-100 上生成的图像，取得了出人意料的出色表现。具体来说，我们的

表*3*。 CIFAR-100 上干净样本及 AutoAttack（自动攻击）下的测试准确率（%）。当结果优于采用相同架构的最强基线时，我们将结果标为**粗体**；当结果在该威胁模型下达到新的最先进水平时，则加下划线。我们没有像 Pang 等人(2022年) †那样使用 CutMix。在批量大小相同的情况下，我们方法每个训练轮次的训练时间与不使用生成数据的基线相当（参见第3节中的“训练时间”段落）。

数据集	架构	方法	生成	批次	训练轮次†	干净样本	AA
CIFAR-100 ($\ell_\infty, \epsilon = 8/255$)	WRN-34-10	Wu等人(2020年)	✗	128	200	60.38	28.86
	WRN-70-16	Gowal 等人(2020年)	✗	512	200	60.86	30.03
	WRN-34-10	Sehwag 等人(2022年)	1M	128	200	65.90	31.20
	WRN-28-10	Pang 等人 （2022年）	1M	512	400	62.08	31.40
		Rebuffi 等人 （2021年）	1M	1024	800	62.41	32.06
		Ours	1M	512	400	**68.06**	**35.65**
			50M	2048	1600	**72.58**	**38.83**
	WRN-70-16	Pang 等人 （2022年）	1M	512	400	63.99	33.65
		Rebuffi 等人 （2021年）	1M	1024	800	63.56	34.64
		Ours	1M	512	400	**70.21**	**38.69**
			50M	1024	2000	**75.22**	**42.67**

表 *4*。 SVHN 和 TinyImageNet 上干净样本图像以及在 AutoAttack（自动攻击，AA）下的测试准确率（%）。我们按照表 3 中的标注突出显示结果。此处“比例”表示原始数据与生成数据比例。所有结果均采用 WRN-28-10 模型架构。按照 Pang 等人 （2022年）的方法，我们未应用 CutMix。 †请注意，Gowal 等人 （2021年）在 ImageNet 上使用类别条件 DDPM 模型（Dhariwal 和 Nichol，2021年），并使用 TinyImageNet 的标签作为类别条件直接生成图像。

数据集	方法	生成	比例	批次	训练轮次	干净样本	AA
SVHN ($\ell_\infty, \epsilon = 8/255$)	Gowal 等人 (2021年)	✗	✗	512	400	92.87	56.83
	Gowal 等人 (2021年)	1M	0.4	1024	800	94.15	60.90
	Rebuffi 等人(2021年)	1M	0.4	1024	800	94.39	61.09
	Ours	1M	0.2	1024	800	**95.19**	**61.85**
		50M	0.2	2048	1600	**95.56**	**64.01**
TinyImageNet ($\ell_\infty, \epsilon = 8/255$)	Gowal 等人 (2021年)	✗	✗	512	400	51.56	21.56
	Ours	1M	0.4	512	400	**53.62**	**23.40**
	Gowal 等人(2021年)† **我们的方法**(ImageNet EDM)	1M	0.3	1024	800	60.95	26.66
		1M	0.2	512	400	**65.19**	**31.30**

最佳 WRN-28-10 模型的干净准确率为 72.58%（+10.17%），鲁棒准确率为 38.83%（+6.77%）；我们的最佳 WRN-70-16 模型的干净准确率为 75.22%（+11.66%），鲁棒准确率为 42.67%（+8.03%）。

表 4 的说明。 我们在 SVHN 和 TinyImageNet 数据集上评估了在（ℓ_∞， $\epsilon = 8/255$）威胁模型下的性能。如图所示，我们的方法显著优于基线方法。我们表现最好的模型使用 5000 万生成数据，在 SVHN 上的鲁棒准确率达到 64.01% $(+2.92\%)$。

我们仅使用 TinyImageNet 的训练集训练 EDM 模型，以生成 100 万条数据。我们的 WRN-28-10 模型在干净准确率和鲁棒准确率方面分别提升了 $+2.06\%$和 $+1.84\%$。值得注意的是，Gowal 等人(2021年) 使用在 ImageNet 上预训练的类别条件 DDPM（Dhariwal & Nichol，2021年）在 TinyImageNet 上生成数据，因为 TinyImageNet 数据集是 ImageNet 的子集。为确保公平比较，我们使用 EDM 提供的在 ImageNet 上预训练的检查点，并使用类别条件生成 TinyImageNet 中指定类别的图像。提升幅度十分显著：我们的模型在批量大小和训练轮次均小于 65。19% $(+4.24\%)$ 的干净准确率，以及 31。30% $(+4.64\%)$ 的鲁棒准确率，而 Gowal 等人 (2021 年)。这些结果证明了生成数据在多个数据集上增强模型鲁棒性的有效性。

5. 生成数据如何影响鲁棒性

Rice等人（2020年）首先观察到对抗训练中的鲁棒过拟合现象：在特定训练轮次之后，例如学习率衰减后不久，测试鲁棒损失开始上升。鲁棒过拟合的成因仍有争议（Pang 等人，2022年），但一种广为接受的观点认为，数据集规模不足以实现鲁棒泛化（Schmidt 等人，2018年）。当训练集大幅扩充，使用大量外部数据（Carmon 等人，2019年；Alayrac 等人，2019年）或合成

表5。 在 CIFAR-10 上，采用（$\ell_\infty, \epsilon = 8/255$）威胁模型训练不同**训练轮数**时的测试准确率（%）。WRN-28-10 模型的批量大小为 2048，原始数据与生成数据比例为 0。2。模型在验证集上的 PGD-40 准确率于“最佳训练轮次”达到最高。“较早”和“最后”分别表示最佳训练轮次和最后训练轮次时的测试性能。“差值”表示“较早”和“最后”之间的准确率差距。

生成	训练轮次	最佳训练轮次	干净样本			PGD-40			AA		
			较早	Last	Diff	较早	Last	Diff	较早	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

数据（Gowal 等人，2021年）时，干净样本准确率和鲁棒准确率均有显著提升。本节全面研究在对抗训练中使用生成图像时，训练细节对鲁棒过拟合和模型性能的影响。除非另有说明，实验均在 CIFAR-10 数据集和 WRN-28-10 模型上进行，模型训练了400 轮，批量大小为512 ，原始数据与生成数据比例为0.3。

5.1. 较早停止与训练轮数

在标准设定下，一系列研究（Zhang 等人，2017年；Belkin 等人，2019年）发现，深度学习模型在实践中并不会出现过拟合，也就是说，测试损失会随着训练损失一同下降，因此通常会尽可能长时间地进行训练。然而，对于对抗训练，Rice等人（2020年）揭示了鲁棒过拟合现象：鲁棒准确率在测试集上迅速下降，而在训练集上则持续上升。若没有生成数据，增加训练轮数并不能保证性能提升。因此，早停成为对抗训练过程中的默认选项：跟踪留出验证集上的鲁棒准确率，并选择验证鲁棒性最佳的检查点。

我们在 CIFAR-10 上使用 EDM 生成的 2000 万张图像开展实验，研究训练轮次和早停如何影响鲁棒性能，并使用了充足的图像。结果如表 5所示。我们还提供了未使用生成数据时的结果，以便更好地比较，并得出以下结论：

- 如先前所观察到的，在不使用生成数据时，早停是有效的（Rice等人，2020年）。早停在训练过程的初期便会触发（第86训练轮次/400训练轮次；第88训练轮次/400训练轮次）。在测试集上，干净样本准确率和鲁棒准确率都会下降，训练时间越长，性能越差。
- 使用生成数据时，早停的重要性较低。表现最佳的模型似乎出现在训练结束时，这意味着提前停止训练不会带来显著提升。“差值”变小，表明充足的训练数据能够有效缓解鲁棒过拟合。
- 使用 2000 万张生成图像时，训练过程越长，模型表现越好。令人意外的是，训练轮次较短会导致鲁棒欠拟合（“差值”为正）。当模型使用充足的数据进行训练时，结果表明，尽可能长时间地训练有助于提升鲁棒性能。

我们将早停视为一种默认技巧，这与先前的研究一致（Pang 等人，2021年），因为它在原始数据集上效果显著，在大数据集上的表现也相当。实现细节请参阅附录 A 。

5.2. 生成数据量

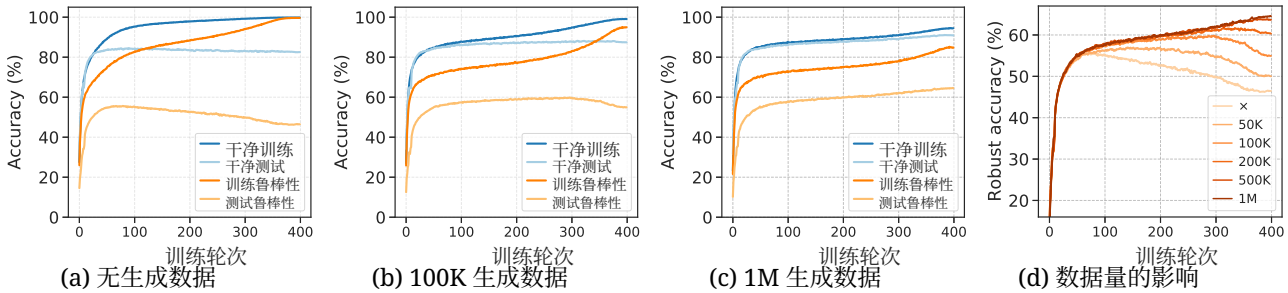
与原始训练集相比，我们可以使用生成模型采样得到更多图像。根据Gowal 等人(2021年)的研究，DDPM 生成的图像越多，鲁棒泛化差距越小。Rebuffi 等人(2021年)成功地使用 DDPM 生成固定规模（100 万张）的数据，避免了鲁棒过拟合，并在学习率下降后实现了稳定的性能。本文探讨生成数据的规模如何影响鲁棒过拟合。结果如图2 和表6所示。

图 2 （a、b、c）展示了不同数量的 EDM 生成数据下，训练集和测试集上的干净样本准确率与鲁棒准确率。不同数据规模下的更多结果见附录 B.5。研究结果表明：

- 我们可以看到，在未使用生成数据时，会出现严重的鲁棒过拟合现象（图2 (a)，表格中的“✗”（表6）。经过一定训练轮次后，干净测试准确率

表6。 在 CIFAR-10 上采用不同**生成数据量**进行训练时，在 $(\ell_\infty, \epsilon = 8/255)$ 威胁模型下的测试准确率（%）。模型在验证集的“最佳训练轮次”时取得最高的 PGD-40 准确率。“最佳”指训练期间达到的最高准确率；“最后”指最后一个训练轮次的测试性能。“差值”表示“最佳”与“最后”之间的差距。由于运行自动攻击耗时较长，我们将“最佳训练轮次”时的自动攻击准确率视为“最佳”。

生成	最佳训练轮次	干净样本			PGD-40			AA		
		Best	Last	Diff	Best	Last	Diff	Best	Last	Diff
✗	91	84.55	82.59	−1.96	55.66	46.47	−9.19	54.37	45.29	−9.08
50K	171	86.15	85.47	−0.68	56.96	50.02	−6.94	55.71	48.85	−6.86
100K	274	88.20	87.47	−0.73	59.85	54.95	−4.90	58.85	53.42	−5.43
200K	365	89.71	89.48	−0.23	61.69	60.32	−1.37	59.91	59.11	−0.80
500K	395	90.76	90.58	−0.18	63.85	63.69	−0.16	62.76	62.77	+0.01
1M	394	91.13	90.89	−0.24	64.67	64.50	−0.17	63.35	63.50	+0.15
5M	395	91.15	90.93	−0.22	64.88	64.88	0	64.05	64.05	0
10M	396	**91.25**	**91.18**	−0.07	65.03	64.96	−0.07	64.19	64.28	+0.09
20M	399	91.17	91.07	−0.10	65.21	65.13	−0.08	64.27	64.16	−0.11
50M	395	91.24	91.15	−0.09	**65.35**	**65.23**	−0.12	**64.53**	**64.51**	−0.02



(a) 无生成数据 (b) 100K 生成数据 (c) 1M 生成数据 (d) 数据量的影响

图 2. 使用（a）无生成数据、（b）10 万条生成数据和（c）100 万条生成数据进行对抗训练时，干净样本准确率和 PGD 鲁棒准确率。（d）展示了使用不同数量生成数据进行对抗训练时的 PGD 测试鲁棒性。

略有下降，而鲁棒准确率的“差值”则很大。在最后一个训练轮次，训练集与测试集鲁棒准确率之间的泛化差距接近 60%。

• 生成数据有助于缩小干净样本准确率和鲁棒准确率的泛化差距。在最后几个训练轮次中，不使用生成数据时，训练损失趋近于零。随着生成数据规模增大，训练准确率下降，而测试准确率提高。结果表明，EDM 生成的图像中包含难以进行鲁棒分类的样本，这有利于提升模型的鲁棒性。

• 如图2 （d）所示，随着生成数据规模的增加，鲁棒过拟合得到了缓解。在生成了 500K 张图像后，继续添加生成图像不再带来显著改善。在表6中，生成 500K 张图像后的“差值”变得很小，且“最佳训练轮次”出现在最后几个训练轮次中。这是意料之中的，因为模型容量不足以利用全部生成数据。因此，我们使用大型模型（WRN-70-16）给出了 SOTA 结果。更长的训练轮次也有助于模型在充足数据上的收敛，如表5所示。

5.3. 数据增强

已有研究表明，数据增强通过增加训练数据的数量和多样性，可以提升标准训练的泛化能力。令人意外的是，先前几乎所有仅通过数据增强来防止鲁棒过拟合的尝试都失败了（Rice等人，2020年；Wu等人，2020年；Tack等人，2022年）。Rebuffi等人（2021年）观察到，将数据增强与权重平均相结合可以提升鲁棒准确率，但在使用外部数据（例如 80M Tiny Images数据集）时效果较差。尽管Gowal 等人（2021年）报告称，Cut Mix 与使用生成数据兼容，但Pang 等人（2022年）的初步实验表明，CutMix 的有效性可能取决于具体的实现方式。

为此，我们考察了多种数据增强方法，并检验它们在使用生成数据进行对抗训练时的效果。常见数据增强（He等人，2016年）用于图像分类任务，包括在图像的每条边缘进行填充、再裁剪回原始尺寸，以及进行水平翻转。Cutout（Devries & Taylor，2017年）会随机丢弃输入图像中的一个区域。CutMix（Yun等人，2019年）会随机用另一张图像替换图像中的部分区域。AutoAugment（Cubuk等人，2019年）和 RandAugment（Cubuk等人，2020年）

表7. 在 CIFAR-10 上，使用 WRN-28-10 和 1M EDM 生成数据，在（$\ell_\infty, \epsilon = 8/255$）威胁模型下采用不同**增强方法**时的测试准确率（%）。

方法	干净样本	PGD-40	AA
常见	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

结合多种图像变换（如颜色、旋转和 Cutout），以找到最优组合。IDBH（Li& Spratling，2023年）是近期专为对抗训练设计的数据增强方案；与上述增强方法相比，在不使用额外数据的情况下，它取得了最佳的鲁棒性能。

我们将常见数据增强视为基线，因为这是对抗训练研究默认采用的设置。使用 1M EDM 生成数据时，表格7 展示了各种数据增强方法的表现。在鲁棒准确率（PGD-40 和 AutoAttack）方面，没有任何增强方法优于常见数据增强。Cutout 和 IDBH 的表现略胜于其他方法。需要指出的是，IDBH 本是为对抗训练设计的，但在使用生成数据的设置下同样失效。在干净准确率方面，Cutout 和 AutoAugment 略优于常见数据增强。

总而言之，基于规则的（Cutout 和 CutMix）以及基于策略的（AutoAugment、RandAugment 和 IDBH）数据增强似乎不太能有效提升鲁棒性，尤其是在使用生成数据时。我们的实证结果与先前研究相矛盾，表明数据增强对对抗训练的效果可能取决于具体实现。因此，我们遵循Pang 等人（2022年）。

5.4. 生成数据的质量

采样步数是扩散模型中的一个关键超参数，控制着生成质量和速度。因此，我们使用不同的采样步数生成数据，以探究生成数据的质量如何影响模型性能。我们通过计算 5 万张生成图像与原始训练图像之间的 FID 分数来评估质量（Heusel 等人，2017年）。在附录B.3中，我们研究了不同采样器和 EDM 公式对模型鲁棒性的影响。

在监督学习中，生成建模有无条件和类别条件两种范式。大量实证证据（Brock 等，2019年；Dhariwal 和 Nichol，2021年）表明，类别条件生成模型更易于训练，且其 FID 低于无条件模型，因为它们利用了数据标签。我们在类别条件和无条件设置下生成数据，但以略有不同的方式提供伪标签，以研究类别条件对对抗训练的影响。有关更多信息，请参阅附录A。我们使用 EDM 生成 100 万张图像，结果汇总于表8。

表8。在 CIFAR-10 上，使用 WRN-28-10 和 100 万 EDM 生成数据时，不同扩散模型**采样步数**在（$\ell_\infty, \epsilon = 8/255$）威胁模型下的测试准确率（%）和 FID。这里 ↓ 表示“越低越好”。

	Step	FID ↓	干净样本	PGD-40	AA
类别条件	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
无条件	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

我们发现，生成数据的 FID 较低会带来较高的干净样本准确率和鲁棒准确率。附录B.3中的结果也得出了相同的结论。FID 是一种常用指标，用于比较生成图像的分布与真实数据分布。较低的 FID 表明生成数据分布与真实数据分布之间的差异较小。结果表明，通过使生成数据更接近真实数据，我们可以提高模型的鲁棒性。

类别条件生成始终优于无条件生成，具有更低的 FID 和更好的鲁棒性能。使用 20 个采样步数时，两种设置都能实现最低的 FID 和最佳性能。因此，在 CIFAR-10 上的实验中，我们采用类别条件生成，使用 20 个采样步数以及 EDM 提供的检查点。[3] 第4 节中基线方法所需的额外数据由 DDPM 生成，其 FID 为3.28。在 CIFAR-100/SVHN 数据集上，我们自行训练模型，并选择经过 25个采样步数后 FID 最佳的模型（CIFAR-100 为 2.09，SVHN 为 1.39，见附录 B.2）。相比之下，DDPM 在 CIFAR-100 和 SVHN 上的 FID 分别为5.58 和4.89 （Gowal 等人，2021年）。FID 的大幅提升显著改善了第 4 节中基线方法的性能。

[3] https://github.com/NVlabs/edm

表 9. 在 CIFAR-10 上，采用（ℓ_∞，$\epsilon = 8/255$）威胁模型时，不同**批量大小**（左）、**标签平滑（LS）**（中）以及 β **TRADES 中的**参数值下的测试准确率（%）（右）。

批次 Size	干净样本	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**

LS	干净样本	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

β	干净样本	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

6. 敏感性分析

本节测试 CIFAR-10 上基本训练超参数的敏感性。使用 EDM 生成的 1M 条数据，训练 WRN-28-10 模型 400 个训练轮次。除非另有说明，默认批量大小为 512。

批次大小。在标准设置中，批次大小是影响模型在大规模数据集上性能的关键参数（Goyal 等人，2017年）。在没有外部数据或生成数据的对抗设置中，批次大小通常设为 128 或 256。Pang 等人（2021年）研究了从 64 到512的广泛批次大小范围，并发现 128 对于 CIFAR-10 是最优的。为评估批次大小对充足数据的影响，我们使用 5M 生成图像训练模型，并在表 9（左）中比较其在五种不同批次大小下的性能。可以观察到，最大的批次大小 2048取得了最佳结果。这表明，在使用充足训练数据和固定初始学习率时，较大的批次大小能够提升鲁棒性能。应用线性缩放规则时，可能存在一个最优批次大小（Pang 等人，2021年）。更大的批次大小需要额外的 GPU 内存，但能够更快地遍历数据集。为取得第4节中的最佳结果，我们根据模型大小和所使用的 GPU 数量增大批次大小。对于 WRN-28-10，我们在 4× A100 GPU 上选择 1024 的批次大小；对于 WRN-70-16，我们在 8×A100 GPU 上选择该批次大小。

标签平滑。对于标准训练，标签平滑（LS）（Szegedy 等人，2016年）可改善标准泛化能力并缓解过度自信问题（Hein 等人，2019年），但无法阻止自适应攻击绕过模型（Tram`er 等人，2020年）。在对抗设置中，LS 被用于在不使用外部数据或生成数据的情况下防止鲁棒过拟合（Stutz 等人，2020年；Chen 等人，2021年）。利用 1 M 生成数据，我们进一步评估 LS 对 AT 的影响。根据表9（中间）的结果，LS 设为0。1 时，干净准确率和鲁棒准确率均有所提升（干净样本+0.72%，自动攻击 +0.52%）。LS 设为0。2 时，干净准确率还可小幅提升，但代价是鲁棒性下降。这与先前研究的发现一致，研究（Jiang 等人，2020年；Pang 等人，2021年），过度使用 LS（0。3 和0。4）会降低模型性能。这也是标签过度平滑所致，它会导致输出 logits 中的信息丢失（M¨uller 等人，2019年）。为获得最佳鲁棒性能，我们在所有实验中均将 LS 设为 $= 0.1$。

.....的影响 β。在 TRADES（Zhang 等人，2019b）框架中，β 是一个重要的超参数，用于控制鲁棒性与准确率之间的权衡（如附录2式中的A）。随着正则化参数 β 增大，干净准确率会降低，而鲁棒准确率会提高，正如 Zhang 等人（2019b）所观察到的。相比之下，使用 1M EDM 生成的数据时，模型的鲁棒性会随着 β取值较大而下降。当 $\beta = 5$ 时，鲁棒性最佳，而较小的 β 仍有助于提升干净准确率。为获得最高的鲁棒准确率，在 CIFAR-10/CIFAR-100 上采用 $\beta = 5$。β在 SVHN 和 TinyImageNet 上分别设为6 和8 。

7. 讨论

扩散模型已证明其在对抗训练和对抗净化中的有效性；然而，研究如何在对抗领域中更高效地利用扩散模型至关重要。目前，即使对于 CIFAR-10 这类小型数据集，对抗训练也需要数百万条生成数据，这使训练阶段效率低下；对抗净化则需要通过扩散模型进行数十次前向过程，这使推理阶段效率低下。我们的工作推动了对抗训练模型最佳性能的极限，但未来研究仍有许多关于学习效率的问题值得探索。

致谢

本研究受到国家自然科学基金委员会资助（项目编号：61976161）以及中央高校基本科研业务费专项资金资助（项目编号：2042022rc0016）。

参考文献

Alayrac, J.、Uesato, J.、Huang, P.、Fawzi, A.、Stanforth, R. 和 Kohli, P. 提升对抗鲁棒性是否需要标签? 载于神经信息处理系统进展（*NeurIPS*），第 12192–12202 页，2019年。

Andriushchenko, M. 和 Flammarion, N. 理解并改进快速对抗训练。载于神经信息处理系统进展（*NeurIPS*），2020年。

Bao, F.、Li, C.、Zhu, J. 和 Zhang, B. Analytic-dpm：扩散概率模型中最优反向方差的解析估计。载于国际学习表征会议（*ICLR*），2022年。

Belkin, M.、Hsu, D.、Ma, S. 和 Mandal, S. 调和现代机器学习实践与经典的偏差-方差权衡。美国国家科学院院刊，116(32):15849–15854，2019年。

Brendel, W.、Rauber, J.、Kurakin, A.、Papernot, N.、Veliqi, B.、Mohanty, S. P.、Laurent, F.、Salath´e, M.、Bethge, M.、Yu, Y. 等。对抗视觉挑战赛。载于*NeurIPS 2018* 竞赛，第129–153页。施普林格，2020年。

Brock, A., Donahue, J., and Simonyan, K. 用于高保真自然图像合成的大规模生成对抗网络训练。发表于国际学习表征会议（*ICLR*），2019年。

Carlini, N., Tramer, F., Kolter, J. Z., et al. （经认证！！）免费获得对抗鲁棒性！*arXiv* 预印本 *arXiv:2206.10550*，2022年。

Carmon, Y., Raghunathan, A., Schmidt, L., Duchi, J.C., and Liang, P. 无标签数据可提升对抗鲁棒性。发表于神经信息处理系统进展（*NeurIPS*），第 11190–11201 页，2019年。

Chen, K., Chen, Y., Zhou, H., Mao, X., Li, Y., He, Y., Xue, H., Zhang, W., and Yu, N. 自监督对抗训练。发表于*IEEE* 国际声学、语音与信号处理会议（*ICASSP*），第 2218–2222 页。IEEE，2020年。

Chen, T., Liu, S., Chang, S., Cheng, Y., Amini, L. 和 Wang, Z. 对抗鲁棒性：从自监督预训练到微调。载于计算机视觉与模式识别会议（*CVPR*），第 699–708 页，2020年。

Chen, T.、Zhang, Z.、Liu, S.、Chang, S. 和 Wang, Z. 恰当学习平滑化或可缓解鲁棒过拟合。发表于国际学习表征会议（*ICLR*），2021年。

C$_{\text{roce}}$, F. and Hein, M. Reliabl对抗luation of ad鲁棒 ial性，采用多种多样的无参数攻击进行集成。载于国际机器学习会议（*ICML*），第 119 卷，第 2206–2216 页，2020年。Croce, F., Andriushchenko, M., Sehwag, V., Debenedetti, E., Flammarion, N., Chiang, M., Mittal, P. 和 Hein, M. Robustbench：标准化的对抗鲁棒性基准。载于神经信息处理系统进展（*NeurIPS*），2021年。

Croce, F., Gowal, S., Brunner, T., Shelhamer, E., Hein, M. 和 Cemgil, T. 评估自适应测试时防御的对抗鲁棒性。载于国际机器学习会议（*ICML*），2022年。

Cubuk, E. D., Zoph, B., Man´e, D., Vasudevan, V., 和 Le, Q. V. Autoaugment：从数据中学习增强策略。发表于 *IEEE*计算机视觉与模式识别会议 *(CVPR)*，第 113–123 页，2019年。

Cubuk, E.D., Zoph, B., Shlens, J., 和 Le, Q. Randaugment：在缩小搜索空间的情况下实现实用的自动数据增强。发表于 神经信息处理系统进展 *(NeurIPS)*，2020年。

Deng, Z., Dong, Y., Pang, T., Su, H., 和 Zhu, J. 用于稳健深度学习的对抗分布训练。发表于 神经信息处理系统进展 *(NeurIPS)*，2020年。

Devries, T. 和 Taylor, G. W. 使用 Cutout 改进卷积神经网络的正则化。*CoRR*，abs/1708.04552，2017年。

Dhariwal, P. and Nichol, A. Q. 扩散模型在图像合成方面胜过生成对抗网络. 收录于神经信息处理系统进展 *(NeurIPS)*，第 8780–8794 页，2021.

Dong, Y., Fu, Q.-A., Yang, X., Pang, T., Su, H., Xiao, Z., and Zhu, J. 图像分类对抗鲁棒性基准测试. 收录于*IEEE*计算机视觉与模式识别会议 *(CVPR)*，第 321–331 页，2020年.

Gao, Y., Shumailov, I., Fawaz, K., and Papernot, N. 关于随机预处理防御的局限性. 收录于神经信息处理系统进展 *(NeurIPS)*，2022年.

Goodfellow, I. J., Shlens, J., and Szegedy, C. 对抗样本的解释与利用. 收录于国际学习表征会议 *(ICLR)*，2015.

Gowal, S., Qin, C., Uesato, J., Mann, T.A., 和 Kohli, P. 揭示针对范数受限对抗样本的对抗训练的局限性。*CoRR*，abs/2010.03593，2020年。

Gowal, S., Rebuffi, S., Wiles, O., Stimberg, F., Calian, D. A., 和 Mann, T. A. 使用生成数据提升鲁棒性。载于 神经信息处理系统进展（*NeurIPS*），第 4218–4233 页，2021年。

Goyal, P., Doll´ar, P., Girshick, R. B., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., and He, K. 精确的大批次 SGD：在 1 小时内训练 ImageNet。*CoRR*, abs/1706.02677, 2017年。

He, K., Zhang, X., Ren, S., and Sun, J. 用于图像识别的深度残差学习。发表于*IEEE*计算机视觉与模式识别会议（*CVPR*），第 770–778 页，2016年。

Hein, M., Andriushchenko, M., and Bitterwolf, J. 为什么 ReLU 网络会对远离训练数据的样本给出高置信度预测，以及如何缓解这一问题。发表于*IEEE*计算机视觉与模式识别会议（*CVPR*），第 41–50 页，2019年。

Hendrycks, D. 和 Dietterich, T. 对常见损坏和扰动下的神经网络鲁棒性进行基准测试。载于 国际学习表征会议（*ICLR*），2019年。

Hendrycks, D. 和 Gimpel, K. 高斯误差线性单元（GELU）。*CoRR*，abs/1606.08415，2016年。

Hendrycks, D.、Lee, K. 和 Mazeika, M. 使用预训练可以提升模型鲁棒性和不确定性。载于 国际机器学习会议（*ICML*），2019年。

Heusel, M.、Ramsauer, H.、Unterthiner, T.、Nessler, B. 和 Hochreiter, S. 采用双时间尺度更新规则训练的生成对抗网络会收敛到局部纳什均衡。载于 神经信息处理系统进展（*NeurIPS*），第 6626–6637 页，2017年。

Hingun, N.、Sitawarin, C.、Li, J. 和 Wagner, D. REAP：大规模真实场景对抗补丁基准。*arXiv* 预印本 *arXiv:2212.05680*，2022年。

Ho, J., Jain, A., and Abbeel, P. 去噪扩散概率模型。发表于神经信息处理系统进展（*NeurIPS*），2020年。

Hsiung, L., Tsai, Y.-Y., Chen, P.-Y., and Ho, T.-Y. Carben：复合对抗鲁棒性基准。*arXiv*预印本 *arXiv:2207.07797*，2022年。

Izmailov, P., Podoprikhin, D., Garipov, T., Vetrov, D. P., and Wilson, A. G. 权重平均带来更宽的最优点和更好的泛化。发表于人工智能不确定性会议（*UAI*），第876–885页，2018年。

Jiang, H.、Chen, Z.、Shi, Y.、Dai, B. 和 Zhao, T. 通过学习攻击来学习防御。*arXiv* 预印本 *arXiv:1811.01213*，2018年。

Jiang, L.、Ma, X.、Weng, Z.、Bailey, J. 和 Jiang, Y. 梯度不平衡：高估对抗鲁棒性的新原因。*CoRR*, abs/2006.13726，2020年。Kang, Q.、Song, Y.、Ding, Q. 和 Tay, W. P. 具有李雅普诺夫稳定平衡点的稳定神经常微分方程，用于防御对抗攻击。载于神经信息处理系统进展（*NeurIPS*），2021年。Karras, T.、Aittala, M.、Aila, T. 和 Laine, S. 阐明基于扩散的生成模型的设计空间。载于神经信息处理系统进展（*NeurIPS*），2022年。

Kingma, D., Salimans, T., Poole, B., and Ho, J. 变分扩散模型。发表于神经信息处理系统进展（*NeurIPS*），2021年。

Krizhevsky, A. 和 Hinton, G. 从微小图像中学习多层特征。技术报告，Citeseer，2009年。

Kurakin, A.、Goodfellow, I.、Bengio, S.、Dong, Y.、Liao, F.、Liang, M.、Pang, T.、Zhu, J.、Hu, X.、Xie, C. 等。对抗攻击与防御竞赛。*arXiv*预印本 *arXiv:1804.00097*，2018年。

Li, B. 和 Liu, W. WAT：提升对抗训练中的最差类别鲁棒性。载于*AAAI*人工智能会议（*AAAI*），2023年。

Li, B.、Wang, S.、Jana, S. 和 Carin, L. 探索快速对抗训练。*arXiv*预印本 *arXiv:2006.03089*，2020年。

Li, L. 和 Spratling, M. W. 仅靠数据增强即可提升对抗训练效果。载于国际学习表征会议（*ICLR*），2023年。

Li, X., Xin, Z., 和 Liu, W. 通过神经动力系统防御对抗攻击。载于神经信息处理系统进展（*NeurIPS*），2022年。

Li, Z., Xu, J., Zeng, J., Li, L., Zheng, X., Zhang, Q., Chang, K.-W., 和 Hsieh, C.-J. 寻找有效的防御方法：对抗词语替换防御的基准测试。*arXiv* 预印本 *arXiv:2108.12777*，2021年。

Lian, J., Mei, S., Zhang, S., 和 Ma, M. 针对航空目标检测的对抗补丁基准测试。*IEEE* 地球科学与遥感汇刊，60:1–16，2022年。

Liu, G., Khalil, I., 和 Khreishah, A. 使用单步对抗训练防御迭代对抗样本。*arXiv* 预印本 *arXiv:2002.09632*，2020年。

Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., and Zhu, J. Dpm-solver：一种用于扩散概率模型采样的快速常微分方程求解器，可在约 10 步内完成。在神经信息处理系统进展（*NeurIPS*），2022年。

Ma, X., Wang, Z., and Liu, W. 关于鲁棒性与公平性之间的权衡。在神经信息处理系统进展（*NeurIPS*），第 35 卷，页码 26230–26241，2022年。

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., 和 Vladu, A. 探索能够抵御对抗攻击的深度学习模型。载于国际学习表征会议 *(ICLR)*，2018年。

Mao, C., Zhong, Z., Yang, J., Vondrick, C., 和 Ray, B. 用于对抗鲁棒性的度量学习。载于神经信息处理系统进展 *(NeurIPS)*，2019年。

Mu, N. 和 Gilmer, J. Mnist-c：计算机视觉鲁棒性基准。*arXiv*预印本 *arXiv:1906.02337*，2019年。M¨uller, R., Kornblith, S., 和 Hinton, G. E. 标签平滑何时有帮助？载于神经信息处理系统进展 *(NeurIPS)*，第 4696–4705 页，2019年。Najafi, A., Maeda, S.-i., Koyama, M., 和 Miyato, T. 从不完整数据中学习时对对抗扰动的鲁棒性。载于神经信息处理系统进展 *(NeurIPS)*，2019年。Naseer, M., Khan, S., Hayat, M., Khan, F. S., 和 Porikli, F. 一种用于对抗鲁棒性的自监督方法。载于计算机视觉与模式识别会议 *(CVPR)*，2020年。Nesterov, Y. E. 一种求解凸规划问题的方法，其收敛速率为 $o(1/k^2)$。载于苏联科学院报告，第 269 卷，第 543–547 页，1983年。Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., 和 Ng, A. 通过无监督特征学习识别自然图像中的数字。载于*NeurIPS*深度学习与无监督特征学习研讨会，2011年。Nie, W., Guo, B., Huang, Y., Xiao, C., Vahdat, A., 和 Anandkumar, A. 用于对抗净化的扩散模型。载于国际机器学习会议 *(ICML)*，2022年。Pang, T., Xu, K., Du, C., Chen, N., 和 Zhu, J. 通过促进集成多样性提升对抗鲁棒性。载于国际机器学习会议 *(ICML)*，2019年。Pang, T., Xu, K., Dong, Y., Du, C., Chen, N., 和 Zhu, J. 重新思考用于对抗鲁棒性的 Softmax交叉熵损失。载于国际学习表征会议 *(ICLR)*，2020年。

Pang, T., Yang, X., Dong, Y., Xu, K., Su, H., and Zhu, J. 利用超球嵌入增强对抗训练。载于神经信息处理系统进展（*NeurIPS*），2020年。Pang, T., Yang, X., Dong, Y., Su, H., and Zhu, J. 对抗训练技巧集。载于国际学习表征会议（*ICLR*），2021年。Pang, T., Lin, M., Yang, X., Zhu, J., and Yan, S. 通过（恰当的）定义，鲁棒性与准确率可以兼容。载于国际机器学习会议（*ICML*），第162卷，第17258–17277页，2022年。Pintor, M., Angioni, D., Sotgiu, A., Demetrio, L., Demontis, A., Biggio, B., and Roli, F. ImageNet-Patch：用于基准测试机器学习抵御对抗补丁鲁棒性的数据集。模式识别，134：109064，2023年。Rade, R. and Moosavi-Dezfooli, S.-M. 基于辅助模型的对抗训练：通过缩小过大的间隔，实现更佳的准确率与鲁棒性权衡。载于*ICML 2021* 对抗机器学习研讨会，2021年。Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. 利用 CLIP 潜变量进行分层文本条件图像生成。*arXiv* 预印本 *arXiv:2204.06125*，2022年。Rebuffi, S., Gowal, S., Calian, D. A., Stimberg, F., Wiles, O., and Mann, T. A. 改进数据增强以提升对抗鲁棒性。*CoRR*，abs/2103.01946，2021年。Rice, L., Wong, E., and Kolter, J. Z. 对抗鲁棒深度学习中的过拟合。载于国际机器学习会议（*ICML*），2020年。Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. 利用潜在扩散模型进行高分辨率图像合成。载于*IEEE*计算机视觉与模式识别会议（*CVPR*），2022年。Schmidt, L., Santurkar, S., Tsipras, D., Talwar, K., and Madry, A. 对抗鲁棒泛化需要更多数据。载于神经信息处理系统进展（*NeurIPS*），第5019–5031页，2018年。Sehwag, V., Mahloujifar, S., Handina, T., Dai, S., Xiang, C., Chiang, M., and Mittal, P. 鲁棒学习与生成模型相遇：代理分布能否提升对抗鲁棒性？载于国际学习表征会议（*ICLR*），2022年。Shafahi, A., Najibi, M., Ghiasi, A., Xu, Z., Dickerson, J., Studer, C., Davis, L. S., Taylor, G., and Goldstein, T. 免费进行对抗训练！载于神经信息处理系统进展（*NeurIPS*），2019年。

Smith, L. N. and Topin, N. 超收敛：利用大学习率快速训练神经网络。载于面向多域作战应用的人工智能与机器学习，第11006卷，第369–386页，2019年。

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., 和 Ganguli, S. 使用非平衡热力学进行深度无监督学习。在国际机器学习会议（*ICML*），2015。

Song, J., Meng, C., 和 Ermon, S. 去噪扩散隐式模型。在国际学习表征会议（*ICLR*），2021a。

Song, Y. 和 Ermon, S. 通过估计数据分布的梯度进行生成建模。在神经信息处理系统进展（*NeurIPS*），2019年。

Song, Y. 和 Ermon, S. 改进基于分数的生成模型训练技术。在神经信息处理系统进展（*NeurIPS*），2020年。

Song, Y., Kim, T., Nowozin, S., Ermon, S., 和 Kushman, N. Pixeldefend：利用生成模型理解并防御对抗样本。在国际学习表征会议（*ICLR*），2018年。

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Er- mon, S., and Poole, B. 通过随机微分方程进行基于得分的生成建模。发表于国际学习表征会议 *(ICLR)*，2021b。

Stutz, D., Hein, M., and Schiele, B. 解耦对抗鲁棒性与泛化能力。发表于*IEEE*计算机视觉与模式识别会议 *(CVPR)*，2019年。

Stutz, D., Hein, M., and Schiele, B. 置信度校准对抗训练：泛化至未见攻击。发表于国际机器学习会议 *(ICML)*，第119卷，第9155–9166页，2020年。

Sun, J., Mehra, A., Kailkhura, B., Chen, P.-Y., Hendrycks, D., Hamm, J., and Mao, Z. M. 认证对抗防御应对分布外损坏：对鲁棒性和简单基线方法进行基准测试。*arXiv*预印本 *arXiv:2112.00659*，2021年。

Szegedy, C.、Vanhoucke, V.、Ioffe, S.、Shlens, J. 和 Wojna, Z. 重新思考用于计算机视觉的 Inception架构。载于 *IEEE*计算机视觉与模式识别会议 *(CVPR)*，第 2818–2826 页，2016年。

Tack, J., Yu, S., Jeong, J., Kim, M., Hwang, S. J. 和 Shin, J. 面向对抗鲁棒性的一致性正则化。载于 *AAAI*人工智能会议（*AAAI*），第 8414–8422 页，2022年。

Tang, S., Gong, R., Wang, Y., Liu, A., Wang, J., Chen, X., Yu, F., Liu, X., Song, D., Yuille, A. 等。Robus- tart：对架构设计和训练技术的鲁棒性进行基准测试。*arXiv* 预印本 *arXiv:2109.05211*，2021年。

Tram`er, F., Kurakin, A., Papernot, N., Boneh, D. 和 Mc- Daniel, P. 集成对抗训练：攻击与防御。载于国际学习表征会议（*ICLR*），2018年。

Tram`er, F.、Carlini, N.、Brendel, W. 和 Madry, A. 针对对抗样本防御的自适应攻击。载于神经信息处理系统进展 *(NeurIPS)*，2020年。

Vivek B, S. 和 Venkatesh Babu, R. 采用丢弃调度的单步对抗训练。载于 *IEEE*计算机视觉与模式识别会议 *(CVPR)*，2020年。

Wang, B.、Xu, C.、Wang, S.、Gan, Z.、Cheng, Y.、Gao, J.、Awadallah, A. H. 和 Li, B. 对抗胶水：用于语言模型鲁棒性评估的多任务基准。*arXiv* 预印本 *arXiv:2111.02840*，2021年。

Wang, H. 和 Yu, C.-N. 使用对抗网络实现鲁棒深度学习的直接方法。发表于国际学习表征会议（*ICLR*），2019年。

Wang, J., Lyu, Z., Lin, D., Dai, B. 和 Fu, H. 用于对抗净化的引导式扩散模型。*arXiv* 预印本 *arXiv:2205.14969*，2022年。

Wang, Y., Zou, D., Yi, J., Bailey, J., Ma, X. 和 Gu, Q. 提升对抗鲁棒性需要重新审视误分类样本。发表于国际学习表征会议（*ICLR*），2020年。

Wang, Z. 和 Liu, W. 对比学习的鲁棒性验证。发表于国际机器学习会议（*ICML*），第162卷，第22865–22883页，2022年。

Wong, E., Rice, L. 和 Kolter, J. Z. 快速胜过免费：重新审视对抗训练。发表于国际学习表征会议（*ICLR*），2020年。

Wu, D.、Xia, S. 和 Wang, Y. 对抗权重扰动有助于实现鲁棒泛化。载于神经信息处理系统进展（*NeurIPS*），2020年。

Xiao, C.、Chen, Z.、Jin, K.、Wang, J.、Nie, W.、Liu, M.、Anandkumar, A.、Li, B. 和 Song, D. Densepure：理解扩散模型以实现对抗鲁棒性。*arXiv*预印本 *arXiv:2211.00322*，2022年。

Xu, C., Ding, W., Lyu, W., Liu, Z., Wang, S., He, Y., Hu, H., Zhao, D., 和 Li, B. Safebench：用于自动驾驶车辆安全评估的基准测试平台。

发表于 神经信息处理系统进展（*NeurIPS*）数据集与基准测试*Track*，2022年。

Xu, J. 和 Liu, W. 关于鲁棒多分类可学习性。载于神经信息处理系统进展（*NeurIPS*），2022年。Yang, Z.、Pang, T. 和 Liu, Y. 深度均衡模型的对抗鲁棒性细究。载于神经信息处理系统进展（*NeurIPS*），2022年。Yoon, J.、Hwang, S. J. 和 Lee, J. 基于分数的生成模型的对抗净化。载于国际机器学习会议（*ICML*），2021年。Yun, S.、Han, D.、Chun, S.、Oh, S. J.、Yoo, Y. 和 Choe, J. Cutmix：利用可定位特征训练强分类器的正则化策略。载于国际计算机视觉会议（*ICCV*），第6022–6031页，2019年。Zagoruyko, S. 和 Komodakis, N. 宽残差网络。载于英国机器视觉会议（*BMVC*），2016年。Zhai, R.、Cai, T.、He, D.、Dan, C.、He, K.、Hopcroft, J. 和 Wang, L. 对抗鲁棒泛化只需要更多无标签数据。*arXiv*预印本 *arXiv:1906.00555*，2019年。Zhang, C.、Bengio, S.、Hardt, M.、Recht, B. 和 Vinyals, O. 理解深度学习需要重新思考泛化。载于国际学习表征会议（*ICLR*），2017年。Zhang, D.、Zhang, T.、Lu, Y.、Zhu, Z. 和 Dong, B. 只传播一次：通过极大值原理加速对抗训练。载于神经信息处理系统进展（*NeurIPS*），2019a。Zhang, H.、Yu, Y.、Jiao, J.、Xing, E. P.、Ghaoui, L. E. 和 Jordan, M. I. 鲁棒性与准确率之间具有理论依据的权衡。载于国际机器学习会议（*ICML*），第97卷，第7472–7482页，2019b。Zhang, J.、Xu, X.、Han, B.、Niu, G.、Cui, L.、Sugiyama, M. 和 Kankanhalli, M. S. 不会导致训练失败的攻击会增强对抗学习。载于国际机器学习会议（*ICML*），第119卷，第11278–11287页，2020年。Zou, X. 和 Liu, W. 对抗对比学习的泛化界。机器学习研究期刊，24(114):1–54，2023年。

A. 技术细节

对抗训练。令 $\|\cdot\|_p$ 表示 ℓ_p 范数，例如，$\|\cdot\|_2$ 和 $\|\cdot\|_\infty$ 分别表示欧几里得范数 ℓ_2 和无穷范数ℓ_∞。$\mathcal{B}_p(x,\epsilon) := \{x' \mid \|x'-x\|_p \le \epsilon\}$ 表示输入 x' 被限制在 ℓ_p 球内，其中 ϵ 是最大扰动约束。Madry 等人（2018年）将对抗训练形式化为一个极小极大优化问题：

$$\arg\min_{\theta} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \left[\max_{x'\in\mathcal{B}_p(x,\epsilon)} \mathcal{L}(f_\theta(x'),y) \right], \tag{1}$$

其中，$\mathcal{D}$ 是由样本对 x 及其对应标签 y构成的数据分布，$f_\theta(\cdot)$是权重为 θ的神经网络分类器，而 $\mathcal{L}$ 是损失函数。内层优化寻找使损失最大化的对抗样本 x'，外层优化则最小化对抗样本上的损失，以更新网络参数 θ。

标准对抗训练的一种典型变体是 TRADES（Zhang 等人，2019b），本文将其用作对抗训练框架。作者表明，干净样本准确率与鲁棒准确率之间存在权衡，并将公式（1）分解为干净样本目标和鲁棒目标。TRADES 使用一个平衡超参数将这两个目标结合起来，以控制这种权衡：

$$\arg\min_{\theta} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \left[\mathrm{CE}(f_\theta(x),y) + \beta \cdot \max_{x'\in\mathcal{B}_p(x,\epsilon)} \mathrm{KL}(f_\theta(x)\|f_\theta(x')) \right], \tag{2}$$

其中，CE 表示标准交叉熵损失，$\mathrm{KL}(\cdot\|\cdot)$ 表示 Kullback–Leibler 散度，而 β 是用于控制权衡的超参数。在第 6节中，我们研究使用生成数据时 β 的敏感性。

PGD 攻击。投影梯度下降（PGD）（Madry 等人，2018年）是求解公式（1）和（2）中内层最大化问题的一种常用技术。令 $\mathrm{sign}(\boldsymbol{a})$ 表示 $\boldsymbol{a}$ 的符号。x_0 是在干净输入 x 的邻域 $\mathcal{B}_p(x,\epsilon)$ 中经过随机扰动的样本，那么 PGD 会通过多次梯度上升迭代构造对抗样本 K，其形式化定义如下：

$$x_k = \mathrm{clip}_{x,\epsilon}(x_{k-1} + \alpha \cdot \mathrm{sign}(\nabla_{x_{k-1}} \mathcal{L}(x_{k-1},y))), \tag{3}$$

其中，x_k 表示第 k 步的对抗样本，$\mathrm{clip}_{x,\epsilon}(x')$ 是裁剪函数，用于将 x' 投影回 $\mathcal{B}_p(x,\epsilon)$，而$\alpha$ 是步长。我们将执行 K 步的这一内层优化过程称为 PGD-K。

对于对抗训练中的对手，我们使用 PGD-10 攻击，并采用以下超参数：对于 ℓ_∞ 鲁棒模型，在 CIFAR-10/CIFAR-100/TinyImageNet 上，扰动大小为 $\epsilon = 8/255$，步长为 $\alpha = 2/255$；在 SVHN 上，步长为 $\alpha = 1.25/255$；对于 ℓ_2 鲁棒模型，在 CIFAR-10 上，扰动大小为 $\epsilon = 128/255$，步长为 $\alpha = 32/255$。

生成数据。这里我们展示 CIFAR-10/CIFAR-100 上生成 100 万张图像的更多细节。对于无条件生成，我们按照 Carmon 等人 (2019年); Rebuffi 等人 (2021年) 的做法，使用预训练 WRN-28-10 模型生成伪标签。该模型在 CIFAR-10 训练集上采用标准方法训练，干净测试准确率达到 96.15%。对于 CIFAR-100，WRN-28-10 模型的干净测试准确率达到 80.47%。随后，我们从无条件 EDM 中采样 500 万张图像，并使用预训练 WRN-28-10 模型提供的最高置信度为每张图像评分。我们将置信度最高的类别视为伪标签。在每个类别中，我们为 CIFAR-10 实验选取得分最高的前 10 万张图像（CIFAR-100 则选取前 1 万张）。

与无条件生成略有不同，类别条件 EDM 可以直接生成属于某个类别的 100 万个样本；因此，伪标签由类别条件直接确定。对于每个类别，我们分别为 CIFAR-10 和 CIFAR-100 实验生成 50 万张和 5 万张图像。同样，我们使用预训练 WRN-28-10 模型为每张图像评分，并为每个类别选取得分最高的前 20% 图像。

对于 SVHN 和 TinyImageNet 数据集生成数据时，或当 CIFAR- 10/CIFAR-100 的生成数据量超过 1M 时，我们直接使用类别条件泛化提供伪标签。每个类别的图像数量相同，以保持数据平衡。采样步数设置为 20，适用于 CIFAR-10（第 5.4 节）；25，适用于 CIFAR-100/SVHN（附录 B.2）；40，适用于 TinyImageNet（遵循 Karras 等人 (2022年)）。

数据集。CIFAR-10 和 CIFAR-100（Krizhevsky 与 Hinton，2009年）分别包含 50K 张训练图像和 10K 张测试图像，共有 10 类和 100 类。所有 CIFAR 图像的分辨率均为 $32\times32\times3$（宽度、高度、RGB 通道）。SVHN（Netzer 等人，2011年）包含 73,257 张训练图像和 26,032 张测试图像（$0\sim9$ 小尺寸裁剪数字，共 10 类）。TinyImageNet[4]包含 100K 张训练图像和 10K 张测试图像。图像分辨率为 $64\times64\times3$，共 200 类。

[4]http://cs231n.stanford.edu/tiny-imagenet-200.zip

更多设置详情。我们遵循先前的研究（Zagoruyko 和 Komodakis，2016年），使用 WideResNet（WRN）（Madry 等人，2018年；Rice等人，2020年；Rebuffi 等人，2021年），这些研究采用了 WRN 系列的变体。大多数实验使用 WRN-28-10（深度为 28，倍数为 10）进行，其参数量为 36M。WRN-28-10 上的实验由四块 NVIDIA A100 SXM4 40GB GPU 并行处理。为了评估大量生成数据对大型网络的影响，我们还在第 4 节使用 WRN-70-16，其参数量为 267M。我们使用八块 A100 GPU 训练 WRN-70-16。

遵循 Rice等人 (2020年的做法，并结合第 5节中的讨论，我们将早停作为默认技巧。我们将训练集中的前 1024 张图像划分出来，作为固定的验证集。在每个对抗训练轮次中，我们通过评估验证集在 PGD-40攻击下的鲁棒准确率，选取最佳检查点。

B. 附加实验

B.1. 原始数据与生成数据比例

原始数据与生成数据比例是指训练批次中原始图像与生成图像的混合比例，例如，比例为 0.3 表示每加入 3 张原始图像，就加入 7 张生成图像。我们使用 1M EDM 生成数据，研究该比例如何影响性能。我们在图3中总结了结果。干净准确率和鲁棒准确率均在比例为 0.3 时达到最佳结果。当生成数据多于 1M 时，我们选择 0.2 的比例以获得更好的性能（表10）。我们还观察到，使用 1M EDM 生成数据的性能优于使用 50k 原始 CIFAR-10 训练集的性能，这与 Rebuffi 等人（2021年）的结果一致。结果表明，只要生成模型能够生成高质量数据，生成图像就能提升鲁棒性。

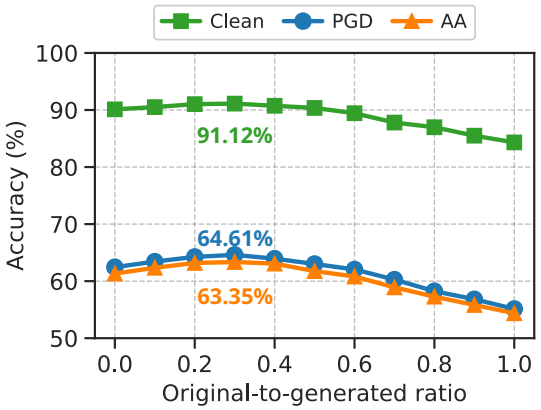


图*3*。干净准确率以及针对 PGD-40 和自动攻击（AA）的鲁棒准确率与原始数据和生成数据比例的关系（0 表示仅使用生成图像，1 表示仅使用 CIFAR-10 训练集）。我们使用 1M 生成数据，在 CIFAR-10 上针对 $(\ell_\infty, \epsilon = 8/255)$ 训练 WRN-28-10 模型。

表10. 在 CIFAR- 10 上，不同比例的**原始数据与生成数据比例**下，针对 $(\ell_\infty, \epsilon = 8/255)$ 的测试准确率（%）。WRN-28-10 模型使用 500 万个 EDM 生成数据进行训练。0.2 的结果始终优于 0.3。

批次	训练轮次	比例	干净样本	PGD-40	AA
512	400	0.2	91.15	64.97	64.25
		0.3	91.07	64.88	64.05
1024	800	0.2	91.87	66.43	65.53
		0.3	91.72	66.43	65.40
2048	1600	0.2	92.16	67.47	66.34
		0.3	91.88	67.19	66.29

B.2. CIFAR-100 和 SVHN 数据集的 FID

我们训练了自己的 EDM 模型，为 CIFAR-100 和 SVHN 实验生成图像。这些模型仅使用 CIFAR-100/SVHN 训练集进行训练。表11展示了在不同采样步数下，生成数据与 CIFAR-100/SVHN 训练集之间的 FID 分数。采样步数为 25 时，FID 表现最佳。

表 11. 使用不同**扩散模型采样步数** 生成的 50K EDM 数据与 CIFAR-100/SVHN 训练集之间的 Fr´echet inception distance (FID)。

Step	5	10	15	20	25	30	35	40
CIFAR-100	24.540	3.054	2.191	2.103	**2.090**	2.092	2.095	2.097
SVHN	125.765	5.458	1.661	1.405	**1.393**	1.410	1.428	1.445

B.3. 扩散模型实现细节的消融研究

我们提供了使用各种常微分方程采样器的扩散模型在 CIFAR-10 上的结果。我们使用 Song 等人(2021年b) 提供的无条件扩散模型检查点，并按照 Karras 等人(2022年) 进行消融研究。Song 等人(2021年b) 在常微分方程采样器中使用欧拉方法（即 DDIM 求解器，表12 （左）第 1 行）；而 Karras 等人(2022年) 发现，海伦二阶方法（即 EDM 求解器，第 2 行和第 3 行）能取得更好的结果。

为确保比较公平，由于 DDIM 求解器所需时间仅为 EDM 求解器的一半，我们采用其两倍的采样步数。表 12 （左）详细说明了生成 500 万张图像所需的时间，其中选取 100 万张图像用于训练。表 12 （左）表明，在生成时间相同的情况下，EDM 求解器能显著改善 FID；此外，EDM 求解器还能提升对抗训练中的干净样本准确率和鲁棒准确率。通过改进 EDM 超参数的选择，还能进一步提升模型的鲁棒性能。

我们还提供了方差保持（VP）和方差爆炸（VE）扩散模型的结果，这两种模型最初分别受到 DDPM（Ho 等人，2020年）和噪声条件得分网络（Song 与 Ermon，2019年）的启发。在 EDM 的实现中，VP 和 VE 模型的架构有所不同，分别采用 DDPM++ 和 NCSN++ 。主论文的实验均采用 VP 公式。表 12 （右）表明，VE 的 FID 与 VP 相当，这与以往的研究结果一致（Karras 等人，2022年）。两种公式得到的模型鲁棒性能相近。

表 12. 测试准确率（%） 在（$\ell_\infty, \epsilon = 8/255$）威胁模型下，使用不同的**采样器** （左）和**EDM 公式** （右） n CIFAR-10。采样步数在表（右）中设为 20。

采样器	Step	时间（小时）	FID	干净样本	PGD-40	AA
Song 等人(2021b)	40	3.75	10.06	86.97	61.49	60.53
+Heun 和 EDM t_i	20	3.62	4.03	87.82	62.17	61.29
+EDM $\sigma(t)$ & $s(t)$	20	3.62	2.98	88.09	62.36	61.46

	FID	干净样本	PGD-40	AA
VP	1.824	91.12	64.61	63.35
VE	1.832	91.11	64.53	63.29

B.4. 计算时间

我们提供了类别条件和无条件 EDM 使用不同采样步数生成 500 万张图像的运行时间。生成过程使用四块 A100 GPU 完成。如表 13所示，无条件生成的速度略快于类别条件生成，但稳健性能较低（表 8）。

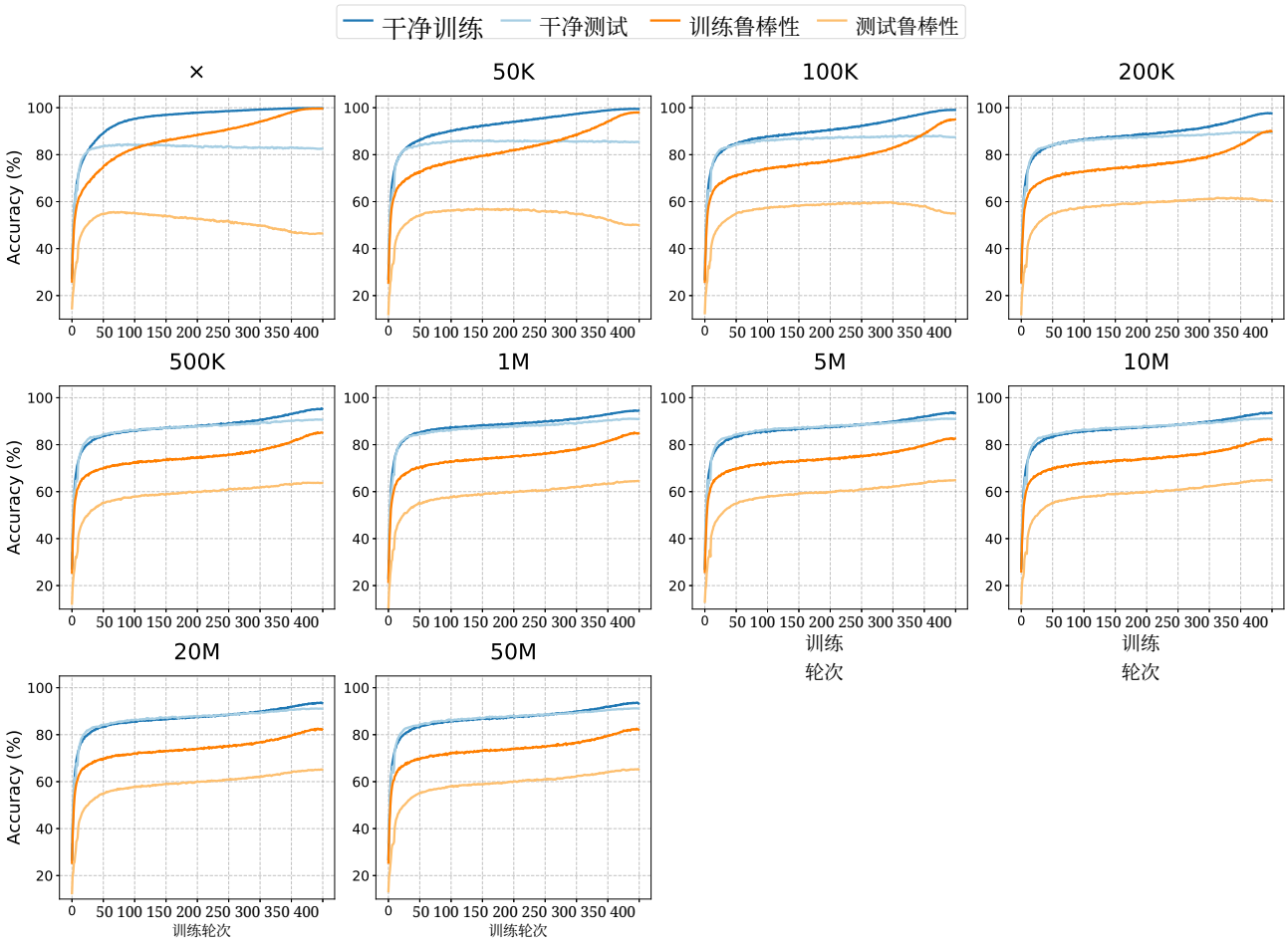
在对抗训练中，WRN-28-10 和 WRN-70-16 分别使用 4 块和 8 块 A100 GPU 并行处理。使用批量大小为 2048 的设置，在 WRN-28-10 模型上训练一个轮次平均需要 3.45 分钟。使用批量大小为 1024 的设置，在 WRN-70-16 模型上训练一个轮次平均需要 9.93 分钟。

表 13. 类别条件和无条件 EDM 生成 500 万条数据所需的时间（小时）。

Step	5	10	15	20	25	30	35	40
类别条件	2.59	5.37	8.13	10.92	13.72	16.47	19.26	22.03
无条件	2.52	5.30	8.05	10.81	13.57	16.40	19.18	21.95

B.5. 生成数据量

图 4 展示了使用不同数量生成数据时的干净样本准确率和 PGD 鲁棒准确率。随着生成数据量的增加，鲁棒过拟合显著减轻。在生成 50 万张图像后，继续增加生成图像已无法进一步缩小干净样本准确率和鲁棒准确率的泛化差距。这是意料之中的，因为模型容量过低，无法充分利用所有这些生成数据。



*Fi*图 *4.* 使用不同生成数据量进行对抗训练时的干净样本准确率和 PGD 鲁棒准确率 .