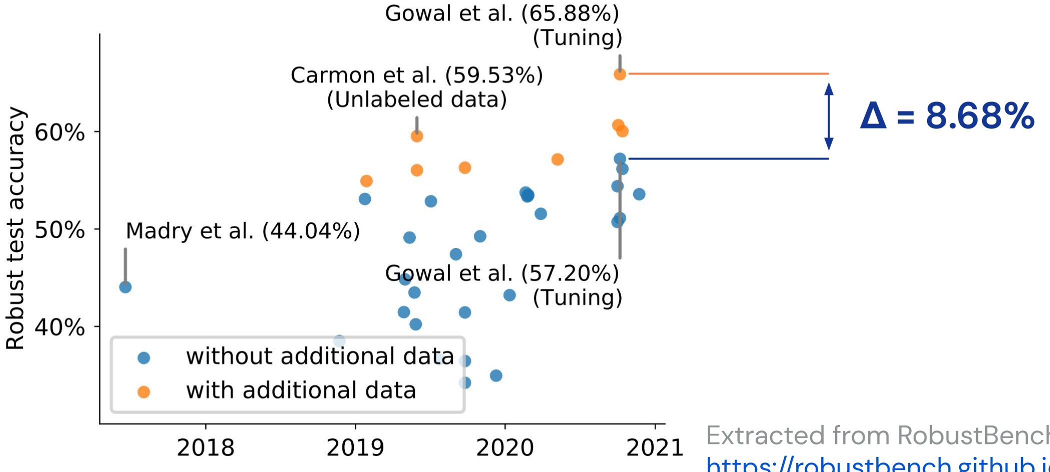
DeepMind

# Improving Robustness using Generated Data

Sven Gowal*, Sylvestre-Alvise Rebuffi*, Olivia Wiles,
Florian Stimberg, Dan A. Calian and Timothy Mann

# Motivation

- Robustness to adversarial perturbations requires substantially larger datasets [1].
- As such, many works [2, 3] use additional (unlabeled) data to improve robustness [4].




Extracted from RobustBench: [https://robustbench.github.io/](https://robustbench.github.io/)


[1] L. Schmidt et al., "Adversarially Robust Generalization Requires More Data," 2018.
[2] Y. Carmon et al., "Unlabeled data improves adversarial robustness," 2019.
[3] J. Uesato et al., "Are labels required for improving adversarial robustness?," 2019.
[4] F. Croce et al., "RobustBench: a standardized adversarial robustness benchmark," 2020.

DeepMind

# ~~Data augmentations?~~

DeepMind

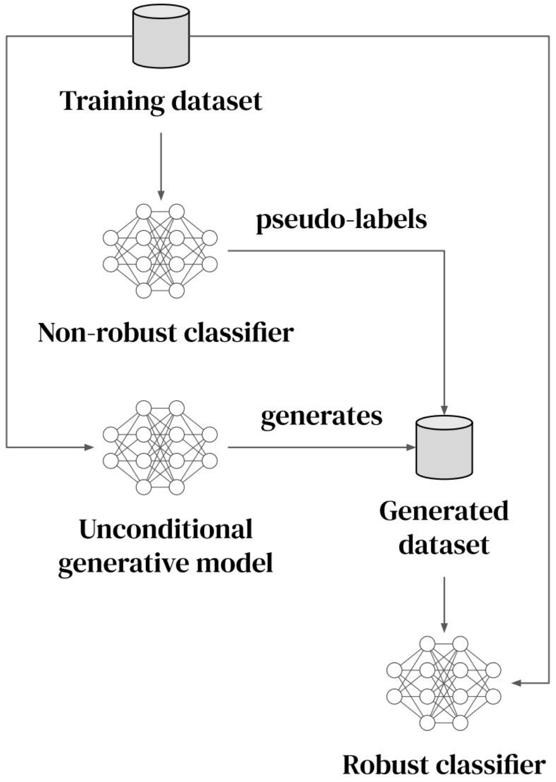
To improve **robust generalization**, it is critical to use additional training samples that are **diverse** and that **complement** the original training set

# Contributions

- We demonstrate that it is possible to use **low-quality random inputs to improve robustness** on CIFAR-10 against $L^\infty$ perturbations of size $\varepsilon = 8/255$.
- We describe **3 sufficient conditions** that explain this phenomenon and elaborate on the intricate relationship between generated data quality and classifier capacity.
- We leverage higher quality generated inputs (using generative models solely trained on the original data), and study three recent generative models: DDPM [5], StyleGAN2 [6], VD-VAE [7] and BigGAN [8].
- We show that images generated by the DDPM allow us to reach a robust accuracy of 66.10% on CIFAR-10 (**improvement of +8.96% upon SOTA***). The method generalizes to CIFAR-100, SVHN and TinyImageNet.

[5] J. Ho et al., “Denoising Diffusion Probabilistic Models,” 2020.
[6] T. Karras et al., “Analyzing and Improving the Image Quality of StyleGAN,” 2020.
[7] R. Child, “Very Deep VAEs Generalize Autoregressive Models and Can Outperform Them on Images,” 2021.
[8] A. Brock et al., “Large Scale GAN Training for High Fidelity Natural Image Synthesis,” 2019.
* Without using additional external data.

# Method

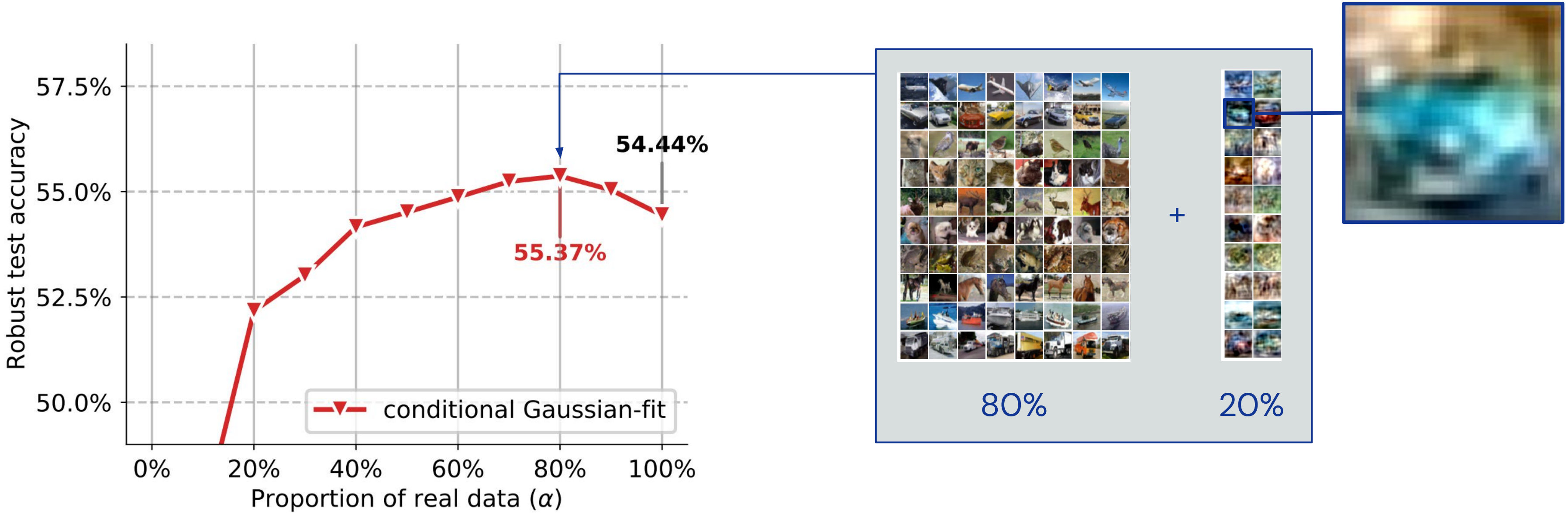



The method is general (beyond lp-norm) if:

- The non-robust classifier is accurate
- The generative model produces realistic inputs **OR** the robust classifier has enough capacity.

# Motivation

We take random samples generated by a conditional Gaussian fitted over the CIFAR-10 training set. We add these sample in different proportions while training.




Generated images only ⟷ Original training images only

# Sufficient conditions

Robustness can be improved if:

1 Accurate pseudo-labeling (i.e., labeling generated samples with high accuracy)

2 Generated and real data distributions are close

OR

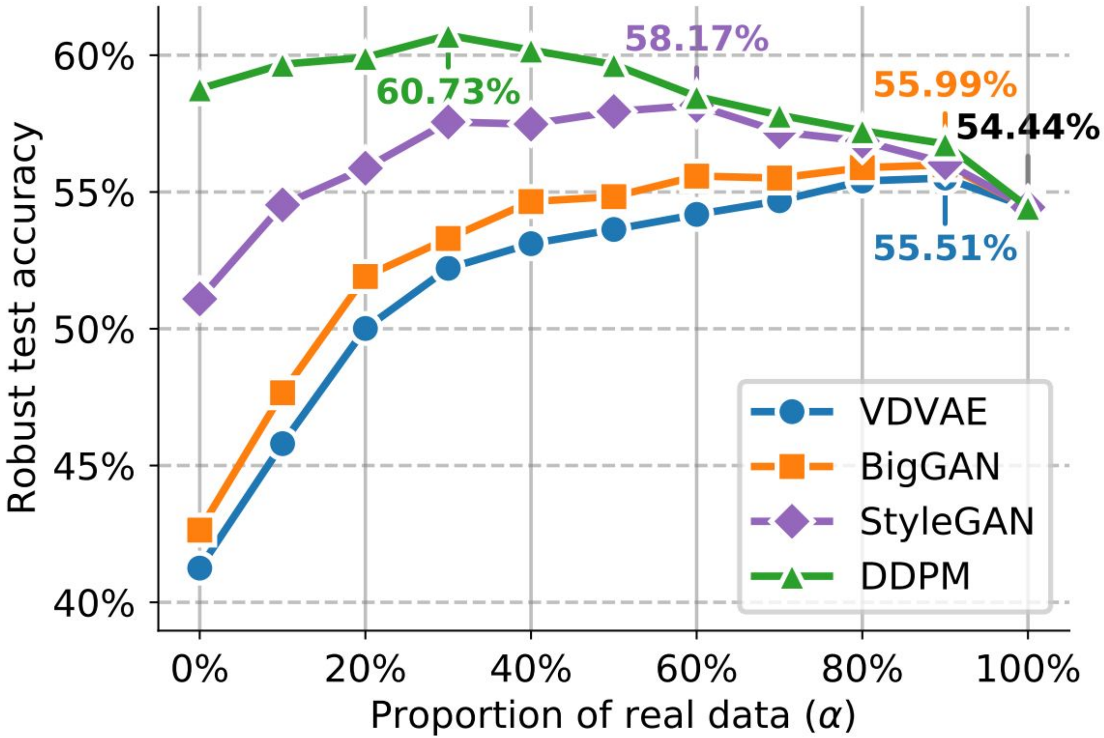
2 Adversarial attacks are unlikely (i.e., random samples are not frequently mis-classified by the pseudo-labeling classifier)

3 Generated samples should cover the manifold of real-images (i.e., there is non-zero chance of sampling a real images)

# Generated data is complementary

| Setup | Complementarity | | | Coverage | |
|---|---|---|---|---|---|
| | Train | Test | Self | Train | Test |
| *mixup* | 90.34% | 3.91% | 5.75% | 90.43% | 45.61% |
| Class-conditional Gaussian-fit | 0.13% | 0.22% | 99.65% | 12.36% | 12.24% |
| VDVAE | 11.97% | 12.14% | 75.89% | 34.20% | 33.76% |
| BigGAN | 14.97% | 14.81% | 70.22% | 38.86% | 39.06% |
| StyleGAN2 | 28.13% | 27.22% | 44.65% | 50.16% | 48.29% |
| DDPM | 29.29% | 29.17% | 41.54% | 49.07% | 49.10% |

# Better samples yield improved robustness (CIFAR-10, WRN-28-10)




Generated images only ⟷ Original training images only

# Results

| MODEL | DATASET | NORM | CLEAN | ROBUST | |
|---|---|---|---|---|---|
| Wu et al. [75] (WRN-34-10) | CIFAR-10 | $\ell_\infty$ | 85.36% | 56.17% | +15.68% |
| Gowal et al. [30] (WRN-70-16) | | | 85.29% | 57.14% | |
| Ours (DDPM) (WRN-28-10) | | | 85.97% | 60.73% | |
| Ours (DDPM) (WRN-70-16) | | | 86.94% | 63.58% | |
| Ours (100M DDPM)* (WRN-70-16) | | | **88.74%** | **66.10%** | |
| Wu et al. [75] (WRN-34-10) | CIFAR-10 | $\ell_2$ | 88.51% | 73.66% | +5.11% |
| Gowal et al. [30] (WRN-70-16) | | | **90.90%** | 74.50% | |
| Ours (DDPM) (WRN-28-10) | | | 90.24% | 77.37% | |
| Ours (DDPM) (WRN-70-16) | | | 90.83% | **78.31%** | |
| Cui et al. [20] (WRN-34-10) | CIFAR-100 | $\ell_\infty$ | 60.64% | 29.33% | +11.52% |
| Gowal et al. [30] (WRN-70-16) | | | **60.86%** | 30.03% | |
| Ours (DDPM) (WRN-28-10) | | | 59.18% | 30.81% | |
| Ours (DDPM) (WRN-70-16) | | | 60.46% | **33.49%** | |
| Ours (without DDPM) (WRN-28-10) | SVHN | $\ell_\infty$ | 92.87% | 56.83% | +7.16% |
| Ours (DDPM) (WRN-28-10) | | | **94.15%** | **60.90%** | |
| Ours (without DDPM) (WRN-28-10) | TINYIMAGENET | $\ell_\infty$ | 51.56% | 21.56% | +23.65% |
| Ours (DDPM) (WRN-28-10) | | | **60.95%** | **26.66%** | |

DeepMind

- More experiments are in the paper:
  - https://openreview.net/forum?id=0NXUSlb6oEu
- Code, data and pre-trained models are available online.
  - [JAX] https://github.com/deepmind/deepmind-research/tree/master/adversarial_robustness
  - [PyTorch] https://github.com/imrahulr/adversarial_robustness_pytorch (kindly reproduced by Rahul Rade)

# Thank you!

sgowal@deepmind.com