GREAT SCORE: GLOBAL ROBUSTNESS EVALUATION OF ADVERSARIAL PERTURBATION USING GENERATIVE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Current studies on adversarial robustness mainly focus on aggregating *local* robustness results from a set of data samples to evaluate and rank different models. However, the local statistics may not well represent the true *global* robustness of the underlying unknown data distribution. To address this challenge, this paper makes the first attempt to present a new framework, called *GREAT Score*, for global robustness evaluation of adversarial perturbation using generative models. Formally, GREAT Score carries the physical meaning of a global statistic capturing a mean certified attack-proof perturbation level over all samples drawn from a generative model. For finite-sample evaluation, we also derive a probabilistic guarantee on the sample complexity and the difference between the sample mean and the true mean. GREAT Score has several advantages: (1) Robustness evaluations using GREAT Score are efficient and scalable to large models, by sparing the need of running adversarial attacks. In particular, we show high correlation and significantly reduced computation cost of GREAT Score when compared to the attack-based model ranking on RobustBench (Croce et al., 2021). (2) The use of generative models facilitates the approximation of the unknown data distribution. In our ablation study with different generative adversarial networks (GANs), we observe consistency between global robustness evaluation and the quality of GANs. (3) GREAT Score can be used for remote auditing of privacy-sensitive black-box models, as demonstrated by our robustness evaluation on several online facial recognition services.

1 INTRODUCTION

Adversarial robustness is the study of model performance in the worst-case scenario, which is a key element in trustworthy machine learning. Without further remediation, state-of-the-art machine learning models, especially neural networks, are known to be overly sensitive to small human-imperceptible perturbations to data inputs (Goodfellow et al., 2014b). Such a property of over-sensitivity could be exploited by bad actors to craft adversarial perturbations leading to prediction-evasive adversarial examples.

Given a threat model specifying the knowledge of the target machine learning model (e.g., white-box or black-box model access) and the setting of plausible adversarial interventions (e.g., norm-bounded input perturbations), the methodology for adversarial robustness evaluation can be divided into two categories: *attack-dependent* and *attack-independent*. Attack-dependent approaches aim to devise the strongest possible attack and use it for performance assessment. A typical example is Auto-Attack (Croce & Hein, 2020), a state-of-the-art attack based on an ensemble of advanced white-box and black-box adversarial perturbation methods. On the other hand, attack-independent approaches aim to develop a certified or estimated score for adversarial robustness, reflecting a quantifiable level of attack-proof certificate. Typical examples include neural network verification techniques (Wong & Kolter, 2018; Zhang et al., 2018), certified defenses such as randomized smoothing (Cohen et al., 2019), and local Lipschitz constant estimation (Weng et al., 2018).

Despite a plethora of adversarial robustness evaluation methods, current studies primarily focus on aggregating *local* robustness results from a set of data samples. However, the sampling process of

these test samples could be biased and unrepresentative of the true *global* robustness of the underlying data distribution, resulting in the risk of incorrect or biased robustness benchmarks. For instance, we find that when assessing the ranking of Imagenet models through Robustbench (Croce et al., 2020), using AutoAttack (Croce & Hein, 2020) with 10,000 randomly selected samples (the default choice) with 100 independent trials results in an unstable ranking coefficient of 0.907 ± 0.0256 when compared to that of the entire 50,000 test samples. This outcome affirms that AutoAttack's model ranking has notable variations with an undersampled or underrepresented test dataset.

An ideal situation is when the data distribution is transparent and one can draw an unlimited number of samples from the true distribution for reliable robustness evaluation. But in reality, the data distribution is unknown and difficult to characterize. In addition to lacking rigorous global robustness evaluation, many attack-independent methods are limited to the white-box setting, requiring detailed knowledge about the target model (e.g., model parameters and architecture) such as input gradients and internal data representations for robustness evaluation. Moreover, state-of-the-art attack-dependent and attack-independent methods often face the issue of scalability to large models and data volumes due to excessive complexity, such as the computational costs in iterative gradient computation and layer-wise interval bound propagation and relaxation (Katz et al., 2017; Gowal et al., 2019).

To address the aforementioned challenges including (i) lack of proper global adversarial robustness evaluation, (ii) limitation to white-box settings, and (iii) computational inefficiency, in this paper we present a novel attack-independent evaluation framework called *GREAT Score*, which is short for global robustness evaluation of adversarial perturbation using generative models. We tackle challenge (i) by using a generative model such as a generative adversarial network (GAN) (Goodfellow et al., 2014a; 2020) or a diffusion model (Ho et al., 2020) as a proxy of the true unknown data distribution. Formally, GREAT score is defined as the mean of a certified lower bound on minimal adversarial perturbation over the data sampling distribution of a generative model, which represents the global distribution-wise adversarial robustness with respect to the generative model in use. It entails a global statistic capturing the mean certified attack-proof perturbation level over all samples from a generative model. For finite-sample evaluation, we also derive a probabilistic guarantee quantifying the sample complexity and the difference between the sample mean and true mean.

For challenge (ii), our derivation of GREAT score leads to a neat closed-form solution that only requires data forward-passing and accessing the model outputs, which applies to any black-box classifiers giving class prediction confidence scores as model output. Moreover, as a byproduct of using generative models, our adversarial robustness evaluation procedure is executed with only synthetically generated data instead of real data, which is particularly appealing to privacy-aware robustness assessment schemes, e.g., remote robustness evaluation or auditing by a third party with restricted access to data and model. We will present how GREAT score can be used to assess the robustness of online black-box facial recognition models. Finally, for challenge (iii), GREAT score is applicable to any off-the-self generative models so that we do not take the training cost of generative models into consideration. Furthermore, the computation of GREAT score is lightweight because it scales linearly with the number of data samples used for evaluation, and each data sample only requires one forward pass through the model to obtain the final predictions.

We highlight our main contributions as follows:

- We present GREAT Score as a novel framework for deriving a global statistic representative of the distribution-wise robustness to adversarial perturbation, based on an off-the-shelf generative model for approximating the data generation process.
- Theoretically, we show that GREAT Score corresponds to a mean certified attack-proof level of $\mathcal{L}_2$-norm bounded input perturbation over the sampling distribution of a generative model (Theorem 1). We further develop a formal probabilistic guarantee on the quality of using the sample mean as GREAT Score with a finite number of samples from generative models (Theorem 2).
- We evaluate the effectiveness of GREAT Score on all neural network models on RobustBench (Croce et al., 2020) (the largest adversarial robustness benchmark), with a total of 17 models on CIFAR-10 and 5 models on ImageNet. We show that the model ranking of GREAT score is highly aligned with that of the original ranking on RobustBench using AutoAttack (Croce & Hein, 2020), while GREAT Score significantly reduces the computation time. Specifically, on CIFAR-10 the computation complexity can be reduced by up to 2,000 times. The results suggest that GREAT score is a competitive and computationally-efficient alternative for adversarial robustness evaluation.

- As a demonstration of GREAT Score's capability for remote robustness evaluation of access-limited systems, we show how Great Score can audit several online black-box facial recognition APIs.

2 Background and Related Works

Adversarial Attack and Defense. Adversarial attacks aim to generate examples that can evade classifier predictions in classification tasks. In principle, adversarial examples can be crafted by small perturbations to a native data sample, where the level of perturbation is measured by different $\mathcal{L}_p$ norms (Szegedy et al., 2014; Carlini & Wagner, 2017; Chen et al., 2018). The procedure of finding adversarial perturbation within a perturbation level is often formulated as a constrained optimization problem, which can be solved by algorithms such as projected gradient descent (PGD) (Madry et al., 2018). The state-of-the-art adversarial attack is the Auto-Attack (Croce & Hein, 2020), which uses an ensemble of white-box and black-box attacks. There are many methods (defenses) to improve adversarial robustness. A popular approach is adversarial training (Madry et al., 2018), which generates adversarial perturbation during model training for improved robustness. One common evaluation metric for adversarial robustness is robust accuracy, which is defined as the accuracy of correct classification under adversarial attacks, evaluated on a set of data samples. RobustBench (Croce & Hein, 2020) is the largest-scale standardized benchmark that ranks the models using robust accuracy against Auto-Attack on test sets from image classification datasets such as CIFAR-10 .

Generative Models. Statistically speaking, let X denote the observable variable and let Y denote the corresponding label, the learning objective for a generative model is to model the conditional probability distribution $P(X \mid Y)$. Among all the generative models, GANs have gained a lot of attention in recent years due to their capability to generate realistic high-quality images (Goodfellow et al., 2020). The principle of training GANs is based on the formulation of a two-player zero-sum min-max game to learn the high-dimension data distribution. Eventually, these two players reach the Nash equilibrium that D is unable to further discriminate real data versus generated samples. This adversarial learning methodology aids in obtaining high-quality generative models. In practice, the generator $G(\cdot)$ takes a random vector z (i.e., a latent code) as input, which is generated from a zero-mean isotropic Gaussian distribution denoted as $z \sim \mathcal{N}(0, I)$, where I means an identity matrix. Conditional GANs refer to the conditional generator $G(\cdot|Y)$ given a class label Y. In addition to GAN, diffusion models (DMs) are also gaining popularity. DMs consist of two stages: the forward diffusion process and the reverse diffusion process. In the forward process, the input data is gradually perturbed by Gaussian Noises and becomes an isotropic Gaussian distribution eventually. In the reverse process, DMs reverse the forward process and implement a sampling process from Gaussian noises to reconstruct the true samples by solving a stochastic differential equation. In our proposed framework, we use off-the-shelf (conditional) GANs and DMs (e.g., DDPM (Ho et al., 2020)) that are publicly available as our generative models.

Formal Local Robustness Guarantee and Estimation. Given a data sample x, a formal local robustness guarantee refers to a certified range on its perturbation level such that within which the top-1 class prediction of a model will remain unchanged (Hein & Andriushchenko, 2017). In $\mathcal{L}_p$-norm ($p \geq 1$) bounded perturbations centered at x, such a guarantee is often called a certified radius r such that any perturbation δ to x within this radius (i.e., $\|\delta\|_p \leq r$) will have the same top-1 class prediction as x. Therefore, the model is said to be provably locally robust (i.e., attack-proof) to any perturbations within the certified radius r. By definition, the certified radius of x is also a lower bound on the minimal perturbation required to flip the model prediction.

Among all the related works on attack-independent local robustness evaluations, the CLEVER framework proposed in (Weng et al., 2018) is the closest to our study. The authors in (Weng et al., 2018) derived a closed-form of certified local radius involving the maximum local Lipschitz constant of the model output with respect to the data input around a neighborhood of a data sample x. They then proposed to use extreme value theory to estimate such a constant and use it to obtain a local robustness score, which is not a certified local radius. Our proposed GREAT score has major differences from (Weng et al., 2018) in that our focus is on global robustness evaluation, and our GREAT score is the mean of a certified radius over the sampling distribution of a generative model. In addition, for every generated sample, our local estimate gives a certified radius.

Notations. All the main notations used in the paper are summarized in Appendix 6.1.

3 GREAT SCORE: METHODOLOGY AND ALGORITHMS

3.1 TRUE GLOBAL ROBUSTNESS AND CERTIFIED ESTIMATE

Let $f = [f_1, \ldots, f_K] : \mathbb{R}^d \to \mathbb{R}^K$ denote a fixed K-way classifier with flattened data input of dimension d, (x, y) denote a pair of data sample x and its corresponding groundtruth label $y \in \{1, \ldots, K\}$, P denote the true data distribution which in practice is unknown, and $\Delta_{\min}(x)$ denote the minimal perturbation of a sample-label pair $(x, y) \sim P$ causing the change of the top-1 class prediction such that $\arg\max_{k\in\{1,\ldots,K\}} f_k(x + \Delta_{\min}(x)) \neq \arg\max_{k\in\{1,\ldots,K\}} f_k(x)$. Note that if the model f makes an incorrect prediction on x, i.e., $y \neq \arg\max_{k\in\{1,\ldots,K\}} f_k(x)$, then we define $\Delta_{\min}(x) = 0$. This means the model is originally subject to prediction evasion on x even without any perturbation. A higher $\Delta_{\min}(x)$ means better local robustness of f on x.

The following statement defines the true global robustness of a classifier f based on the probability density function $p(\cdot)$ of the underlying data distribution P.

Definition 1 (True global robustness **w.r.t.** P)**.** *The true global robustness of a classifier f with respect to a data distribution P is defined as:*

$$\Omega(f) = \mathbb{E}_{x\sim P}[\Delta_{min}(x)] = \int_{x\sim P} \Delta_{\min}(x)p(x)dx \tag{1}$$

Unless the probability density function of P and every local minimal perturbation are known, the exact value of the true global robustness cannot be computed. An alternative is to estimate such a quantity. Extending Definition 1, let $g(x)$ be a local robustness statistic. Then the corresponding global robustness estimate is defined as

$$\widehat{\Omega}(f) = \mathbb{E}_{x\sim P}[g(x)] = \int_{x\sim P} g(x)p(x)dx \tag{2}$$

Furthermore, if one can prove that $g(x)$ is a valid lower bound on $\Delta_{min}(x)$ such that $g(x) \leq \Delta_{min}(x)$, $\forall x$, then the estimate $\widehat{\Omega}(f)$ is said to be a certified lower bound on the true global robustness with respect to P, and larger $\widehat{\Omega}(f)$ will imply better true global robustness. In what follows, we will formally introduce our proposed GREAT Score and show that it is a certified estimate of the lower bound on the true robustness with respect to the data-generating distribution learned by a generative model.

3.2 USING GMS TO EVALUATE GLOBAL ROBUSTNESS

Recall that a generative model (GM) takes a random vector $z \sim \mathcal{N}(0, I)$ sampled from a zero-mean isotropic Gaussian distribution as input to generate a data sample $G(z)$. In what follows, we present our first main theorem that establishes a certified lower bound $\widehat{\Omega}(f)$ on the true global robustness of a classifier f measured by the data distribution given by $G(\cdot)$.

Without loss of generality, we assume that all data inputs are confined in the scaled data range $[0, 1]^d$, where d is the size of any flattened data input. The K-way classifier $f : [0, 1]^d \mapsto \mathbb{R}^K$ takes a data sample x as input and outputs a K-dimensional vector $f(x) = [f_1(x), \ldots, f_K(x)]$ indicating the likelihood of its prediction on x over K classes, where the top-1 class prediction is defined as $\hat{y} = \arg\max_{k=\{1,\ldots,K\}} f_k(x)$. We further denote c as the groundtruth class of x. Therefore, if $\hat{y} \neq c$, then the classifier is said to make a wrong top-1 prediction. When considering the adversarial robustness on a wrongly classified sample x, we define the minimal perturbation for altering model prediction as $\Delta_{\min}(x) = 0$. The intuition is that an attacker does not need to take any action to make the sample x evade the correct prediction by f, and therefore the required minimal adversarial perturbation level is 0 (i.e., zero robustness).

Given a generated data sample $G(z)$, we now formally define a local robustness score function as

$$g\left(G(z)\right) = \sqrt{\frac{\pi}{2}} \cdot \max\{f_c(G(z)) - \max_{k\in\{1,\ldots,K\}, k\neq c} f_k(G(z)), 0\} \tag{3}$$

The scalar $\sqrt{\pi/2}$ is a constant associated with the sampling Gaussian distribution of G, which will be apparent in later analysis. We further offer several insights into understanding the physical meaning of the considered local robustness score in (3): (i) The inner term $f_c(G(z)) -$

$\max_{k\in\{1,\dots,K\},k\neq c} f_k(G(z))$ represents the gap in the likelihood of model prediction between the correct class c and the most likely class other than c. A positive and larger value of this gap reflects higher confidence of the correct prediction and thus better robustness. (ii) Following (i), a negative gap means the model is making an incorrect prediction, and thus the outer term $\max\{\text{gap}, 0\} = 0$, which corresponds to zero robustness.

Next, we use the local robustness score g defined in (3) to formally state our theorem on establishing a certified lower bound on the true global robustness and the proof sketch.

Theorem 1 (certified global robustness estimate)**.** *Let $f : [0,1]^d \mapsto [0,1]^K$ be a K-way classifier and let $f_k(\cdot)$ be the predicted likelihood of class k, with c denoting the groundtruth class. Given a generator G such that it generates a sample $G(z)$ with $z \sim \mathcal{N}(0, I)$. Define $g(G(z)) = \sqrt{\frac{\pi}{2}} \cdot \max\{f_c(G(z)) - \max_{k\in\{1,\dots,K\},k\neq c} f_k(G(z)), 0\}$. Then the global robustness estimate of f evaluated with $\mathcal{L}_2$-norm bounded perturbations, defined as $\widehat{\Omega}(f) = \mathbb{E}_{z\sim\mathcal{N}(0,I)}[g(G(z))]$, is a certified lower bound of the true global robustness $\Omega(f)$ with respect to G.*

The complete proof is given in Appendix 6.4.

3.3 Probabilistic Guarantee on Sample Mean

As defined in Theorem 1, the global robustness estimate $\widehat{\Omega}(f) = \mathbb{E}_{z\sim\mathcal{N}(0,I)}[g(G(z))]$ is the mean of the local robustness score function introduced in (3) evaluated through a generator G and its sampling distribution. In practice, one can use a finite number of samples $\{G(z_i|y_i)\}_{i=1}^n$ generated from a conditional generator $G(\cdot|y)$ to estimate $\widehat{\Omega}(f)$, where y denotes a class label and it is also an input parameter to the conditional generator. The simplest estimator of $\widehat{\Omega}(f)$ is the sample mean, defined as

$$\widehat{\Omega}_S(f) = \frac{1}{n}\sum_{i=1}^{n} g(G(z_i|y_i)) \tag{4}$$

In what follows, we present our second main theorem to deliver a probabilistic guarantee on the sample complexity to achieve ϵ difference between the sample mean $\widehat{\Omega}_S(f)$ and the true mean $\widehat{\Omega}(f)$.

Theorem 2 (probabilistic guarantee on sample mean)**.** *Let f be a K-way classifier with its outputs bounded by $[0,1]^K$ and let e denote the natural base. For any $\epsilon, \delta > 0$, if the sample size $n \geq \frac{32e\cdot\log(2/\delta)}{\epsilon^2}$, then with probability at least $1-\delta$, the sample mean $\widehat{\Omega}_S(f)$ is ϵ-close to the true mean $\widehat{\Omega}(f)$. That is, $|\widehat{\Omega}_S(f) - \widehat{\Omega}(f)| \leq \epsilon$.*

The complete proof is given in Appendix 6.5. The proof is built on a concentration inequality in (Maurer & Pontil, 2021). It is worth noting that the bounded output assumption of the classifier f in Theorem 2 can be easily satisfied by applying a normalization layer at the final model output, such as the softmax function or the element-wise sigmoid function.

3.4 Algorithm and Computational Complexity

To conclude this section, the Algorithm 1 in Appendix 6.6 summarizes the procedure of computing GREAT Score using the sample mean estimator. It can be seen that the computation complexity of GREAT Score is linear in the number of generated samples N_S, and for each sample, the computation of the statistic g defined in (3) only requires drawing a sample from the generator G and taking a forward pass to the classifier f to obtain the model predictions on each class. As a byproduct, GREAT Score applies to the setting when the classifier f is a black-box model, meaning only the model outputs are observable by an evaluator.

3.5 Calibrated GREAT Score

In cases when one has additional knowledge of adversarial examples on a set of images from a generative model, e.g., successful adversarial perturbations (an upper bound on the minimal perturbation of each sample) returned by any norm-minimization adversarial attack method such as

the CW attack (Carlini & Wagner, 2017), The CW attack employs two loss terms, classification loss and distance metric, to generate adversarial examples. See Appendix 6.7 for details. we can further "calibrate" the Great Score with respect to the available perturbations. Moreover, since Theorem 1 informs some design choices on the model output layer, as long as the model output is a non-negative K-dimensional vector $f \in [0,1]^K$ reflecting the prediction confidence over K classes, we will incorporate such flexibility in the calibration process.

Specifically, we use calibration in the model ranking setup where there are M models $\{f^{(j)}\}_{j=1}^M$ for evaluation, and each model (indexed by j) has a set of known perturbations $\{\delta_i^{(j)}\}_{i=1}^N$ on a common set of N image-label pairs $\{x_i, y_i\}_{i=1}^N$ from the same generative model. We further consider four different model output layer designs (that are attached to the model logits): (i) $\mathsf{sigmoid}(\cdot|T_1)$: sigmoid with temperature T_1, (ii) $\mathsf{softmax}(\cdot|T_2)$: softmax with temperature T_2, (iii) $\mathsf{sigmoid}(\mathsf{softmax}(\cdot|T_2 = 1)|T_1)$: sigmoid with temperature after softmax, and (iv) $\mathsf{softmax}(\mathsf{sigmoid}(\cdot|T_1 = 1)|T_2)$: softmax with temperature after sigmoid. Finally, let $\{\widehat{\Omega}_S(f^{(j)})\}_{j=1}^M$ denote the Great Score computed based on $\{x_i, y_i\}_{i=1}^N$ for each model. We calibrate Great Score by optimizing some rank statistics (e.g., the Spearman's rank correlation coefficient) over the temperature parameter by comparing the ranking consistency between $\{\widehat{\Omega}_S(f^{(j)})\}_{j=1}^M$ and $\{\delta_i^{(j)}\}_{i=1}^N$. In our experiments, we find that setting (iv) gives the best result and use it as the default setup for calibration, as detailed in Appendix 6.8.

4 Experimental Results

4.1 Experiment Setup

Datasets and Models. We conduct our experiment on several datasets including CIFAR-10 (Krizhevsky et al., 2009), ImageNet-1K (Deng et al., 2009) and CelebA-HQ (Karras et al., 2018)/CelebA (Liu et al., 2015). For neural network models, we use the available models on RobustBench (Croce et al., 2020) (see more details in the next paragraph), which includes 17/5 models on CIFAR-10/ImageNet, correspondingly. We also use several off-the-shelf GANs and diffusion models (DMs) trained on CIFAR-10 and ImageNet for computing GREAT Score in an ablation study (we defer the model details to later paragraphs).

Summary of Classifiers on RobustBench. The RobustBench (RobustBench) is to-date the largest benchmark for robustness evaluation with publicly accessible neural network models submitted by contributors. RobustBench uses the default test dataset from several standard image classification tasks, such as CIFAR-10 and ImageNet-1K, to run Auto-Attack (Croce & Hein, 2020) and report the resulting accuracy with $\mathcal{L}_2$-norm and $\mathcal{L}_\infty$-norm perturbations (i.e., the robust accuracy – RA) as a metric for adversarial robustness. Even under one perturbation type, it is not easy to make a direct and fair comparison among all submitted models on RobustBench because they often differ by the training scheme, network architecture, as well as the usage of additional real and/or synthetic data. To make a meaningful comparison with GREAT Score, we select all non-trivial models (having non-zero RA) submitted to the CIFAR-10 and ImageNet-1K benchmarks and evaluated with $\mathcal{L}_2$-norm perturbation with a fixed perturbation level of 0.5 using Auto-Attack. We list the model names in Table 1 and provide their descriptions in Appendix 6.9.

GANs and DMs. We used off-the-shelf GAN models provided by StudioGAN (Minguk Kang & Park, 2021), a library containing released GAN models. StudioGAN also reports the Inception Score (IS) to rank the model quality. We use the GAN model with the highest IS value as our default GAN for GREAT Score, which are StyleGAN2 (Karras et al., 2020)/ BigGAN (Brock et al., 2019) for CIFAR-10 /ImageNet with IS = $10.477/99.705$, respectively. For the ablation study of using different generative models in GREAT Score (Section 4.4), we also use the following GAN/DM models: LSGAN (Mao et al., 2017), GGAN (Lim & Ye, 2017), SAGAN (Zhang et al., 2019), SNGAN (Miyato et al., 2018), DDPM (Ho et al., 2020) and StyleGAN2 (Karras et al., 2020).

GREAT Score implementation. The implementation follows Algorithm 1 in Appendix 6.6 with a sigmoid/softmax function on the logits of the CIFAR-10/ImageNet classifier to ensure the model output of each dimension is within $[0,1]$, as implied by Theorem 1. As ImageNet-1K has 1000 classes, applying sigmoid will make the robustness score function in (3) degenerate. We use softmax instead. 500 samples drawn from a generative model were used for computing GREAT Score.

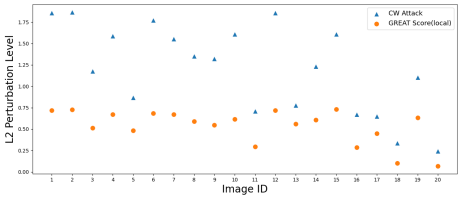

Figure 1: Comparison of local GREAT Score and CW attack in $\mathcal{L}_2$ perturbation on CIFAR-10 with Rebuffi_extra model (Rebuffi et al., 2021). The x-axis is the image id. The result shows the local GREAT Score is indeed a lower bound of the perturbation level found by CW attack.

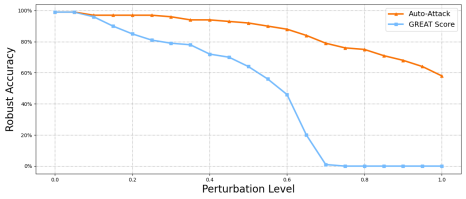

Figure 2: Cumulative robust accuracy (RA) with varying $\mathcal{L}_2$ perturbation level using 500 samples. Note that GREAT Score gives a certified RA for attack-proof robustness.

Comparative methods. We compare the effectiveness of GREAT Score in two objectives: robustness ranking (global robustness) and per-sample perturbation. For the former, we compare the RA reported in RobustBench on the test dataset (named RobustBench Accuracy) as well as the RA of Auto-Attack on the generated data samples (named AutoAttack Accuracy). For the latter, we report the RA of Auto-Attack in $\mathcal{L}_2$-norm with a fixed perturbation level of 0.5.

Evaluation metrics. For robustness ranking, we report Spearman's rank correlation coefficient between two sets of model rankings (e.g., GREAT Score v.s. RobustBench Accuracy). A value closer to 1 means higher consistency. Robust accuracy refers to the fraction of correctly classified samples against adversarial perturbations.

Calibration Method. We run $\mathcal{L}_2$-norm CW attack (Carlini & Wagner, 2017) (with learning rate 0.005 and 200 iterations) on each generated data sample to find the minimal adversarial perturbation. Then, we use grid search in the range [0,2] with an interval of 0.00001 to find temperature value maximizing the Spearmans' rank correlation coefficient between Great Score and CW attack distortion.

Compute Resources. All our experiments were run on a GTX 2080 Ti GPU with 12GB RAM.

4.2 Local and Global Robustness Analysis

Recall from Theorem 1 that the local robustness score proposed in (3) gives a certified perturbation level for generated samples from a generative model. To verify this claim, we randomly select 20 generated images on CIFAR-10 and compare their local certified perturbation level to the perturbation found by the CW attack (Carlini & Wagner, 2017) using the Rebuffi_extra model (Rebuffi et al., 2021). Rebuffi et al. (Rebuffi et al., 2021) proposed a fixing data augmentation method such as using CutMix (Yun et al., 2019) and GANs to prevent over-fitting. Figure 1 shows the perturbation level of local GREAT Score in (3) and that of the corresponding CW attack per sample. We can see that the local GREAT Score is a lower bound of CW attack, as the CW attack finds a successful adversarial perturbation that is no smaller than the minimal perturbation $\Delta_{\min}$ (i.e., an over-estimation). The true $\Delta_{\min}$ value lies between these lower and upper bounds.

In Figure 2, we compare the cumulative robust accuracy (RA) of GREAT Score and Auto-Attack over 500 samples by sweeping the $\mathcal{L}_2$ perturbation level from 0 to 1 with a 0.05 increment for Auto-Attack. The cumulative RA of GREAT Score at a perturbation level r means the fraction of samples with local GREAT scores greater than r, which gives an attack-proof guarantee that no attacks can achieve a lower RA at the same perturbation level. We see that the trend of attack-independent certified robustness (GREAT Score) is similar to that of empirical attacks (Auto-Attack). The gap between our certified curve and the empirical curve of AutoAttack does not necessarily indicate inferiority, it could mean that there exist undiscovered adversarial examples at higher perturbation radii.

Table 1 compares the global robustness statistics of the 17 grouped CIFAR-10 models on RobustBench for uncalibrated and calibrated versions respectively, in terms of the GREAT Score and the average distortion of CW attack, which again verifies GREAT Score is a certified lower bound on the true global robustness (see its definition in Section 3.1), while any attack with 100% attack success rate only gives an upper bound on the true global robustness. We also observe that calibration can indeed enlarge the Great Score and tighten its gap to the distortion of CW attack.

Table 1: Comparison of (Calibrated) GREAT Score v.s. minimal distortion found by CW attack (Carlini & Wagner, 2017) on CIFAR-10. The results are averaged over 500 samples from StyleGAN2.

Model Name	RobustBench Accuracy(%)	AutoAttack Accuracy(%)	GREAT Score	Calibrated GREAT Score	CW Distortion
Rebuffi_extra Rebuffi et al. (2021)	82.32	87.20	0.507	1.216	1.859
Gowal_extra Gowal et al. (2020)	80.53	85.60	0.534	1.213	1.324
Rebuffi_70_ddpm Rebuffi et al. (2021)	80.42	90.60	0.451	1.208	1.943
Rebuffi_28_ddpm Rebuffi et al. (2021)	78.80	90.00	0.424	1.214	1.796
Augustin_WRN_extra Augustin et al. (2020)	78.79	86.20	0.525	1.206	1.340
Sehwag Sehwag et al. (2021)	77.24	89.20	0.227	1.143	1.392
Augustin_WRN Augustin et al. (2020)	76.25	86.40	0.583	1.206	1.332
Rade Rade & Moosavi-Dezfooli (2021)	76.15	86.60	0.413	1.200	1.486
Rebuffi_R18Rebuffi et al. (2021)	75.86	87.60	0.369	1.210	1.413
Gowal Gowal et al. (2020)	74.50	86.40	0.124	1.116	1.253
Sehwag_R18 Sehwag et al. (2021)	74.41	88.60	0.236	1.135	1.343
Wu2020Adversarial Wu et al. (2020)	73.66	84.60	0.128	1.110	1.369
Augustin2020Adversarial Augustin et al. (2020)	72.91	85.20	0.569	1.199	1.285
Engstrom2019Robustness Engstrom et al. (2019)	69.24	82.20	0.160	1.020	1.084
Rice2020Overfitting Rice et al. (2020)	67.68	81.80	0.152	1.040	1.097
Rony2019Decoupling Rony et al. (2019)	66.44	79.20	0.275	1.101	1.165
Ding2020MMA Ding et al. (2018)	66.09	77.60	0.112	0.909	1.095

Table 2: Spearman's rank correlation coefficient on CIFAR-10 using GREAT Score, RobustBench (with test set), and Auto-Attack (with generated samples).

	Uncalibrated	Calibrated
GREAT Score vs. RobustBench Correlation	0.6618	0.8971
GREAT Score vs. AutoAttack Correlation	0.3690	0.6941
RobustBench vs. AutoAttack Correlation	0.7296	0.7296

4.3 Model Ranking on CIFAR-10 and ImageNet

Following the experiment setup in Section 4.1, we compare the model ranking on CIFAR-10 using GREAT Score (evaluated with generated samples), RobustBench (evaluated with Auto-Attack on the test set), and Auto-Attack (evaluated with Auto-Attack on generated samples). Table 2 presents their mutual rank correlation (higher value means more aligned ranking) with calibrated and uncalibrated versions. We note that there is an innate discrepancy between Spearman's rank correlation coefficient (way below 1) of RobustBench v.s. Auto-Attack, which means Auto-Attack will give inconsistent model rankings when evaluated on different data samples. In addition, GREAT Score measures *classification margin*, while AutoAttack measures *accuracy* under a fixed perturbation budget ϵ. AutoAttack's ranking will change if we use different ϵ values. E.g., comparing the ranking of $\epsilon = 0.3$ and $\epsilon = 0.7$ on 10000 CIFAR-10 test images for AutoAttack, the Spearman's correlation is only 0.9485. Therefore, we note that GREAT Score and AutoAttack are *complementary* evaluation metrics and they don't need to match perfectly. Despite their discrepancy, before calibration, the correlation between GREAT Score and RobustBench yields a similar value. With calibration, there is a significant improvement in rank correlation between Great Score to Robustbench and Auto-Attack, respectively.

Table 3 presents the global robustness statistics of these three methods on ImageNet. We observe almost perfect ranking alignment between GREAT Score and RobustBench, with their Spearman's rank correlation coefficient being 0.8, which is higher than that of Auto-Attack and RobustBench (0.6). These results suggest that GREAT Score is a useful metric for *margin-based* robustness evaluation.

4.4 Ablation Study and Run-time Analysis

Ablation study on GANs and DMs. Evaluating on CIFAR-10, Figure 3 compares the inception score (IS) and the Spearman's rank correlation coefficient between GREAT Score and RobustBench on five GANs and DDPM. One can observe that models with higher IS attain better ranking consistency.

Run-time analysis. Figure 4 compares the run-time efficiency of GREAT Score over Auto-Attack on the same 500 generated CIFAR-10 images. We show the ratio of their average per-sample run-time (wall clock time of GREAT Score/Auto-Attack is reported in Appendix 6.11) and observe around 800-2000 times improvement, validating the computational efficiency of GREAT Score.

Sample Complexity and GREAT Score. In Appendix 6.12, we report the mean and variance of GREAT Score with a varying number of generated data samples. The results show that the statistics of GREAT Score are quite stable even with a small number of data samples (i.e., $\geq$500).

Table 3: Robustness evaluation on ImageNet using GREAT Score, RobustBench (with test set), and Auto Attack (with generated samples). The Spearman's rank correlation coefficient for GREAT score v.s. RobustBench and Auto-Attack v.s. RobustBench is 0.9 and 0.872, respectively.

Model Name	RobustBench Accuracy (%)	AutoAttack Accuracy (%)	GREAT Score
Trans1 Salman et al. (2020)	38.14	30.4	0.504
Trans2 Salman et al. (2020)	34.96	25.8	0.443
LIBRARY Engstrom et al. (2019)	29.22	30.6	0.449
Fast Wong et al. (2020)	26.24	19.2	0.273
Trans3 Salman et al. (2020)	25.32	19.6	0.275

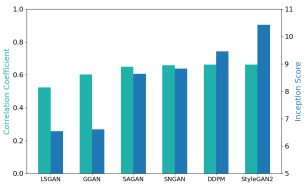

Figure 3: Comparison of Inception Score and Spearman's rank correlation to RobustBench using GREAT Score with different GANs.

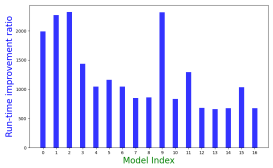

Figure 4: Run-time improvement (GREAT Score over Auto-Attack) on 500 generated CIFAR-10 images.

Table 4: Group-wise and overall robustness evaluation for online gender classification APIs over 500 generated samples (per group).

Online API Name	Old	Young	With Eyeglasses	Without Eyeglasses	Total
BetaFace	0.950	0.662	0.547	0.973	0.783
Inferdo	0.707	0.487	0.458	0.669	0.580
ARSA-Technology	1.033	0.958	0.739	1.082	0.953
DEEPFACE	0.979	0.774	0.763	0.969	0.872
Baidu	1.097	1.029	0.931	1.134	1.048
Luxand	1.091	0.912	0.673	1.010	0.944

Table 5: GREAT Score v.s. accuracy under square attack (Andriushchenko et al., 2020).

DEEPFACE	Old	Young	With Eyeglasses	Without Eyeglasses
Square Attack	84.40%	72.60%	65.80%	89.00%
GREAT Score	0.979	0.774	0.763	0.969

4.5 Evaluation on Online Facial Recognition APIs

To demonstrate Great Score enables robustness evaluation of black-box models that only provide model inference outcomes based on date inputs, we use synthetically generated face images with hidden attributes to evaluate six online face recognition APIs for gender classification. It is worth noting that Great Score is suited for privacy-sensitive assessment because it only uses synthetic face images for evaluation and does not require using real face images.

We use an off-the-shelf face image generator InterFaceGAN (Shen et al., 2020) trained on CelebA-HQ dataset (Karras et al., 2018), which can generate controllable high-quality face images with the choice of attributions such as eyeglasses, age, and expression. We generate four different groups (attributes) of face images for evaluation: Old, Young, With Eyeglasses, and Without Eyeglasses. For annotating the ground truth gender labels of the generated images, we use the gender predictions from the FAN classifier (He et al.). In total, 500 gender-labeled face images are generated for each group. Appendix 6.14 shows some examples of the generated images for each group.

We evaluate the GREAT Score on six online APIs for gender classification: BetaFace (BetaFace), Inferdo (Inferdo), Arsa-Technology (Arsa-Technology), DeepFace (Serengil & Ozpinar, 2021), Baidu (Baidu) and Luxand (Luxand). These APIs are "black-box" models to end users or an external model auditor because the model details are not revealed and only the model inference results returned by APIs (prediction probabilities on Male/Female) are provided.

Finally, we upload these images to the aforementioned online APIs and calculate the GREAT Score using the returned prediction results. Table 4 displays the group-level and overall GREAT Score results. Our evaluation reveals interesting observations. For instance, APIs such as BetaFace, Inferno, and DEEPFACE exhibit a large discrepancy for Old v.s. Young, while other APIs have comparable scores. For all APIs, the score of With Eyeglasses is consistently and significantly lower than that of Without Eyeglasses, which suggests that eyeglasses could be a common spurious feature that affects the group-level robustness in gender classification. The analysis demonstrates how Great Score can be used to study the group-level robustness of an access-limited model in a privacy-enhanced manner.

To verify our evaluation, in Table 5 we compare GREAT Score to the black-box square attack (Andriushchenko et al., 2020) with ϵ=2 and # queries=100 on DEEPFACE. For both Age and Eyeglasses groups (Old v.s. Young and W/ v.s. W/O eyeglasses), we see consistently that a higher Great Score (second row) indicates better robust accuracy (%, first row) against square attack.

5 Conclusion

In this paper, we presented GREAT Score, a novel and computation-efficient attack-independent metric for global robustness evaluation against adversarial perturbations. GREAT Score uses an off-the-shelf generative model such as GANs for evaluation and enjoys theoretical guarantees on its estimation of the true global robustness. Its computation is lightweight and scalable because it only requires accessing the model predictions on the generated data samples. Our extensive experimental results on CIFAR-10 and ImageNet also verified high consistency between GREAT Score and the attack-based model ranking on RobustBench, demonstrating that GREAT Score can be used as an efficient alternative for robustness benchmarks. We also demonstrated the novel use of Great Score for the robustness evaluation of online facial recognition APIs.

Limitations. One limitation could be that our framework of global adversarial robustness evaluation using generative models is centered on $\mathcal{L}_2$-norm based perturbations. This limitation could be addressed if the Stein's Lemma can be extended for other $\mathcal{L}_p$ norms.

REFERENCES

Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square attack: a query-efficient black-box adversarial attack via random search. In *European Conference on Computer Vision*, pp. 484–501. Springer, 2020. 9

Arsa-Technology. Arsa API. https://rapidapi.com/arsa-technology-arsa-technology-default/api/face-recognition18. 9

Maximilian Augustin, Alexander Meinke, and Matthias Hein. Adversarial robustness on in-and out-distribution improves explainability. In *European Conference on Computer Vision*, pp. 228–245. Springer, 2020. 8, 19, 20, 23

Baidu. Baidu API. https://console.bce.baidu.com/. 9

BetaFace. BetaFace API. https://rapidapi.com/betaface/api/face-recognition. 9

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=B1xsqj09Fm. 6

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pp. 39–57. IEEE, 2017. 3, 6, 7, 8

Pin-Yu Chen, Yash Sharma, Huan Zhang, Jinfeng Yi, and Cho-Jui Hsieh. EAD: elastic-net attacks to deep neural networks via adversarial examples. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 10–17, 2018. 3

Jeremy M Cohen, Elan Rosenfeld, and J Zico Kolter. Certified adversarial robustness via randomized smoothing. *International Conference on Machine Learning*, 2019. 1

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, pp. 2206–2216. PMLR, 2020. 1, 2, 3, 6

Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. Robustbench: a standardized adversarial robustness benchmark. *arXiv preprint arXiv:2010.09670*, 2020. 2, 6

Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. Robustbench: a standardized adversarial robustness benchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. URL https://openreview.net/forum?id=SSKZPJCt7B. 1

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009. 6

Gavin Weiguang Ding, Yash Sharma, Kry Yik Chau Lui, and Ruitong Huang. Mma training: Direct input space margin maximization through adversarial training. *arXiv preprint arXiv:1812.02637*, 2018. 8, 19, 20, 23

Logan Engstrom, Andrew Ilyas, Hadi Salman, Shibani Santurkar, and Dimitris Tsipras. Robustness (python library), 2019. URL https://github.com/MadryLab/robustness. 8, 20, 23

Mahyar Fazlyab, Alexander Robey, Hamed Hassani, Manfred Morari, and George Pappas. Efficient and accurate estimation of lipschitz constants for deep neural networks. *Advances in Neural Information Processing Systems*, 32, 2019. 14

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014a. 2

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144, 2020. 2, 3

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014b. 1

Sven Gowal, Krishnamurthy Dj Dvijotham, Robert Stanforth, Rudy Bunel, Chongli Qin, Jonathan Uesato, Relja Arandjelovic, Timothy Mann, and Pushmeet Kohli. Scalable verified training for provably robust image classification. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4842–4851, 2019. 2

Sven Gowal, Chongli Qin, Jonathan Uesato, Timothy Mann, and Pushmeet Kohli. Uncovering the limits of adversarial training against norm-bounded adversarial examples. *arXiv preprint arXiv:2010.03593*, 2020. 8, 19, 20, 23

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016a. 19

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pp. 630–645. Springer, 2016b. 19

Keke He, Yanwei Fu, Wuhao Zhang, Chengjie Wang, Yu-Gang Jiang, Feiyue Huang, and Xiangyang Xue. Harnessing synthesized abstraction images to improve facial attribute recognition. 9

Matthias Hein and Maksym Andriushchenko. Formal guarantees on the robustness of a classifier against adversarial manipulation. *arXiv preprint arXiv:1705.08475*, 2017. 3

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 2, 3, 6

Inferdo. Inferdo API. https://rapidapi.com/inferdo/api/face-detection6. 9

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018. 6, 9

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8110–8119, 2020. 6

Guy Katz, Clark Barrett, David L Dill, Kyle Julian, and Mykel J Kochenderfer. Reluplex: An efficient smt solver for verifying deep neural networks. In *International Conference on Computer Aided Verification*, pp. 97–117. Springer, 2017. 2

Alex Krizhevsky et al. Learning multiple layers of features from tiny images. 2009. 6

Klas Leino, Zifan Wang, and Matt Fredrikson. Globally-robust neural networks. In *International Conference on Machine Learning*, pp. 6212–6222. PMLR, 2021. 14

Tengyuan Liang. How well generative adversarial networks learn distributions. *The Journal of Machine Learning Research*, 22(1):10366–10406, 2021. 14

Jae Hyun Lim and Jong Chul Ye. Geometric gan. *arXiv preprint arXiv:1705.02894*, 2017. 6

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015. 6

Luxand. Luxand API. https://dashboard.luxand.cloud/. 9

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *International Conference on Learning Representations*, 2018. 3

Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 2794–2802, 2017. 6

Andreas Maurer and Massimiliano Pontil. Some hoeffding-and bernstein-type concentration inequalities. *arXiv preprint arXiv:2102.06304*, 2021. 5, 16

Minsu Cho Minguk Kang, Woohyeon Shim and Jaesik Park. Rebooting ACGAN: Auxiliary Classifier GANs with Stable Training. 2021. 6

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018. 6

Remigijus Paulavičius and Julius Žilinskas. Analysis of different norms and corresponding lipschitz constants for global optimization. *Technological and Economic Development of Economy*, 12(4):301–306, 2006. 15

Rahul Rade and Seyed-Mohsen Moosavi-Dezfooli. Helper-based adversarial training: Reducing excessive margin to achieve a better accuracy vs. robustness trade-off. In *ICML 2021 Workshop on Adversarial Machine Learning*, 2021. 8, 19, 20, 23

Sylvestre-Alvise Rebuffi, Sven Gowal, Dan A Calian, Florian Stimberg, Olivia Wiles, and Timothy Mann. Fixing data augmentation to improve adversarial robustness. *arXiv preprint arXiv:2103.01946*, 2021. 7, 8, 19, 20, 23

Leslie Rice, Eric Wong, and Zico Kolter. Overfitting in adversarially robust deep learning. In *International Conference on Machine Learning*, pp. 8093–8104. PMLR, 2020. 8, 19, 20, 23

RobustBench. Robustbench. https://robustbench.github.io/. 6

Jérôme Rony, Luiz G Hafemann, Luiz S Oliveira, Ismail Ben Ayed, Robert Sabourin, and Eric Granger. Decoupling direction and norm for efficient gradient-based l2 adversarial attacks and defenses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4322–4330, 2019. 8, 19, 20, 23

Wenjie Ruan, Min Wu, Youcheng Sun, Xiaowei Huang, Daniel Kroening, and Marta Kwiatkowska. Global robustness evaluation of deep neural networks with provable guarantees for the l_0 norm. *arXiv preprint arXiv:1804.05805*, 2018. 14

Hadi Salman, Andrew Ilyas, Logan Engstrom, Ashish Kapoor, and Aleksander Madry. Do adversarially robust imagenet models transfer better? *Advances in Neural Information Processing Systems*, 33:3533–3545, 2020. 8, 19

Nicolas Schreuder, Victor-Emmanuel Brunel, and Arnak Dalalyan. Statistical guarantees for generative models without domination. In *Algorithmic Learning Theory*, pp. 1051–1071. PMLR, 2021. 14

Vikash Sehwag, Saeed Mahloujifar, Tinashe Handina, Sihui Dai, Chong Xiang, Mung Chiang, and Prateek Mittal. Robust learning meets generative models: Can proxy distributions improve adversarial robustness? *arXiv preprint arXiv:2104.09425*, 2021. 8, 19, 20, 23

Sefik Ilkin Serengil and Alper Ozpinar. Hyperextended lightface: A facial attribute analysis framework. In *2021 International Conference on Engineering and Emerging Technologies (ICEET)*, pp. 1–4. IEEE, 2021. doi: 10.1109/ICEET53442.2021.9659697. URL https://doi.org/10.1109/ICEET53442.2021.9659697. 9

Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *CVPR*, 2020. 9

Charles M Stein. Estimation of the mean of a multivariate normal distribution. *The annals of Statistics*, pp. 1135–1151, 1981. 15

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *International Conference on Learning Representations*, 2014. 3

Tsui-Wei Weng, Huan Zhang, Pin-Yu Chen, Jinfeng Yi, Dong Su, Yupeng Gao, Cho-Jui Hsieh, and Luca Daniel. Evaluating the robustness of neural networks: An extreme value theory approach. *arXiv preprint arXiv:1801.10578*, 2018. 1, 3, 14, 15

Eric Wong and Zico Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *International Conference on Machine Learning*, pp. 5286–5295, 2018. 1

Eric Wong, Leslie Rice, and J Zico Kolter. Fast is better than free: Revisiting adversarial training. *arXiv preprint arXiv:2001.03994*, 2020. 8, 19

Dongxian Wu, Shu-Tao Xia, and Yisen Wang. Adversarial weight perturbation helps robust generalization. *Advances in Neural Information Processing Systems*, 33:2958–2969, 2020. 8, 19, 20, 23

Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6023–6032, 2019. 7, 19

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *BMVC*, 2016. 19

Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *International conference on machine learning*, pp. 7354–7363. PMLR, 2019. 6

Huan Zhang, Tsui-Wei Weng, Pin-Yu Chen, Cho-Jui Hsieh, and Luca Daniel. Efficient neural network robustness certification with general activation functions. In *Advances in Neural Information Processing Systems*, pp. 4944–4953, 2018. 1

6 APPENDIX

6.1 NOTATIONS

Table 6: Main notations used in this paper

Notation	Description
d	dimensionality of the input vector
K	number of output classes
$f : \mathbb{R}^d \to \mathbb{R}^K$	neural network classifier
$x \in \mathbb{R}^d$	data sample
y	groundtruth class label
$\delta \in \mathbb{R}^d$	input perturbation
$\|\delta\|_p$	$\mathcal{L}_p$ norm of perturbation, $p \geq 1$
$\Delta_{\min}$	minimum adversarial perturbation
G	(conditional) generative model
$z \sim \mathcal{N}(0, I)$	latent vector sampled from Gaussian distribution
g	robustness score function defined in (3)
$\Omega(f)/\widehat{\Omega}(f)$	true/estimated global robustness defined in Section 3.1

6.2 MORE MOTIVATIONS ON USING GENERATIVE MODELS FOR ROBUSTNESS EVALUATION

We emphasize the necessity of generative models using the points below.

1. *Global robustness assessment requires a GM*. The major focus and novelty of our study are to evaluate the global robustness with respect to the underlying true data distribution, and we propose to use a GM as a proxy. We argue that such a proxy is necessary to evaluate global robustness unless the true data distribution is known.
2. *GAN can provably match data distribution.* Recent works such as (Schreuder et al., 2021) and (Liang, 2021) have proved the convergence rate of approaching the true data distribution for a family of GANs under certain conditions. This will benefit global robustness evaluation (see Figure 3 for ablations on GAN variants).
3. *Privacy-sensitive remote model auditing.* As shown in Sec 4.5, synthetic data from generative models can facilitate the robustness evaluation of privacy-sensitive models.

6.3 RELATED WORKS FOR GLOBAL ROBUSTNESS EVALUATION FOR DEEP NEURAL NETWORKS.

There are some works studying "global robustness", while their contexts and scopes are different than ours. In (Ruan et al., 2018), the global robustness is defined as the expectation of the maximal certified radius of $\mathcal{L}_0$-norm over a test dataset. Ours is not limited to a test set, and we take the novel perspective of the entire data distribution and use a generative model to define and evaluate global robustness. The other line of works considers deriving and computing the global Lipschitz constant of the classifier as a global certificate of robustness guarantee, as it quantifies the maximal change of the classifier with respect to the entire input space (Leino et al., 2021). The computation can be converted as a semidefinite program (SDP) (Fazlyab et al., 2019). However, the computation of SDP is expensive and hard to scale to larger neural networks. Our method does not require computing the global Lipschitz constant, and our computation is as simple as data forward pass for model inference.

6.4 PROOF OF THEOREM 1

In this section, we will give detailed proof for the certified global robustness estimate in Theorem 1. The proof contains three parts: (i) derive the local robustness certificate; (ii) derive the closed-form global Lipschitz constant; and (iii) prove the proposed global robustness estimate is a lower bound on the true global robustness.

We provide a proof sketch below:

1. We use the local robustness certificate developed in (Weng et al., 2018), which shows an expression of a certified (attack-proof) $\mathcal{L}_p$-norm bounded perturbation for any $p \geq 1$. The certificate is a

function of the gap between the best and second-best class predictions, as well as a local Lipschitz constant associated with the gap function.

2. We use Stein's Lemma (Stein, 1981) which states that the mean of a measurable function integrated over a zero-mean isotropic Gaussian distribution has a closed-form global Lipschitz constant in the $\mathcal{L}_2$-norm. This result helps avoid the computation of the local Lipschitz constant in Step 1 for global robustness evaluation using generative models.

3. We use the results from Steps 1 and 2 to prove that the proposed global robustness estimate $\widehat{\Omega}(f)$ is a lower bound on the true global robustness $\Omega(f)$ with respect to G.

A.2.1 Local robustness certificate In this part, we use the local robustness certificate in (Weng et al., 2018) to show an expression for local robustness certificate consisting of a gap function in model output and a local Lipschitz constant. The first lemma formally defines Lipschitz continuity and the second lemme introduces the the local robustness certificate in (Weng et al., 2018).

Lemma 1 (Lipschitz continuity in Gradient Form ((Paulavičius & Žilinskas, 2006))). *Let $\boldsymbol{S} \subset \boldsymbol{R}^d$ be a convex bound closed set and let $f : \boldsymbol{S} \to \boldsymbol{R}$ be a continuously differentiable function on an open set containing $\boldsymbol{S}$. Then f is a Lipschitz continuous function if the following inequality holds for any $x, y \in \boldsymbol{S}$:*

$$|f(x) - f(y)| \leq L_q \|x - y\|_p \tag{5}$$

where $L_q = \max_{x \in \boldsymbol{S}} \|\nabla f(x)\|_q$: is the corresponding Lipschitz constant, and $\nabla f(x) = (\frac{\partial f}{\partial x_1}, ... \frac{\partial f}{\partial x_d})^\top$ is the gradient of the function f(x), and $1/q + 1/p = 1$, $p \geq 1, q \leq \infty$.

We say f is L_q-continuous in $\mathcal{L}_p$ norm if (5) is satisfied.

Lemma 2 (Formal guarantee on lower bound for untargeted attack of Theorem 3.2 in (Weng et al., 2018)). *Let $x_0 \in \boldsymbol{R}^d$ and $f : \boldsymbol{R}^d \to \boldsymbol{R}^K$ be a multi-class classifier, and f_i be the i-th output of f. For untargeted attack, to ensure that the adversarial examples can not be found for each class, for all $\delta \in \boldsymbol{R}^d$, the lower bound of minumum distortion can be expressed by:*

$$\|\delta\|_p \leq \min_{i \neq m} \frac{f_m(x_0) - f_i(x_0)}{L_q^i} \tag{6}$$

where $m = \arg\max_{i \in \{1,...,K\}} f_i(x_0)$, $1/q + 1/p = 1$, $p \geq 1, q \leq \infty$, and L_q^i is the Lipschitz constant for the function $f_m(x) - f_i(x)$ in L_q norm.

A.2.2 Proof of closed-form global Lipschitz constant in the L_2-norm over Gaussian distribution In this part, we present two lemmas towards developing the global Lipschitz constant of a function smoothed by a Gaussian distribution.

Lemma 3 (Stein's lemma (Stein, 1981)). *Given a soft classifier $F : \boldsymbol{R}^d \to \boldsymbol{P}$, where $\boldsymbol{P}$ is the space of probability distributions over classes. The associated smooth classifier with parameter $\sigma \geq 0$ is defined as:*

$$\bar{F} := (F * \mathcal{N}(0, \sigma^2 I))(x) = \mathbb{E}_{\delta \sim \mathcal{N}(0, \sigma^2 I)}[F(x + \delta)] \tag{7}$$

Then, $\bar{F}$ is differentiable, and moreover,

$$\nabla \bar{F} = \frac{1}{\sigma^2} \mathbb{E}_{\delta \sim \mathcal{N}(0, \sigma^2 I)}[\delta \cdot F(x + \delta)] \tag{8}$$

In a lecture note[1], Li used Stein's Lemma (Stein, 1981) to prove the following lemma:

Lemma 4 (Proof of global Lipschitz constant). *Let $\sigma \geq 0$, let $h : \mathbb{R}^d \to [0, 1]$ be measurable, and let $H = h * \mathcal{N}(0, \sigma^2 I)$. Then H is $\sqrt{\frac{2}{\pi \sigma^2}}$ – continuous in L_2 norm*

[1] https://jerryzli.github.io/robust-ml-fall19/lec14.pdf

A.2.3 Proof of the proposed global robustness estimate $\widehat{\Omega}(f)$ is a lower bound on the true global robustness $\Omega(f)$ with respect to G Recall that we assume a generative model $G(\cdot)$ generates a sample $G(z)$ with $z \sim \mathcal{N}(0, I)$. Following the form of Lemma 2 (but ignoring the local Lipschitz constant), let

$$g'(G(z)) = \max\{f_c(G(z)) - \max_{k\in\{1,\dots,K\},k\neq c} f_k(G(z)), 0\} \tag{9}$$

denote the gap in the model likelihood of the correct class c and the most likely class other than c of a given classifier f, where the gap is defined to be 0 if the model makes an incorrect top-1 class prediction on $G(z)$. Then, using Lemma 4 with g', we define

$$\mathbb{E}_{z\sim\mathcal{N}(0,I)}[g'(G(z))] = g' * \mathcal{N}(0, I) \tag{10}$$

and thus $\mathbb{E}_{z\sim\mathcal{N}(0,I)}[g'(G(z))]$ has a Lipschitz constant $\sqrt{\frac{2}{\pi}}$ in $\mathcal{L}_2$ norm. This implies that for any input perturbation δ,

$$|\mathbb{E}_{z\sim\mathcal{N}(0,I)}[g'(G(z)+\delta)] - \mathbb{E}_{z\sim\mathcal{N}(0,I)}[g'(G(z))]| \tag{11}$$

$$\leq \sqrt{\frac{2}{\pi}} \cdot \|\delta\|_2 \tag{12}$$

and therefore

$$\mathbb{E}_{z\sim\mathcal{N}(0,I)}[g'(G(z)+\delta)] \tag{13}$$

$$\geq \mathbb{E}_{z\sim\mathcal{N}(0,I)}[g'(G(z))] - \sqrt{\frac{2}{\pi}} \cdot \|\delta\|_2 \tag{14}$$

Note that if the right-hand side of (13) is greater than zero, this will imply the classifier attains a nontrivial positive mean gap with respect to the generative model. This condition holds for any δ satisfying $\|\delta\|_2 < \sqrt{\frac{\pi}{2}} \cdot \mathbb{E}_{z\sim\mathcal{N}(0,I)}[g'(G(z))]$. Note that by definition any minimum perturbation on $G(z)$ will be no smaller than $\sqrt{\frac{\pi}{2}} \cdot \mathbb{E}_{z\sim\mathcal{N}(0,I)}[g'(G(z))]$ as it will make $g'(G(z)) = 0$ almost surely. Therefore, by defining $g = \sqrt{\frac{\pi}{2}} \cdot g'$, we conclude that the global robustness estimate $\widehat{\Omega}(f)$ in (2) using the proposed local robustness score g defined in (3) is a certified lower bound on the true global robustness $\Omega(f)$ with respect to G.

6.5 Proof of Theorem 2

To prove Theorem 2, we first define some notations as follows, with a slight abuse of the notation f as a generic function in this part. For a vector of independent random variables $X = (X_1 ..., X_n)$, define $X' = (X_1' ..., X_n')$ to be i.i.d. to $\mathbf{X}$, $x = (x_1, ..., x_n) \in \mathbf{X}$, and the sub-exponential norms $\|\cdot\|_{\psi_2}$ for any random variable Z as

$$\|Z\|_{\psi_2} = \sup_{p\geq 1} \frac{\|Z\|_p}{\sqrt{p}} \tag{15}$$

Let $f : X^n \mapsto \mathbf{R}$. We further define the k-th centered conditional version of f as :

$$f_k(X) = f(X) - \mathbb{E}[f(X)|X_1, ..., X_{k-1}, X_{k+1}, ...X_n] \tag{16}$$

Lemma 5 (Concentration inequality from Theorem 3.1 in (Maurer & Pontil, 2021))**.** *Let $f : X^n \mapsto \boldsymbol{R}$ and $X = (X_1, \dots, X_n)$ be a vector of independent random variables with values in a space $\mathbb{X}$. Then for any $t > 0$ we have*

$$Pr(f(X) - E[f(X')] > t) \leq \exp\left(\frac{-t^2}{32e\left\|\sum_k \|f_k(X)\|_{\psi_2}^2\right\|_\infty}\right) \tag{17}$$

Recall that we aim to derive a probabilistic guarantee on the sample mean of the local robustness score in (3) from a K-way classifier with its outputs bounded by $[0,1]^K$. Following the definition of g (for simplicity, ignoring the constant $\sqrt{\pi/2}$), the sample mean f can be expressed as:

$$f(X) = \frac{1}{n}\sum_{i=1}^{n} g(X_i) \tag{18}$$

where $X_i \sim \mathcal{N}(0, I)$.

Following the definition of (16),

$$f_k(X) = f(X) - \mathbb{E}[f(X)|X_1, ..., X_{k-1}, X_{k+1}, ...X_n] \tag{19}$$

$$= \frac{1}{n}[g(X_k) - g(X_k^{'})] \leq \frac{1}{n} \tag{20}$$

This implies $f_k(X)$ is bounded by $\frac{1}{n}$, i.e., $\|f_k(X)\|_\infty \leq \frac{1}{n}$, and also $\|f_k(X)\|_{\psi_2} \leq \frac{1}{n}$.

Squaring over $\|f_k(X)\|_{\psi_2}$ gives

$$\|f_k(X)\|_{\psi_2}^2 \leq \frac{1}{n^2} \tag{21}$$

As a result,

$$\left\| \sum_k \|f_k(X)\|_{\psi_2}^2 \right\|_\infty \leq n \cdot \frac{1}{n^2} = \frac{1}{n} \tag{22}$$

Divide both side of (22) and multiply with $\frac{-t^2}{32e}$ gives:

$$\frac{-t^2}{32e \left\| \sum_k \|f_k(X)\|_{\psi_2}^2 \right\|_\infty} \leq \frac{-t^2 n}{32e} \tag{23}$$

Take exponential function over both side of (23) gives

$$\exp\left(\frac{-t^2}{32e \left\| \sum_k \|f_k(X)\|_{\psi_2}^2 \right\|_\infty} \right) \leq \exp\left(\frac{-t^2 n}{32e} \right) \tag{24}$$

Recall Lemma 5, since this bound holds on both sides of the central mean, we rewrite it as:

$$\text{Prob}(\left|f(X) - \mathbb{E}[f(X')]\right| > t) \leq 2\exp\left(\frac{-t^2}{32e \left\| \sum_k \|f_k(X)\|_{\psi_2}^2 \right\|_\infty} \right) \tag{25}$$

Hence to ensure that given a statistical tolerance $\epsilon > 0$ with δ as the maximum outage probability, i.e., Prob$(|f(X) - E[f(X')|] > \epsilon) \leq \delta$, we have

$$2 \cdot \exp\left(\frac{-\epsilon^2}{32e \left\| \sum_k \|f_k(X)\|_{\psi_2}^2 \right\|_\infty} \right) \leq 2\exp\left(\frac{-\epsilon^2 n}{32e} \right) \tag{26}$$

$$\leq \delta \tag{27}$$

Finally, (26) implies that the sample complexity to reach the (ϵ, δ) condition is $n \geq \frac{32e \cdot \log(2/\delta)}{\epsilon^2}$.

Algorithm 1: GREAT Score Computation

Input: K-way classifier $f(\cdot)$, conditional generator $G(\cdot)$, local score function $g(\cdot)$ defined in (3), number of generated samples N_S

Output: GREAT Score $\widehat{\Omega}_S(f)$

for $i \leftarrow 1$ *to* N_S **do**

Randomly select a class label $y \in \{1, 2, \ldots, K\}$

Sample $z \sim \mathcal{N}(0, I)$ from a Gaussian distribution and generate a sample $G(z|y)$ with class y

Pass $G(z|y)$ into the model f and get the prediction for each class $\{f_k(G(z|y))\}_{k=1}^{K}$

Record the statistic

$$g^{(i)}(G(z|y)) = \sqrt{\frac{\pi}{2}} \cdot \max\{f_y(G(z|y)) - \max_{k \in \{1,\ldots,K\},\ k \neq y} f_k(G(z|y)), 0\}$$

end

$\widehat{\Omega}_S(f) \leftarrow$Compute the sample mean of $\{g^{(i)}\}_{i=1}^{N_S}$

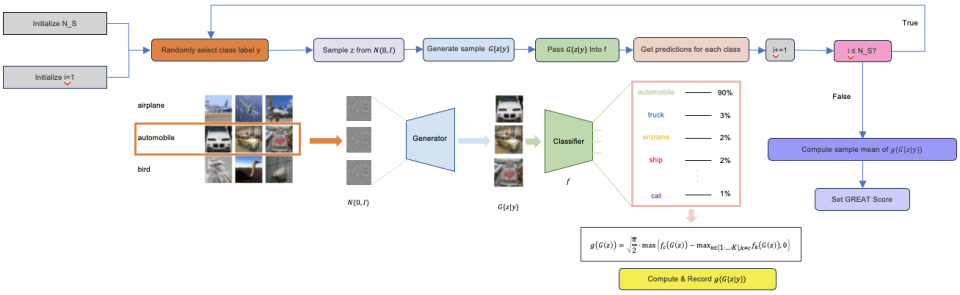

Figure 5: The Flow Chart of GREAT Score.

6.6 Algorithm and Flowchart of GREAT Score

Figure 5 shows the flow chart of Algorithm 1.

6.7 Comparison between CW Attack and GREAT Score

We provide a detailed comparison of the time complexity between GREAT Score and CW Attack.

The time complexity of the GREAT Score algorithm is determined by the number of iterations (generated samples) in the loop, which is denoted as N_S. Within each iteration, the algorithm performs operations such as random selection, sampling from a Gaussian distribution, generating samples, and predicting class labels using the classifier. We assume these operations have constant time complexity I and absorb them in the big O notation. Additionally, the algorithm computes the sample mean of the recorded statistics, which involves summing and dividing the values. As there are N_S values to sum and divide, this step has a time complexity of $O(N_S)$. Therefore, the overall time complexity of the algorithm can be approximated as $O(N_S \cdot I)$.

Using our nation, consider a K-way classifier f. Let x be a data sample and y be its top-1 classification label. Denote δ as the adversarial perturbation. The untargeted CW Attack (L_2 norm) solves the following optimization objective:

$$\delta^* = \arg\min_{\delta}(\|\delta\|_2^2 + \alpha \cdot \max\{f_y(x+\delta) - \max_{k \in \{1,\ldots,K\}, k \neq y} f_k(x+\delta), 0\}) \tag{28}$$

where $f_k(\cdot)$ is the prediction of the k-th class, and $\alpha > 0$ is a hyperparameter.

For CW attack, the optimization process iteratively finds the adversarial perturbation. The number of iterations required depends on factors such as the desired level of attack success and the convergence criteria. Each iteration involves computing gradients, updating variables, and evaluating the objective function. It also involves a hyperparameterα search stage to adjust the weighted loss function.

Table 7: Spearman's rank correlation coeffienct on CIFAR-10 using GREAT Score, RobustBench (with test set), and Auto-Attack (with generated samples) with different calibration methods.

	GREAT Score v.s. RobustBench Correlation	GREAT Score v.s. AutoAttack Correlation	RobustBench v.s AutoAttack Correlation
softmax with temperature	-0.5024	-0.5334	0.7296
sigmoid with temperature	0.7083	0.3641	0.7296
sigmoid with temperature after softmax	-0.2525	-0.2722	0.7296
softmax with temperature after sigmoid	0.8971	0.6941	0.7296

Specifically, let B be the complexity of backpropagation, T_g be the number of iterative optimizations, and T_b be the number of binary search steps for α. The dominant computation complexity of CW attack for N_S samples is in the order of $O(N_S \cdot T_g \cdot T_b \cdot B)$. Normally, T_g is set to 1000, and T_b is set to 9. Therefore, CW attack algorithm is much more time-consuming than GREAT Score.

6.8 Best Calibration Coefficient on different activation methods

Table 7 shows the best ranking coefficient we achieved on each calibration option for CIFAR-10. Among all these four calibration choices, we found that Sigmoid then Temperature Softmax achieves the best result.

6.9 Detailed descriptions of the Models

We provide the detail description for classifiers on RobustBench in what follows. The classfiers for CIFAR-10 are mentioned first and the last paragraph provides descriptions for ImageNet classifiers.
• *Rebuffi et al. (Rebuffi et al., 2021):* Rebuffi et al. (Rebuffi et al., 2021) proposed a fixing data augmentation method such as using CutMix (Yun et al., 2019) and GANs to prevent over-fitting. There are 4 models recorded in (Rebuffi et al., 2021): Rebuffi_extra uses extra data from Tiny ImageNet in training, while Rebuffi_70_ddpm uses synthetic data from DDPM. Rebuffi_70_ddpm/Rebuffi_28_ddpm/Rebuffi_R18 varies in the network architecture. They use WideResNet-70-16 (Zagoruyko & Komodakis, 2016)/WideResNet-28-10 (Zagoruyko & Komodakis, 2016)/PreActResNet-18 (He et al., 2016b).
• *Gowal et al.(Gowal et al., 2020):* Gowal et al. (Gowal et al., 2020) studied various training settings such as training losses, model sizes, and model weight averaging. Gowal_extra differs from Gowal in using extra data from Tiny ImageNet for training.
• *Augustin et al.(Augustin et al., 2020):* Augustin et al.(Augustin et al., 2020) proposed RATIO, which trains with an out-Of-distribution dataset. Augustin_WRN_extra uses the out-of-distribution data samples for training while Augustin_WRN does not.
• *SehWag et al. (Sehwag et al., 2021):* SehWag et al. (Sehwag et al., 2021) found that a proxy distribution containing extra data can help to improve the robust accuracy. Sehwag/Sehwag_R18 uses WideResNet-34-10 (Zagoruyko & Komodakis, 2016)/ResNet-18 (He et al., 2016a), respectively.
• *Rade et al. (Rade & Moosavi-Dezfooli, 2021):* Rade (Rade & Moosavi-Dezfooli, 2021) incorporates wrongly labeled data samples for training.
• *Wu et al. (Wu et al., 2020):* Wu2020Adversarial (Wu et al., 2020) regularizes weight loss landscape.
• *LIBRARY:* Engstrom2019Robustness [2] is a package used to train and evaluate the robustness of neural network.
• *Rice et al. (Rice et al., 2020):* Rice2020Overfitting (Rice et al., 2020) uses early stopping in reduce over-fitting during training.
• *Rony et al. (Rony et al., 2019):* Rony2019Decoupling (Rony et al., 2019) generates gradient-based attacks for robust training.
• *Ding et al. (Rony et al., 2019):* Ding2020MMA (Ding et al., 2018) enables adaptive selection of perturbation level during training.

For the 5 ImageNet models, Trans (Salman et al., 2020) incorporates transfer learning with adversarial training. Its model variants Trans1/Trans2/Trans3 use WideResNet-50-2 (Zagoruyko & Komodakis, 2016)/ResNet-50 (He et al., 2016a)/ResNet-18 (He et al., 2016a). LIBRARY means using the package mentioned in Group of other models to train on ImageNet. Fast (Wong et al., 2020) means fast adversarial training. There is no $\mathcal{L}_2$-norm benchmark for ImageNet on RobustBench, so we use the $\mathcal{L}_\infty$-norm benchmark.

[2] https://github.com/MadryLab/robustness

Table 8: Group-wise time efficiency evaluation on CIFAR-10 using GREAT Score and Auto-Attack (with 500 generated samples).

Model Name	GREAT Score(Per Sample)(s)	AutoAttack(Per Sample)(s)
Rebuffi_extra (Rebuffi et al., 2021)	0.038	60.872
Gowal_extra (Gowal et al., 2020)	0.034	59.586
Rebuffi_70_ddpm (Rebuffi et al., 2021)	0.034	61.3362
Rebuffi_28_ddpm (Rebuffi et al., 2021)	0.011	10.3828
Augustin_WRN_extra (Augustin et al., 2020)	0.013	10.096
Sehwag (Sehwag et al., 2021)	0.011	10.3662
Augustin_WRN (Augustin et al., 2020)	0.011	10.1056
Rade (Rade & Moosavi-Dezfooli, 2021)	0.007	4.4114
Rebuffi_R18(Rebuffi et al., 2021)	0.008	4.4644
Gowal (Gowal et al., 2020)	0.034	60.746
Sehwag_R18 (Sehwag et al., 2021)	0.007	3.8652
Wu2020Adversarial (Wu et al., 2020)	0.012	10.9826
Augustin2020Adversarial (Augustin et al., 2020)	0.014	6.9148
Engstrom2019Robustness (Engstrom et al., 2019)	0.012	6.6462
Rice2020Overfitting (Rice et al., 2020)	0.007	3.5776
Rony2019Decoupling (Rony et al., 2019)	0.010	8.5834
Ding2020MMA (Ding et al., 2018)	0.008	3.6194

6.10 Approximation Error and Sample Complexity

Figure 6 presents the sample complexity as analyzed in Theorem 2 with varying approximation error (ϵ) and three confidence parameters (δ) for quantifying the difference between the sample mean and the true mean for global robustness estimation. As expected, smaller δ or smaller ϵ will lead to higher sample complexity.

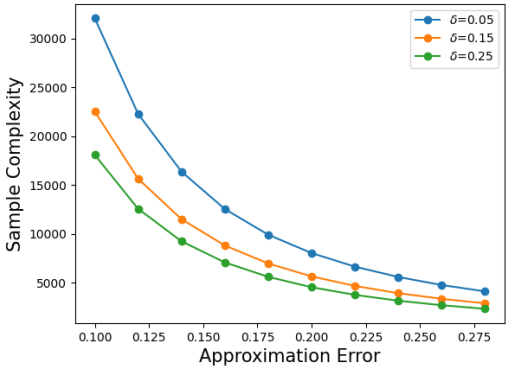

Figure 6: The relationship between the approximation error (ϵ) and sample complexity in Theorem 2, with three different confidence levels: $\delta = \{5, 15, 25\}\%$.

6.11 Complete Run-time Results

The complete run-time results of Figure 4 are given in Table 8:

6.12 Sample Complexity and GREAT Score

Figure 7 reports the mean and variance of GREAT Score with a varying number of generated data samples using CIFAR-10 and the Rebuffi_extra model, ranging from 500 to 10000 with 500 increment. Figure 8 reports the mean and variance of GREAT Score ranging from 50 to 1000 with 50 increment. The results show that the statistics of GREAT Score are quite stable even with a small number of data samples.

6.13 GREAT Score Evaluation on the Original Test Samples of CIFAR-10

Besides evaluating the GREAT Score on the generated samples from GAN, we also run the evaluation process on 500 test samples of CIFAR-10. Table 9 shows the evaluated Great score.

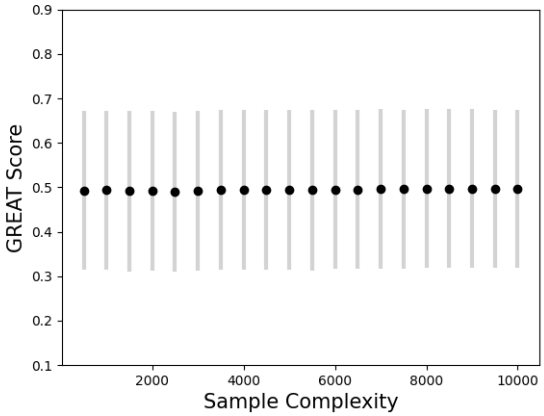

Figure 7: The relation of GREAT Score and sample complexity using CIFAR-10 and Rebuffi_extra model over (500-10000) range. The data points refer to the mean value for GREAT Score, and the error bars refers to the standard derivation for GREAT Score.

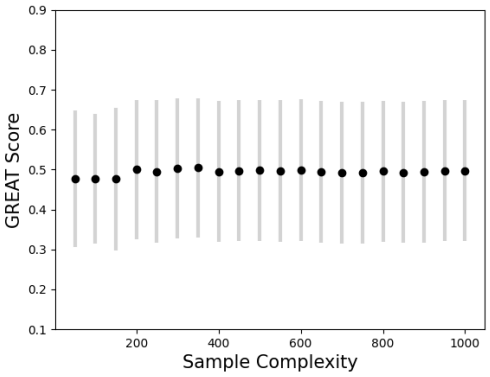

Figure 8: The relation of GREAT Score and sample complexity using CIFAR-10 and Rebuffi_extra model over (50-1000) range. The data points refer to the mean value for GREAT Score, and the error bars refers to the standard derivation for GREAT Score.

6.14 Generated Images from Facial GAN Models

We show the generated images from four groups in what follows.

Figure 9: Generated Images for old subgroup.

Figure 10: Generated Images for young subgroup.

Figure 11: Generated Images for with-eyeglasses subgroup.

Figure 12: Generated Images for without-eyeglasses subgroup.

Table 9: GREAT Score on CIFAR-10. The results are averaged over 500 original test samples.

Model Name	RobustBench Accuracy(%)	AutoAttack Accuracy(%)	GREAT Score	Test Samples GREAT Score
Rebuffi_extra (Rebuffi et al., 2021)	82.32	87.20	0.507	0.465
Gowal_extra (Gowal et al., 2020)	80.53	85.60	0.534	0.481
Rebuffi_70_ddpm (Rebuffi et al., 2021)	80.42	90.60	0.451	0.377
Rebuffi_28_ddpm (Rebuffi et al., 2021)	78.80	90.00	0.424	0.344
Augustin_WRN_extra (Augustin et al., 2020)	78.79	86.20	0.525	0.525
Sehwag (Sehwag et al., 2021)	77.24	89.20	0.227	0.227
Augustin_WRN (Augustin et al., 2020)	76.25	86.40	0.583	0.489
Rade (Rade & Moosavi-Dezfooli, 2021)	76.15	86.60	0.413	0.331
Rebuffi_R18(Rebuffi et al., 2021)	75.86	87.60	0.369	0.297
Gowal (Gowal et al., 2020)	74.50	86.40	0.124	0.109
Sehwag_R18 (Sehwag et al., 2021)	74.41	88.60	0.236	0.176
Wu2020Adversarial (Wu et al., 2020)	73.66	84.60	0.128	0.106
Augustin2020Adversarial (Augustin et al., 2020)	72.91	85.20	0.569	0.493
Engstrom2019Robustness (Engstrom et al., 2019)	69.24	82.20	0.160	0.127
Rice2020Overfitting (Rice et al., 2020)	67.68	81.80	0.152	0.120
Rony2019Decoupling (Rony et al., 2019)	66.44	79.20	0.275	0.221
Ding2020MMA (Ding et al., 2018)	66.09	77.60	0.112	0.08