Diffusion Denoised Smoothing for Certified and Adversarial Robust Out-Of-Distribution Detection*

Nicola Franco[1,*], Daniel Korth[1], Jeanette Miriam Lorenz[1], Karsten Roscher[1] and Stephan Günnemann[2]

[1]Fraunhofer Institute for Cognitive Systems IKS, Munich, Germany
[2]Dept. of Computer Science & Munich Data Science Institute, Technical Univ. of Munich, Germany

Abstract

As the use of machine learning continues to expand, the importance of ensuring its safety cannot be overstated. A key concern in this regard is the ability to identify whether a given sample is from the training distribution, or is an "Out-Of-Distribution" (OOD) sample. In addition, adversaries can manipulate OOD samples in ways that lead a classifier to make a confident prediction. In this study, we present a novel approach for certifying the robustness of OOD detection within a ℓ_2-norm around the input, regardless of network architecture and without the need for specific components or additional training. Further, we improve current techniques for detecting adversarial attacks on OOD samples, while providing high levels of certified and adversarial robustness on in-distribution samples. The average of all OOD detection metrics on CIFAR10/100 shows an increase of $\sim 13\%/5\%$ relative to previous approaches. Code: https://github.com/FraunhoferIKS/distro

Keywords

Robust Machine Learning, Robustness Certificates, Out-Of-Distribution, Randomized Smoothing

1. Introduction & Related Work

Although recent advances in Machine Learning (ML) demonstrate its validity in a wide range of applications, its use in safety-critical conditions remains challenging. Since the appearance of unexpected low robustness to natural [1] and adversarial [2] perturbations to the input data, several types of defenses have been proposed along the years. Two main branches of defenses exist: *empirical* [3] and *certified* [4], which aim at *improving* or *assuring* the robustness of the prediction in the vicinity of the input, respectively. *Certified* defenses might give the inaccurate impression that robustness makes ML systems ready for deployment in safety-critical applications. Unfortunately, further issues lie also beyond robustness, including the lack of guarantees for Out-Of-Distribution (OOD) data, the lack of fairness, or the lack of explainability [5].

OOD Detection. With Maximum Softmax Probability (MSP) [6] as a baseline method, OOD detection aims to identify inputs that fall outside the scope of the training distribution. Outlier Exposure (OE) [7] trains models to differentiate between in-distribution (ID) and out-of-distribution (OOD) samples. Recent approaches include Virtual Outlier Synthesis (VOS) [8] and LogitNorm [9]. VOS adaptively synthesizes virtual outliers, while LogitNorm normalizes the logit vector to reduce overconfidence, using thresholding for OOD detection.

Adversarial OOD Detection. Other lines of research [10, 11, 12], focus on providing low confidence for OOD data when perturbed with adversarial noise. Hein et al. [10] show that ReLU networks can have arbitrarily high confidence for data that is *far enough* from the training distribution. Additionally, they propose ACET [10], an adversarial training method to enforce low confidence on OOD data, but at the cost of decreased ID accuracy. ATOM [12] addresses this issue by using outlier mining techniques to automatically select a diverse set of OOD samples from a large pool of potential OOD samples.

Guaranteed OOD Detection. Recent studies like Bitterwolf et al. [13], Meinke et al. [14] bring forth ℓ_∞-norm certified robustness for OOD data with a simple but effective method: Interval Bound Propagation (IBP) [15]. GOOD [13] proposes a training approach using IBP, but it produce loose bounds, impacting accuracy. While ProoD [14] combines a certified discriminator and OE model, achieving state-of-the-art performance but with practical limitations: low certified accuracy, reliance on external datasets, and reduced scalability due to IBP's impact on larger models.

In this study, we propose a novel technique for certifying OOD detection within the ℓ_2-norm of the input sample, without requiring the use of binary discriminators or specific training. This enables us to establish a guar-

AISafety-SafeRL 2023 Workshop (IJCAI), August 19–21, 2023, Macao, SAR, China

* This work was funded by the Bavarian Ministry for Economic Affairs, Regional Development and Energy.

*Corresponding author.
nicola.franco@iks.fraunhofer.de (N. Franco);
daniel.korth@iks.fraunhofer.de (D. Korth);
jeanette.miriam.lorenz@iks.fraunhofer.de (J. M. Lorenz);
karsten.roscher@iks.fraunhofer.de (K. Roscher);
s.guennemann@tum.de (S. Günnemann)

Table 1
Comparison between this work and previous methods in terms of ID and OOD robustness properties. In this case, the ✓indicates that property was provided in the work. While (✓) indicates that the property is actually lower than expected.

Methods	In-Distribution (ID) Accuracy			Out-Of-Distribution (OOD) Detection				
	Clean	Adversarial ℓ_∞	Certified ℓ_2	Clean	Adversarial ℓ_∞	Certified ℓ_∞	 ℓ_2	Asymptotic underconfidence
- **Standard**								
OE [7]	✓			✓				
VOS [8]	✓			✓				
LogitNorm [9]	✓			✓				
- **Adversarial**								
ACET [10]	(✓)	✓		✓	(✓)			
ATOM [12]	(✓)			✓	(✓)			
- **Guaranteed**								
GOOD [13]					✓	✓		✓
ProoD [14]	✓			✓	✓	✓		✓
DISTRO (Our)	✓	✓	✓	✓	✓	✓	✓	✓

anteed upper bound on the classifier's confidence within a defined region surrounding the input. Unlike before, certified robust OOD detection can now be computed for standard OOD detection approaches. Additionally, we incorporate a diffusion denoiser [16, 17], which recovers the perturbed images and returns high quality denoised inputs. This leads to better levels of both adversarial and certified robustness for ID and OOD data. This work and previous methods are compared in Table 1.

In summary, our contributions are:

- A novel technique to robustly certify the confidence of any classifier within an ℓ_2-norm on OOD data. This technique can be applied to any architecture and does not require additional components, even though it has higher computational costs compared to previous approaches.
- A method named DISTRO: **DI**ffusion denoised **S**mooT**h**ing for **R**obust **O**OD detection. This method incorporates a diffusion denoiser model to improve the detection of adversarial and certified OOD samples, while providing high adversarial and certified accuracy for ID data.

2. Background

We define a *hard* classifier as a function $f : \mathbb{R}^d \to \mathcal{Y}$ which maps input samples $x \in \mathbb{R}^d$ to output $y \in \mathcal{Y}$, where $\mathcal{Y} = \{1, \ldots, K\}$ is the discrete set of K classes. Additionally, we introduce a *soft* version $F : \mathbb{R}^d \to \mathbb{P}(\mathcal{Y})$ of f, where $\mathbb{P}(\mathcal{Y})$ is the set of probability distributions over $\mathcal{Y}$. It is possible to convert any soft classifier F into a hard classifier f by mapping $f(x) = \arg\max_{y\in\mathcal{Y}} F(x)_y$. Additionally, we define as $\mathcal{N}(0, 1)$ the standard Gaussian distribution, as $\Phi(x)$ the Gaussian CDF and as $\Phi^{-1}(x)$ its inverse (or quantile).

Robustness Certificates. Even though an adversarially-trained network is resilient to attacks created during training, it can still be susceptible to unseen new attacks. To overcome this problem, certified defenses formally guarantee the stability of the prediction in a neighbourhood of the input. In other words, a neural network f is certifiably robust for the input $x \in \mathbb{R}^d$, if the prediction for all perturbed versions $\tilde{x}$ remains unchanged such that $\|\tilde{x} - x\|_p \leq \epsilon$, where $\|\cdot\|_p$ is the ℓ_p-norm around x of size $\epsilon > 0$.

Randomized Smoothing. This robustness verification method [4] computes the ℓ_2-norm certificates around an input sample x by counting which class is most likely to be returned when x is perturbed by isotropic Gaussian noise. Formally, given a *soft* classifier F, randomized smoothing considers a *smooth* version of F defined as:

$$G(x) \overset{\text{def}}{=} \underset{\delta\sim\mathcal{N}(0,\sigma^2 I)}{\mathbb{E}} \left[F(x + \delta)\right], \tag{1}$$

where $\sigma > 0$ represents the standard deviation. As previously, we define the hard version of $G(x)$ as $g(x) = \arg\max_{y\in\mathcal{Y}} G(x)_y$. Cohen et al. [4] demonstrated that G is robust to perturbations of radius R, where the radius R is defined as the difference in probabilities between the most likely class and the second most likely class. A more general interpretation is given by Yang et al. [18].

Lemma 2.1. *[Yang et al. [18]] Given a smoothed classifier G defined as in Equation 1, such that $G(x) = (G(x)_1, \ldots, G(x)_K)$ is a vector of probabilities that G assigns to each class $1, \ldots, K$. Suppose G predicts class c on input x, and the probability is $p = max_{y\in\mathcal{Y}} G(x)_y > 1/2$, then G continues to predict class c when x is perturbed by any δ with:*

$$\|\delta\|_2 < \sigma\Phi^{-1}(p).$$

One should consider p as the probability that the smoothed classifier will assign to the predicted class rather than any other. As a consequence, if $p > 1/2$, it will continue to do so even if the input is perturbed by Gaussian noise of magnitude smaller than the radius $R = \sigma\Phi^{-1}(p)$.

Salman et al. [19] show that randomized smoothing can postprocess the network to make it locally Lipschitz continuous. The connection between randomized smoothing and Lipschitz continuity is provided in the following lemma, which offers an analytical form of the gradient of a smooth function.

Lemma 2.2. *[Stein [20]] Let $\sigma > 0$, let $h : \mathbb{R}^d \to \mathbb{R}$ be measurable, and let $H(x) = \mathbb{E}_{\delta\sim\mathcal{N}(0,\sigma^2 I)}[h(x+\delta)]$. Then H is differentiable, and moreover:*

$$\nabla H(x) = \frac{1}{\sigma^2} \mathop{\mathbb{E}}_{\delta\sim\mathcal{N}(0,\sigma^2 I)} \left[\delta \cdot h(x+\delta)\right].$$

The smoothed function H is also known as the *Weierstrass transform* of h, and a classical property of the Weierstrass transform is its induced smoothness.

Diffusion Denoised Smoothing. In a nutshell, forward diffusion involves adding Gaussian noise to an image until it produces an isotropic Gaussian distribution with a large variance. Denoising diffusion probabilistic models work by learning how to reverse this process. In formal terms, given an input sample $x \in \mathbb{R}^d$, a diffusion model selects a predetermined *timestep* $t \in \mathbb{N}^+$ and samples a noisy image x_t as follows:

$$x_t \overset{\text{def}}{=} \sqrt{\alpha_t} \cdot x + \sqrt{1-\alpha_t} \cdot \mathcal{N}(0, I), \tag{2}$$

where the amount of noise to be added to the image is determined by a constant called α_t derived from t.

As Salman et al. [21] suggested, denoising Gaussian pertubed images leads to out-of-the-box certified robustness for plain models. Following this trend Carlini et al. [17] make use of a diffusion model as one-shot denoiser achieving state-of-the-art performances. The minor proposed adjustment held in the estimation of t, computed such that $\frac{1-\alpha_t}{\alpha_t} = \sigma^2$. Additionally, the perturbed version $\tilde{x} = x + \delta$ is scaled by $\sqrt{\alpha_t}$, to match the noise model of Equation 2.

3. Certified Robust OOD Detection

This section explains how using local Lipschitz continuity, achieved through smoothing the classifier with Gaussian noise, can guarantee the detection of OOD samples within a ℓ_2-sphere around the input.

Preliminaries. To determine how well a classifier distinguishes between ID and OOD samples, it is common to threshold the confidence level and to calculate the area under the receiver operating characteristic curve (AUROC or AUC). Formally, let us consider a function[1] $h \in \mathbb{R}^d \to \mathbb{R}$, the AUC is defined as:

$$\mathrm{AUC}_h(\mathcal{D}_{in}, \mathcal{D}_{out}) = \mathop{\mathbb{E}}_{\substack{x\sim\mathcal{D}_{in},\\ z\sim\mathcal{D}_{out}}} \left[\mathbb{1}_{h(x)>h(z)}\right],$$

where $\mathcal{D}_{in}, \mathcal{D}_{out}$ are ID and OOD data sets, respectively, and $\mathbb{1}$ returns 1 if the argument is true and 0 otherwise. A number of prior works [11, 13, 12, 14] also investigated the worst-case AUC (WCAUC), which is defined as the lowest AUC attainable when every OOD sample is perturbed so that the highest level of confidence is achieved within a specific threat model. Specifically, the WCAUC is defined as:

$$\mathrm{WCAUC}_h(\mathcal{D}_{in}, \mathcal{D}_{out}) = \mathop{\mathbb{E}}_{\substack{x\sim\mathcal{D}_{in},\\ z\sim\mathcal{D}_{out}}} \left[\mathbb{1}_{h(x) > \max\limits_{\|\tilde{z}-z\|_p \le \epsilon} h(\tilde{z})}\right].$$

Due to the intractable nature of the maximization problem, we can compute upper or lower bounds only, i.e. $\underline{h}(z) \le \max_{\|\tilde{z}-z\|_p \le \epsilon} h(\tilde{z}) \le \bar{h}(z)$. The lower bound $\underline{h}(z)$ is typically calculated using projected gradient methods [2, 23] and named Adversarial AUC (AAUC) (upper bound of WCAUC). In the context of ℓ_∞-norm, the upper bound $\bar{h}(z)$, called Guaranteed AUC (GAUC) (lower bound of WCAUC), is computed using IBP in Bitterwolf et al. [13] and Meinke et al. [14].

Here, we propose a method for computing the upper bound of any classifier without the need for special training or modifications. Thus, the main theorem for an ℓ_2-norm robustly certified upper bound is stated.

Theorem 3.1. *Let $F : \mathbb{R}^d \to \mathbb{P}(\mathcal{Y})$ be any soft classifier and G be its associated smooth classifier as defined in Equation 1, with $\sigma > 0$. If $p = \max_{y\in\mathcal{Y}} G(x)_y > 1/2$, then, we have that:*

$$\max_{y\in\mathcal{Y}} G(x+\delta)_y \le \sqrt{\frac{2}{\pi}}\Phi^{-1}(p) + p, \tag{3}$$

for every $\|\delta\|_2 < \sigma\Phi^{-1}(p)$.

Proof. As a prerequisite to proving the theorem, we need to know the analytic form of the gradient of a smoothed function given in Lemma 2.2. Let us consider the soft classifier $F(x) : \mathbb{R}^d \to \mathbb{P}(\mathcal{Y})$, and its smooth version $G(x) = \mathbb{E}_{\delta\sim\mathcal{N}(0,\sigma^2 I)}[F(x+\delta)]$, with $\sigma > 0$. Since F its a measurable function, we consider the *Weierstrauss* transform of F (which coincide with the *smooth* version of F):

$$\mathop{\mathbb{E}}_{\delta\sim\mathcal{N}(0,\sigma^2 I)} \left[F(x+\delta)\right] = \left(F * \mathcal{N}(0, \sigma^2 I)\right)(x),$$

where $*$ denotes the convolution operator. Thus, $G(x)$ is differentiable and from Lemma 2.2 we have:

$$\nabla G(x) = \frac{1}{\sigma^2} \mathop{\mathbb{E}}_{\delta\sim\mathcal{N}(0,\sigma^2 I)} \left[\delta \cdot h(x+\delta)\right].$$

[1] e.g. the Maximum Softmax Probability [6], or the Energy function [22].

Since $F : \mathbb{R}^d \to [0, 1]$ and ℓ_2 is self-dual, it is sufficient to show that the gradients of G are bounded in ℓ_2. From Lemma 2.2, for any unit vector $v \in \mathbb{R}^d$ we have that $|\langle v, \nabla G(x)\rangle|$ is equal to:

$$\left|\frac{1}{(2\pi\sigma^2)^{d/2}} \int_{\mathbb{R}^d} F(t) \left\langle v, \frac{t-x}{\sigma^2}\right\rangle e^{\left(-\frac{1}{2\sigma^2}\|x-t\|_2^2\right)} dt\right|,$$
$$\leq \frac{1}{(2\pi\sigma^2)^{d/2}} \int_{\mathbb{R}^d} \left|\left\langle v, \frac{t-x}{\sigma^2}\right\rangle\right| e^{\left(-\frac{1}{2\sigma^2}\|x-t\|_2^2\right)} dt,$$

where we make use of the triangle inequality and know that F is bounded by 1. Given that projections of Gaussians are Gaussians and from the classical integration of the Gaussian density, we obtain:

$$\frac{1}{(2\pi\sigma^2)^{d/2}} \int_{\mathbb{R}^d} \left|\left\langle v, \frac{t-x}{\sigma^2}\right\rangle\right| e^{\left(-\frac{1}{2\sigma^2}\|x-t\|_2^2\right)} dt,$$
$$= \frac{1}{\sigma^2} \mathop{\mathbb{E}}_{Z\sim\mathcal{N}(0,\sigma^2)} [|Z|] = \sqrt{\frac{2}{\pi\sigma^2}},$$

where we consider the supremum over all unit vectors v. Since, we know that $G(x)$ is $\sqrt{\frac{2}{\pi\sigma^2}}$-Lipschitz in ℓ_2, it is possible to use the Lipschitz constant to bound the difference between $G(x + \delta)$ and $G(x)$ for any value of δ, with $\|\delta\|_2 < \sigma\Phi^{-1}(p)$, where $p = \max_{y\in\mathcal{Y}} G(x)_y$. Formally:

$$|G(x+\delta)| \leq \sqrt{\frac{2}{\pi\sigma^2}} \|\delta\|_2 + |G(x)|,$$

where we make use of the reverse triangle inequality. Since $G(x) : \mathbb{R}^d \to [0, 1]$, we can assume $|G(x)| = G(x)$, and moreover:

$$\max_{y\in\mathcal{Y}} G(x+\delta)_y \leq \sqrt{\frac{2}{\pi}} \Phi^{-1}(p) + \max_{y\in\mathcal{Y}} G(x)_y.$$

□

In other words, if the smooth classifier assigns the most likely class more than half the time, it is locally Lipschitz continuous in x, and its maximum prediction is bounded within a radius smaller than $R = \sqrt{\frac{2}{\pi}}\Phi^{-1}(p)$.

Discussion

While this theorem provides some advantages, it is important to note a couple of its limitations. One of the main limitations is that the upper bound of the smooth classifier G only applies to G and not to the original classifier F. As a result, the guarantee only applies to G, and its robustness at a given input point x cannot be precisely evaluated or certified. To overcome this, Monte Carlo algorithms can be used to approximate these evaluations with high probability [4].

Another limitation is that the guarantees provided by this theorem are only probabilistic in practice. Therefore, a hypothesis test [24] should be used to avoid making predictions with low confidence. As with randomized smoothing [4], a large number of samples must be generated in order to achieve high levels of confidence in the certification radius. However, generating these samples can be computationally expensive for complex models.

Despite these limitations, the theorem provides a novel way of calculating the upper bound of any classifier, without the need for special training or modification. Additionally, we provide a tighter certificate compared to previous approaches [13, 14], as they used IBP. This can be useful for evaluating the certified robustness of a broader category of standard OOD detection methods as well as larger models, where IBP bounds explode in size and make them unusable [25].

4. DISTRO: DIffusion denoised SmooThing for Robust OOD detection

In this section, we present our method. Essentially, it combines three techniques: (i) a diffusion denoiser, (ii) a standard OOD detector, and (iii) a certified binary discriminator. Each component of this method is designed to overcome a specific problem of ordinary classifiers, as they are not robust to adversarial attacks, either ID or OOD, and do not detect OOD inputs well.

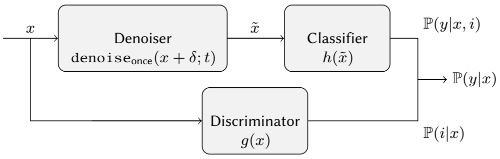
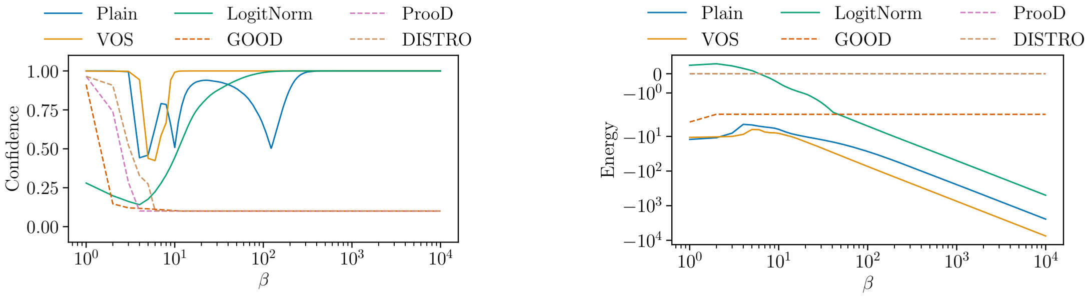


Figure 1: Overview of DISTRO.

In Figure 1, we show an overview of DISTRO. First, a diffusion denoiser is employed before the classifier itself to provide robustness against ID attacks. As a result, adversarial noise introduced by the attack is mitigated by the denoiser. This technique has already been proven to be very efficient and does not affect clean accuracy [17].

Secondly, numerous post-hoc OOD detection methods exist. The most straightforward being MSP [6], which can be added to the image classifier without retraining or fine-tuning. Alternatively, standard OOD detection methods, such as OE [7], VOS [8] or LogitNorm [9], could also replace the classifier. Thirdly, to make the model more robust to OOD adversarial attacks, we add a binary discriminator to the model that is trained to be certifiably robust against OOD attacks. Additionally, this discrimi-



(a) MSP

(b) Energy

Figure 2: Asymptotic confidence as: (a) MSP [6] and (b) Energy [22], for several OOD detection models divided into two categories: *standard* (continuous line) and *guaranteed* (dashed line).

nator is combined with the OOD detection method from (ii) which is necessary to have the property of asymptotic underconfidence for far-OOD inputs.

Configuration. This method does not require any new technical knowledge. We begin by making the assumption that OOD samples are unrelated and thus maximally un-informative to the ID data. Thus, for every class $y \in \mathcal{Y}$, the conditional distribution on the input x is given as:

$$\mathbb{P}(y|x) = \mathbb{P}(y|x,i)\mathbb{P}(i|x) + \frac{1}{K}(1 - \mathbb{P}(i|x)), \quad (4)$$

where $\mathbb{P}(i|x)$ is the conditional distribution representing the probability that x is part of the ID, while $\mathbb{P}(y|x,i)$ is the conditional distribution representing the ID. Similarly to Meinke et al. [14], we assign independent models to each distribution:

- $\mathbb{P}(y|x,i) = h(\texttt{denoise}_{\text{once}}(x + \delta; t))$, where $h : \mathbb{R}^d \to [0,1]$ is the confidence of the main classifier $F(x)$, and $\tilde{x} = \texttt{denoise}_{\text{once}}(x + \delta; t)$ represents one single step of denoising operation with $\delta \sim \mathcal{N}(0, \sigma^2 I)$.
- $\mathbb{P}(i|x) = \frac{1}{1+e^{-g(x)}}$, where $g : \mathbb{R}^d \to \mathbb{R}$ refers to a binary discriminator trained in a certified robust manner based on an ℓ_∞-threat model as in Bitterwolf et al. [13], Meinke et al. [14].

As can be seen, the denoiser is the main addition. The one-step denoiser $\texttt{denoise}_{\text{once}}$ estimates the fully denoised image x from the current timestep t. Then it computes the average between the denoised image and the noisy image from the previous timestep. As discussed in Carlini et al. [17], multiple applications of the denoiser will only destroy information about x. Denoising with iterative steps essentially transfers the classification task to the denoiser, which can determine how the image should be filled. For these reason, we apply only a single step of denoising.

Asymptotic Underconfidence. Here, we show that by coupling a classifier trained to be OOD aware with a diffusion denoiser and running a certified discriminator in parallel, we can guarantee asymptotic underconfidence for data *far enough* from the training distribution.

To obtain asymptotic underconfidence of the joint classifier, we consider $\mathbb{P}(y|x,i) \leq 1$ and rewrite Equation 4 as follows:

$$\mathbb{P}(y|x) \leq \frac{K-1}{K}\mathbb{P}(i|x) + \frac{1}{K}. \quad (5)$$

Since the right term only depends on $\mathbb{P}(i|x)$, we just need to assure that $\lim_{\beta\to\infty} \mathbb{P}(i|\beta x) \to 0$. If we employ a certified binary discriminator, trained with IBP on OOD data, as descibed in Meinke et al. [14], to compute $\mathbb{P}(i|x)$, we achieve asymptotic underconfidence independently of the main classifier. Readers are referred to Meinke et al. [14] for a more detailed explanation.

Empirical Evaluation. In Figure 2, we show an empirical evaluation of the asymptotic confidence for standard and robust OOD detection methods[2]. In this test, we consider a single ID sample x and multiply by a scalar β. In Figure 2a we plot the MSP [6] as confidence, while in Figure 2b we plot the Energy [22] for increasing values of $\beta > 0$. In the context of MSP, we observe that standard OOD detection methods are asymptotically overconfident, after a small drop, whereas certified methods such as GOOD [13], ProoD [14] and DISTRO converge to $1/K$. On the other hand, for Energy as β increases, VOS [8], LogitNorm [9], and Plain models asymptotically decrease, whereas GOOD [13], ProoD (Meinke at al., 2022), and DISTRO remain stable.

As a result, underconfidence can be easily obtained when using an energy score instead of MSP, regardless of whether it is on a plain or OOD aware model. However, asymptotic underconfidence does not necessarily imply that the model will perform better in detecting OOD

[2] the models are described in section 5.

samples since all inputs are usually normalized to some range (e.g. [0, 1] or [-1, 1]). Thus the choice of MSP over the energy function is directly related to the possibility of certified robustness for OOD samples.

5. Experiments

In this section, DISTRO is evaluated for a variety of robust ID and OOD tests and is compared to previous approaches. As baseline, we consider the pre-trained models[3] from Meinke et al. [14]. The normal trained (**Plain**) and outlier exposure (**OE**) [7] models share the same ResNet18 [26] architecture and hyperparameters as **ProoD** [14]. **GOOD** [13] uses a 'XL' convolutional neural network. Additionally, we evaluate the pretrained DenseNet101 [27] models for **ATOM** [12] and **ACET** [10]; and the standard OOD detection methods: **VOS**[4] [8] and **LogitNorm**[5] [9] with the pretrained WideResNet40 [28] models provided in the respective works. We consider **DDS** [17] with a pre-trained diffusion model[6] from Nichol and Dhariwal [16] in front of the OE classifier. With **DISTRO**, we incorporate the same pre-trained diffusion model of DDS before the main classifier of ProoD, and maintain its discriminator. The diffusion models have been used with the settings described in Carlini et al. [17]. In the context of ℓ_∞, we set $\sigma = \sqrt{d} \cdot \epsilon$.

We evaluate all methods on the standard datasets CIFAR10/100 [29] as ID. For the OOD detection evaluation we consider the following set of datasets: CIFAR100/10, SVHN [30], LSUN [31] cropped (LSUN_CR) and resized (LSUN_RS), TinyImageNet [32] cropped (TinyImageNet_CR), Textures [33] and synthetic (Gaussian and Uniform) noise distributions. We use a random but fixed subset of 1000 images for all datasets considered as a test for OOD. For ID, we consider the entire dataset. We run all our experiments on a single NVIDIA A100.

5.1. In-Distribution Results

Here, we compare clean, adversarial, and certified accuracy for ID samples. Adversarial accuracy is evaluated with AutoAttack [23] for ℓ_∞-norm attacks of budget $\epsilon \in \{2/255, 8/255\}$. We ran the standard version of AutoAttack without additional hyper-parameters. Certified accuracy is evaluated for ℓ_2-norm robustness of deviation $\sigma \in \{0.12, 0.25\}$. To this end, random smoothing is performed on 10'000 Gaussian distributed samples around the input with a failure probability of 0.001. All $R > 0$ are considered for the certified accuracy. In the context of DISTRO and DDS we run 100 evaluation of the entire test set of CIFAR10 to estimate the clean accuracy and report the average. Further, we ran AutoAttack in both *rand* and *standard* modes, and considered the lowest results for DISTRO and DDS.

Table 2
ID Accuracy: Results of clean, adversarial and certified accuracy (%) on the CIFAR10 test set. The grayed-out models have an accuracy drop greater than 3% relative to the model with the highest accuracy.

Method	Clean	Adversarial (ℓ_∞)		Certified (ℓ_2)	
		$\epsilon = 2/255$	$\epsilon = 8/255$	$\sigma = 0.12$	$\sigma = 0.25$
Plain*	95.01	2.16	0.00	28.14	14.17
OE*	95.53	1.97	0.00	31.48	10.88
VOS†	94.62	2.24	0.00	13.13	10.02
LogitNorm‡	94.48	2.65	0.00	12.53	10.25
ATOM*	92.33	0.00	0.00	0.00	0.00
ACET*	91.49	69.01	6.04	57.13	12.48
GOOD^*_{80}	90.13	11.65	0.23	17.33	10.31
ProoD* $\Delta = 3$	95.46	2.69	0.00	33.92	13.50
DDS	**95.55**	72.97	24.09	82.26	64.58
DISTRO (our)	95.47	**73.34**	**27.14**	**82.77**	**65.63**

∗ Pre-trained models from Meinke et al. [14], † Pre-trained from Du et al. [8],
‡ Pre-trained from Wei et al. [9].

In Table 2, we show the results. As expected, Plain and OE are not robust to adversarial attacks. This applies to ProoD as well, since OE is its primary classifier. Similarly, standard OOD detection methods, as LogitNorm and VOS, show poor robustness for ID data. GOOD demonstrates better results than ProoD for adversarial attacks and worse in terms of certified accuracy. Suprisingly, ACET reveals strong adversarial and certified accuracy despite of its reduced clean accuracy. Meanwhile, ATOM results in zero for all tests since any slight perturbation of the input triggers the last neuron used for OOD detection.

Discussion

It is clear that diffusion models can enhance adversarial and certified robustness while maintaining high clean accuracy. As diffusion introduces variance into gradient estimators, standard attacks become much less effective. Nevertheless, robustness accuracy of diffusion models varies over different runs for the same input, so it should be defined differently from deterministic accuracy, e.g. as expectation. Luckily, one-shot diffusion introduces such a tiny variance that throughout a few of runs, our results were similar.

5.2. Evaluation Metrics

To discriminate between ID and OOD samples, we use the confidence of the classifier, i.e. MSP [6]. Traditionally, the following metrics are used to evaluate the OOD detection performance: (i) false positive rate (FPR95) of OODs when ID samples have a 95% true positive rate; (ii) the area under the receiver operating characteristic curve

[3] https://github.com/AlexMeinke/Provable-OOD-Detection
[4] https://github.com/deeplearning-wisc/vos
[5] https://github.com/hongxin001/logitnorm_ood
[6] https://github.com/openai/improved-diffusion

(AUROC or AUC); and (iii) the area under the precision-call curve (AUPR). In order to determine robustness, we compare adversarial (AAUC, AAUPR, AFPR) and guaranteed (GAUC, GAUPR, GFPR) versions of the previous metrics. For the adversarial metrics, we use the settings in Meinke et al. [14] to ensure a fair comparison.

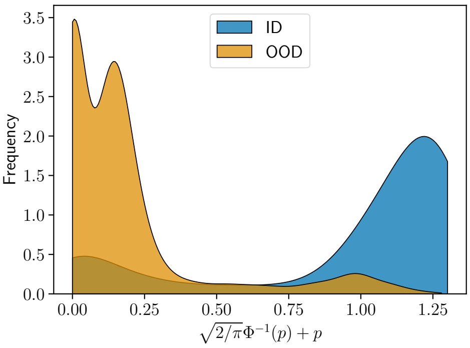


Figure 3: Kernel density estimation (bandwidth = 1) of the distribution of certified smooth ($\sigma = 0.12$) scores for DISTRO on ID (CIFAR10) and OOD (all other datasets) samples.

Guaranteed. The guaranteed metrics (GAUC, GAUPR and GFPR) are computed for ℓ_2 and ℓ_∞ norms robustness certificates. Similarly to Meinke et al. [14], the ℓ_∞-norm is obtained with IBP only on OOD data. On the other hand, the ℓ_2-norm is computed with Theorem 3.1 on both ID and OOD data. Similarly to subsection 5.1, we sampled 10'000 Gaussian data points around the input with a deviation $\sigma = 0.12$. Since, the certified bound is only probabilistic in practice, we ran a binomial proportion confidence test [34] with failure probability of 0.001. We have assigned a score of 0 to all samples that fail to be certified, i.e. with $p \leq 1/2$. The Lipschitz continuity does not hold in the case of non-certified samples, therefore we are unable to bound the score. To ensure a fair comparison, we decided to compute the ℓ_2-norm GAUC on both ID and OOD datasets.

In Figure 3, we plot the normalized frequency of occurrences of the certified upper bound ($\sqrt{2/\pi} \cdot \Phi^{-1}(p) + p$) for ID versus OOD data of DISTRO. We observe that OOD data tend to peak close to zero, while ID data are spread out with larger values. This suggests that a large radius is more likely to be associated with ID data versus OOD samples. As a result, robustly certifying the detection of OOD samples becomes more feasible.

5.3. Out-Of-Distribution Results

Here, we describe the results shown in Table 3. As previously, we grayed-out models with an accuracy drop greater than 3% with respect to the model with highest accuracy. The objective of this choice is to prioritize clean ID accuracy over all other metrics. A comparison of the remaining metrics is then made on an equal basis. Despite this, there is no direct comparison between the GAUC of ℓ_2 and ℓ_∞ norms. This is primarily due to the fact that the guaranteed upper bound of ℓ_∞ is computed only for OOD data, whereas ℓ_2 is computed for both (ID & OOD). Additionally, we choose any radius $R > 0$ for ℓ_2, while for ℓ_∞, ϵ is fixed to 0.01[7].

We observe that the performances of LogitNorm and VOS on clean AUC, AUPR and FPR are suboptimal. The reason for this is that we are evaluating MSP [6] instead of the suggested normalization [9] and energy [8] functions for LogitNorm and VOS, respectively. To ensure a fair comparison we decided to standardize the output function across all models. On CIFAR100, only the most effective methods of CIFAR10 have been tested.

Outcomes. In light of these considerations, we note that OE achieved the highest clean AUC, AUPR, and FPR. In case of AAUC, ACET shows the best results for CIFAR10. While ATOM achieves close to optimal performance for the guaranteed ℓ_2-norm AUC, AUPR and FPR. Both methods are trained adversarially on outliers, which makes them more robust on OOD data, but at the expense of a reduced clean accuracy.

Similarly to the ID results, DISTRO demonstrates the potential benefits of diffusion models to augment the model robustness in terms of ℓ_2-norm guaranteed and adversarial AUCs. Although there is a slight decrease in ℓ_∞-norm GAUC, GAUPR and GFPR, which could likely be suppressed by fine tuning the classifier in conjunction with the denoiser. In Table 4, we average all the metrics of Table 3 for CIFAR10 (including clean ID accuracy). Surprisingly, ATOM shows similar results as ProoD and GOOD. This can be related to the high certification radius obtained for GAUC of ℓ_2-norm.

Table 4: Overall average between the metrics of Table 3 for CIFAR10/100 (C-10, C-100).

Method	Average	
	C-10	C-100
Plain	44.02	34.48
OE	50.12	40.42
VOS	38.60	-
LogitNorm	46.31	-
ACET	59.64	41.86
ATOM	64.79	54.38
GOOD_{80}	64.74	-
ProoD $\Delta = 3$	64.09	52.51
DISTRO (our)	**77.08**	**59.95**

5.3.1. Similar Model Capacity

Here, we outline the configurations and results of Table 5. Each technique is evaluated using the same architecture, acknowledging that the results from Table 3 do not depend just on the performance of the method, but also on the robustness of the model and the specific

[7] This problem can be addressed by considering $R \geq \sqrt{d} \cdot \epsilon$.

Table 3

Robust OOD detection. We consider the following metrics: clean top-1 accuracy on CIFAR10/100 test sets, clean AUC, guaranteed (GAUC), adversarial AUC (AAUC), clean AUPR, guaranteed AUPR (GAUPR), adversarial AUPR (AAUPR), clean FPR95% (FPR), guaranteed FPR95% (GFPR) and adversarial FPR95% (AFPR). Averaging was performed on a variety of OOD datasets. We consider MSP [6] for all methods and metrics (with temperature $T = 1$). The guaranteed ℓ_2-norm is computed for $\sigma = 0.12$ for all $R > 0$, while the adversarial and guaranteed ℓ_∞-norm are computed for $\epsilon = 0.01$. The grayed-out models have an accuracy drop greater than 3% relative to the model with the highest accuracy. **Bold** numbers are superior results.

ID: CIFAR10	Acc.	AUC↑	GAUC↑		AAUC↑	AUPR↑	GAUPR↑		AAUPR↑	FPR↓	GFPR↓		AFPR↓
			ℓ_2	ℓ_∞	ℓ_∞		ℓ_2	ℓ_∞	ℓ_∞		ℓ_2	ℓ_∞	ℓ_∞
- **Standard**													
Plain*	95.01	94.56	48.86	0.00	24.52	99.42	60.05	0.00	82.30	35.72	100.0	100.0	96.72
OE*	**95.53**	**98.78**	46.88	0.00	37.91	**99.87**	63.08	0.00	84.49	**4.71**	100.0	100.0	70.26
VOS†	94.62	90.82	30.13	0.00	20.62	99.15	41.62	0.00	81.80	61.66	94.10	100.0	100.0
LogitNorm‡	94.48	96.71	40.73	0.00	39.76	99.64	49.31	0.00	86.47	13.95	100.0	100.0	91.10
- **Adversarial**													
ACET*	91.48	97.24	60.21	0.00	93.01	99.68	76.22	0.00	99.16	13.82	95.65	100.0	32.15
ATOM*	92.33	98.82	97.15	0.00	44.65	99.86	95.51	0.00	85.74	4.14	5.04	100.0	62.65
- **Guaranteed**													
GOOD^*_{80}	90.13	93.12	36.45	57.52	78.11	99.22	52.31	89.54	95.19	30.00	100.0	72.45	47.55
ProoD*$\Delta = 3$	95.46	98.72	52.36	**59.56**	64.22	**99.87**	66.53	**93.89**	94.52	5.49	100.0	100.0	86.49
DISTRO (our)	95.47	98.72	**88.97**	59.53	**83.24**	**99.87**	**92.75**	**93.89**	**97.32**	5.29	**67.86**	100.0	**34.56**
ID: CIFAR100	Acc.	AUC↑	GAUC↑		AAUC↑	AUPR↑	GAUPR↑		AAUPR↑	FPR↓	GFPR↓		AFPR↓
			ℓ_2	ℓ_∞	ℓ_∞		ℓ_2	ℓ_∞	ℓ_∞		ℓ_2	ℓ_∞	ℓ_∞
- **Standard**													
Plain*	**77.38**	81.60	30.63	0.00	16.98	97.84	45.10	0.00	81.27	82.52	100.0	100.0	100.0
OE*	77.28	90.41	39.87	0.00	22.79	98.90	49.46	0.00	81.96	47.49	100.0	100.0	87.74
- **Adversarial**													
ACET*	74.47	90.27	36.36	0.00	27.68	98.84	43.50	0.00	82.60	44.11	**90.41**	100.0	74.99
ATOM*	71.73	91.72	84.38	0.00	31.52	98.88	79.95	0.00	83.36	30.81	30.09	100.0	73.69
- **Guaranteed**													
ProoD*$\Delta = 1$	76.79	**90.90**	42.83	**37.67**	43.81	**98.91**	50.90	**89.66**	90.46	42.12	100.0	100.0	97.11
DISTRO (our)	76.78	90.89	**59.39**	37.53	**62.77**	98.90	**69.41**	89.63	**93.59**	**40.94**	100.0	100.0	**58.58**

$*$ Pre-trained models from Meinke et al. [14], † Pre-trained from Du et al. [8], ‡ Pre-trained from Wei et al. [9].

OOD dataset utilized. Therefore we retrain all presented methods using a ResNet18 [26] architecture for CIFAR10 and CIFAR100 respectively. For methods that require an additional OOD dataset for training, such as OE [7], ACET [10], ATOM [12], ProoD [14] and DISTRO, we use the same subset of OpenImages [35] containing 50'000 images. Furthermore, we consider an input normalization of 0.5 across all dimensions for both mean and standard deviation. In addition, we attempt to be as minimally intrusive as possible when it comes to the default training procedure.

For Plain, OE and LogitNorm we run the implementation[8] from Yang et al. [36] and leave the hyperparameters unchanged. Similarly for ACET and ATOM, we only change the model architecture and normalization and run both implementations from ATOM[9]. Lastly, we train ProoD[10] from Meinke et al. [14] using their training configuration files, where the discriminator is trained for 1000 epochs and the bias shift (Δ) is 3/1 for CIFAR10/100, respectively.

Discussion

It is evident that the ℓ_2-norm GAUC (and GAUPR) diverge from zero when standard OOD detection models are considered. This illustrates the potential of the ℓ_2-norm to provide certified OOD detection for any method and architecture. Consequently, it facilitates the experimental evaluation of new robust OOD detection algorithms (both adversarial and certified).

As a side note, the one-shot denoiser appears to improve robustness certification metrics while not compromising clean metrics, such as AUC. In some cases, it also appears to be slightly better, even though the denoising process should produce images that are as similar as possible to those considered during training. This is because a single shot of denoising does not compromise the OOD sample or generate an allucinated one. Additionally, one-shot denoising introduces so little variance that in this benchmark, the results were similar across multiple runs.

6. Conclusion

Current OOD robustness certification relies on external discriminators or loose certification mechanisms [14].

[8] https://github.com/Jingkang50/OpenOOD

[9] https://github.com/jfc43/informative-outlier-mining

[10] https://github.com/AlexMeinke/Provable-OOD-Detection

Table 5
Robust OOD detection with ResNet18. We consider the following metrics: clean top-1 accuracy on `CIFAR10/100` test sets, clean AUC, guaranteed (GAUC), adversarial AUC (AAUC), clean AUPR, guaranteed AUPR (GAUPR), adversarial AUPR (AAUPR), clean FPR95% (FPR), guaranteed FPR95% (GFPR) and adversarial FPR95% (AFPR). Averaging was performed on a variety of OOD datasets. We consider MSP [6] for all methods and metrics (with temperature $T = 1$). The guaranteed ℓ_2-norm is computed for $\sigma = 0.12$ for all $R > 0$, while the adversarial and guaranteed ℓ_∞-norm are computed for $\epsilon = 0.01$. The grayed-out models have an accuracy drop greater than 3% relative to the model with the highest accuracy. **Bold** numbers are superior results.

ID: CIFAR10	Acc.	AUC↑	GAUC↑		AAUC↑	AUPR↑	GAUPR↑		AAUPR↑	FPR↓	GFPR↓		AFPR↓
			ℓ_2	ℓ_∞	ℓ_∞		ℓ_2	ℓ_∞	ℓ_∞		ℓ_2	ℓ_∞	ℓ_∞
Plain	94.32	92.28	35.81	0.00	23.71	99.00	46.83	0.00	82.00	40.21	93.56	100.0	98.88
LogitNorm	94.71	95.58	34.19	0.00	35.00	99.54	49.63	0.00	85.14	33.06	95.12	100.0	92.20
OE	92.41	97.35	50.56	0.00	37.95	99.71	62.25	0.00	85.51	13.44	100.0	100.0	74.91
ACET	93.66	**97.86**	37.45	0.00	65.21	**99.75**	50.26	0.00	91.99	8.94	100.0	100.0	**50.29**
ATOM	91.90	98.12	97.98	97.63	62.79	99.78	98.16	99.78	91.49	8.7	9.42	0.00	51.56
ProoD	**95.20**	96.91	44.95	**63.44**	64.61	99.63	60.27	**94.37**	94.42	**16.03**	100.0	**91.90**	78.22
DISTRO (our)	**95.20**	96.80	**86.63**	59.86	**71.70**	99.62	**90.80**	93.78	**95.72**	16.55	**66.88**	99.96	67.59

ID: CIFAR100	Acc.	AUC↑	GAUC↑		AAUC↑	AUPR↑	GAUPR↑		AAUPR↑	FPR↓	GFPR↓		AFPR↓
			ℓ_2	ℓ_∞	ℓ_∞		ℓ_2	ℓ_∞	ℓ_∞		ℓ_2	ℓ_∞	ℓ_∞
Plain	77.54	84.50	38.11	0.00	24.17	98.16	44.96	0.00	82.32	67.61	100.0	100.0	98.04
LogitNorm	76.25	84.06	40.93	0.00	47.64	98.04	46.80	0.00	87.25	73.70	100.0	100.0	87.98
OE	75.84	88.96	38.90	0.00	17.90	**98.72**	48.82	0.00	81.43	49.61	100.0	100.0	99.41
ACET	73.71	95.65	42.03	0.00	52.49	99.44	48.54	0.00	89.23	13.96	100.0	100.0	60.39
ProoD	**77.77**	**89.47**	40.72	**37.68**	49.16	98.66	49.97	**89.66**	91.08	**40.44**	100.0	100.0	84.15
DISTRO (our)	77.73	88.90	**55.57**	29.71	**51.89**	98.60	**67.62**	87.44	**91.71**	43.24	100.0	100.0	**79.34**

We propose an alternative using randomized smoothing [4] for ℓ_2-norm certificates, applicable to any classifier without specific requirements or training. In comparison with previously proposed ℓ_∞-norm GAUC, standard approaches for OOD detection show non-zero results for guaranteed ℓ_2-norm AUC and AUPR. Unfortunately, a large number of samples derived around the input must be propagated through the network, increasing computational costs. Additionally, we propose a method combining three techniques: diffusion denoising for noise removal, an OOD detection method, and a certified binary discriminator. This combination improves OOD robustness detection by around 13%/5% on CIFAR10/100 datasets compared to earlier approaches.

References

[1] D. Hendrycks, T. Dietterich, Benchmarking neural network robustness to common corruptions and perturbations, in: International Conference on Learning Representations, 2018.

[2] N. Carlini, D. Wagner, Towards evaluating the robustness of neural networks, in: 2017 IEEE Symposium on Security and Privacy (SP), 2017, pp. 39–57.

[3] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, A. Vladu, Towards deep learning models resistant to adversarial attacks, in: International Conference on Learning Representations, 2018.

[4] J. Cohen, E. Rosenfeld, Z. Kolter, Certified adversarial robustness via randomized smoothing, in: ICML, PMLR, 2019, pp. 1310–1320.

[5] A. Paleyes, R.-G. Urma, N. D. Lawrence, Challenges in deploying machine learning: a survey of case studies, ACM Computing Surveys 55 (2022) 1–29.

[6] D. Hendrycks, K. Gimpel, A baseline for detecting misclassified and out-of-distribution examples in neural networks, in: International Conference on Learning Representations, 2017.

[7] D. Hendrycks, M. Mazeika, T. G. Dietterich, Deep anomaly detection with outlier exposure, in: 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019, OpenReview.net, 2019.

[8] X. Du, Z. Wang, M. Cai, Y. Li, Vos: Learning what you don't know by virtual outlier synthesis, in: International Conference on Learning Representations, 2021.

[9] H. Wei, R. Xie, H. Cheng, L. Feng, B. An, Y. Li, Mitigating neural network overconfidence with logit normalization, in: K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, S. Sabato (Eds.), Proceedings of the 39th International Conference on Machine Learning, volume 162 of *Proceedings of Machine Learning Research*, PMLR, 2022, pp. 23631–23644.

[10] M. Hein, M. Andriushchenko, J. Bitterwolf, Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem, in: CVPR 2019, Long Beach, CA, USA, June 16-20, 2019, Computer Vision Foundation / IEEE, 2019, pp. 41–50.

[11] A. Meinke, M. Hein, Towards neural networks that provably know when they don't know, in: 8th Inter-

national Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020, OpenReview.net, 2020.

[12] J. Chen, Y. Li, X. Wu, Y. Liang, S. Jha, Atom: Robustifying out-of-distribution detection using outlier mining, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2021, pp. 430–445.

[13] J. Bitterwolf, A. Meinke, M. Hein, Certifiably adversarially robust detection of out-of-distribution data, in: Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, 2020.

[14] A. Meinke, J. Bitterwolf, M. Hein, Provably robust detection of out-of-distribution data (almost) for free, in: NeurIPS, 2022.

[15] M. Mirman, T. Gehr, M. Vechev, Differentiable abstract interpretation for provably robust neural networks, in: ICML, PMLR, 2018, pp. 3578–3586.

[16] A. Q. Nichol, P. Dhariwal, Improved denoising diffusion probabilistic models, in: International Conference on Machine Learning, PMLR, 2021, pp. 8162–8171.

[17] N. Carlini, F. Tramer, K. D. Dvijotham, L. Rice, M. Sun, J. Z. Kolter, (certified!!) adversarial robustness for free!, in: The Eleventh International Conference on Learning Representations, 2023.

[18] G. Yang, T. Duan, J. E. Hu, H. Salman, I. Razenshteyn, J. Li, Randomized smoothing of all shapes and sizes, in: ICML, PMLR, 2020, pp. 10693–10705.

[19] H. Salman, J. Li, I. Razenshteyn, P. Zhang, H. Zhang, S. Bubeck, G. Yang, Provably robust deep learning via adversarially trained smoothed classifiers, in: H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 32, Curran Associates, Inc., 2019.

[20] C. M. Stein, Estimation of the mean of a multivariate normal distribution, The annals of Statistics (1981) 1135–1151.

[21] H. Salman, M. Sun, G. Yang, A. Kapoor, J. Z. Kolter, Denoised smoothing: A provable defense for pretrained classifiers, Advances in Neural Information Processing Systems 33 (2020) 21945–21957.

[22] W. Liu, X. Wang, J. Owens, Y. Li, Energy-based out-of-distribution detection, Advances in Neural Information Processing Systems 33 (2020) 21464–21475.

[23] F. Croce, M. Hein, Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks, in: H. D. III, A. Singh (Eds.), Proceedings of the 37th International Conference on Machine Learning, volume 119 of *Proceedings of Machine Learning Research*, PMLR, 2020, pp. 2206–2216.

[24] K. Hung, W. Fithian, Rank verification for exponential families, The Annals of Statistics 47 (2019) 758–782.

[25] N. Jovanovic, M. Balunovic, M. Baader, M. T. Vechev, Certified defenses: Why tighter relaxations may hurt training?, CoRR abs/2102.06700 (2021).

[26] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[27] G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4700–4708.

[28] S. Zagoruyko, N. Komodakis, Wide residual networks, in: British Machine Vision Conference 2016, British Machine Vision Association, 2016.

[29] A. Krizhevsky, V. Nair, G. Hinton, Cifar-10 and cifar-100 datasets, URl: https://www. cs. toronto. edu/kriz/cifar. html 6 (2009) 1.

[30] P. Sermanet, S. Chintala, Y. LeCun, Convolutional neural networks applied to house numbers digit classification, in: Proceedings of the 21st international conference on pattern recognition (ICPR2012), 2012, pp. 3288–3291.

[31] F. Yu, A. Seff, Y. Zhang, S. Song, T. Funkhouser, J. Xiao, Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop, arXiv preprint arXiv:1506.03365 (2015).

[32] Y. Le, X. Yang, Tiny imagenet visual recognition challenge, CS 231N 7 (2015) 3.

[33] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , A. Vedaldi, Describing textures in the wild, in: Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2014.

[34] L. D. Brown, T. T. Cai, A. DasGupta, Interval Estimation for a Binomial Proportion, Statistical Science 16 (2001) 101 – 133.

[35] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, A. Kolesnikov, et al., The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale, International Journal of Computer Vision 128 (2020) 1956–1981.

[36] J. Yang, P. Wang, D. Zou, Z. Zhou, K. Ding, W. PENG, H. Wang, G. Chen, B. Li, Y. Sun, X. Du, K. Zhou, W. Zhang, D. Hendrycks, Y. Li, Z. Liu, OpenOOD: Benchmarking generalized out-of-distribution detection, in: Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track, 2022.