Certified but Fooled! Breaking Certified Defences with Ghost Certificates

Quoc Viet Vo[1], Tashreque Mohammed Haq[1], Paul Montague[3], Tamas Abraham[3], Ehsan Abbasnejad[2], Damith C. Ranasinghe[1]

[1]University of Adelaide
[2]Monash University
[3]Defence Science and Technology Group

quocviet.vo@adelaide.edu.au, tashrequemohammed.haq@adelaide.edu.au, paul.montague@defence.gov.au, tamas.abraham@defence.gov.au, ehsan.abbasnejad@monash.edu, damith.ranasinghe@adelaide.edu.au

Abstract

Certified defenses promise provable robustness guarantees. We study the malicious exploitation of probabilistic certification frameworks to better understand the limits of guarantee provisions. Now, the objective is to not only mislead a classifier, but also manipulate the certification process to generate a robustness guarantee for an adversarial input—*certificate spoofing*. A recent study in ICLR demonstrated that crafting large perturbations can shift inputs far into regions capable of generating a certificate for an incorrect class. Our study investigates if perturbations needed to cause a misclassification and yet coax a certified model into issuing a deceptive, large robustness radius for a target class can still be made *small* and *imperceptible*. We explore the idea of region-focused adversarial examples to craft imperceptible perturbations, spoof certificates and achieve certification radii larger than the source class—*ghost certificates*. Extensive evaluations with the `ImageNet` demonstrate the ability to effectively bypass state-of-the-art certified defenses such as Densepure. Our work underscores the need to better understand the limits of robustness certification methods.

Introduction

Deep neural networks (DNNs) are vulnerable to adversarial examples—carefully crafted perturbations to manipulate inputs to coerce incorrect model decisions whilst remaining imperceptible to human observers (Biggio et al. 2013; Szegedy et al. 2014; Papernot et al. 2017; Carlini and Wagner 2017b; Madry et al. 2018; Brendel, Rauber, and Bethge 2018). In response, various empirical defenses such as adversarial training and preprocessing inputs are proposed (Dalvi et al. 2004; Papernot et al. 2016; Buckman et al. 2018; Guo et al. 2018; Samangouei, Kabkab, and Chellappa 2018; Shafahi et al. 2019; Wong, Rice, and Kolter 2020; Rebuffi et al. 2021; Doan et al. 2022). But, demonstrating empirical robustness alone does not facilitate reasoning about robustness guarantees, and strong adaptive attacks can bypass defenses (Carlini and Wagner 2017a; Athalye, Carlini, and Wagner 2018; Tramer et al. 2020; Croce and Hein 2020; Wong, Rice, and Kolter 2020). So, certified robustness has emerged to provide provable lower bounds on model accuracy under bounded perturbations.

Conceptually, certification methods provide both a *model* for a task to generate a prediction and a *verifier* for generating a certificate guaranteeing an input image is not an adversarial example under a predefined threat model. Existing certification methods in the vision domain predominantly focus on l_2 or l_∞-bounded threat models, effectively ensuring that all inputs within an ϵ-ball neighbourhood of a given image are consistently classified under the same label. We focus on the probabilistic *Randomized Smoothing* (Cohen, Rosenfeld, and Kolter 2019; Zhang et al. 2019) frameworks offering scalable certification for tasks.

Unlike conventional attacks—eg., PGD (Madry et al. 2018), Wasserstein (Levine and Feizi 2020), and semantic attacks (Shahin Shamsabadi, Sanchez-Matilla, and Cavallaro 2020; Bhattad et al. 2020)—we explore the adversarial exploitation of certified models that *both* mislead the classifier and falsely receive robustness certificates. If attackers can generate an adversarial example that remains imperceptible and semantically consistent, while ensuring neighbouring images within a certification radius share an incorrect label—a verifier is fooled with a spoofed certificate (Ghiasi, Shafahi, and Goldstein 2020).

Our Study. Spoofing attacks introduce a new dimension to the adversarial threat landscape by targeting *both* the classification and certification processes. To investigate the assurances provided by certified defences, we introduce a new certificate spoofing mechanism that systematically undermines classification and certification mechanisms.

> Our idea is to explore *if* region-based manipulation of inputs to constrain translation in the input-space—a means to preserve input semantics whilst minimizing manipulation—can still shift inputs in the latent-space just far enough to achieve a misclassified label from a model and yet, a *large* radius certificate from a verifier.

To achieve our idea, we construct a malicious objective to induce high-probability of misclassified neighborhoods to yield certified radii—beyond just fooling classifiers.

Contributions and Findings

- We propose a new algorithm for an adversarial attack to spoof robustness certificates. Unlike the prior study's constrained manipulation of an input image to seek certifiable adversarials, we consider a region-based approach

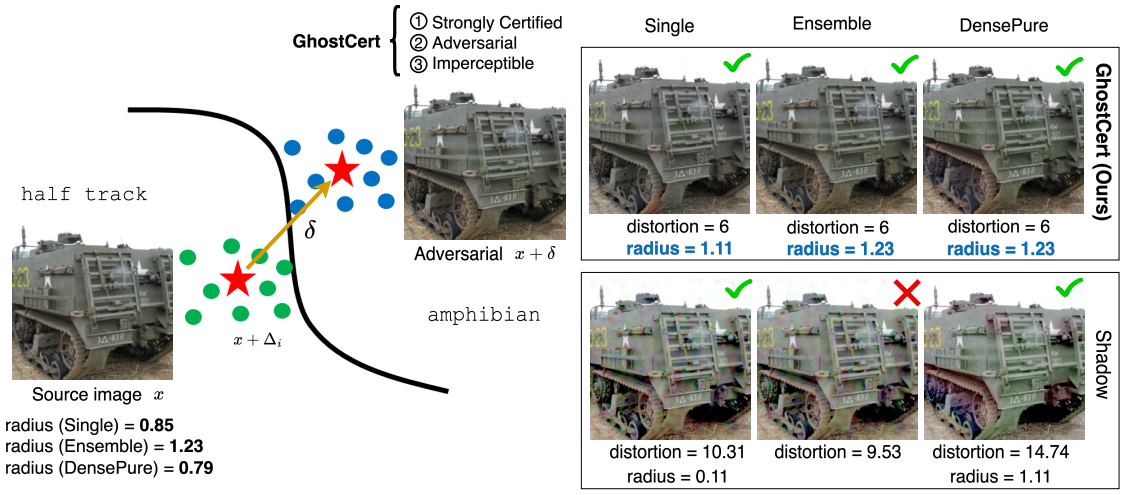


Figure 1: Overview of our attack formulation. For a given source image (`half track`) and certification radii, we show the corresponding adversarial examples created by our attack GhostCert and Shadow Attack in ICLR (Ghiasi, Shafahi, and Goldstein 2020) against three certified defense methods: Randomized Smoothing (with Resnet50), Smoothed Ensemble, and DensePure (Diffusion based denoiser & Transformer under Randomized Smoothing). Shadow fails to generate a spoofed certificate (✗) for Smoothed Ensemble even with a *larger* distortion. GhostCert generates more *natural-looking* adversarials across all three defenses while achieving misclassification with: i) higher spoofed certification radii; and ii) significantly lower l_2 norms ($||\delta||_2$) compared to the Shadow Attack (*Adversarial* and *Imperceptible*). GhostCert results also surpass the certification radii of the source image (*Strongly Certified*)—see Fig. 7 for results of a user-study on imperceptibility. **Code**: https://github.com/ghostcert

to select areas for perturbation. This enables the perturbation to remain *imperceptible* but more efficacious and evasive because the region selection considers *natural* image boundaries and *salient* regions.

- We provide a rigorous evaluation of our attack algorithm, dubbed GhostCert, with the large-scale `ImageNet` task. To advance the prior evaluation of untargeted attacks, we construct *targetted* attacks, including a significant effort to evaluate with a state-of-the-art certification method based on diffusion models (DensePure).
- We raise awareness of weaknesses in current robustness certification methods & discuss the implications of attacks targeting both classification and certification.

Importantly, while our new exploit demonstrates that existing certification methods can be exploited with stealth, the attack does not invalidate the certificates produced—the assertion made by a certificate that an input is not an adversarial example in the chosen bounded norm is still correct. *Rather, our attack is a cautionary tale*.

Key Takeaways: Safe deployment of systems with certifiably robust models should use certificates as an indicator of label correctness with caution. A strongly certified sample input (with a large certification radius) does not necessarily imply correctness nor that a potential manipulation to spoof a certificate will lead to easily visible evidence. We confirm, even targetted attacks (certification for a target label chosen by an attacker) are possible. But, as expected, are harder. The state-of-the-art denoiser-based method remains the most effective certification method, even when compared to model ensembling under randomised smoothing. We hope the study helps reveal and deepen understanding of flaws in certified defenses for adversarial robustness.

Related Work

Adversarial attack algorithms can launch powerful attacks like Projected Gradient Descent (PGD) to craft and apply imperceptible perturbations to inputs to mislead or hijack the decision of deep learning models (Szegedy et al. 2014; Papernot et al. 2017; Carlini and Wagner 2017b; Madry et al. 2018; Athalye, Carlini, and Wagner 2018).

The recent Shadow Attack in (Ghiasi, Shafahi, and Goldstein 2020) exposes a weakness in a certified defences. The attack generates large perturbations in the input-space to move an image *far* from a class boundary to a region capable of generating a fake certificate with a large radius. The attack augments PGD to constrain semantic changes and perturbations with three penalties: i) to force the perturbation δ to have small total variation to attempt to appear smooth and natural; ii) to limit the perturbation δ by constraining the change in the mean of each color channel; and iii) to promote perturbations that assume similar values in each colour channel to suppress extreme or dramatic colour changes. *We devise a simpler attack and demonstrate that such large perturbations are not necessary to break a certified defence*.

Preliminaries on Scalable Certification

Complete certification methods guarantee finding adversarial examples if they exist, but are limited to small datasets and simple models due to scalability constraints (Weng et al. 2018). Complete methods are typically restricted to specific architectures and struggle with large-scale tasks (Cohen, Rosenfeld, and Kolter 2019; Hayes 2020) like

ImageNet (Deng et al. 2009a). Incomplete approaches, including deterministic (Lyu et al. 2020; Levine and Feizi 2020) and probabilistic methods, can certify the lower bound of model performance under certain ℓ_p norm attacks or abstain from deciding (Li, Xie, and Li 2023). Lecuyer et al. (2019) introduced the incomplete method—*randomized smoothing*—using Gaussian and Laplace noise. This approach offers a general and scalable approach for certification on large scale tasks such as ImageNet. The approach provides a non-trivial probabilistic robustness guarantee. Subsequent studies tightened the bound (Cohen, Rosenfeld, and Kolter 2019) and further improved certified performance by integrating adversarial training (Salman et al. 2019), consistency regularization (Jeong and Shin 2020), and model ensembling (Horváth et al. 2022).

Randomized Smoothing. Consider a classification problem from $\boldsymbol{x} \in \mathbb{R}^d$ to classes $\mathcal{Y}$. As introduced in (Cohen, Rosenfeld, and Kolter 2019), randomized smoothing is a method for constructing a new, "smoothed" classifier g from an arbitrary base classifier f. When queried by $\boldsymbol{x}$, the smoothed classifier g returns what the base classifier f is most likely to return when $\boldsymbol{x}$ is perturbed by noise $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I)$:

$$g(\boldsymbol{x}) = \arg\max_{c \in \mathcal{Y}} \mathbb{P}(f(\boldsymbol{x} + \boldsymbol{\varepsilon}) = c) \tag{1}$$

The noise level σ is a hyperparameter of the smoothed classifier g which controls a robustness/accuracy tradeoff; it does not change with the input $\boldsymbol{x}$.

Smoothed Ensemble. Recent work (Horváth et al. 2022) has introduced several advancements to enhance Randomized Smoothing. For a set of k classifiers $\{f^l : \mathbb{R}^d \to \mathbb{R}^m\}_{l=1}^k$, a soft-ensemble $\bar{f}$ is constructed by averaging the logits $\bar{f}(x) = \frac{1}{k}\sum_{l=1}^k f^l(x)$, where, $f^l(x)$ are the pre-softmax outputs. With $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$, the smoothed ensemble can be formulated as follows:

$$g_e(\boldsymbol{x}) = \arg\max_{c \in \mathcal{Y}} \mathbb{P}(\bar{f}(\boldsymbol{x} + \boldsymbol{\varepsilon}) = c) \tag{2}$$

Denoised Smoothing. An alternative approach to obtaining a provably robust classifier *without retraining* the underlying model has been proposed through the idea of *denoised smoothing* (Salman et al. 2020). Unlike prior work that primarily focus on training classifiers to withstand Gaussian perturbations—often using Gaussian noise augmentation (Cohen, Rosenfeld, and Kolter 2019) or adversarial training (Salman et al. 2019)—this method leaves the pre-trained classifier unchanged. Building on this idea, several recent methods have explored the use of diffusion-based denoiser such as DiffusionDenoisedSmoothing (DDS) (Carlini et al. 2023), DensePure (Xiao et al. 2023) or DiffSmooth (Zhang et al. 2023). Denoised Smoothing essentially augments the base classifier f with a denoiser $D_\theta : \mathbb{R}^d \to \mathbb{R}^d$ to form a new base classifier defined as $f \circ D_\theta : \mathbb{R}^d \to \mathcal{Y}$. Assuming the denoiser D_θ is effective at removing Gaussian noise, this setup is configured to classify well under Gaussian perturbation of its inputs. Formally, with $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I)$, this procedure can be defined as follow:

$$g_d(\boldsymbol{x}) = \arg\max_{c \in \mathcal{Y}} \mathbb{P}\big[f(D_\theta(\boldsymbol{x} + \boldsymbol{\varepsilon})) = c\big] \tag{3}$$

Certifiable Robustness and Abstaining. (Cohen, Rosenfeld, and Kolter 2019) introduces an analytic form of certifiable robustness that provides a formal guarantee for smoothed classifiers. Concretely, the prediction of a smoothed classifier remains unchanged within a bounded ℓ_2-norm region defined by a certified radius R.

$$R = \frac{\sigma}{2}\left[\Phi^{-1}(p_A) - \Phi^{-1}(p_B)\right], \tag{4}$$

where p_A, p_B are the probabilities of the top class c_A and the runner-up class c_B respectively, Φ^{-1} is the inverse Gaussian cumulative distribution function (CDF). Based on Theorem 1 in (Cohen, Rosenfeld, and Kolter 2019), the smoothed classifier $g(\boldsymbol{x})$ always returns c_A and certified radius $\sigma\Phi^{-1}(\underline{p_A})$ if the lower bound of the probability of the top class $\underline{p_A}$ exceeds 0.5. Otherwise, the smoothed classifier abstains from making a prediction. To prevent an adversary from manipulating the smoothed classifier to abstain at a high rate, a margin μ is introduced as follows:

$$\mathbb{P}(f(\boldsymbol{x} + \boldsymbol{\delta} + \boldsymbol{\varepsilon}) = c_A) - \mathbb{P}(f(\boldsymbol{x} + \boldsymbol{\delta} + \boldsymbol{\varepsilon}) = c_B) \geq \mu \tag{5}$$

Proposed Method

Threat Model. We consider a white-box threat model for attacks on a target DNN (deep neural network), where adversaries have full access to the model's architecture, parameters and the noise level used by the smoothed classifier.

Problem Formulation. Given a neural network f, an input x, and the ground truth label y, we define an adversarial attack as an optimization problem searching for a perturbation δ that maximizes the loss function L. The objective is to generate adversarial examples misleading a smoothed classifier while ensuring the perturbation remains imperceptible.

$$\max_{\delta} \sum_{i=1}^{N} L(f(x + \varepsilon_i + \delta), y) \text{ s.t. } \|\delta\|_2 \leq \epsilon, \tag{6}$$

where $L(\cdot, \cdot)$ is the loss function, δ denotes the adversarial perturbation, ϵ is the perturbation budget, ε_i represents isotropic Gaussian noise applied to the input, N is the number of noisy samples used to approximate the probability mass in Randomized Smoothing. The formulation of this optimization allows to find an adversarial perturbation δ that consistently forces the smoothed model to misclassify the adversarial example $\boldsymbol{x} + \boldsymbol{\delta}$.

When the base model is an ensemble of k classifiers $\{f^l : \mathbb{R}^d \to \mathbb{R}^m\}_{l=1}^k$ and the resulting classifier is $\bar{f}$, the problem can be formulated as:

$$\max_{\delta} \sum_{i=1}^{N} L(\bar{f}(x + \varepsilon_i + \delta), y) \text{ s.t. } \|\delta\|_2 \leq \epsilon, \tag{7}$$

Similarly, when the base model f_θ is subjected to a denoiser D_θ, the new base model becomes $f \circ D_\theta$ and therefore, the problem formulation changes to:

$$\max_{\delta} \sum_{i=1}^{N} L(f(D_\theta(x + \varepsilon_i + \delta)), y) \text{ s.t. } \|\delta\|_2 \leq \epsilon. \tag{8}$$

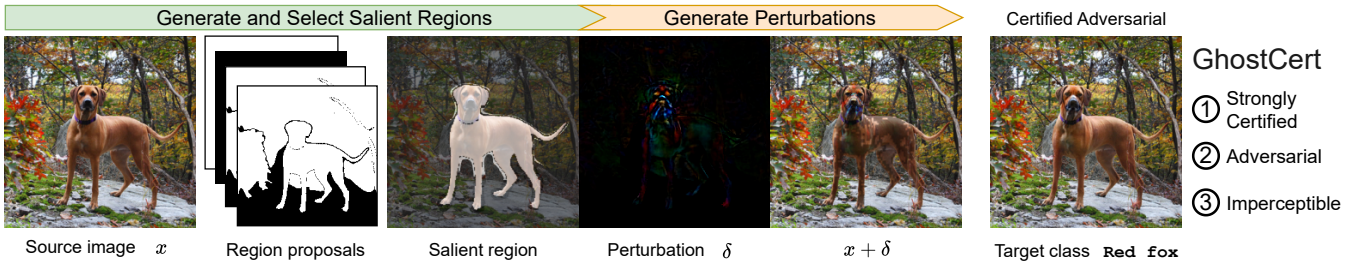


Figure 2: A pictorial illustration of GhostCert. Starting from the source image x with label **Rhodesian ridgeback** and given a target label **Red fox**, region proposal are evaluated to select regions for manipulation considering salient features important for classification decisions. The idea is to preserve semantics whilst minimising distortions. Then, crafting perturbations constrained to the salient regions, δ, yields the adversarial $x + \delta$ misclassified as a **Red fox** while being strongly certified with imperceptible visual differences to the source image x.

Method Intuition

To solve these attack optimization problems, several gradient-based methods such as Fast Gradient Sign Method (FGSM) (Goodfellow, Shlens, and Szegedy 2015), Basic Iterative Method (BIM) (Kurakin, Goodfellow, and Bengio 2017) can be employed. However, Projected Gradient Descent (PGD) (Madry et al. 2018) has emerged and provides superior performance for white-box attacks due to its ability to navigate complex loss surfaces through iterative small steps gradient ascent. Therefore, we adopt PGD in our study to navigate toward the adversarial solution.

While traditional adversarial attacks successfully employ a global perturbation, they have not leveraged saliency information within natural images to enhance the imperceptibility of adversarial perturbation. Recent empirical observations in (Vo, Abbasnejad, and Ranasinghe 2022) demonstrate that, when searching for perturbations to inputs in black-box settings, they tend to concentrate within salient regions of an image, although the attack *does not explicitly target* these regions. This suggests that more effective perturbations could be crafted by manipulating the salient regions.

To identify salient regions with Convolutional-based models, GradCAM (Selvaraju et al. 2017) represents a natural choice due to its widespread adoption in highlighting decision-critical areas. However, GradCAM's gradient-based saliency maps produce amorphous regions that disregard the natural structural boundaries inherent in images. This results in perturbations producing unnatural artifacts and compromising semantic coherence. To overcome this fundamental limitation, we introduce a new notion—*salient-region masks*—that combines GradCAM's gradient-driven saliency information with semantic segmentation boundaries derived from the Segment Anything Model (SAM) (Kirillov et al. 2023). Similarly, for transformer-based models, Attention maps provide an alternative to GradCAM, directly revealing which spatial locations the model prioritizes during classification. This integration ensures that perturbations maintain natural-looking boundaries while maintaining focus on salient regions. Our core hypothesis posits that constraining adversarial perturbations to these semantically-coherent salient regions will yield adversarial examples with imperceptibility and preserved semantic meaning.

GhostCert Attack Algorithm

A pictorial illustration of how GhostCert produces the adversarially perturbed image leading to misclassification and certificate spoofing while being visually imperceptible is shown in Figure 2. By using standard image segmentation techniques to generate region proposals, combined with saliency analysis, to select regions defined by natural image boundaries, we aim to generate more natural-looking adversarial examples, that are strongly certified (large spoofed certification radius, often higher or comparable with the source image). Owing to their imperceptibility objective, we refer to these as ghost certificates and dub our method GhostCert.

Generate Salient-Region Mask. Let $\mathcal{S} \in [0,1]^{H \times W}$ be the saliency map generated by GradCAM or Attention depending on the model under attack, and let $\mathcal{M} = \{M_1, M_2, \ldots, M_n\}$ denote the set of binary segmentation masks produced by the SAM model, each with area greater than 300 pixels. Additionally, let U be a binary unmask candidate mask representing pixels not covered by any of the segmentation masks in $\mathcal{M}$. For each mask $M_i \in \mathcal{M}$, the saliency overlap score is defined as:

$$\text{score}(M_i) = \frac{\sum_{x,y} M_i(x,y) \cdot \mathcal{S}(x,y)}{\sum_{x,y} M_i(x,y) + \sum_{x,y} \mathcal{S}(x,y)}. \quad (9)$$

Similarly, an overlap score for the unmask candidate U is also obtained. Let $\mathcal{T} \subset \mathcal{M} \cup \{U\}$ be the set of top-k masks selected based on the highest scores. The final combined mask—*salient-region mask*—m is then obtained by summing the top-k masks:

$$m = \sum_{M \in \mathcal{T}} M. \quad (10)$$

This salient-region mask m highlights the most salient and semantically meaningful regions, guided by both segmentation and GradCAM/attention-based saliency.

Generate Perturbations. Considering an untargeted attack setting, for every image a batch of noisy images is generated, with each image in the batch subjected to random Gaussian noise of standard deviation σ. The perturbation δ, initially

zero, is added to the batch of noisy images, and the following problem is optimized.

$$\max_{\delta} \sum_{i=1}^{N} L\left(f_\theta\left(x + \Delta_i + \delta \odot m\right), y\right) \quad \text{s.t. } \|\delta \odot m\|_2 \leq \epsilon, \tag{11}$$

where $L(\cdot, \cdot)$ is the loss function. In this work, we use cross-entropy loss. $\odot$ denotes element-wise multiplication, δ is the adversarial perturbation applied to the input, Δ_i represents the random Gaussian noise of standard deviation σ, m is the selective region or mask to perturb, ϵ is the perturbation budget. This selective perturbation, when added to the source image x, while bounded by ϵ produces an adversarial image with visually less perceptible changes. Similarly, when the attack setting is targeted, a target label y_{target} is provided instead of y since the targeted attack aims to fool the model f_θ to predict the class label of image x to be y_{target}.

Attack pipeline. We codify the attack in Algorithm 1. Figure 3 shows samples from successful attacks with GhostCert and the prior Shadow Attack.

Algorithm 1 GhostCert

Require: Input image x, ground truth label y, target label y_{target} (if targeted), noise Δ_i, mask m, step size λ, maximum distortion ϵ, attack type (targeted or untargeted)

Initialize $\delta \leftarrow 0$
for $i = 1$ to N **do**
  if attack is targeted **then**
    $g \leftarrow -\nabla_\delta L(f_\theta(x + \Delta_i + \delta), y_{\text{target}})$
  else
    $g \leftarrow \nabla_\delta L(f_\theta(x + \Delta_i + \delta), y)$
  end if
  $\delta \leftarrow \delta + \lambda \cdot \frac{g}{\|g\|_2}$ ▷ Gradient ascent/descent step
  $\delta \leftarrow \left(\epsilon \frac{\delta}{\|\delta\|_2}\right) \odot m$ ▷ Projection and mask step
end for
return δ

Experiments and Evaluations

Dataset(s). We used the large-scale `ImageNet` (Deng et al. 2009b) validation set for experiments as in Shadow Attack. For each base model and σ combination, under Randomized Smoothing, 100 correctly classified images and their certification radii were identified (notably Shadow Attack used 50 samples). This ensures the attacks were carried out considering only images that were correctly certified by the models under Randomized Smoothing. For each set of 100 images, on average, there were 94 to 97 different class labels, suggesting a low bias towards a particular class label.

Defended models and attacks. We employed three certified defenses: i) Randomized Smoothing (RS) with Resnet50 (Single model) (Cohen, Rosenfeld, and Kolter 2019); and ii) Ensemble of three consistency-trained Resnet50 models (Horváth et al. 2022) under RS (with $\boldsymbol{\sigma} = 0.25, 0.5, 1.0$); and iii) diffusion-based denoiser prepended to a BEiT large Patch16 512 transformer under RS (DensePure) (with $\boldsymbol{\sigma} = 0.25, 0.5$) (Xiao et al. 2023). The ensemble was included as part of the defended models because it is well established that, in terms of certified robustness, an ensemble of consistency-trained models outperforms a single Resnet50, yielding higher certified accuracy under Randomized Smoothing.

To evaluate, three attacks were carried out—GhostCert, current state-of-the-art Shadow Attack, and our modified version Shadow Attack to bound the distortion limit to compare with GhostCert referred to as Shadow Attack (bounded).

Performance Metrics. ***Attack Success Rate (ASR)***. In the untargeted case, it is the proportion of samples misclassified by the defending model. In the targeted case, it is the proportion of samples classified by the defending model as the target class and assigned a certificate (radius)—if the defence abstains from making a prediction, these were recorded as ***Denial of Service (DoS)*** attacks. ***Spoofing Radius***. This is the certification radius calculated for samples that have been successfully misclassified.

Evaluation Protocol. For both untargeted and targeted settings, ASR and the average Spoofing Radius were reported across all defense methods—Single, Ensemble, and DensePure. For GhostCert (Ours) and Bounded Shadow Attack [1], the perturbation budgets (ϵ) were 2, 4, 6, 8, and 10. In the targeted setting, due to computational constraints, only one target label was selected per source image by sequentially searching for the next image with a different label. This process helps eliminate bias in target label selection. Additionally, DoS results were reported for targeted attacks.

Attacking a (Single) and (Ensemble) Classifier Under Randomised Smoothing

Attack Success Rate (ASR). (***Single***) For both untargeted and targeted attacks, GhostCert consistently achieves significantly higher ASR than both bounded and unbounded shadow attacks across all noise levels ($\sigma = 0.25, 0.5, 1.0$) as demonstrated in Figure 4 and Figure 5.

(***Ensemble***) For untargeted attacks, at $\sigma = 0.25$ GhostCert achieves $\approx 100\%$ ASR while shadow attack and its variant plateau at 40% across different ($\sigma = 0.25, 0.5, 1.0$) as shown in Figure 4. The ensemble defense significantly impacts shadow attack and its variant but has minimal effect on GhostCert's effectiveness. For targeted attacks, at $\sigma = 0.25$, GhostCert achieves an ASR of over 80% while the shadow variants max out at just over 20%. At higher noise levels ($\sigma = 0.5, 1.0$), while the ASR is not as high across all attacks, GhostCert still produces significantly larger ASR than the shadow attacks as shown in Figure 5.

Overall, the performance gap is most pronounced at higher noise levels, where GhostCert maintains effectiveness while shadow attacks degrade significantly.

Spoofed Radii. (***Single***) For untargeted attacks, GhostCert generates substantially larger spoofed radii compared to the shadow baselines, particularly evident at $\sigma = 0.5, 1.0$. Notably, GhostCert's spoofed radii consistently exceed or approach the source certified radius (*orange* dashed line), indicating effective circumvention of the defense mechanism. For targeted attacks, GhostCert consistently achieves larger

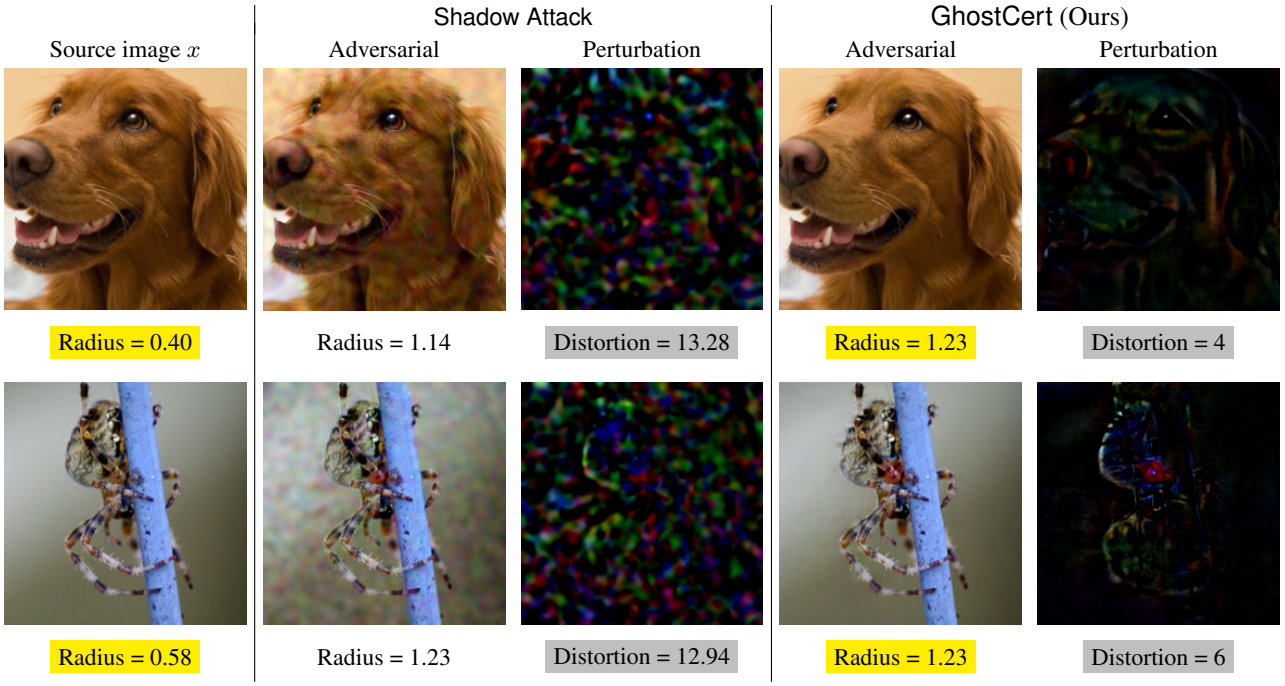


Figure 3: Illustrative examples of successful attacks by GhostCert are presented. For each case, we display the adversarial image and its corresponding perturbation generated by both the Shadow attack and our method, GhostCert. The results clearly show that GhostCert produces strongly certified adversarial examples with perturbations that are more visually imperceptible than those from the Shadow attack, while also achieving higher spoofed certification radii at lower l_2 norms ($||\delta||_2$).

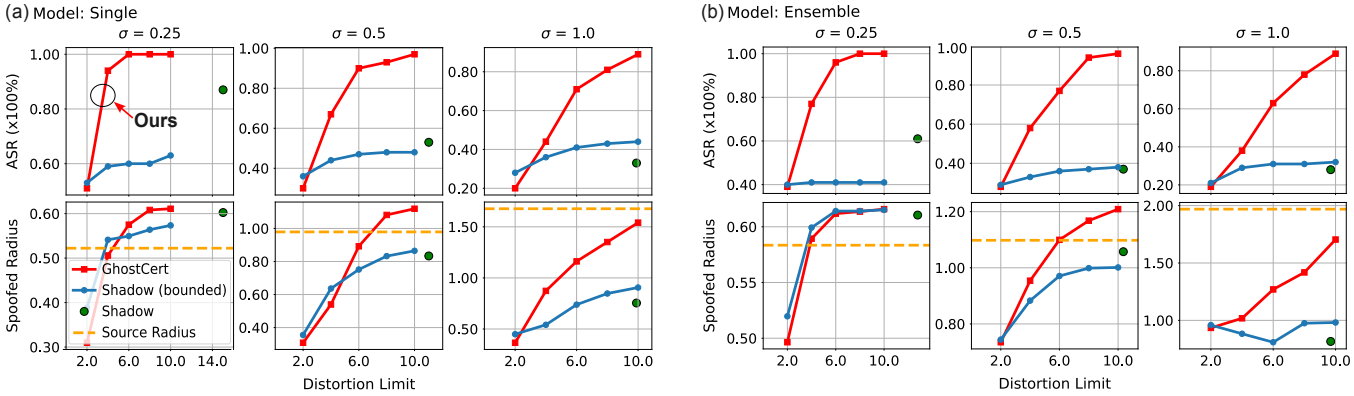


Figure 4: Comparing ASR and spoofed radii for three attacks in *untargeted* settings against (a) single ResNet-50 under Randomized Smoothing (RS) and (b) an ensemble of three consistency ResNet-50 models under RS vs. distortion $\|\delta\|_2$ budgets.

spoofed radii than its Shadow Attack counterparts.

(***Ensemble***) In the untargeted setting, GhostCert consistently generates larger spoofed radii than shadow baselines across all perturbation budgets. At $\sigma = 0.5$, GhostCert's spoofed radii approach or exceed the source radius (orange dashed line), indicating strongly certified defense bypass. At $\sigma = 1.0$, while all methods fall below the original radius due to stronger noise, GhostCert maintains a substantial advantage. In a targeted setting, in cases where there is significant ASR to report, GhostCert consistently generates larger spoofed radii than shadow baselines. In cases where the shadow variants generate similar spoofed radii, the ASR is negligibly low.

DoS Success. Interestingly, certification methods can abstain from making a decision. Failing to certify robustness for an input results in the verifier abstaining instead of making a potentially incorrect or vulnerable prediction. Table 1 reports abstain results observed in targetted attacks. When the given perturbation budget is inadequate for spoofing a certification (indicated by low ASR in Fig. 5), GhostCert input samples lead to a higher DoS success. Importantly, when the ASR for GhostCert is similar to Shadow Attack, the DoS success for our attack is generally higher than both shadow variants. This indicates the region-based perturbations are

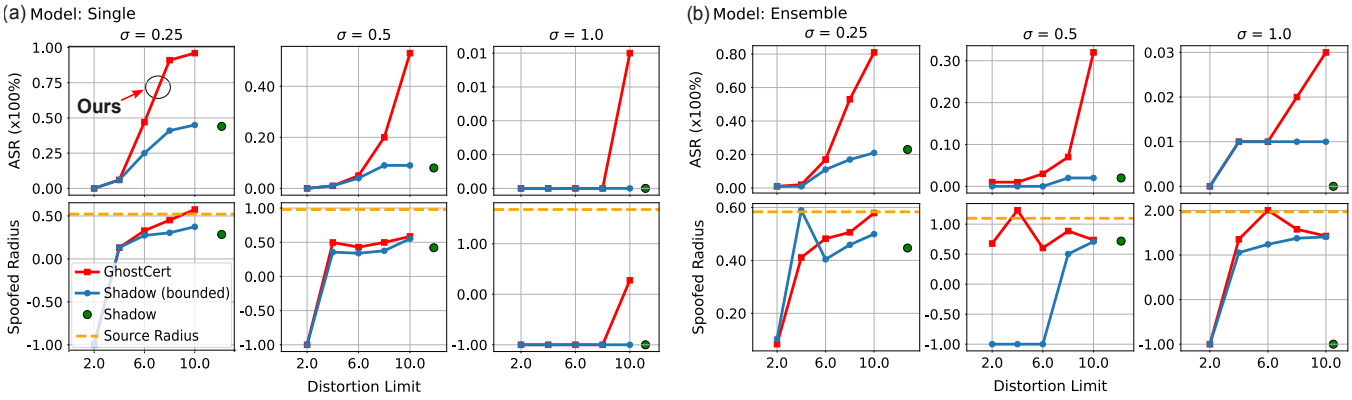


Figure 5: Comparing ASR and spoofed radii for three attacks in a *targeted* setting against (a) single ResNet-50 under Randomized Smoothing (RS) and (b) an ensemble of three consistency ResNet-50 models under RS vs. distortion $\|\delta\|_2$ budgets.

$\|\delta\|_2$	Shadow (σ)			Shadow (Bounded) (σ)			GhostCert (σ)		
	0.25	0.5	1.0	0.25	0.5	1.0	0.25	0.5	1.0
2				14	8	9	8	84	8
4				31	18	18	20	68	12
6	25	27	34	25	29	23	16	52	21
8				24	29	35	4	38	34
10				21	38	39	2	45	45

Table 1: DoS (Abstain) attack success (%) (Single).

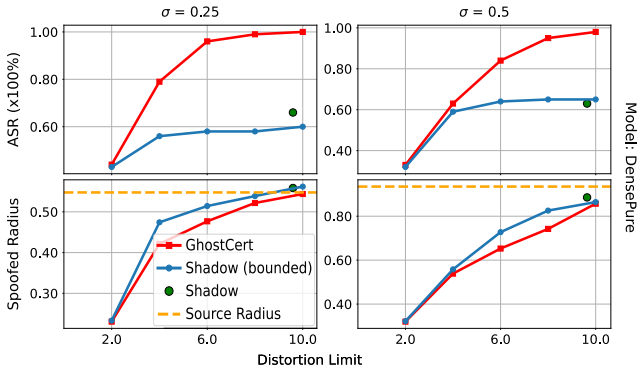
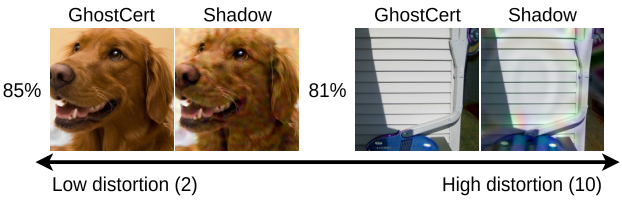


Figure 6: Comparing ASR and spoofed radii between three attacks in an *untargeted* setting against DensePure.

more effective but the adversarial crafted is near a decision boundary and the ϵ-bound it too large to spoof a certificate.

Attacking Denoised Smoothing (DensePure)

Attack Success Rate (ASR). GhostCert demonstrates superior and consistent performance across both noise levels as shown in Figure 6. At ($\sigma = 0.25$) GhostCert significantly outperforms shadow attacks and its variant across different perturbation budgets. GhostCert maintains $30 - 100\%$ success rates while shadow method and its variant remain constrained at $30 - 65\%$. *Spoofed Radii.* Across pertubations budgets, GhostCert achieves spoofed certification radii that are slightly lower or comparable with Shadow Attack but consistently maintains a significantly higher ASR.



Distortion budget $\|\delta\|_2$	2	4	6	8	10
Avg. % Selecting GhostCert	74	69	58	60	62

Additional images and full results are in the Appendix.

Figure 7: Naturalism/imperceptibility of the adversarial images generated by GhostCert vs. Shadow Attack across minimum & maximum distortion budgets. GhostCert images were consistently perceived as *more* natural looking.

Evaluation of Imperceptibility: User Study

A user study compare the perceptual realism of adversarial images generated by GhostCert and Shadow Attack. For each distortion level ($\|\delta\|_2$), 10 successful image pairs (1 as control) were presented to workers on Amazon Mechanical Turk, with the display order of images from Shadow Attack and GhostCert randomized. To reduce noise, responses from workers who answered randomly or spent less than a minute on the task were excluded. The results from nearly valid participants for each distortion level are shown in Figure 7. GhostCert consistently produced images rated as more natural across both high/low distortion levels.

Conclusion

We show region-based input manipulations preserve semantics while subtly shifting inputs to cause misclassification and yet receive large-radius certificates. Our method, GhostCert, outperforms the state-of-the-art Shadow Attack attack by achieving higher success in both misclassification and certificate spoofing, while producing more natural-looking, imperceptible adversarials. Our findings urge caution in using certification frameworks and encourage further research into certification methods and attack vectors.

References

Athalye, A.; Carlini, N.; and Wagner, D. 2018. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International Conference on Machine Learning (ICML)*.

Bhattad, A.; Chong, M. J.; Liang, K.; Li, B.; and Forsyth, D. A. 2020. Unrestricted Adversarial Examples via Semantic Manipulation. In *International Conference on Learning Representations*.

Biggio, B.; Corona, I.; Maiorca, D.; Nelson, B.; Šrndić, N.; Laskov, P.; Giacinto, G.; and Roli, F. 2013. Evasion Attacks against Machine Learning at Test Time. In Blockeel, H.; Kersting, K.; Nijssen, S.; and Železný, F., eds., *Machine Learning and Knowledge Discovery in Databases*, 387–402. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN 978-3-642-40994-3.

Brendel, W.; Rauber, J.; and Bethge, M. 2018. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. In *International Conference on Learning Recognitions (ICLR)*.

Buckman, J.; Roy, A.; Raffel, C.; and Goodfellow, I. 2018. Thermometer Encoding: One Hot Way To Resist Adversarial Examples. In *International Conference on Learning Representations*.

Carlini, N.; Tramèr, F.; Dvijotham, K. D.; and Kolter, J. Z. 2023. (Certified!!) Adversarial Robustness for Free!

Carlini, N.; and Wagner, D. 2017a. Adversarial Examples Are Not Easily Detected: Bypassing Ten Detection Methods. arXiv:1705.07263.

Carlini, N.; and Wagner, D. 2017b. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy (SSP)*.

Cohen, J. M.; Rosenfeld, E.; and Kolter, J. Z. 2019. Certified Adversarial Robustness via Randomized Smoothing. *arXiv preprint arXiv:1902.02918*.

Croce, F.; and Hein, M. 2020. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks.

Dalvi, N.; Domingos, P.; Mausam; Sanghai, S.; and Verma, D. 2004. Adversarial classification. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, 99–108. New York, NY, USA: Association for Computing Machinery. ISBN 1581138881.

Deng, J.; Dong, W.; Socher, R.; Li, L.; Li, K.; and Fei-Fei, L. 2009a. ImageNet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition(CVPR)*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009b. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255.

Doan, B. G.; Abbasnejad, E. M.; Shi, J. Q.; and Ranasinghe, D. C. 2022. Bayesian Learning with Information Gain Provably Bounds Risk for a Robust Adversarial Defense. In *International Conference on Machine Learning (ICML)*.

Ghiasi, A.; Shafahi, A.; and Goldstein, T. 2020. Breaking Certified Defenses: Semantic Adversarial Examples With Spoofed Robustness Certificates. In *International Conference on Learning Representations*.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2015. Explaining and Harnessing Adversarial Examples.

Guo, C.; Rana, M.; Cisse, M.; and van der Maaten, L. 2018. Countering Adversarial Images using Input Transformations. In *International Conference on Learning Representations*.

Hayes, J. 2020. Extensions and limitations of randomized smoothing for robustness guarantees. arXiv:2006.04208.

Horváth, M. Z.; Mueller, M. N.; Fischer, M.; and Vechev, M. 2022. Boosting Randomized Smoothing with Variance Reduced Classifiers. In *International Conference on Learning Representations*.

Jeong, J.; and Shin, J. 2020. Consistency Regularization for Certified Robustness of Smoothed Classifiers. In *Advances in Neural Information Processing Systems*.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4015–4026.

Kurakin, A.; Goodfellow, I. J.; and Bengio, S. 2017. Adversarial examples in the physical world.

Lecuyer, M.; Atlidakis, V.; Geambasu, R.; Hsu, D.; and Jana, S. 2019. Certified Robustness to Adversarial Examples with Differential Privacy. arXiv:1802.03471.

Levine, A.; and Feizi, S. 2020. Wasserstein Smoothing: Certified Robustness against Wasserstein Adversarial Attacks. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*.

Li, L.; Xie, T.; and Li, B. 2023. SoK: Certified Robustness for Deep Neural Networks . In *2023 IEEE Symposium on Security and Privacy (SP)*.

Lyu, Z.; Ko, C.; Kong, Z.; Wong, N.; Lin, D.; and Daniel, L. 2020. Fastened CROWN: Tightened Neural Network Robustness Certificates. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence*.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2018. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Recognitions (ICLR)*.

Papernot, N.; McDaniel, P.; Goodfellow, I.; Jha, S.; Celik, Z. B.; and Swami, A. 2017. Practical Black-Box Attacks against Machine Learning. *ACM Asia Conference on Computer and Communications Security (ASIA CCS)*.

Papernot, N.; McDaniel, P.; Wu, X.; Jha, S.; and Swami, A. 2016. Distillation as a Defense to Adversarial Perturbations Against Deep Neural Networks. In *2016 IEEE Symposium on Security and Privacy (SP)*, 582–597.

Rebuffi, S.-A.; Gowal, S.; Calian, D. A.; Stimberg, F.; Wiles, O.; and Mann, T. 2021. Fixing Data Augmentation to Improve Adversarial Robustness. arXiv:2103.01946.

Salman, H.; Li, J.; Razenshteyn, I.; Zhang, P.; Zhang, H.; Bubeck, S.; and Yang, G. 2019. Provably robust deep learning via adversarially trained smoothed classifiers. In *Advances in Neural Information Processing Systems*.

Salman, H.; Yang, J. L.; Zhang, H.; Zhang, H.; Hsieh, C.-J.; and Madry, A. 2020. Denoised Smoothing: A Provable Defense for Pretrained Classifiers. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Samangouei, P.; Kabkab, M.; and Chellappa, R. 2018. Defense-GAN: Protecting Classifiers Against Adversarial Attacks Using Generative Models. In *International Conference on Learning Representations*.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 618–626.

Shafahi, A.; Najibi, M.; Ghiasi, A.; Xu, Z.; Dickerson, J.; Studer, C.; Davis, L. S.; Taylor, G.; and Goldstein, T. 2019. *Adversarial Training for Free!*

Shahin Shamsabadi, A.; Sanchez-Matilla, R.; and Cavallaro, A. 2020. ColorFool: Semantic Adversarial Colorization . In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; and Fergus, R. 2014. Intriguing properties of neural networks.

Tramer, F.; Carlini, N.; Brendel, W.; and Madry, A. 2020. On adaptive attacks to adversarial example defenses. In *Advances in Neural Information Processing Systems*.

Vo, V. Q.; Abbasnejad, E.; and Ranasinghe, D. C. 2022. RamBoAttack: A Robust Query Efficient Deep Neural Network Decision Exploit. In *Proceedings of the 2022 Network and Distributed System Security Symposium (NDSS)*.

Weng, T.-W.; Zhang, H.; Chen, H.; Song, Z.; Hsieh, C.-J.; Boning, D.; Dhillon, I. S.; and Daniel, L. 2018. Towards Fast Computation of Certified Robustness for ReLU Networks. arXiv:1804.09699.

Wong, E.; Rice, L.; and Kolter, J. Z. 2020. Fast is better than free: Revisiting adversarial training. arXiv:2001.03994.

Xiao, C.; Chen, Z.; Jin, K.; Wang, J.; Nie, W.; Liu, M.; Anandkumar, A.; Li, B.; and Song, D. 2023. DensePure: Understanding Diffusion Models towards Adversarial Robustness.

Zhang, H.; Chen, H.; Xiao, C.; Li, B.; Boning, D.; and Hsieh, C.-J. 2019. Towards Stable and Efficient Training of Verifiably Robust Neural Networks.

Zhang, J.; Chen, Z.; Zhang, H.; Xiao, C.; and Li, B. 2023. DiffSmooth: Certifiably Robust Learning via Diffusion Models and Local Smoothing. arXiv:2308.14333.