IMPROVING RANDOMIZED SMOOTHING CERTIFICATES THROUGH GUIDING DIFFUSION-BASED DENOISERS

Anonymous authors
Paper under double-blind review

ABSTRACT

This paper tackles the problem of improving certified robustness to adversarial examples in image classification. We focus on randomized smoothing, a scalable certification framework that provides provable robustness guarantees by evaluating a classifier under Gaussian perturbations of its input. While increasing the noise level can yield stronger certificates, it also substantially degrades classification accuracy. Recent diffusion-denoised smoothing methods mitigate this trade-off by denoising perturbed inputs before classification, but the denoising process is typically agnostic to the downstream classifier. In this work, we investigate how classifier-informed guidance can be used to steer diffusion denoising toward regions that are more favorable to certification. We introduce an entropy-guided diffusion smoothing method that encourages confident predictions without prescribing a target class. We further present principled theoretical framework for analyzing our method. This analysis yields sufficient conditions under which increasing the guidance strength improves class separation and, in the binary setting, provably increases the certification radius of diffusion-denoised randomized smoothing.

1 INTRODUCTION

Modern ML models are vulnerable to adversarial examples, i.e., small (often imperceptible) perturbations of their inputs deliberately crafted to alter their outputs. Since the seminal work of (Biggio et al., 2013; Goodfellow et al., 2015), which first exposed this phenomenon in the context of image classification, a substantial body of work has emerged to study this puzzling problem (Tsipras et al., 2019; Meunier et al., 2022; Bubeck & Sellke, 2023; Awasthi et al., 2023; Heredia et al., 2025). More recently, this line of research has extended beyond standard classification settings to large language models, notably through adversarial prompting and jailbreak attacks designed to circumvent model alignment (Carlini et al., 2023a; Qi et al., 2024). Despite extensive efforts, reliably defending ML models against adversarial attacks remains notoriously difficult. One fundamental challenge lies in evaluating a defense: empirical robustness against a particular attack does not guarantee robustness against a stronger, potentially adaptive, adversary. Indeed, many defenses initially thought to be robust were subsequently circumvented under more careful evaluation (Carlini & Wagner, 2017; Carlini et al., 2019; Tramèr et al., 2020). Since finding the strongest possible attack against a given classifier is generally intractable, an important line of research has instead focused on so-called certified defenses. Rather than assessing robustness against a particular attack, these methods provide formal guarantees that the classifier's prediction remains unchanged under any perturbation whose magnitude is below a given threshold (Lécuyer et al., 2018; Wong & Kolter, 2018; Tsuzuku et al., 2018).

Randomized smoothing. Among these approaches, randomized smoothing has emerged as a particularly attractive framework for scalable certified robustness (Lécuyer et al., 2018; Li et al., 2018; Cohen et al., 2019; Yang et al., 2020). Its core idea is rooted in a classical regularization principle saying that convolving a

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

potentially irregular function with a Gaussian measure produces a smoother function whose sensitivity to input perturbations can be controlled. In the classification setting, randomized smoothing applies this principle by computing the probability distribution of a base classifier's predictions under Gaussian perturbations of its input. The resulting mapping (from inputs to distributions over predictions) defines a smoothed classifier whose variation under input perturbations can be explicitly controlled. In particular, if the probability of the most likely prediction is sufficiently separated from that of its strongest competitor, perturbations up to a certain magnitude cannot reverse their ordering. This magnitude can be characterized explicitly in terms of these two probabilities, yielding a certified ℓ_2 radius within which the most likely prediction of the smoothed classifier is guaranteed to remain unchanged.

Improving randomized smoothing with diffusion models. While randomized smoothing provides an appealing route toward scalable certified robustness, its main limitation follows directly from its smoothing mechanism. Increasing the standard deviation of the Gaussian noise generally allows for larger certified radii, but also produces increasingly corrupted inputs on which the underlying classifier becomes less accurate. An important line of research has consequently sought to alleviate this trade-off between classification accuracy and certified robustness (Salman et al., 2019; Jeong et al., 2021; 2023). Recently, diffusion models have emerged as a promising tool for addressing this limitation. Diffusion-denoised smoothing first applies the Gaussian perturbation required by randomized smoothing and subsequently uses a pretrained diffusion model to denoise the corrupted input before classification (Carlini et al., 2023b; Jeong & Shin, 2023; Shpilevskiy et al., 2025). The core idea of this concept is that the inverse process of a diffusion model can map a perturbed observation $X_{\sigma_{\mathrm{sm}}}$ (with noise level σ_{sm}) back into the unperturbed data distribution. Essentially, instead of trying to classify the noisy sample, the idea is to perform classification on a surrogate sample drawn from the denoised posterior distribution μ_z, i.e., the conditional distribution of the denoised sample given $X_{\sigma_{\mathrm{sm}}} = z$. This allows to recover some of the classification accuracy lost when injecting noise for randomized smoothing while leaving the certification framework intact. Existing approaches, however, typically rely on an unguided diffusion process whose objective is simply to bring the corrupted input back toward the data distribution, i.e., have μ_z as close as possible from it. While this can considerably improve classification accuracy, it does not fully exploit the flexibility offered by the diffusion process. For certified classification, recovering a plausible sample is not the only objective: ideally, the denoising trajectory should also favor regions of the data distribution where the downstream classifier makes confident and stable predictions, thereby increasing the likelihood of obtaining a strong certificate.

Contributions. In this paper, instead of sampling μ_z, we propose to sample a law that concentrates on unambiguous reconstructions, but still guaranties that no label supported by the observation is excluded a priori. This motivates the introduction of a tilted target distribution $\mu_z^\Psi \propto \mathrm{e}^\Psi \mu_z$, where Ψ is a bounded potential used to concentrate the original target with respect to confident regions. This naturally suggests using information from the classifier itself to define Ψ and guide the denoising process toward samples that are not only plausible, but also favorable for certification.[1] Sampling the tilted target is arguably challenging as μ_z is only accessible through the backward dynamics, and the normalizing constant of μ_z^Ψ is intractable. However, we note that tilting the endpoint of the backward process is equivalent to updating the score network and perturbing the drift of the diffusion. Hence we address this issue by using a guidance scheme inspired by (Chung et al., 2023; Wu et al., 2023; Cardoso et al., 2024). Our contributions can be summarized as follows.

- **C1** We introduce entropy-guided randomized smoothing, which couples the downstream classifier with a projection step of the perturbed point x into zones that are *both* likely according to the data distribution *and* represents region of low-entropy for the downstream classifier.
- **C2** Under reasonable assumptions on the data distribution and on the potential being used, we prove that the probability of the most likely prediction is nondecreasing in the guidance parameter. In the binary setting,

[1] Concurrent work Shpilevskiy et al. (2025) also considers guided denoising, but uses the observed input itself to steer the reconstruction in order to improve predictive accuracy. In contrast, our guidance is derived from the downstream classifier and explicitly aims to favor regions that are favorable for certification.

we establish a stronger monotonicity result showing that the diffusion-denoised randomized-smoothing certified radius obtained via guided diffusion is also nondecreasing in the guidance parameter, hence demonstrating that guidance provably improves the certificate obtain with diffusion-denoised smoothing.

C3 We evaluate numerically in several settings the proposed method. In a controlled setting we show that the proposed projection into low-entropy regions do achieve bigger separation between modes and avoid zones of transition between models. We also show that this gain of certified accuracy for a given radius translate to image datasets, even though it implies a bigger memory footprint which might be limiting for high-dimensional / complex classifier problems.

Paper organization. The remainder of the paper is organized as follows. Section 2 reviews the necessary background on randomized smoothing and diffusion-based denoising. Section 3 introduces classifier-guided diffusion smoothing and our entropy-based guidance mechanism. Section 4 develops the theoretical foundations of our approach, characterizing the effect of guidance on class probabilities and establishing conditions under which it improves certified robustness. Finally, Section 5 presents our empirical evaluation, and Section 6 concludes the paper with discussions and perspectives. All proofs are postponed to appendix A.

2 Notations and background

Let us consider a classification task with input space $\mathbb{R}^d$ (encoding a set of images) and output space $\mathcal{Y} = \{1, \dots, K\}$. We denote by $\mathrm{p}_{\mathrm{data}}$ the ground-truth data distribution on $\mathbb{R}^d$ and write Δ_K for the probability simplex over $\mathcal{Y}$. For this problem, we define a soft classifier as a mapping $\mathrm{q} : \mathbb{R}^d \to \Delta_K$. This mapping induces a hard classification rule defined as follows:

$$\mathbf{q}(x) := \arg\max_{y \in \mathcal{Y}} \mathrm{q}(y \mid x). \tag{1}$$

Because randomized smoothing and diffusion-denoised smoothing are classifier-agnostic certification strategies, we assume throughout the paper a fixed classifier $\mathrm{q} : \mathbb{R}^d \to \Delta_K$. The decision regions associated with this classifier are given, for all $y \in \mathcal{Y}$, by $\mathcal{C}_y^{\mathrm{q}} := \{x \in \mathbb{R}^d : \ \mathbf{q}(x) = y\}$. Accordingly, the robustness of $\mathbf{q}$ at x is measured by the largest Euclidean ball around x on which the prediction is constant, i.e.,

$$\mathrm{R}_{\mathrm{cert}}(\mathbf{q}; x) := \sup\{\tau \in \mathbb{R}_+ \mid \mathrm{B}_\tau(x) \subset \mathcal{C}^{\mathrm{q}}_{\mathbf{q}(x)}\}\,. \tag{2}$$

In most settings, $\mathrm{R}_{\mathrm{cert}}(\mathbf{q}; x)$ is not available explicitly, and the objective of certification is to introduce classifiers for which a nontrivial lower bound for (2) can be computed with high confidence.

2.1 Standard Randomized smoothing

Randomized smoothing transforms any base classifier into a smoothed classifier whose certified radius can be estimated with arbitrarily high confidence using Monte Carlo sampling. For a given noise level $\sigma_{\mathrm{sm}} > 0$, the randomized smoothing operator over the space of predictive distributions $\Delta_K^{\mathbb{R}^d}$ is defined as the mapping $\mathrm{Q}_{\mathcal{RS}}$ that for any soft classifier $\mathrm{q} \in \Delta_K^{\mathbb{R}^d}$ produces a smooth soft classifier $\mathrm{Q}_{\mathcal{RS}}\left[\sigma_{\mathrm{sm}}; \mathrm{q}\right] \in \Delta_K^{\mathbb{R}^d}$ defined as follows

$$\mathrm{Q}_{\mathcal{RS}}\left[\sigma_{\mathrm{sm}}; \mathrm{q}\right](y \mid x) := \mathbb{E}\left[\mathrm{q}(y \mid Z_x)\right], \quad \text{with } Z_x \sim \mathcal{N}\left(x; \sigma_{\mathrm{sm}}^2 \mathbf{I}_d\right). \tag{3}$$

We can then directly define the associated randomized-smoothed hard classifier as $\mathbf{Q}_{\mathcal{RS}}\left[\sigma_{\mathrm{sm}}; q\right]$ by analogy with (1). For a given input x, let c_A and c_B denote respectively the most likely and second most likely classes under the smoothed prediction distribution:

$$c_A := \arg\max_{y \in \mathcal{Y}} \mathrm{Q}_{\mathcal{RS}}\left[\sigma_{\mathrm{sm}}; \mathrm{q}\right](y \mid x), \qquad c_B := \arg\max_{y \in \mathcal{Y} \setminus \{c_A\}} \mathrm{Q}_{\mathcal{RS}}\left[\sigma_{\mathrm{sm}}; \mathrm{q}\right](y \mid x).$$

Using these notations, previous work (Salman et al., 2019; Zhai et al., 2020; Súkeník et al., 2022) have shown that the gap between the probabilities assigned to these two classes determines a certified ℓ_2 neighborhood around x within which the prediction of the smoothed classifier cannot change. Specifically, they show that

$$\mathrm{R}_{\mathrm{cert}}(\mathbf{Q}_{\mathcal{RS}}\left[\sigma_{\mathrm{sm}}; q\right]; x) \geq \frac{\sigma_{\mathrm{sm}}}{2}\left[\Phi^{-1}\left(\mathrm{Q}_{\mathcal{RS}}\left[\sigma_{\mathrm{sm}}; \mathrm{q}\right](c_A \mid x)\right) - \Phi^{-1}\left(\mathrm{Q}_{\mathcal{RS}}\left[\sigma_{\mathrm{sm}}; \mathrm{q}\right](c_B \mid x)\right)\right],$$

where Φ^{-1} is the inverse cumulative distribution function of the standard Gaussian distribution. Importantly, this result does not require any particular property on the base model q. Designing a certifiably robust classifier therefore reduces to designing a base classifier whose smoothed predictions exhibits a sufficiently large margin for a given noise level σ_{sm}. In general, however, the smoothed soft classifier $\mathrm{Q}_{\mathcal{RS}}\left[\sigma_{\mathrm{sm}}; \mathrm{q}\right]$ cannot be evaluated exactly, since its computation involve an expectation over Gaussian perturbations. Instead, a Monte Carlo procedure is used that either returns an estimate of $\mathbf{Q}_{\mathcal{RS}}\left[\sigma_{\mathrm{sm}}; \mathrm{q}\right](x)$ or abstains when there is insufficient statistical evidence to identify the most likely prediction. The classical sampling and prediction procedures under Gaussian noise are given in Algorithms 1 and 2 (Appendix B). In particular, Algorithm 2 identifies the two most frequently observed classes, with respective counts n_A and n_B, and returns the empirical majority class only when their frequencies are statistically distinguishable. This is assessed using $\textsc{BinomPValue}(n_A, n_A + n_B, 1/2)$, the two-sided binomial-test p-value for the null hypothesis that the two classes have equal probability; otherwise, the procedure abstains.

2.2 Score-based generative models

Score-based generative modeling uses stochastic differential equations (SDEs) (Song et al., 2021; Lim et al., 2023) to transform data into noise and then generate new samples by reversing the noising process. Due to the proximity with the randomized smoothing setting, we focus the presentation on the variance exploding framework and refer the readers to (Song et al., 2021) for the general presentation. We consider a time horizon $T \in \mathbb{R}_+$ and a smooth increasing noising function $\sigma : [0, T] \to [0, T]$. The forward diffusion process is defined for all $t \geq 0$ by

$$\overrightarrow{\mathrm{X}}_t = X_0 + \int_0^t g(t)\,\mathrm{dB}_t, \tag{4}$$

where $X_0 \sim \mathrm{p}_{\mathrm{data}}$ and $(\mathrm{B}_t)_{t\geq 0}$ is a standard Brownian motion on $\mathbb{R}^d$ and $g(t) = (2\sigma_t\dot{\sigma}_t)^{1/2}$. By elementary properties of Brownian motion, $\int_0^t g(t)\mathrm{dB}_t \sim \mathcal{N}\left(0; \sigma_t^2\,\mathrm{I}\right)$ and thus if we denote by p_t the density of $\overrightarrow{\mathrm{X}}_t$ we have that

$$\mathrm{p}_t(x) = \int \psi\left[x_0; \sigma_t^2\mathbf{I}\right](x)\,\mathrm{p}_{\mathrm{data}}(\mathrm{d}x_0) := \mathrm{p}_{\mathrm{data}} * \mathcal{N}\left(0; \sigma_t^2\,\mathrm{I}\right), \tag{5}$$

where $\sigma_t^2 = \int_0^t g(t)^2\mathrm{d}s$ and $\psi\left[m; \Sigma\right](\cdot)$ is the p.d.f. of a Gaussian distribution with mean m and covariance matrix Σ. The time-reversal of (4) is formally given by the solution of the equation

$$\overleftarrow{\mathrm{X}}_t = \overleftarrow{\mathrm{X}}_0 + \int_0^t \bar{g}(s)^2\nabla \log \mathrm{p}_{T-s}(\overleftarrow{\mathrm{X}}_s)\mathrm{d}s + \int_0^t \bar{g}(s)\,\mathrm{d}\tilde{\mathrm{B}}_s, \tag{6}$$

where $\bar{g}(t) = g(T - t)$. For each $(s, t) \in [0, T]^2$ with $s < t$ we define $\overleftarrow{\mathrm{P}}_{t|s}$ to be a regular conditional distribution of $\overleftarrow{\mathrm{X}}_t$ given $\overleftarrow{\mathrm{X}}_s$. The score $\nabla \log \mathrm{p}_t$ is unknown but can be approximated through the learning of a multi-level Gaussian denoiser for $\mathrm{p}_{\mathrm{data}}$. Indeed, by differentiating (5), we obtain Tweedie's formula

$$\mathbb{E}\left[\overrightarrow{\mathrm{X}}_0 \middle| \overrightarrow{\mathrm{X}}_t\right] = \overrightarrow{\mathrm{X}}_t + \sigma_t^2\nabla \log \mathrm{p}_t(\overrightarrow{\mathrm{X}}_t)\,. \tag{7}$$

As the conditional expectation is the minimizer of the MSE for L_2 variables, (7) shows that estimating the score $\nabla \log \mathrm{p}_t$ is equivalent to estimating a denoising network $\mathrm{D}_\theta : \mathbb{R}^d \times [0, \sigma_T] \to \mathbb{R}^d$ that minimizes

$\mathbb{E}\left[\|\mathrm{D}_\theta\left(\mathrm{X}_0+\sigma_t Z,\sigma_t\right)-\mathrm{X}_0\|_2^2\right]$ where $Z\sim\mathcal{N}\left(0;\mathbf{I}_d\right)$. As the score must be learned for all times $t\in[0,T]$, one usually minimizes

$$\mathcal{L}_{\mathrm{denoiser}}(\theta)=\int_0^{\sigma_t}\eta(\varsigma)\mathbb{E}\left[\|\mathrm{D}_\theta\left(\mathrm{X}_0+\varsigma Z,\varsigma\right)-\mathrm{X}_0\|_2^2\right]\mathrm{d}\varsigma\,, \tag{8}$$

and $\eta:[0,\sigma_t]\to\mathbb{R}_+$ is a weighting function. We refer the readers to Karras et al. (2022) for further details on the choice of η and parameterization of the denoiser network. The denoiser D_θ can then be converted to a score approximation by (7) which leads to $\mathrm{s}_\theta(x)=\left(\mathrm{D}_\theta\left(x,\sigma_t\right)-x\right)/\sigma_t^2$.

Score-based models have been extensively studied from a theoretical perspective. The error between $\mathrm{p}_{\mathrm{data}}$ and the approximated distribution is commonly decomposed into (i) mixing error, (ii) discretization error, and (iii) approximation error. Existing works establish sampling guarantees for instance for classical noising process (Chen et al., 2023b;a; Benton et al., 2024; Conforti et al., 2025; Strasman et al., 2025a; 2026) and for Critically Damped Langevin Diffusions (Chen et al., 2023b; Conforti et al., 2025; Strasman et al., 2025b).

The backward equation and randomized smoothing. The backward equation can be seen as a projection of a noisy sample z back into the data distribution. Indeed, given a single noisy observation z at level σ_{sm}, one might integrate (6) from $T-t(\sigma_{\mathrm{sm}})$ to T, where $t(\sigma_{\mathrm{sm}})$ is such that $\sigma_{t(\sigma_{\mathrm{sm}})}=\sigma_{\mathrm{sm}}$. This produces a sample from the denoised posterior

$$\mu_z:=\mathrm{Law}\left(X_0\mid\overrightarrow{X}_{\sigma_{\mathrm{sm}}}=z\right)=\mathrm{Law}\left(\overleftarrow{\mathrm{X}}_T|\overleftarrow{\mathrm{X}}_{T-t(\sigma_{\mathrm{sm}})}=z\right)=\overleftarrow{\mathrm{P}}_{T|T-t(\sigma_{\mathrm{sm}})}(\cdot|z). \tag{9}$$

The distribution μ_z is the law that the (unguided) denoised samples.

3 Guided diffusion-denoised smoothing

Diffusion-denoised smoothing uses a diffusion model to map the noisy observation produced by randomized smoothing back toward the data distribution before classification. More formally, as introduced by Carlini et al. (2023b), it constructs the mapping $z\mapsto\mathbb{E}\left[\mathbf{q}(U_z)\right]$, with $U_z\sim\mu_z$. Its effectiveness therefore depends on how much mass μ_z assigns to regions where the downstream classifier $\mathbf{q}$ is both accurate and confident. However, since μ_z is $\mathbf{q}$-agnostic, there is no reason for it to concentrate its mass on such regions. Our approach builds on this observation. Rather than sampling directly from μ_z, we seek to bias the denoising process toward regions where $\mathbf{q}$ makes confident predictions (while retaining the information contained in the noisy observation). To this end, we introduce a guided diffusion method where the unconditional denoising is adapted at sampling time using external guidance. The goal of guidance is to provide a flexible framework for steering diffusion processes toward a prescribed target distribution.

Guided diffusion-denoising. Let $\Psi:\mathbb{R}^d\to\mathbb{R}$ be a smooth function, referred to as the *tilting potential*, we define the tilted data distribution by

$$\mathrm{p}^\Psi(x):=\exp(\Psi\left(x\right))\,\mathrm{p}_{\mathrm{data}}(x)/\int\exp(\Psi\left(\tilde{x}\right))\,\mathrm{p}_{\mathrm{data}}(\tilde{x})\mathrm{d}\tilde{x}\,.$$

The guided reverse diffusion is simply given by (6) with $\nabla\log\mathrm{p}_t$ replaced by $\nabla\log\mathrm{p}_t^\Psi$ where, for every $t>0$ and every $z\in\mathbb{R}^d$, we have

$$\nabla\log\mathrm{p}_t^\Psi(z)=\nabla\log\mathrm{p}_t(z)+\nabla_z\log G_t\left(z\right)\text{ , and }G_t(z):=\mathbb{E}\left[\mathrm{e}^{\Psi(X_0)}\middle|\overrightarrow{\mathrm{X}}_t=z\right]. \tag{10}$$

The guided reverse diffusion started at $\overleftarrow{\mathrm{X}}_{\sigma_{\mathrm{sm}}}=z$ has terminal law defined as $\mu_z^\Psi(\mathrm{d}u)\propto\mathrm{e}^{\Psi(u)}\mu_z(\mathrm{d}u)$. The target of the guided sampler is not the tilted data distribution $\mathrm{p}_{\mathrm{data}}^\Psi$ but μ_z^Ψ. In fact, the algorithm guides the reconstructions of z to preserve prior information carried by the noisy sample. The ideal guidance term is the

Doob h-transform $\nabla_z \log G_t(z)$, which is often untractable and approximated in practice. We define η_t^Ψ, and $\overleftarrow{\mathrm{P}}^\Psi_{t|s}$ analogously to η_t, and $\overleftarrow{\mathrm{P}}_{t|s}$ using the tilted score $\nabla \log \mathrm{p}_t^\Psi$ instead of $\nabla \log \mathrm{p}_t$. Then, the backward guided diffusion is solution to

$$\overleftarrow{\mathrm{X}}^\Psi_{T|T-t}(z) = z + \int_{T-t}^{T} \bar{g}(s)^2 \nabla \log \mathrm{p}^\Psi_{T-s}(\overleftarrow{\mathrm{X}}^\Psi_{s|T-t}(z))\mathrm{d}s + \int_{T-t}^{T} \bar{g}(s)\,\mathrm{d}\tilde{\mathrm{B}}_s\,.$$

For a given smoothing level σ_{sm} and potential Ψ, analogously to diffusion-denoised randomized smoothing, our scheme constructs the mapping $z \mapsto \mathbb{E}\left[\mathbf{q}(U_z^\Psi)\right]$, with $U_z^\Psi \sim \mu_z^\Psi$. More formally, we define the soft classifier $\mathrm{q}^\Psi_{\sigma_{\mathrm{sm}}}$ such that for any $y \in \mathcal{Y}$ and $z \in \mathbb{R}^d$ we have

$$\mathrm{q}^\Psi_{\sigma_{\mathrm{sm}}}(y|z) = \mathbb{E}\left[\mathbb{1}_{\mathcal{C}_y^{\mathrm{q}}}\left(\overleftarrow{\mathrm{X}}^\Psi_{T|T-t(\sigma_{\mathrm{sm}})}(z)\right)\right]\,. \tag{11}$$

This soft classifier then serves as the base classifier to which we apply randomized smoothing, as defined in (3). When $\Psi \equiv 0$, the mapping $\mathrm{q}^\Psi_{\sigma_{\mathrm{sm}}}$ reduces to the standard diffusion-denoised classifier introduced by Carlini et al. (2023b). This naturally raises the question of whether a non-trivial tilting potential can yield a base classifier more favorable to randomized smoothing than $\mathrm{q}^0_{\sigma_{\mathrm{sm}}}$.

Choice of the tilting potential. It would be natural to aim for a high-confidence prediction for the base (non-perturbed) input x. Unfortunately, we are not allowed to access this input without compromising the guarantees provided by random smoothing. Consequently, we must select a tilting potential that directs the denoising toward a classification region with high confidence without accessing the input x. A logical objective in this context is therefore to target a potential Ψ designed to generate classifier predictions with low entropy (regardless of the prediction being made). A natural choice for such a potential is $\Psi = \lambda\Psi_{\mathrm{ent}}$ with

$$\Psi_{\mathrm{ent}}(z) := -\mathrm{H}(\mathrm{q}(\cdot \mid z)) = \sum_{k=1}^{K} \mathrm{q}(k \mid z) \log \mathrm{q}(k \mid z), \text{ and } \lambda \geq 0. \tag{12}$$

Since Ψ_{ent} is larger when the entropy is smaller, adding its gradient to the reverse dynamics encourages samples to move toward regions where the classifier is confident.

4 Improved certificate for guided diffusion-denoised smoothing

In this section, we study the effect of the guidance parameter λ on the certificate obtained with diffusion-denoised randomized smoothing. As recalled in Section 2, the certified radius depends on the separation between the largest class probabilities of the smoothed classifier. We therefore analyze how these probabilities evolve with λ. We first characterize the effect of guidance on the class probabilities after denoising, and then study their behavior after averaging over the Gaussian perturbation. This yields sufficient conditions under which guidance increases the probability margin between the most likely class and any other of the classes. Finally, in the binary setting, we show that this implies that the certified radius is nondecreasing with λ hence demonstrating that our scheme improves the certified radius provided by diffusion-denoised randomized smoothing. Throughout we consider a reference fixed prediction model q and $\Psi = \lambda\Psi_{\mathrm{ent}}$ with $\lambda \geq 0$. For any $x \in \mathbb{R}^d$ and $y \in \mathcal{Y}$, we defined the probabilities associated to the guided diffusion-denoised randomized smoothing as

$$p_y^\lambda(x) := \mathrm{Q}_{\mathcal{RS}}\left[\sigma_{\mathrm{sm}}; \mathrm{q}^{\lambda\Psi_{\mathrm{ent}}}_{\sigma_{\mathrm{sm}}}\right](y \mid x) = \mathbb{E}\left[\mathrm{q}^{\lambda\Psi_{\mathrm{ent}}}_{\sigma_{\mathrm{sm}}}(y|Z_x)\right], \text{ where } Z_x \sim \mathcal{N}\left(x; \sigma_{\mathrm{sm}}^2 \mathbf{I}_d\right). \tag{13}$$

We start our analysis at the level of a single noisy observation z. Intuitively , increasing λ tilts the denoised posterior toward samples with larger values for Ψ_{ent} , and hence toward regions where the downstream classifier has lower entropy. This does not, however, imply that the probability of every class increases. In fact, guidance redistributes probability mass among the classifier's decision regions. The following lemma

characterizes this redistribution more precisely. Specifically, it establishes that for a fixed noisy observation z, the probability assigned by $\mathrm{q}^{\Psi}_{\sigma_{\mathrm{sm}}}$ to a class y increases with λ when an expected weighted average value of the potential within class y is greater than its average under the denoising distribution.

Lemma 4.1. *Let $z \in \mathbb{R}^d$ and $y \in \mathcal{Y}$ such that $\mathrm{q}^{\Psi}_{\sigma_{\mathrm{sm}}}(y|z) > 0$ we define the expected weighted average value of the potential within class y and its average under the denoising distribution respectively as $m^{\Psi}_y(z) := \mathbb{E}\left[\Psi\left(\overleftarrow{\mathrm{X}}^{\Psi}_{T|T-t(\sigma_{\mathrm{sm}})}(z)\right) \middle| \overleftarrow{\mathrm{X}}^{\Psi}_{T|T-t(\sigma_{\mathrm{sm}})}(z) \in \mathcal{C}^{\mathrm{q}}_y\right]$ and $\bar{m}^{\Psi}(z) = \mathbb{E}\left[\Psi\left(\overleftarrow{\mathrm{X}}^{\Psi}_{T|T-t(\sigma_{\mathrm{sm}})}(z)\right)\right]$. Assume that z is such that for all $\lambda \geq 0$, we have $Z_\lambda(z) := \overleftarrow{\mathrm{P}}_{T|T-t(\sigma_{\mathrm{sm}})} \exp(\lambda \Psi_{\mathrm{ent}})(\mathrm{s}_{t(\sigma_{\mathrm{sm}})} z) < \infty$. Then,*

$$\partial_\lambda \, \mathrm{q}^{\lambda\Psi_{\mathrm{ent}}}_{\sigma_{\mathrm{sm}}}(y|z) = \mathrm{q}^{\lambda\Psi_{\mathrm{ent}}}_{\sigma_{\mathrm{sm}}}(y|z)\left(m^{\lambda\Psi_{\mathrm{ent}}}_y(z) - \bar{m}^{\lambda\Psi_{\mathrm{ent}}}(z)\right).$$

Moreover, for all classes $y, y' \in \mathcal{Y}$,

$$\partial_\lambda \log \frac{\mathrm{q}^{\lambda\Psi_{\mathrm{ent}}}_{\sigma_{\mathrm{sm}}}(y|z)}{\mathrm{q}^{\lambda\Psi_{\mathrm{ent}}}_{\sigma_{\mathrm{sm}}}(y'|z)} = m^{\lambda\Psi_{\mathrm{ent}}}_y(z) - m^{\lambda\Psi_{\mathrm{ent}}}_{y'}(z).$$

Lemma 4.1 characterizes how guidance modifies the output of $\mathrm{q}^{\lambda\Psi_{\mathrm{ent}}}_{\sigma_{\mathrm{sm}}}$ for a fixed noisy observation z. Certification, however, depends on the probabilities $p^\lambda_y(x)$ obtained after averaging over the random variable Z_x. Averaging the identity of Lemma 4.1 yields $\mathbb{E}\left[\mathrm{q}^{\lambda\Psi_{\mathrm{ent}}}_{\sigma_{\mathrm{sm}}}(y|Z_x)\left(m^{\lambda\Psi_{\mathrm{ent}}}_y(Z_x) - \bar{m}^{\lambda\Psi_{\mathrm{ent}}}(Z_x)\right)\right]$. Thus, to study whether guidance increases the separation between the reference class c_A, defined as the most likely class at $\lambda = 0$, and its competitors, it is sufficient to compare their rates of change with respect to λ. This motivates the following assumption.

Assumption 4.2. Let $x \in \mathbb{R}^d$, $Z_x \sim \mathcal{N}\left(x; \sigma^2_{\mathrm{sm}}\mathbf{I}_d\right)$ and use the notations of Lemma 4.1. For every $\lambda \geq 0$ and every $y \in \mathcal{Y}$, define

$$\Gamma^\lambda_y(x) := \mathbb{E}\left[\mathrm{q}^{\lambda\Psi_{\mathrm{ent}}}_{\sigma_{\mathrm{sm}}}(y|Z_x)\left(m^{\lambda\Psi_{\mathrm{ent}}}_y(Z_x) - \bar{m}^{\lambda\Psi_{\mathrm{ent}}}(Z_x)\right)\right] \quad \text{and } c_A := \arg\max_{y\in\mathcal{Y}} p^0_y(x).$$

Then, for every $\lambda \geq 0$ and every $y \neq c_A$, we have $\Gamma^\lambda_{c_A}(x) \geq \Gamma^\lambda_y(x)$.

Assumption 4.2 requires the probability of the reference class c_A to vary with λ at least as favorably as that of any competing class. Equivalently, it requires the derivative of each pairwise margin $p^\lambda_{c_A}(x) - p^\lambda_y(x)$ to be nonnegative. The following proposition shows that this condition is sufficient to preserve the reference class while ensuring that its probability margin over every competitor is nondecreasing.

Proposition 4.3. *Let $x \in \mathbb{R}^d$ and assume that $c_A = \arg\max_{y\in\mathcal{Y}} p^0_y(x)$ is unique and that Assumption 4.2 holds. Then, for all $\lambda \geq 0$ and all $y \neq c_A$, we have*

$$p^\lambda_{c_A}(x) \geq p^\lambda_y(x) \quad \text{and } p^\lambda_{c_A}(x) - p^\lambda_y(x) \geq p^0_{c_A}(x) - p^0_y(x) > 0.$$

Proposition 4.3 shows that, under Assumption 4.2, increasing the guidance strength improves the probability margin between the reference class c_A and every competitor. We now specialize to binary classification, where the two class probabilities sum to one. In this setting, the previous result can be strengthened to show that increasing the guidance strength directly improves the certified radius.

Proposition 4.4. *Assume $|\mathcal{Y}| = 2$ and that Assumption 4.2 holds. Let $x \in \mathbb{R}^d$ and assume that $c_A = \arg\max_{y\in\mathcal{Y}} p^0_y(x)$ is unique. Denote by c_B the remaining class in $\mathcal{Y}$ after removing c_A. Then, the mapping*

$$\lambda \mapsto \frac{\sigma_{\mathrm{sm}}}{2}\left[\Phi^{-1}\left(\mathrm{Q}_{\mathcal{RS}}\left[\sigma_{\mathrm{sm}}; \mathrm{q}^{\lambda\Psi_{\mathrm{ent}}}_{\sigma_{\mathrm{sm}}}\right](c_A \mid x)\right) - \Phi^{-1}\left(\mathrm{Q}_{\mathcal{RS}}\left[\sigma_{\mathrm{sm}}; \mathrm{q}^{\lambda\Psi_{\mathrm{ent}}}_{\sigma_{\mathrm{sm}}}\right](c_B \mid x)\right)\right]$$

is nondecreasing. In other words, increasing λ improves the certified radius for $\mathrm{q}^{\lambda\Psi_{\mathrm{ent}}}_{\sigma_{\mathrm{sm}}}$

Overall, our analysis provides sufficient conditions under which classifier-informed guidance leads to more favorable class probabilities for randomized smoothing, with a direct guarantee on the certified radius in the binary setting. We next investigate empirically how these effects manifest in practice.

5 NUMERICAL EXPERIMENTS

As previously mentioned, the quantity $\nabla_z \log G_t(z)$ from (10) is often untractable if $\mathrm{p}_{\mathrm{data}}$ is not known analytically. This guidance can, however, be implemented in several ways, from straightforward approximation such as in Chung et al. (2023); Song et al. (2023), or using asymptotically correct samplers based on Sequential Monte Carlo (SMC) methods (Wu et al., 2023; Cardoso et al., 2024; Nazemi et al., 2024), or Sequential Langevin Monte Carlo Zhang et al. (2025). Here we focus on the solution provided in (Chung et al., 2023), where given an approximation $\mathrm{s}_\theta(z,t)$ of $\nabla \log \mathrm{p}_t$, the authors propose using

$$\hat{\mathrm{s}}^{\Psi}_\theta(z,t) := \mathrm{s}_\theta(z,t) + \zeta \nabla_z \Psi\left(z + \sigma_t^2\, \mathrm{s}_\theta(z,t)\right)\,, \tag{14}$$

with $\zeta > 0$ an hyperparameter, as an approximation of (10). The overall score is thus tractable using an automatic differentiation tensor library. We will refer to this approximation as DPS. Furthermore, exact integration of (6) are also out of reach, and thus one often employ a given discretization in order to have a tractable simulation algorithm. We use the DDPM(Ho et al. (2020)) discretization. Given $z \in \mathbb{R}^d$, $k \in \mathbb{N}$, $t_0 \leq \cdots \leq t_k$ (usually with $\sigma_{t_k} = \sigma_{\mathrm{sm}}$), we define $\overleftarrow{\mathrm{X}}^{\Psi,\mathrm{Alg}}_0(z) = z$ and for $\ell = 0, \cdots, k-1$,

$$\overleftarrow{\mathrm{X}}^{\Psi,\mathrm{Alg}}_{\ell+1}(z) = \overleftarrow{\mathrm{X}}^{\Psi,\mathrm{Alg}}_{\ell}(z) + \sigma^2_{t_{k-(\ell+1)}}\, \hat{\mathrm{s}}^{\Psi}_\theta(\overleftarrow{\mathrm{X}}^{\Psi,\mathrm{Alg}}_{\ell}(z), t_{k-\ell}) + \frac{\sigma_{t_{k-(\ell+1)}}}{\sigma_{t_{k-\ell}}}(\sigma^2_{t_{k-\ell}} - \sigma^2_{t_{k-(\ell+1)}})^{1/2} Z_{\ell+1}\,, \tag{15}$$

where $(Z_\ell)_{\ell \in \{1,\cdots,k\}}$ is a sequence of mutually independent standard Gaussian variables. As the $\overleftarrow{\mathrm{X}}^{\Psi,\mathrm{Alg}}_k(z)$ is an approximate version of $\overleftarrow{\mathrm{X}}^{\Psi}_{T|T-t_k}(z)$, we will also write it as $\overleftarrow{\mathrm{X}}^{\Psi,\mathrm{Alg}}_{T|T-t_k}(z)$, but one should always remember that there is implicitly a discretization of the interval $[T - t_k, T]$.

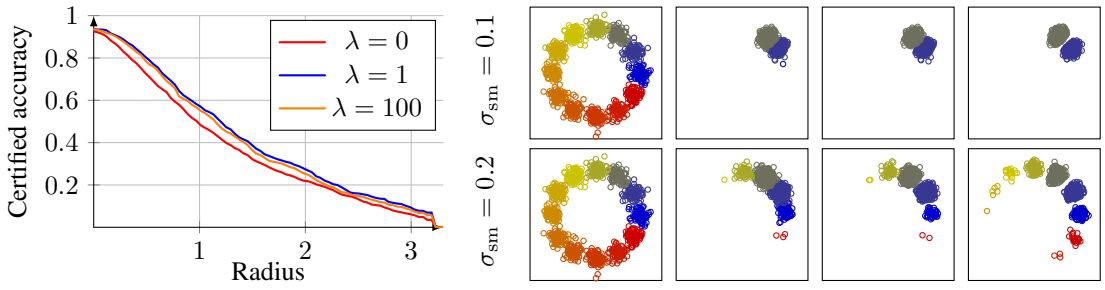


Figure 1: On the left, certified accuracy over several σ_{sm} on the CIFAR-10 dataset using $\Psi = \lambda \Psi_{\mathrm{ent}}$. On the right, illustration of the Gaussian mixture toy problem (first subplot) and of the proposed method with $\Psi = \lambda \Psi_{\mathrm{ent}}$ and values $\lambda = 0, 1, 10$ respectively.

Toy example. We first consider a toy-example where $\mathrm{p}_{\mathrm{data}}$ is a mixture of twelve two-dimensional Gaussian distributions with variance 10^{-2} and whose means $\{\mu_i\}_{i=1}^{12}$ are arranged in an equidistant fashion in the unitary circle. We then consider the *Bayes* classifier for this mixture, namely, the classifier defined by $\mathrm{q}(y \mid x) := \psi\left[\mu_y; 10^{-2}\,\mathrm{I}\right](x) \,/\, \sum_{j=1}^{12} \psi\left[\mu_j; 10^{-2}\,\mathrm{I}\right](x)$. In this case, the score is analytically available and no approximation is needed. We apply the proposed method with $k = 20$ and the discretization chosen using the EDM scheduler Karras et al. (2022) with $\rho = 1$ between 0.002 and σ_{sm}. We choose $x = (\sqrt{2}, \sqrt{2})/2$ and ran the method with $\sigma_{\mathrm{sm}} \in \{0.1, 0.2\}$ and with the potentials $\Psi(x) = \lambda \Psi_{\mathrm{ent}}(x)$ with $\lambda \in \{0, 1, 10\}$. We observe that as σ_{sm} increases, so does the zones of high probability of Law $\left(\overleftarrow{\mathrm{X}}^{\Psi,\mathrm{Alg}}_{T|T-t(\sigma_{\mathrm{sm}})}(Z_x)\right)$. But more importantly, the modes are further separated with increasing λ, meaning that the backward does guide the samples toward high confidence zones of the classifier. We note however that for the biggest value of λ, we

observe modes that were not observed for the case $\lambda = 0$. Thus this is expected, as in the limit $\lambda \to \infty$ all the center of the modes would be in the distribution, this can have unexpected effects on the final performance.

CIFAR-10. In order to test the proposed method using an off-the-shelf classifier, we consider the standard CIFAR-10 dataset and associated classification task Krizhevsky (2009). We considered the same score and classifier networks as in Carlini et al. (2023b) and used a total of 20 interpolation steps using the EDM scheduler with $\rho = 1$ between σ_{sm} and $\sigma_{\min}(0.002)$. We ran the experiments for $\sigma_{\text{sm}} \in \{0.25, 0.5, 0.75, 1.00\}$ and for $\Psi = \lambda\Psi_{\text{ent}}$ with $\lambda \in \{0, 1, 100\}$. The results are presented at Figure 1. Figure 1 shows that increasing λ improves significantly the overall certified accuracy. It also shows that choosing high λ values can have a negative effect. We show in Figure 3 the evolution of the same curves but for each value of σ_{sm}.

Imagenet. We tested the proposed approach in the higher dimensional Imagenet dataset and classification tasks Deng et al. (2009). We apply the proposed method with $\sigma_{\text{sm}} = 3$, different levels of λ for 10 randomly draw test set images. One of the samples is shown in Figure 2 and the others in Figures 4 to 13.

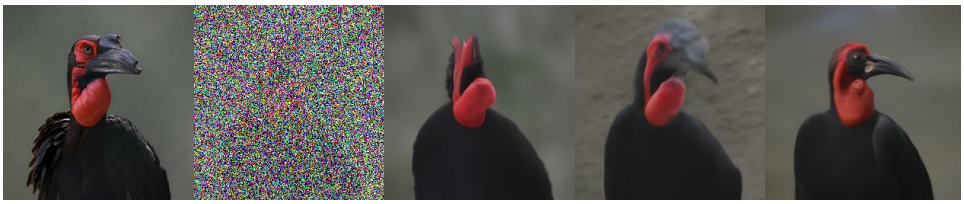


Figure 2: Visualization of the projection procedure applied to a given noisy image from ImageNet.

While the proposed method seems to behave accordingly, we use this dataset to highlight one of the main limitations of the method. The *total sampling time* of the Markov chain defined by (15) is not impacted by choosing $\lambda = 0$ or $\lambda > 0$. However, setting $\lambda > 0$ does impact the *memory footprint* of the procedure, as it now propagates gradient through the classifier *combined* with the score network. This amounts to a considerable extension of the memory requirements, particularly when approximating the expectation in Randomized Smoothing, which usually uses a number of Monte Carlo samples of 10^4. It makes then unreasonable to replicate the experiments from Carlini et al. (2023b) on Imagenet and is thus a limitation of the proposed method.

6 Conclusion & Discussion

In this work, we introduced classifier-guided diffusion denoising as a way to improve diffusion-denoised randomized smoothing. By using the entropy of the downstream classifier as a tilting potential, our approach steers denoised samples toward confident regions that allow to provide improve certificate. We provide both theoretical and empirical evidence for the usefulness of our method. Nevertheless, we note that proposed method requires a considerable augmentation of the available GPU memory to compute the same expectation estimation as the method proposed in Carlini et al. (2023b). While this can be reasonable for some models, it might be a bottleneck when using large architectures. A possible solution for such problem would be to use the entropy of *another* smaller classifier for the guidance part. One would probably need to make sure that both classifiers are aligned or use a distillation-based approach. Another possible solution is to consider learning $\nabla_z \log G_\sigma$ following Bergmeister et al. (2026) instead of using the DPS approximation, which would then bypass gradient calculations and rely only on evaluation of Neural networks. While this would need a training phase, it could lead to considerable lower GPU memory usage for the inference of the randomized smoothing classifiers.

REFERENCES

Pranjal Awasthi, Natalie S. Frank, and Mehryar Mohri. On the existence of the adversarial bayes classifier (extended version), 2023.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEiT: BERT pre-training of image transformers. 2021. URL https://openreview.net/forum?id=p-BhZSz59o4.

Joe Benton, Valentin De Bortoli, Arnaud Doucet, and George Deligiannidis. Nearly d-linear convergence bounds for diffusion models via stochastic localization. In The Twelfth International Conference on Learning Representations, 2024.

Andreas Bergmeister, Stefanie Jegelka, Nikolas Nüsken, Carles Domingo-Enrich, and Jakiw Pidstrigach. Reinforce Adjoint Matching: Scaling RL Post-Training of Diffusion and Flow-Matching Models, May 2026.

Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In Proceedings of the 2013th European Conference on Machine Learning and Knowledge Discovery in Databases - Volume Part III, ECMLPKDD'13, pp. 387–402, Berlin, Heidelberg, 2013. Springer-Verlag. ISBN 9783642409936. doi: 10.1007/978-3-642-40994-3_25. URL https://doi.org/10.1007/978-3-642-40994-3_25.

Sébastien Bubeck and Mark Sellke. A universal law of robustness via isoperimetry. Journal of the ACM, 70 (2):1–18, 2023.

Gabriel Cardoso, Yazid Janati El Idrissi, Sylvain Le Corff, and Eric Moulines. Monte carlo guided denoising diffusion models for bayesian linear inverse problems. In International Conference on Learning Representations, 2024.

Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, pp. 3–14, 2017.

Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian J. Goodfellow, Aleksander Madry, and Alexey Kurakin. On evaluating adversarial robustness. CoRR, abs/1902.06705, 2019. URL http://arxiv.org/abs/1902.06705.

Nicholas Carlini, Milad Nasr, Christopher A. Choquette-Choo, Matthew Jagielski, Irena Gao, Pang Wei Koh, Daphne Ippolito, Florian Tramèr, and Ludwig Schmidt. Are aligned neural networks adversarially aligned? In Thirty-seventh Conference on Neural Information Processing Systems, 2023a. URL https://openreview.net/forum?id=OQQoD8Vc3B.

Nicholas Carlini, Florian Tramer, Krishnamurthy Dj Dvijotham, Leslie Rice, Mingjie Sun, and J Zico Kolter. (certified!!) adversarial robustness for free! In The Eleventh International Conference on Learning Representations, 2023b. URL https://openreview.net/forum?id=JLg5aHHv7j.

Hongrui Chen, Holden Lee, and Jianfeng Lu. Improved analysis of score-based generative modeling: User-friendly bounds under minimal smoothness assumptions. In International Conference on Machine Learning, pp. 4735–4763. PMLR, 2023a.

Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru R. Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions, 2023b.

Hyungjin Chung, Jeongsol Kim, Michael Thompson Mccann, Marc Louis Klasky, and Jong Chul Ye. Diffusion Posterior Sampling for General Noisy Inverse Problems. In The Eleventh International Conference on Learning Representations, 2023.

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In international conference on machine learning, pp. 1310–1320. PMLR, 2019.

Giovanni Conforti, Alain Durmus, and Marta Gentiloni Silveri. Kl convergence guarantees for score diffusion models under minimal data assumptions. SIAM Journal on Mathematics of Data Science, 7(1):86–109, 2025.

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.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. 34:8780–8794, 2021.

Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In International Conference on Learning Representations, 2015.

Lucas Gnecco Heredia, Matteo Sammut, Muni Sreenivas Pydi, Rafael Pinot, Benjamin Négrevergne, and Yann Chevaleyre. Unveiling the role of randomization in multiclass adversarial classification: Insights from graph theory. In Yingzhen Li, Stephan Mandt, Shipra Agrawal, and Mohammad Emtiyaz Khan (eds.), International Conference on Artificial Intelligence and Statistics, AISTATS 2025, Mai Khao, Thailand, 3-5 May 2025, volume 258 of Proceedings of Machine Learning Research, pp. 3682–3690. PMLR, 2025. URL https://proceedings.mlr.press/v258/heredia25a.html.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In Advances in Neural Information Processing Systems, 2020.

Jongheon Jeong and Jinwoo Shin. Multi-scale diffusion denoised smoothing. In Thirty-seventh Conference on Neural Information Processing Systems, 2023. URL https://openreview.net/forum?id=zQ4yraDiRe.

Jongheon Jeong, Sejun Park, Minkyu Kim, Heung-Chang Lee, Doguk Kim, and Jinwoo Shin. Smoothmix: Training confidence-calibrated smoothed classifiers for certified robustness. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), Advances in Neural Information Processing Systems, 2021. URL https://openreview.net/forum?id=nlEQMVBD359.

Jongheon Jeong, Seojin Kim, and Jinwoo Shin. Confidence-aware training of smoothed classifiers for certified robustness. In Brian Williams, Yiling Chen, and Jennifer Neville (eds.), AAAI-23 Technical Tracks 7, Proceedings of the 37th AAAI Conference on Artificial Intelligence, AAAI 2023, pp. 8005–8013. AAAI Press, June 2023. doi: 10.1609/aaai.v37i7.25968. Publisher Copyright: Copyright © 2023, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.; 37th AAAI Conference on Artificial Intelligence, AAAI 2023 ; Conference date: 07-02-2023 Through 14-02-2023.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the Design Space of Diffusion-Based Generative Models. In Proc. NeurIPS, 2022.

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

Mathias Lécuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel J. Hsu, and Suman Sekhar Jana. Certified robustness to adversarial examples with differential privacy. 2019 IEEE Symposium on Security and Privacy (SP), pp. 656–672, 2018. URL https://api.semanticscholar.org/CorpusID:49431481.

Bai Li, Changyou Chen, Wenlin Wang, and Lawrence Carin. Second-order adversarial attack and certifiable robustness. arXiv preprint arXiv:1809.03113, 2018. URL http://arxiv.org/abs/1809.03113.

Sungbin Lim, Eun Bi Yoon, Taehyun Byun, Taewon Kang, Seungwoo Kim, Kyungjae Lee, and Sungjoon Choi. Score-based generative modeling through stochastic evolution equations in hilbert spaces. Advances in neural information processing systems, 36:37799–37812, 2023.

Laurent Meunier, Raphael Ettedgui, Rafael Pinot, Yann Chevaleyre, and Jamal Atif. Towards consistency in adversarial classification. In S. Koyejo et al. (ed.), Advances in Neural Information Processing Systems, volume 35, pp. 8538–8549. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/38d6af46cca4ce1f7d699bf11078cb84-Paper-Conference.pdf.

Amir Nazemi, Mohammad Hadi Sepanj, Nicholas Pellegrino, Chris Czarnecki, and Paul Fieguth. Particle-filtering-based latent diffusion for inverse problems, 2024. URL https://arxiv.org/abs/2408.13868.

Xiangyu Qi, Kaixuan Huang, Ashwinee Panda, Peter Henderson, Mengdi Wang, and Prateek Mittal. Visual adversarial examples jailbreak aligned large language models. In Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence, AAAI'24/IAAI'24/EAAI'24. AAAI Press, 2024. ISBN 978-1-57735-887-9. doi: 10.1609/aaai.v38i19.30150. URL https://doi.org/10.1609/aaai.v38i19.30150.

Hadi Salman, Greg Yang, Jerry Li, Pengchuan Zhang, Huan Zhang, Ilya Razenshteyn, and Sébastien Bubeck. Provably robust deep learning via adversarially trained smoothed classifiers. Curran Associates Inc., Red Hook, NY, USA, 2019.

Frederick Shpilevskiy, Saiyue Lyu, Krishnamurthy Dj Dvijotham, Mathias Lécuyer, and Pierre-André Noël. Adaptive diffusion denoised smoothing : Certified robustness via randomized smoothing with differentially private guided denoising diffusion, 2025. URL https://arxiv.org/abs/2507.08163.

Jiaming Song, Arash Vahdat, Morteza Mardani, and Jan Kautz. Pseudoinverse-guided diffusion models for inverse problems. In International Conference on Learning Representations, 2023.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. International Conference on Learning Representations (ICLR), 2021.

Stanislas Strasman, Antonio Ocello, Claire Boyer, Sylvain Le Corff, and Vincent Lemaire. An analysis of the noise schedule for score-based generative models. Transactions on Machine Learning Research, 2025a. ISSN 2835-8856.

Stanislas Strasman, Sobihan Surendran, Claire Boyer, Sylvain Le Corff, Vincent Lemaire, and Antonio Ocello. Wasserstein convergence of critically damped langevin diffusions. In Advances in Neural Information Processing Systems, 2025b.

Stanislas Strasman, Gabriel Victorino Cardoso, Sylvain Le Corff, Vincent Lemaire, and Antonio Ocello. On forgetting and stability of score-based generative models. arXiv preprint arXiv:2601.21868, 2026.

Peter Súkeník, Aleksei Kuvshinov, and Stephan Günnemann. Intriguing properties of input-dependent randomized smoothing. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), Proceedings of the 39th International Conference on Machine Learning, volume 162 of Proceedings of Machine Learning Research, pp. 20697–20743. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/sukeni-k22a.html.

Florian Tramèr, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. On adaptive attacks to adversarial example defenses. In Conference on Neural Information Processing Systems (NeurIPS), 2020. URL https://arxiv.org/abs/2002.08347.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In International Conference on Learning Representations, 2019. URL https://openreview.net/forum?id=SyxAb30cY7.

Yusuke Tsuzuku, Issei Sato, and Masashi Sugiyama. Lipschitz-margin training: scalable certification of perturbation invariance for deep neural networks. In Proceedings of the 32nd International Conference on Neural Information Processing Systems, NIPS'18, pp. 6542–6551, Red Hook, NY, USA, 2018. Curran Associates Inc.

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.

Luhuan Wu, Brian L. Trippe, Christian A. Naesseth, David M. Blei, and John P. Cunningham. Practical and asymptotically exact conditional sampling in diffusion models. In Advances in Neural Information Processing Systems, 2023.

Greg Yang, Tony Duan, J. Edward Hu, Hadi Salman, Ilya Razenshteyn, and Jerry Li. Randomized smoothing of all shapes and sizes. In Hal Daumé III and Aarti Singh (eds.), Proceedings of the 37th International Conference on Machine Learning, volume 119 of Proceedings of Machine Learning Research, pp. 10693–10705. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/yang20c.html.

Runtian Zhai, Chen Dan, Di He, Huan Zhang, Boqing Gong, Pradeep Ravikumar, Cho-Jui Hsieh, and Liwei Wang. Macer: Attack-free and scalable robust training via maximizing certified radius. In International Conference on Learning Representations, 2020. URL https://openreview.net/forum?id=rJx1Na4Fwr.

Bingliang Zhang, Wenda Chu, Julius Berner, Chenlin Meng, Anima Anandkumar, and Yang Song. Improving Diffusion Inverse Problem Solving with Decoupled Noise Annealing. In 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 20895–20905, 2025. doi: 10.1109/CVPR52734.2025.01946. URL https://ieeexplore.ieee.org/abstract/document/11094844.

A TECHNICAL RESULTS

A.1 PROOF OF LEMMA 4.1

Let $A_y^\lambda(z) = \int_{\mathcal{C}_y^{\mathrm{q}}} \exp(\lambda \Psi_{\mathrm{ent}}(u)) \overleftarrow{\mathrm{P}}_{T|T-t(\sigma_{\mathrm{sm}})}(\mathrm{d}u|\, \mathrm{s}_{t(\sigma_{\mathrm{sm}})}\, z)$. Then, $\mathrm{q}_{\sigma_{\mathrm{sm}}}^{\lambda\Psi_1}(y|z) = A_y^\lambda(z)/Z_\lambda(z)$ and

$$\partial_\lambda A_y^\lambda(z) = \int_{\mathcal{C}_y^{\mathrm{q}}} \Psi_1(u) \exp(\lambda \Psi_{\mathrm{ent}}(u)) \overleftarrow{\mathrm{P}}_{T|T-t(\sigma_{\mathrm{sm}})}(\mathrm{d}x|\, \mathrm{s}_{t(\sigma_{\mathrm{sm}})}\, z) = A_y^\lambda(z) m_y^{\lambda\Psi}(z)\,,$$

and similarly

$$\partial_\lambda Z_\lambda(z) = Z_\lambda(z) \bar{m}^{\lambda\Psi}(z)\,.$$

Therefore,

$$\partial_\lambda \mathrm{q}_{\sigma_{\mathrm{sm}}}^{\lambda\Psi_1}(y|z) = \frac{\partial_\lambda A_y^\lambda(z)}{Z_\lambda(z)} - \frac{A_y^\lambda(z)\partial_\lambda Z_\lambda(z)}{Z_\lambda(z)^2} = \mathrm{q}_{\sigma_{\mathrm{sm}}}^{\lambda\Psi_1}(y|z)\left(m_y^{\lambda\Psi_1}(z) - \bar{m}^{\lambda\Psi_1}(z)\right),$$

which concludes the proof.

A.2 PROOF OF PROPOSITION 4.3

Let $y \neq c_A$. By Lemma 4.1,

$$\partial_\lambda p_y^\lambda(x) = \mathbb{E}\left[\partial_\lambda \mathrm{q}_{\sigma_{\mathrm{sm}}}^{\lambda\Psi_{\mathrm{ent}}}(y|Z_x)\right] = \mathbb{E}\left[\mathrm{q}_{\sigma_{\mathrm{sm}}}^{\lambda\Psi_{\mathrm{ent}}}(y|Z_x)\left(m_y^{\lambda\Psi_{\mathrm{ent}}}(Z_x) - \bar{m}^{\lambda\Psi_{\mathrm{ent}}}(Z_x)\right)\right] = \Gamma_y^\lambda(x).$$

Therefore,

$$\partial_\lambda\left(p_{c_A}^\lambda(x) - p_y^\lambda(x)\right) = \Gamma_{c_A}^\lambda(x) - \Gamma_y^\lambda(x) \geq 0,$$

where the inequality follows from Assumption 4.2. Therefore,

$$p_{c_A}^\lambda(x) - p_y^\lambda(x) \geq p_{c_A}^0(x) - p_y^0(x).$$

Since c_A is the unique maximizer at $\lambda = 0$, $p_{c_A}^0(x) > p_y^0(x)$, which concludes the proof.

A.3 PROOF OF PROPOSITION 4.4

Let us first start with a preliminary Lemma.

Lemma A.1. *Assume $|\mathcal{Y}| = 2$ and that Assumption 4.2 holds. Let $x \in \mathbb{R}^d$ and assume that $c_A = \arg\max_{y\in\mathcal{Y}} p_y^0(x)$ is unique. Denote by c_B the remaining class in $\mathcal{Y}$ after removing c_A. Then for every $\lambda \geq 0$, we have*

$$\mathbb{E}\left[\mathrm{q}_{\sigma_{\mathrm{sm}}}^{\lambda\Psi_{\mathrm{ent}}}(c_A|Z_x)\,\mathrm{q}_{\sigma_{\mathrm{sm}}}^{\lambda\Psi_{\mathrm{ent}}}(c_B|Z_x)\left(m_{c_A}^{\lambda\Psi_{\mathrm{ent}}}(Z_x) - m_{c_B}^{\lambda\Psi_{\mathrm{ent}}}(Z_x)\right)\right] \geq 0\,.$$

Proof. First note that, $\mathcal{Y} = \{c_A, c_B\}$. Hence by definition of $\mathrm{q}_{\sigma_{\mathrm{sm}}}^{\lambda\Psi_{\mathrm{ent}}}$ and $\bar{m}^{\lambda\Psi_{\mathrm{ent}}}$, for any $\lambda \geq 0$ and $z \in \mathbb{R}^d$ we have $\mathrm{q}_{\sigma_{\mathrm{sm}}}^{\lambda\Psi_{\mathrm{ent}}}(c_A|z) + \mathrm{q}_{\sigma_{\mathrm{sm}}}^{\lambda\Psi_{\mathrm{ent}}}(c_B|z) = 1$ and

$$\bar{m}^{\lambda\Psi_{\mathrm{ent}}}(z) = \mathrm{q}_{\sigma_{\mathrm{sm}}}^{\lambda\Psi_{\mathrm{ent}}}(c_A|z) m_{c_A}^{\lambda\Psi_{\mathrm{ent}}}(z) + \mathrm{q}_{\sigma_{\mathrm{sm}}}^{\lambda\Psi_{\mathrm{ent}}}(c_B|z) m_{c_B}^{\lambda\Psi_{\mathrm{ent}}}(z)\,. \tag{16}$$

Therefore, substituting (16) in the definition of $\Gamma_{c_A}^\lambda(x)$ (from Assumption 4.2), we get

$$\begin{aligned}\Gamma_{c_A}^\lambda(x) &:= \mathbb{E}\left[\mathrm{q}_{\sigma_{\mathrm{sm}}}^{\lambda\Psi_{\mathrm{ent}}}(c_A|Z_x)\left(m_{c_A}^{\lambda\Psi_{\mathrm{ent}}}(Z_x) - \mathrm{q}_{\sigma_{\mathrm{sm}}}^{\lambda\Psi_{\mathrm{ent}}}(c_A|Z_x) m_{c_A}^{\lambda\Psi_{\mathrm{ent}}}(Z_x) - \mathrm{q}_{\sigma_{\mathrm{sm}}}^{\lambda\Psi_{\mathrm{ent}}}(c_B|Z_x) m_{c_B}^{\lambda\Psi_{\mathrm{ent}}}(Z_x)\right)\right] \\ &= \mathbb{E}\left[\mathrm{q}_{\sigma_{\mathrm{sm}}}^{\lambda\Psi_{\mathrm{ent}}}(c_A|Z_x)\,\mathrm{q}_{\sigma_{\mathrm{sm}}}^{\lambda\Psi_{\mathrm{ent}}}(c_B|Z_x)\left(m_{c_A}^{\lambda\Psi_{\mathrm{ent}}}(Z_x) - m_{c_B}^{\lambda\Psi_{\mathrm{ent}}}(Z_x)\right)\right]\end{aligned}$$

Similarly, we can show that $\Gamma_{c_B}^\lambda(x) = -\Gamma_{c_A}^\lambda(x)$. Consequently, in the binary setting, Assumption 4.2, which requires $\Gamma_{c_A}^\lambda(x) \geq \Gamma_{c_B}^\lambda(x)$, is equivalent to $\Gamma_{c_A}^\lambda(x) \geq 0$. This concludes the proof. □

We now prove the result using Lemma A.1. In fact, by (16) Lemma 4.1,

$$\begin{aligned}\partial_\lambda \, \mathrm{q}_{\sigma_{\mathrm{sm}}}^{\lambda\Psi_{\mathrm{ent}}}(c_A|z) &= \mathrm{q}_{\sigma_{\mathrm{sm}}}^{\lambda\Psi_{\mathrm{ent}}}(c_A|z)\left(m_{c_A}^{\lambda\Psi_{\mathrm{ent}}}(z) - \bar{m}^{\lambda\Psi_{\mathrm{ent}}}(z)\right) \\ &= \mathrm{q}_{\sigma_{\mathrm{sm}}}^{\lambda\Psi_{\mathrm{ent}}}(c_A|z)\,\mathrm{q}_{\sigma_{\mathrm{sm}}}^{\lambda\Psi_{\mathrm{ent}}}(c_B|z)\left(m_{c_A}^{\lambda\Psi_{\mathrm{ent}}}(z) - m_{c_B}^{\lambda\Psi_{\mathrm{ent}}}(z)\right) .\end{aligned}$$

By (13) and Lemma A.1,

$$\partial_\lambda p_{c_A}^\lambda(x) = \mathbb{E}\left[\mathrm{q}_{\sigma_{\mathrm{sm}}}^{\lambda\Psi_{\mathrm{ent}}}(c_A|Z_x)\,\mathrm{q}_{\sigma_{\mathrm{sm}}}^{\lambda\Psi_{\mathrm{ent}}}(c_B|Z_x)\left(m_{c_A}^{\lambda\Psi_{\mathrm{ent}}}(Z_x) - m_{c_B}^{\lambda\Psi_{\mathrm{ent}}}(Z_x)\right)\right] \geq 0 .$$

Therefore, $\lambda \mapsto p_{c_A}^\lambda(x)$ is nondecreasing and $\lambda \mapsto p_{c_B}^\lambda(x)$ is nonincreasing. Moreover, since by definition of c_A $p_{c_A}^0(x) > p_{c_B}^0(x)$, we have, for every $\lambda \geq 0$,

$$p_{c_A}^\lambda(x) \geq p_{c_A}^0(x) > p_{c_B}^0(x) \geq p_{c_B}^\lambda(x) .$$

Consequently, $\arg\max_{y\in\mathcal{Y}} p_{c_A}^\lambda(x) = c_A$ for every $\lambda \geq 0$.

In addition, since for all u, $\Phi^{-1}(1-u) = -\Phi^{-1}(u)$, $\mathsf{R}_\lambda(x) = \sigma_{\mathrm{sm}}\Phi^{-1}(p_{c_A}^\lambda(x))$. For all λ such that $p_{c_A}^\lambda(x) \in (0,1)$,

$$\partial_\lambda \mathsf{R}_\lambda(x) = \sigma_{\mathrm{sm}} \frac{\partial_\lambda p_{c_A}^\lambda(x)}{\phi\left(\Phi^{-1}\left(p_{c_A}^\lambda(x)\right)\right)} .$$

It follows that $\partial_\lambda \mathsf{R}_\lambda(x) \geq 0$, which concludes the proof.

B ALGORITHMS MISSING FROM THE MAIN DOCUMENT

Algorithm 1 Sampling the guided base classifier under Gaussian noise

function SAMPLEGUIDED($h_{\theta,\lambda,\varphi}, x, n, \sigma_{\mathrm{sm}}$)
  counts $\leftarrow (0, \ldots, 0) \in \{0, \ldots, n\}^K$
  for $i = 1, \ldots, n$ **do**
    Sample $\epsilon_i \sim \mathcal{N}(0, \sigma_{\mathrm{sm}}^2 \mathbf{I}_d)$
    $y_i \leftarrow h_{\theta,\lambda,\varphi}(x + \epsilon_i)$
    counts$[y_i] \leftarrow$ counts$[y_i] + 1$
  end for
  return counts
end function

Algorithm 2 Prediction with the guided smoothed classifier

function PREDICTGUIDED($h_{\theta,\lambda,\varphi}, x, n, \sigma_{\mathrm{sm}}, \alpha$)
  counts $\leftarrow$ SAMPLEGUIDED($h_{\theta,\lambda,\varphi}, x, n, \sigma_{\mathrm{sm}}$)
  $\hat{c}_A, \hat{c}_B \leftarrow$ indices of the two largest entries of counts
  $n_A \leftarrow$ counts$[\hat{c}_A]$
  $n_B \leftarrow$ counts$[\hat{c}_B]$
  $p_{\mathrm{val}} \leftarrow$ BINOMPVALUE($n_A, n_A + n_B, 1/2$)
  if $p_{\mathrm{val}} \leq \alpha$ **then**
    return $\hat{c}_A$
  else
    return ABSTAIN
  end if
end function

C FURTHER DETAILS FOR THE NUMERICAL EXPERIMENTS.

C.1 ARCHITECTURE DETAILS

We follow for both CIFAR10 and ImageNet datasets the denoisers and the classifiers used in Carlini et al. (2023b). Therefore, for CIFAR-10 we used the denoiser available at hugging face under the tag ‘google/ddpm-cifar-10-32‘ and a ViT classifier finetuned to CIFAR-10 from a model trained on ImageNet and available at Hugging Face under the tag ‘aaraki / vit-base-patch16-224-in21k-finetuned-cifar10‘. For Imagenet, we used the unconditional denoiser network proposed in Dhariwal & Nichol (2021) and the ”beit-large-patch16-512” classifier proposed in Bao et al. (2021).

C.2 PLOTS FOR CIFAR 10

We show in Figure 3 the evolution of certified accuracy for a given radius for all configurations tested in the CIFAR 10 case.

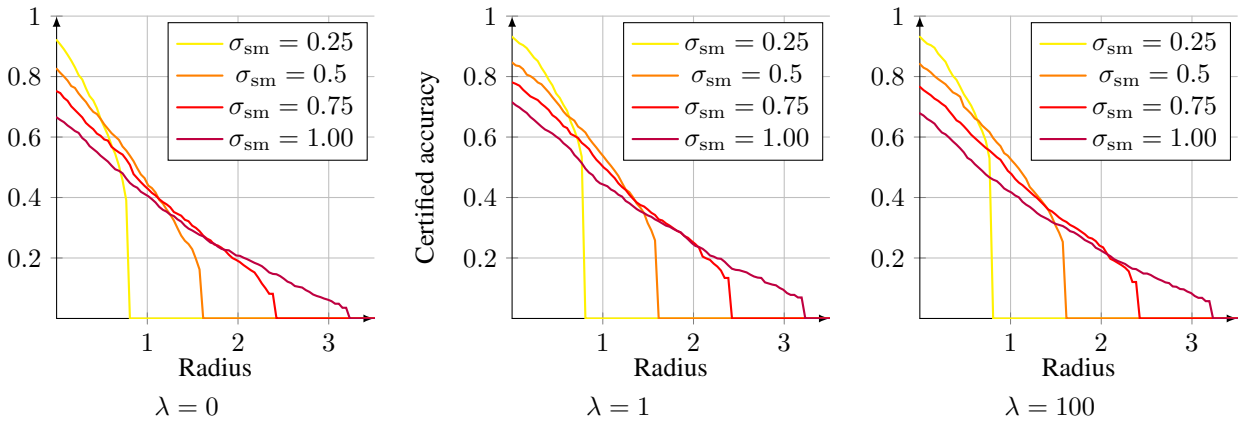


Figure 3: CIFAR 10 with $\Psi = \lambda\Psi_{\text{ent}}$ for different σ_{sm} and λ.

C.3 PLOTS IMAGENET

We selected 10 random images $\{X_0^1, \cdots, X_0^{10}\}$ in the test set of ImageNet and generated for each image 4 noised images $\overrightarrow{\mathrm{X}}^{i,j}_{t(\sigma_{\text{sm}})} = X_0^i + \sigma_{\text{sm}} Z^j$ with $\sigma_{\text{sm}} = 3$. We then sample reconstructions according to $\overleftarrow{\mathrm{P}}^{\Psi}_{T|T-t(\sigma_{\text{sm}})}(\overrightarrow{\mathrm{X}}^{i,j}_{t(\sigma_{\text{sm}})})$ for $\Psi = \lambda\Psi_{\text{ent}}$ with $\lambda \in \{0, 1, 10\}$. The results are shown in Figures 4 to 13.



Figure 4: Visualization of Imagenet on other images



Figure 5: Visualization of Imagenet on other images

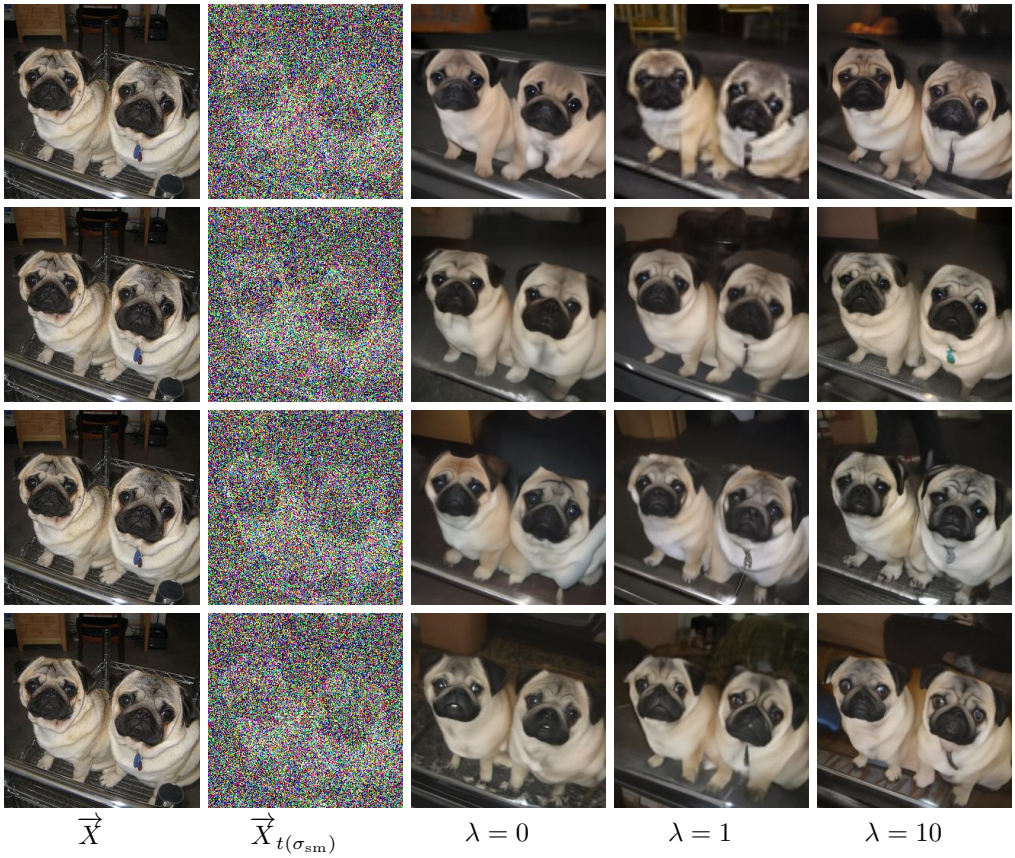


Figure 6: Visualization of Imagenet on other images

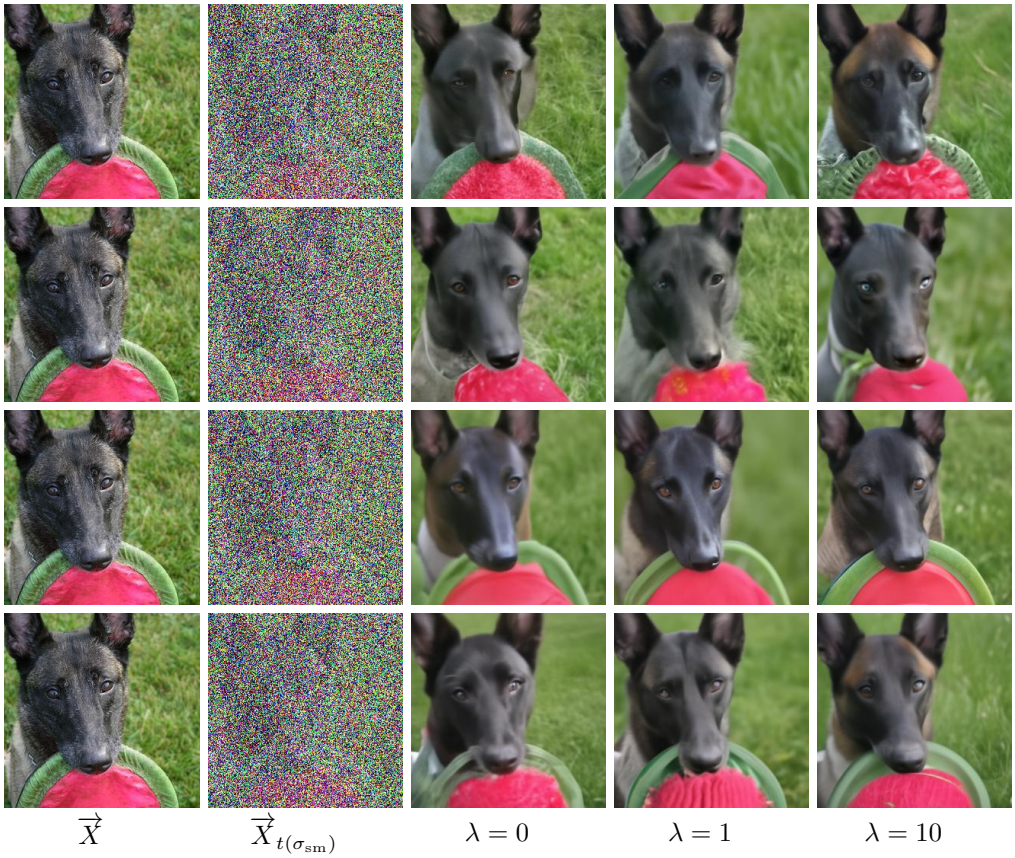


Figure 7: Visualization of Imagenet on other images

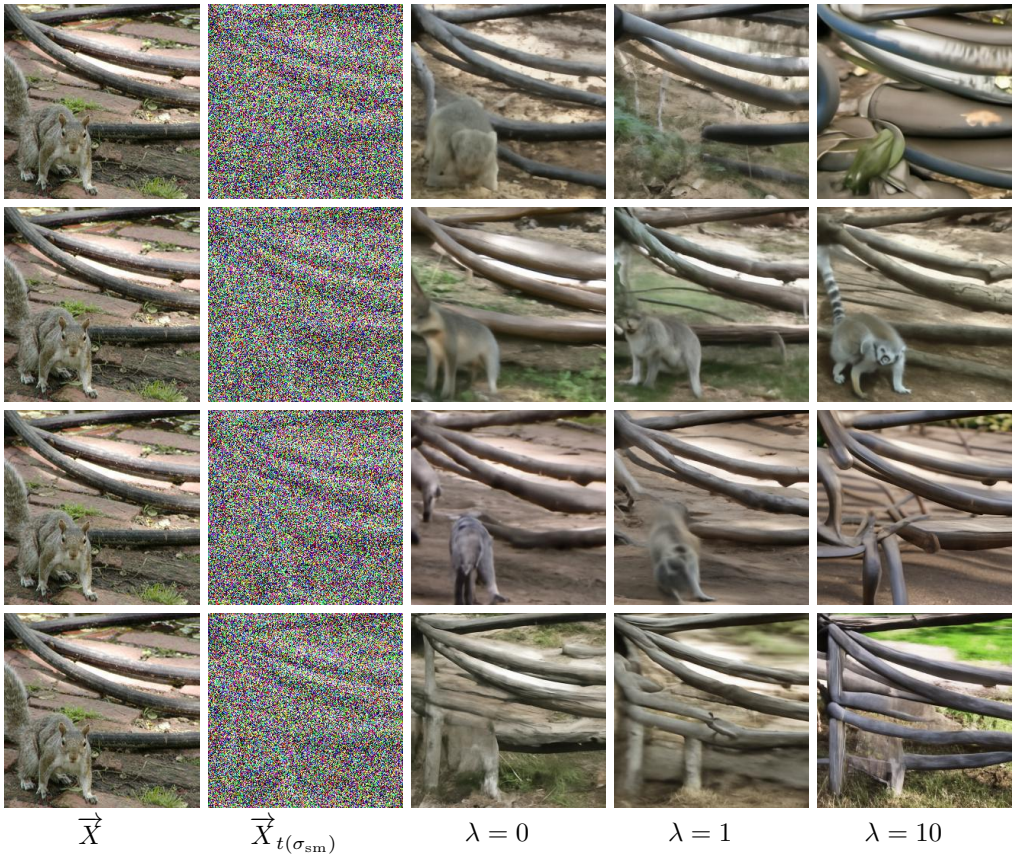


Figure 8: Visualization of Imagenet on other images

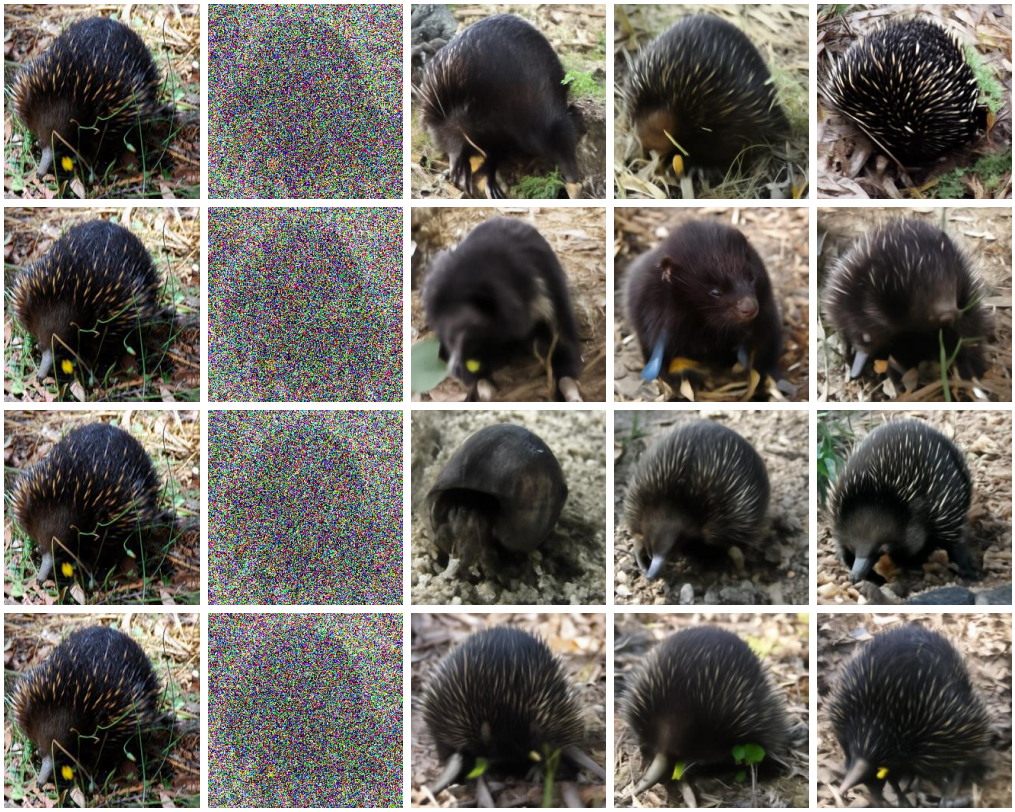


Figure 9: Visualization of Imagenet on other images



Figure 10: Visualization of Imagenet on other images

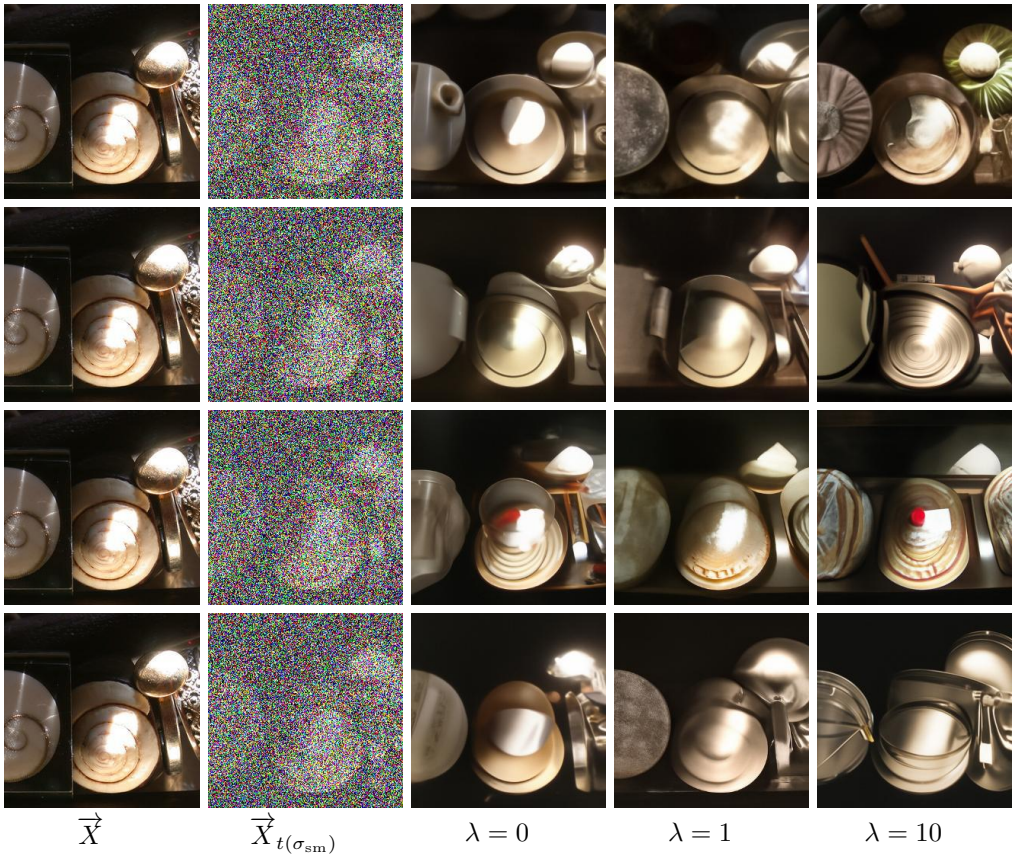


Figure 11: Visualization of Imagenet on other images

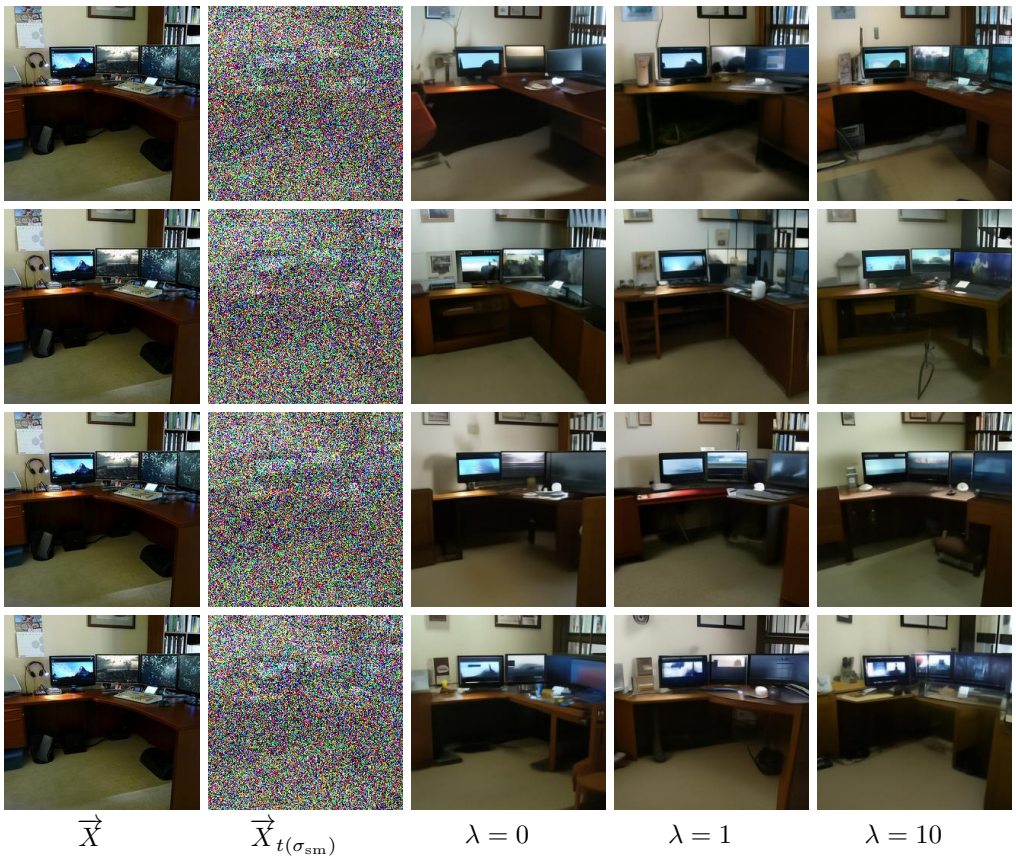


Figure 12: Visualization of Imagenet on other images

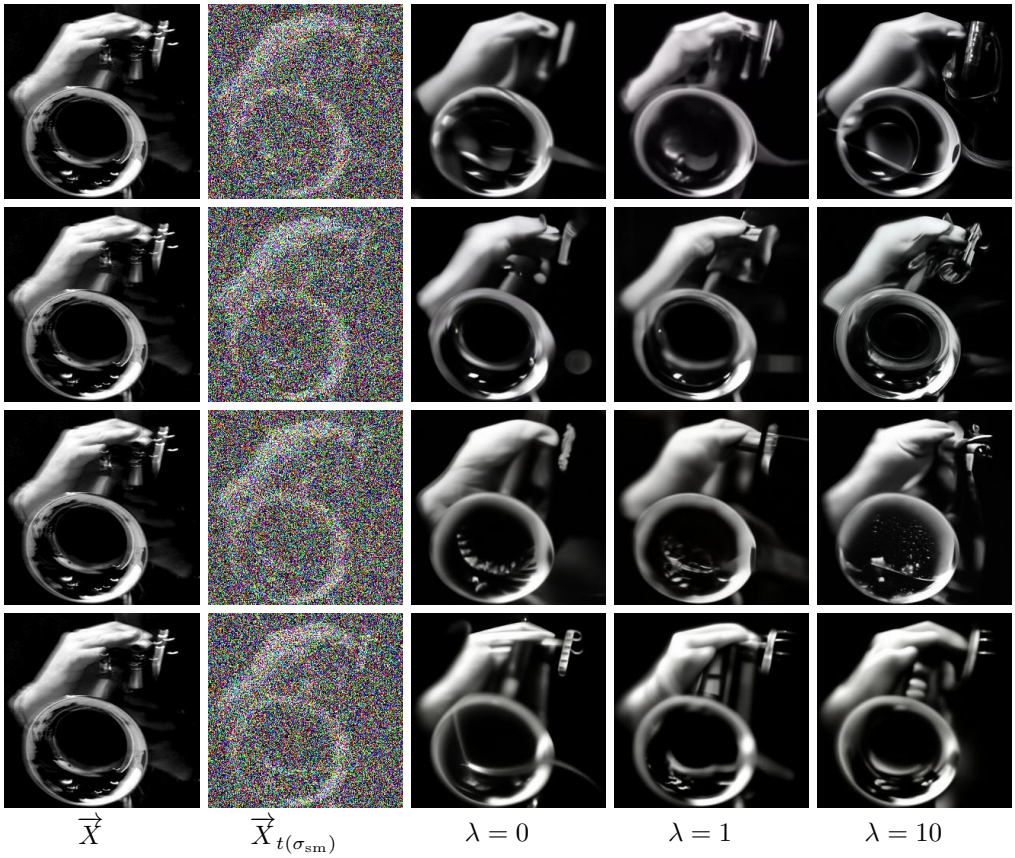


Figure 13: Visualization of Imagenet on other images