Towards Better Certified Segmentation via Diffusion Models

Othmane Laousy[1] **Alexandre Araujo**[2] **Guillaume Chassagnon**[3] **Marie-Pierre Revel**[3] **Siddharth Garg**[2]
Farshad Khorrami[2] **Maria Vakalopoulou**[1]

[1]MICS, CentraleSupélec, Paris-Saclay University, Inria Saclay, France
[2]New York University, NY, USA
[3]Paris Cité University, France

Abstract

The robustness of image segmentation has been an important research topic in the past few years as segmentation models have reached production-level accuracy. However, like classification models, segmentation models can be vulnerable to adversarial perturbations, which hinders their use in critical-decision systems like healthcare or autonomous driving. Recently, randomized smoothing has been proposed to certify segmentation predictions by adding Gaussian noise to the input to obtain theoretical guarantees. However, this method exhibits a trade-off between the amount of added noise and the level of certification achieved. In this paper, we address the problem of certifying segmentation prediction using a combination of randomized smoothing and diffusion models. Our experiments show that combining randomized smoothing and diffusion models significantly improves certified robustness, with results indicating a mean improvement of 21 points in accuracy compared to previous state-of-the-art methods on Pascal-Context and Cityscapes public datasets. Our method is independent of the selected segmentation model and does not need any additional specialized training procedure.

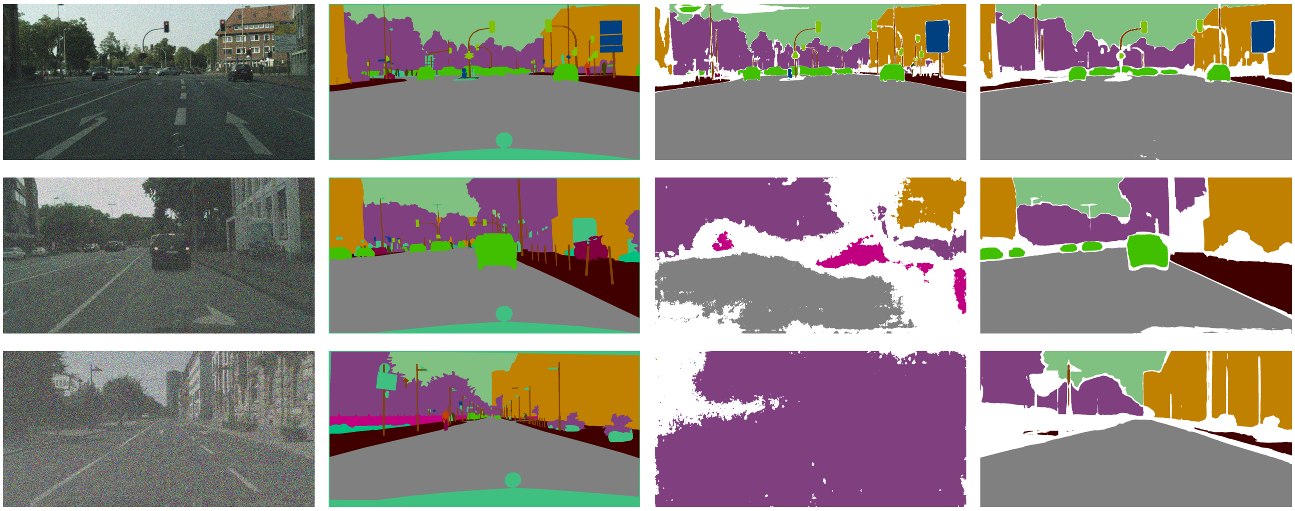
1 INTRODUCTION

Neural networks have been known to be vulnerable to adversarial perturbations (Szegedy et al., 2013; Madry et al., 2018; Goodfellow et al., 2014; Carlini and Wagner, 2017), *i.e.*, imperceptible variations of natural examples, crafted to deliberately mislead the models. In recent years, significant efforts have been made to develop certified defenses that guarantee a specified level of robustness against adversarial inputs within a certain radius. (*e.g.*, 1-Lipschitz Networks (Trockman and Kolter, 2021; Meunier et al., 2022; Araujo et al., 2023), bound propagation (Gowal et al., 2018; Huang et al., 2021), randomized smoothing (Li et al., 2019a; Cohen et al., 2019; Salman et al., 2019)). Although most defenses focus on classification tasks, in this paper, we focus on certifying segmentation models and argue that certified segmentation is an even more pressing issue as these models are already used in critical systems such as healthcare and autonomous vehicles.

Randomized smoothing has emerged as the leading technique for certified robustness due to its scalability and model-agnostic properties. It consists in applying a convolution between a base classifier and a Gaussian distribution, enabling the method to handle large input sizes (*e.g.*, ImageNet, Pascal-Context, Cityscapes), while providing state-of-the-art certified accuracy. However, this technique exhibits a trade-off between adding enough noise for certification and preserving the input's semantic information for accurate predictions. In fact, several impossibility results from an information-theory perspective have been introduced (Kumar et al., 2020; Blum et al., 2020; Yang et al., 2020) and inherently limit randomized smoothing from providing large certified radii. Nevertheless, recent works, both theoretical (Ettedgui et al., 2022; Mohapatra et al., 2020) and empirical (Salman et al., 2020; Carlini et al., 2023), have explored potential solutions to this trade-off. To address the issue of removed information due to noise injection, several works, in the context of classification tasks, have proposed methods to *denoise* the input after the noise injection step (Salman et al., 2020; Carlini et al., 2023). While Salman et al. (2020) trained their own denoiser models on Gaussian noise for the specific task of certified robustness, Carlini et al. (2023) extended the work of Salman et al. (2020) by using off-the-shelf *Denoising Diffusion Probabilistic Models* (Sohl-Dickstein et al., 2015; Ho et al., 2020; Nichol and Dhariwal, 2021), a form of generative models that takes a random Gaussian noise and generates a real-world image.

In this paper, we build upon previous work on certified robustness to improve certified segmentation extending the

Accepted for the 39th *Conference on Uncertainty in Artificial Intelligence* (UAI 2023).



(a) Image with noise (b) Ground truth segmentation (c) SEGCERTIFY segmentation (d) Our segmentation

Figure 1: Examples of our approach (DENOISECERTIFY) compared to SEGCERTIFY proposed by Fischer et al. (2021) on the Cityscapes dataset. From left to right: (a) the initial image with added noise, (b) the ground truth segmentation, (c) the abstained segmentation obtained with SEGCERTIFY, (d) the abstained segmentation obtained with DENOISECERTIFY (ours). Each row corresponds to a noise level, from top to bottom: $\sigma = 0.25, 0.5$ and 1.0.

work of Fischer et al. (2021) and Carlini et al. (2023). We present a comprehensive set of experiments on PASCAL-Context (Mottaghi et al., 2014) and Cityscapes (Cordts et al., 2016) datasets and successfully achieve state-of-the-art results on certified robustness for segmentation tasks. Our results show that combining randomized smoothing and diffusion models significantly improves certified robustness, with a mean increase of 21 points in accuracy and 14 points in mIoU when compared to previous methods. Our main contributions are summarized as follows:

- First, we build upon the work of Fischer et al. (2021) and Carlini et al. (2023) and propose for the first time, a certified segmentation approach leveraging diffusion models. Through a series of experiments, we demonstrate that incorporating a denoiser in conjunction with a segmentation model that has been *trained with noise injection* presents certain trade-offs in the certified accuracy achieved, depending on the variance of the noise.

- Second, we further improve certified accuracy by combining off-the-shelf diffusion and state-of-the-art segmentation models allowing us to reach state-of-the-art results for certified segmentation.

- Third, we propose an in-depth analysis through a series of experiments on the use of noise during training as well as the generalization of denoising diffusion models with respect to image resolution and data distribution.

2 RELATED WORK

Adversarial Attacks & Certified Defenses. Since the discovery of adversarial examples (Szegedy et al., 2013), a wealth of work focused on devising attacks (Goodfellow et al., 2014; Kurakin et al., 2018; Carlini and Wagner, 2017; Croce and Hein, 2020, 2021) and defenses (Goodfellow et al., 2014; Madry et al., 2018; Pinot et al., 2019; Araujo et al., 2020, 2021), leading to an ongoing back-and-forth battle. Most of these defenses relied on smoothing the local neighborhood around each point, resulting in very small gradients on which attacks were based. However, it has become apparent that many of the empirical defenses that have been created could be circumvented with stronger attacks (Athalye et al., 2018).

This false sense of security and the persistent cat-and-mouse game called for *certified defenses* that provide provable robustness guarantees. In recent years, mainly two types of certified defenses have been proposed. The first approach provides robustness guarantees based on the Lipschitz constant of the networks and their margin (*i.e.*, the difference between the highest and second highest logits). This connection was introduced by Tsuzuku et al. (2018) and opened an important research direction in the design and training of 1-Lipschitz neural networks (Miyato et al., 2018; Farnia et al., 2019; Li et al., 2019b; Trockman and Kolter, 2021; Singla and Feizi, 2021; Yu et al., 2022; Meunier et al., 2022; Prach and Lampert, 2022; Xu et al., 2022; Araujo et al., 2023). Although this approach offers fast certificate computation, it suffers from important drawbacks. Indeed, due to the strict constraint on the networks and reduced expressivity, 1-Lipschitz neural networks offer a reduced natural and certified accuracy and do not scale to large datasets (*e.g.*, ImageNet, Pascal-Context, Cityscapes). On the other hand, a second approach called Randomized Smoothing leverages randomization. This method, introduced by Lecuyer et al. (2019) and further improved by Li et al. (2019a); Cohen et al. (2019) and Salman et al. (2019), consists in convolving the function with a Gaussian probability distribution during the inference phase. The desirable property of a smooth classifier is ensuring that the prediction is constant within an ℓ_2 ball around any input.

Diffusion models. Diffusion probabilistic models have been introduced by Sohl-Dickstein et al. (2015), and further refined by Ho et al. (2020) and Nichol and Dhariwal (2021). The goal was to design a generative Markov chain that transforms a known distribution (*e.g.*, Gaussian) into a target (data) distribution using a diffusion process. However, instead of using a Markov chain to evaluate the model, they defined the probabilistic model as the endpoint of the Markov chain. Subsequently, this methodology was refined and applied for producing high-quality samples, such as images, as demonstrated by Ho et al. (2020) and Nichol and Dhariwal (2021). The results indicated that this type of model can generate better images in comparison to other methods and also demonstrated a connection with denoising. Recently, diffusion probabilistic models have been applied successfully in the context of certified robustness for classification tasks where a diffusion model is used as a first step to denoise inputs for randomized smoothing (Carlini et al., 2023).

Certified Segmentation. Deep neural networks trained for segmentation tasks have been shown to be vulnerable to adversarial attacks (Xie et al., 2017; Arnab et al., 2018; Xiang et al., 2019; He et al., 2019; Kang et al., 2020). In this context, Fischer et al. (2021) use the work of Cohen et al. (2019) and propose a method to certify segmentation with randomized smoothing for norm-bounded perturbations. Other lines of work investigate certified robustness for structured outputs, for example, Kumar and Goldstein (2021) proposed a procedure based on randomized smoothing to find the minimum enclosing ball in the output space and Yatsura et al. (2023) introduced a method called *demasked smoothing* to defend against adversarial patch attacks for semantic segmentation tasks.

In this paper, we build upon the work of Fischer et al. (2021) and Carlini et al. (2023) and introduce, for the first time, a randomized smoothing approach with a denoising step in the context of certified semantic segmentation.

3 BACKGROUND

In this section, we review the necessary background on randomized smoothing and on certified segmentation.

3.1 ADVERSARIAL ATTACKS & RANDOMIZED SMOOTHING FOR CLASSIFICATION

We first introduce adversarial attacks and randomized smoothing in the setting they have been introduced, *i.e.*, for a classification task. We will generalize it to the segmentation task in the next paragraph.

Let $\mathcal{X} \subset \mathbb{R}^d$ and $\mathcal{Y} = \{1, \dots, K\}$ be the input space and target space respectively with K denoting the number of classes. Let us denote a classifier $f : \mathcal{X} \to \mathcal{Y}$ (*e.g.*, a neural network) such that for a given couple input-label $(x, y) \in \mathcal{X} \times \mathcal{Y}$, we say the classifier f correctly classifies x if: $f(x) = y$. An adversarial attack is a small norm-bounded perturbation $\delta \in \mathbb{R}^d$ with $\|\delta\|_2 \leq \epsilon$ such that:

$$f(x + \delta) \neq y. \tag{1}$$

Randomized smoothing, introduced in Cohen et al. (2019), considers a *smooth* version of the classifier f, such that:

$$g(x) = \arg\max_c \mathbb{P}_{\eta \sim \mathcal{N}(0,\sigma^2\mathbf{I})} \left[f(x + \eta) = c \right] . \tag{2}$$

To compute the probability in Equation 2, Cohen et al. (2019) proposed a Monte-Carlo approach where the prediction is computed from a small number of samples, *i.e.*, n_0, with a majority vote and a lower-bound on the certified radius computed with a higher number of samples, *i.e.*, n. A benefit of using the smooth classifier g is obtaining a certified radius of robustness for each data point, thus determining a certified level of accuracy within a specified attack 'budget' ϵ. More formally, Cohen et al. introduced the following theorem:

Theorem 1 (From Cohen et al. (2019)). *Suppose $y \in \mathcal{Y}$, let*

$$p_y = \mathbb{P}_{\eta \sim \mathcal{N}(0,\sigma^2\mathbf{I})} \left[f(x + \eta) = y \right] \tag{3}$$

and let $\underline{p_y}$ the lower bound of p_y computed via Monte-Carlo sampling. Let $\overline{p_{\neg y}} = 1 - \underline{p_y}$, then, if

$$\mathbb{P}_\eta \left[f(x + \eta) = y \right] \geq \underline{p_y} \geq \overline{p_{\neg y}} \geq \max_{c \neq y} \mathbb{P}_\eta \left[f(x + \eta) = c \right] , \tag{4}$$

then $g(x + \delta) = y$ for all δ satisfying $\|\delta\|_2 \leq R$ with $R := \sigma\Phi^{-1}(\underline{p_y})$ and Φ is the cumulative distribution function of the standard Gaussian distribution.

To properly approximate the probability p_y with a confidence interval, Cohen et al. (2019) proposed a procedure which samples n realizations of $\eta \sim \mathcal{N}(0, \sigma^2\mathbf{I})$ and computes $f(x+\eta)$. From these n realizations, a vector of counts for each class in $\mathcal{Y}$ is computed and these counts are then used to estimate the probability p_y and the radius R with confidence $1 - \alpha$ with $\alpha \in [0, 1]$. If the confidence level is not reached (for example, the number of samples is not enough), the procedure will abstain.

Algorithm 1 Predict & Certify by Fischer et al. (2021)

```
function SegCertify($g, \sigma, x, n, n_0, \delta, \alpha$)
    $\texttt{cnts}_1^0, \ldots, \texttt{cnts}_N^0 \leftarrow$ Sample($g, x, n_0, \sigma$)
    $\texttt{cnts}_1, \ldots, \texttt{cnts}_N \leftarrow$ Sample($g, x, n, \sigma$)
    for $i \leftarrow \{1, \ldots, N\}$:
        $\hat{c}_i \leftarrow$ top index in $\texttt{cnts}_i^0$
        $n_i \leftarrow \texttt{cnts}_i[\hat{c}_i]$
        $pv_i \leftarrow$BinPValue($n_i, n, \leq, \delta$)
    $r_1, \ldots, r_N \leftarrow$ FwerControl($\alpha, pv_1, \ldots, pv_N$)
    for $i \leftarrow \{1, \ldots, N\}$:
        if $\neg r_i$: $\hat{c}_i \leftarrow \varnothing$
    $R \leftarrow \sigma\Phi^{-1}(\delta)$
    return $\hat{c}_1, \ldots, \hat{c}_N, R$
```

Algorithm 2 Sample Function

```
function Sample($g, x, n, \sigma$)
    $\texttt{cnts} \leftarrow []$
    for $0$ to $n - 1$ do
        $t^\star, \beta_{t^\star} \leftarrow \texttt{computeTimestep}(\sigma)$
        $x_{t^\star} \leftarrow \sqrt{\beta_{t^\star}}(x + \mathcal{N}(0, \sigma^2\mathbf{I}))$
        $y \leftarrow g(\texttt{denoise}(x_{t^\star}; t^\star))$
        $\texttt{cnts}_y \leftarrow \texttt{cnts}_y + 1$
    return $\texttt{cnts}$

function computeTimestep($\sigma$)
    $t^\star \leftarrow$ find $t$ s.t. $\frac{1-\beta_t}{\beta_t} = \sigma^2$
    return $t^\star, \beta_{t^\star}$
```

3.2 CERTIFIED SEGMENTATION VIA RANDOMIZED SMOOTHING

Fischer et al. (2021) extended the work of Cohen et al. (2019) for segmentation tasks and presented the first approach for certified segmentation. To perform image segmentation, each pixel in an image is assigned a segmentation class. This can be seen as a type of classification task, but instead of predicting the content of the entire image, the goal is to predict the class of each individual pixel. In this setting, the target space corresponds to regions/categories to segment (*e.g.*, cars, roads, pedestrians, etc.), and the classifier $f : \mathbb{R}^d \rightarrow \mathcal{Y}^d$ outputs a class for each pixel and classifies each component individually. It is relatively straightforward to extend the certification algorithm proposed by Cohen et al. (2019) for the segmentation task. Nevertheless, Fischer et al. (2021) identified two primary challenges with the method. First, given that the certified radius of a particular region will be the minimum radius over the entire region, the algorithm may report an extremely low certified radius based on only a few *bad* pixels. Second, since Cohen et al.'s certification algorithm is applied to each pixel separately, and the certification is only valid with a probability of $1 - \alpha$, considering the entire region and applying the union bound could significantly reduce the overall confidence in the certificate. To address the first challenge and limit the impact on bad pixels on the overall result, Fischer et al. (2021) proposed a simple solution which consists in defining a threshold $\tau \in [\frac{1}{2}, 1]$ and instead of checking $\underline{p_y} > \frac{1}{2}$, they advise $\underline{p_y} > \tau$. To account for the multiple testing problem, *i.e.*, low confidence due to the union bound of the entire region, Fischer et al. (2021) introduce the `FwerControl` function used in Algorithm 1 which is based on the Holm-Bonferroni method (Holm, 1979), and performs multiple-testing correction. Conceptually, the idea is to control the type I error (rejecting the null hypothesis when it is actually true) while reducing type II errors (not rejecting the null hypothesis when it is false). Now that we have reviewed randomized smoothing for classification and segmentation tasks, we will present how it is possible to improve upon the current state-of-the-art with diffusion models.

4 CERTIFIED SEGMENTATION VIA DIFFUSION MODELS

To prevent a distribution shift when using randomized smoothing for inference, it is common practice to train networks with noise injection (Cohen et al., 2019). However, from an information theory perspective, randomized smoothing has inherent trade-offs and limitations. While adding noise during training can enhance the certified accuracy of models compared to those trained without noise, it may also lower the model's natural accuracy, as the variance of the noise decreases the information present in the input. These limitations have led to a series of no-go results for randomized smoothing (Blum et al., 2020; Hayes, 2020; Kumar et al., 2020; Yang et al., 2020; Mohapatra et al., 2021; Wu et al., 2021; Ettedgui et al., 2022), suggesting that achieving high certified accuracy may be challenging due to the significant variance that must be introduced in the input. Consequently, the destruction of information due to noise can result in information loss, potentially leading to a useless classifier.

To address this important limitation of randomized smoothing, Salman et al. (2020) have investigated denoising the input before giving it to the classifier. The idea is to use trained neural networks to *reconstruct* the removed information of the image due to the noise. This process has two main advantages: it mitigates the no-go results of randomized smoothing since the destroyed information is "reconstructed" by the denoiser and it does not involve training the classifier with noise mitigating the reduced natural accuracy of training with noise injection. Of course, in this new setting, the quality of the denoiser will matter. Salman et al. (2020) were able to boost the certified accuracy by up to 33% on the ImageNet dataset with respect to previous state-of-the-art defenses.

Table 1: Segmentation results of DenoiseCertify (ours) and SegCertify proposed by Fischer et al. (2021) on both Cityscapes and Pascal-Context datasets. Two network architectures were used for both pipelines, HRNet trained with noise and ViT trained without noise. We also report the performance of HRNet trained without noise. For each dataset, we used the same 100 images with $n_0 = 10, n = 100, \alpha = 0.001$ and $\tau = 0.75$. Results are certified at radius R, acc. being the mean per-pixel accuracy, mIoU the mean intersection over union and %⊘ the mean percentage of pixel abstentions on all images.

Model	**Architecture**	**Trained with noise**	σ	R	**Cityscapes**			**Pascal Context**		
					Acc	mIoU	%⊘	Acc.	mIoU	%⊘
Non-robust	HRNet	✗	0.00	0.00	0.97	0.81	0.00	0.77	0.42	0.00
SegCertify	HRNet	✓	0.00	0.00	0.91	0.57	0.00	0.53	0.18	0.00
		✓	0.25	0.17	0.88	0.59	0.11	0.55	0.22	0.22
		✓	0.50	0.34	0.34	0.06	0.40	0.17	0.03	0.41
		✓	1.00	0.67	0.06	0.00	0.31	0.08	0.00	0.13
DenoiseCertify **(ours)**	Diffusion +HRNet	✓	0.25	0.17	0.70	0.32	0.26	0.47	0.17	0.27
		✓	0.50	0.34	0.55	0.21	0.41	0.42	0.15	0.46
		✓	1.00	0.67	0.36	0.09	0.60	0.15	0.04	0.77
	Diffusion +ViT	✗	0.00	0.00	0.94	0.67	0.00	0.85	0.58	0.00
		✗	0.25	0.17	0.77	0.41	0.24	0.67	0.48	0.28
		✗	0.50	0.34	0.65	0.28	0.36	0.54	0.32	0.40
		✗	1.00	0.67	0.47	0.15	0.53	0.28	0.15	0.62

Carlini et al. (2023) go even further and propose to use state-of-the-art Diffusion Probabilistic Models (DPM) to perform the denoising step. With this approach, they were able to further improve the state of the art by up to 14% on the ImageNet dataset. Denoising diffusion probabilistic models, which have been introduced by Sohl-Dickstein et al. (2015) and further improved by Ho et al. (2020) and Nichol and Dhariwal (2021), are a class of generative models and have shown to beat Generative Adversarial Networks (GANs) (Goodfellow et al., 2020) on image synthesis. Conceptually, the training of these models consists in adding noise at each step of the diffusion process until purely random noise is reached. The reverse process then starts from random noise and generates a new image from the data distribution. Carlini et al. (2023) proposes a procedure to use these models for *denoising* instead of generating new images. The idea is to start the reverse process with a noisy image instead of Gaussian noise in order for the DPM to output an image from the initial data distribution that resembles the original image. As explained by Carlini et al. (2023), to use the DPM in the context of randomized smoothing, one needs to convert the noise added for randomized smoothing, *i.e.*, $x_{\text{rs}} = x + \tau$ with $\tau \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$ to the specific step in the diffusion process: $x_{\text{DPM}} = \sqrt{\beta_t} x + \tau(1 - \beta_t)$ where β_t denotes a constant from the timestamp t that controls the amount of noise added to the image during the diffusion process. For more details on how to compute the timestamp t, one can refer to Section 3 of Carlini et al. (2023). We provide in Algorithm 2 an updated version of the algorithm to compute the samples for the Predict & Certify function of Fischer et al. (2021).

Pipeline. Our pipeline starts by passing the image through a Denoising Diffusion Probabilistic Model (DDPM) and then calling a semantic segmentation model for prediction. For both components, we use an off-the-shelf model made publicly available. To denoise images, we use the class unconditional DDPM from Dhariwal and Nichol (2021). This 552M-parameter denoiser has been trained on ImageNet and performs very well on images from both Cityscapes and Pascal-Context. For segmentation, we use two model architectures with different training strategies. First, we test on High-resolution networks, HRNet from Wang et al. (2020), trained in two different ways. The *non-robust* HRNet has been trained with natural images and the *base model* is an HRNet trained with a Gaussian noise of $\sigma = 0.25$. The second architecture we use is the Vision Transformer Adapter for Dense Predictions, ViT from Chen et al. (2023), that was trained only on natural images. We use the 568M-parameter model trained on Pascal-Context and the 571M-parameter model trained on Cityscapes. Both models were reported to provide state-of-the-art accuracy and mean intersection over union (mIoU) on the task of semantic segmentation. Our code is provided at: `https://github.com/othmanela/certified_segmentation`

5 EXPERIMENTS

We evaluate our method on a set of experiments with multiple approaches. First, we compare our technique with SegCertify, the state-of-the-art introduced by Fischer et al. (2021). Then, we set new state-of-the-art results using off-the-shelf models. We name our method DenoiseCer-

Table 2: Performance of SegCertify and DenoiseCertify (ours) on an off-the-shelf HRNet model trained without Gaussian noise. Scale corresponds to the image sizing scale used as input to the segmentation model (*e.g.*, a scale of 0.5 on cityscapes will resize the images to 512×1024). Accuracy, mean intersection over union (mIoU) and percentage of abstentions (%$\oslash$) are certified given a noise level σ and radius R. All results are provided with Holm correction.

Scale	**Model**	σ	R	**Cityscapes**			**Pascal Context**		
				Acc.	mIoU	%$\oslash$	Acc.	mIoU	%$\oslash$
0.25	SegCertify	0.25	0.17	0.34	0.05	0.29	0.18	0.05	0.67
		0.50	0.34	0.18	0.01	0.14	0.07	0.01	0.70
		1.00	0.67	0.18	0.01	0.06	0.02	0.00	0.58
	DenoiseCertify	0.25	0.17	0.78	0.41	0.22	0.45	0.19	0.28
		0.50	0.34	0.68	0.29	0.32	0.37	0.11	0.45
		1.00	0.67	0.46	0.15	0.54	0.11	0.03	0.74
0.50	SegCertify	0.25	0.17	0.48	0.07	0.19	0.34	0.13	0.49
		0.50	0.34	0.19	0.01	0.15	0.12	0.03	0.62
		1.00	0.67	0.17	0.01	0.11	0.03	0.00	0.43
	DenoiseCertify	0.25	0.17	0.74	0.37	0.26	0.56	0.23	0.27
		0.50	0.34	0.60	0.22	0.40	0.43	0.18	0.46
		1.00	0.67	0.32	0.11	0.67	0.12	0.04	0.64
1.00	SegCertify	0.25	0.17	0.18	0.01	0.08	0.31	0.10	0.52
		0.50	0.34	0.17	0.01	0.08	0.08	0.01	0.45
		1.00	0.67	0.01	0.00	0.98	0.01	0.02	0.49
	DenoiseCertify	0.25	0.17	0.58	0.26	0.42	0.47	0.22	0.27
		0.50	0.34	0.42	0.15	0.55	0.36	0.10	0.53
		1.00	0.67	0.24	0.70	0.70	0.10	0.02	0.73

tify.

Datasets. All of our experiments are performed on the task of semantic image segmentation on Pascal-Context and Cityscapes datasets, two very common datasets for this task. Pascal-Context (Mottaghi et al., 2014) consists of an extension of the Pascal-VOC (Everingham et al., 2015) dataset with all of the image pixels annotated. There are 60 classes (59 foreground and 1 background). Typical evaluation strategies use either all of the 60 classes or the 59 foreground classes only. We evaluate here on the 59 foreground classes in order to have a fair comparison with SegCertify. The Cityscapes dataset (Cordts et al., 2016) contains high resolution 1024×2048 images of diverse street scenes from 50 different cities. The images are annotated in 30 classes but only 19 of them are used for evaluation. Similar to SegCertify, we evaluate our method on the same 100 images set from both datasets. We use every 5th image on the Cityscapes dataset and every 51st on Pascal.

DenoiseCertify on models trained with noise. We start first by comparing DenoiseCertify with SegCertify. The state-of-the-art certification results proposed by the latter were obtained with an HRNet trained with a Gaussian noise of $\sigma = 0.25$. A comparison of both methods is provided in the first two sections of Table 1. We notice that DenoiseCertify outperforms SegCertify for all sigmas except for 0.25. In fact, for $\sigma = 0.5$ the accuracy jumps from 0.34 to 0.55 and the mIoU from 0.06 to 0.21 which corresponds to an increase of 61% and 250% respectively. This gives us an idea of the power of denoising diffusion models when used to certify segmentation models. Since our pipeline contains an added denoising step, we note an increase in the reported runtime in seconds. On the largest images of the dataset (1024×2048), the runtime increases from 92.69 to 131.42 seconds with HRNet, which per image is minor. We did not perform any optimization on the code to make our pipeline faster. With more engineering, the runtime can be optimized further. Also, we believe that the gain in performance easily outweighs the increase in runtime. For $\sigma = 1.0$, it appears from Table 1 that SegCertify has a lower number of abstentions than DenoiseCertify. However, looking at the segmentations it looks like SegCertify predicts a large number of pixels in the image with the wrong class. An example is provided in the last row of Figure 1. For $\sigma = 0.25$, SegCertify outperforms our technique but this may be due to two main reasons. First, the model we are using was trained with a Gaussian noise of 0.25. Thus, it is performing best when provided with images with the same level of noise. In the next section, we show that given the right model, DenoiseCertify outperforms SegCertify. Second, one of the limitations of using

Table 3: Comparison of two denoising strategies on the DENOISECERTIFY pipeline. All of the reported results use a Vision Transformer (ViT) segmentation model with a scale of 1. For the denoising diffusion model, we use the same timestep t^* found with the COMPUTETIMESTEP function presented in Algorithm 2.

Denoise Method	σ	R	**Cityscapes**			**Pascal Context**		
			Acc.	mIoU	%⊘	Acc.	mIoU	%⊘
Denoise single step	0.00	0.00	0.94	0.67	0.00	0.85	0.58	0.00
	0.25	0.17	0.77	0.41	0.24	0.67	0.48	0.28
	0.50	0.34	0.65	0.28	0.36	0.54	0.32	0.40
	1.00	0.67	0.47	0.15	0.53	0.28	0.15	0.62
Denoise multi-step	0.00	0.00	0.89	0.39	0.00	0.80	0.49	0.00
	0.25	0.17	0.70	0.27	0.33	0.61	0.34	0.34
	0.50	0.34	0.52	0.15	0.50	0.46	0.24	0.49
	1.00	0.67	0.29	0.06	0.73	0.14	0.07	0.75

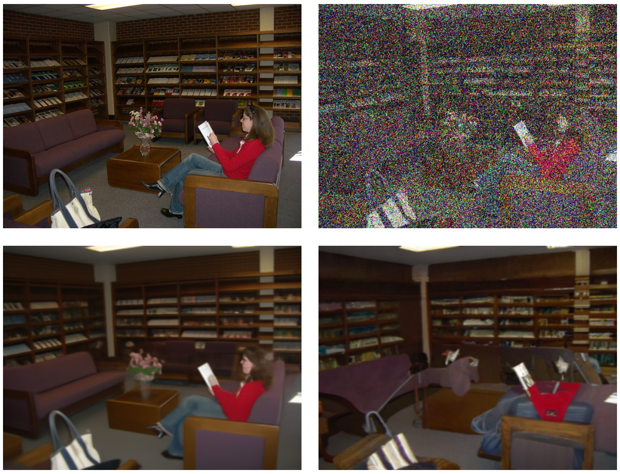


Figure 2: Qualitative results of the performance of a denoising diffusion model on Pascal Context images. Top row from left to right: ground truth and ground truth with a noise of $\sigma = 1.0$. Bottom row: single-step denoised image and multistep denoised image.

the off-the-shelf denoiser provided by Dhariwal and Nichol (2021) is rescaling the images to 256×256. Therefore scaling them back to their original size may decrease the image quality, especially when using high-resolution images like Cityscapes. We perform experiments with multiple scales and report them in the subsequent section.

DENOISECERTIFY on non-robust models. Here we use an HRNet model that was trained on natural images only without any introduction of Gaussian noise. We compare our performance with SEGCERTIFY and report our results in Table 2. Focusing on a scale of 1, we notice that SEGCERTIFY achieves a poor performance. In fact, as σ increases to values > 0.25 the mIoU and accuracy end up becoming 0. This is an expected result since models trained on natural images are very sensitive to Gaussian noise. However, those types of models perfectly suit our methodology, as we denoise images, we are able to use off-the-shelf segmentation models and achieve a much better prediction. This introduces a paradigm shift as we no longer require training robust deep learning models that need highly engineered strategies and that also degrade the natural accuracy significantly. As reported in Table 1, when comparing the first two rows, the non-robust HRNet accuracy drops from 0.97 to 0.91 and the decrease is more significant for the mIoU, going from 0.81 to 0.57. Therefore, our technique allows us to limit the drop in performance that traditional models used to suffer from while keeping strong certification guarantees.

Impact of image scales on performance. One of the limitations of using off-the-shelf models is having to comply with their restrictions. The unconditional DDPM we are using only takes as input images of size 256×256. We thus have to downscale the input images and upscale them back to their original size for prediction. As stated above, this is the main limitation of our method. But, since semantic segmentation models can be invoked with multiple scales, we can use them to predict at a given scale and then upsample the output probabilities back to the original size of the image. This also has the advantage of providing faster predictions. As an example, at a scale of 0.5 for Cityscapes, we downsample the images to 256×256 in order to call the DDPM, the denoised image is then reshaped to 512×1024 and serves as input to the segmentation model. The output probabilities of the segmentation model are then upsampled to their original size (1024×2048) to be compared with the ground truth. We always perform the certification on the original size of the image in order to follow the same strategy as SEGCERTIFY and perform a fair comparison. The performance of both methods with multiple scales is reported in Table 2. Examples of denoised and upscaled images are provided in Figure 3. For Cityscapes, we notice that the smaller the scale, the better the performance. In fact, the accuracy jumps from 0.58 to 0.74, and 0.78 for

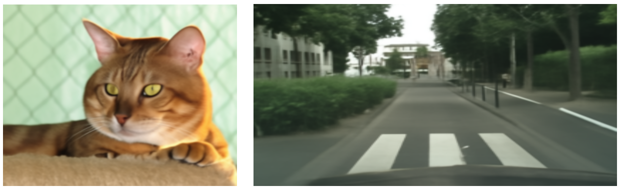


Figure 3: Examples of denoised and upscaled images from the Denoising Diffusion Model. On the left, a Pascal-Context image denoised on size 256×256 and upscaled to 373×480. On the right, a Cityscapes image denoised on 256×256 and upscaled to 1024×2048.

scale values of 1, 0.5, and 0.25 respectively. DENOISECERTIFY performs best for a scale of 0.25, corresponding to a Cityscapes image size of 256×512 which is very close to the output of the DDPM. Therefore, the rescaling does not impact the details and overall quality of the denoised image. The same happens for the Pascal-Context dataset, the best performance is obtained for a scale of 0.5 which corresponds to images of size 240×240, again very close to the 256×256 DDPM output. Training DDPMs on images of higher resolution would be another way to circumvent this limitation. Also, with the improvement of powerful techniques that rely on the denoising backbone, our approach would still be able to leverage the resources made available. We believe that having a denoiser that is also able to upscale images to very high resolutions would allow us to improve our results even further.

DENOISECERTIFY on state-of-the-art segmentation models. So far we have discussed how DENOISECERTIFY performs on both a robust and non-robust HRNet model. We have empirically shown that it achieves the best results on standard deep learning-based segmentation models. Going a step further, we can leverage the power of Vision Transformers which have been reported to be more robust to attacks (Mao et al., 2022), but also give state-of-the-art results on semantic segmentation tasks (Chen et al., 2023). In this section, we use the ViT Adapter trained with natural images and report the results in the last section of Table 1. When comparing with the results of Table 2 on the same scale of 1, we notice that the ViT model provides a considerable increase. For the lowest $\sigma = 0.25$, the accuracy and mIoU are respectively boosted to 0.77 and 0.41 compared to 0.58 and 0.26 previously. This empirically proves the points of Mao et al. (2022) and would even encourage us to make the assumption that we would be able to obtain higher certification results with stronger transformer models. Overall, DENOISECERTIFY combined with a ViT achieves state-of-the-art semantic segmentation certification results for Pascal-Context and Cityscapes.

Generalization of Denoising Diffusion Models. As stated above, we use an off-the-shelf denoising diffusion model that was trained on ImageNet. We set the number of channels to 256 and apply a linear scheduler with 1000 steps. Qualitative results are provided in Figure 2. We clearly notice that the DDPM is able the denoise the image with the highest level of noise ($\sigma = 1.0$) while keeping all of the information of the picture. Therefore, it is important to state that diffusion models generalize well to datasets they were not trained on. Pascal-Context and Cityscapes are the first two examples. Future work will involve testing DDPMs on images from other distributions (*e.g.*, the medical domain).

Multistep Denoising Diffusion Model. Denoising diffusion probabilistic models have been introduced as a class of generative models that beat GANs on image synthesis (Dhariwal and Nichol, 2021). Starting with Gaussian noise, each step of the DDPM consists in denoising an input image at timestep t to a marginally less noisy image at timestep $t-1$. The complete diffusion process is an iterative procedure starting from t^* until $t = 0$. Programmatically, each call to the denoiser d at timestep t performs two actions; it predicts the completely denoised image and returns the average between the estimated denoised image and the noisy image of timestep $t - 1$. We conduct experiments on the two possible denoising strategies. The top section of Table 3 reports the results of a single-step denoised image prediction from the class unconditional DDPM. The bottom section of Table 3 on the other hand reports results of a multiple-step denoising strategy going from t^* until $t = 0$ iteratively on the same class unconditional DDPM. Both use the ViT as the segmentation model. From the presented results, it is clear that the single-step denoiser performs better than the multi-step one in terms of accuracy, mIoU, percentage of abstentions, and runtime. This shows that denoising the image in a single shot is better than repeatedly denoising it multiple times. Intuitively, since the DDPMs are generative models at heart, they will tend to behave as such when denoising an image multiple times. Therefore, the output image at $t = 0$ may have lost a lot of its original information or may even end up from a different distribution. Qualitative results from Figure 2 support this claim as we can clearly notice that elements of the image were removed in the multi-step approach (the flower pot, as well as the reader disappeared, and the shape of the furniture changed). Another advantage of single-step denoising is the runtime efficiency. Instead of having to call the denoiser multiple times passing the outputted image at each timestep t, the denoiser is only called once (*e.g.*, For $\sigma = 1.0$ a denoiser with linear scheduling will be called 258 times compared to a single time with the first scheme). This represents a nonnegligible advantage of the single-shot denoising since we are using multiple calls to the denoiser for each image in order to obtain the certificate. We deduce that denoising diffusion models are powerful but should be used accordingly. In our case, we would like to leverage the denoising properties of the DDPM more than their generative properties. Thus, a single-step denoising strategy should be adopted.

6 CONCLUSION

We present the first work on certified semantic segmentation that leverages denoising diffusion probabilistic models and vision transformers. We conduct a comprehensive set of experiments on Pascal-Context (Mottaghi et al., 2014) and Cityscapes (Cordts et al., 2016) datasets and show that our method achieves state-of-the-art results on certified robustness for semantic segmentation tasks. We were able to achieve significant improvements in accuracy and mIoU using off-the-shelf models that are not trained or fine-tuned for robustness. This work provides a new direction for certified image segmentation with *off-the-shelf* models. However, an interesting direction would be to explore task-specific training. For instance, in the context of certified segmentation, Salman et al. (2019) improved upon the work of Cohen et al. (2019) by training classifiers with noise injection and adversarial training. It would be straightforward to extend this approach to certified segmentation with our DENOISECERTIFY procedure by adversarially training a classifier or a diffusion model. Although computationally expensive, this method may lead to further improvements. Moreover, we have seen that the diffusion model is able to generalize to Pascal-Context and Cityscapes datasets. A promising future direction would be to investigate the generalization of this model for denoising medical images and provide certified segmentation for critical healthcare applications.

Acknowledgements

This work was granted access to the HPC resources of IDRIS under the allocation 2023-AD011013308R1 made by GENCI.

References

Alexandre Araujo, Laurent Meunier, Rafael Pinot, and Benjamin Negrevergne. Advocating for multiple defense strategies against adversarial examples. In *ECML PKDD 2020 Workshops 2020*, pages 165–177. Springer, 2020.

Alexandre Araujo, Benjamin Negrevergne, Yann Chevaleyre, and Jamal Atif. On lipschitz regularization of convolutional layers using toeplitz matrix theory. *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6661–6669, 2021.

Alexandre Araujo, Aaron J Havens, Blaise Delattre, Alexandre Allauzen, and Bin Hu. A unified algebraic perspective on lipschitz neural networks. In *The Eleventh International Conference on Learning Representations*, 2023.

Anurag Arnab, Ondrej Miksik, and Philip HS Torr. On the robustness of semantic segmentation models to adversarial attacks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 888–897, 2018.

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International conference on machine learning*, pages 274–283. PMLR, 2018.

Avrim Blum, Travis Dick, Naren Manoj, and Hongyang Zhang. Random smoothing might be unable to certify ℓ_∞ robustness for high-dimensional images. *The Journal of Machine Learning Research*, 21(1):8726–8746, 2020.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. Ieee, 2017.

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*, 2023.

Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision transformer adapter for dense predictions. 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*, pages 1310–1320. PMLR, 2019.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

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

Francesco Croce and Matthias Hein. Mind the box: l_1-apgd for sparse adversarial attacks on image classifiers. In *International Conference on Machine Learning*, pages 2201–2211. PMLR, 2021.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

Raphael Ettedgui, Alexandre Araujo, Rafael Pinot, Yann Chevaleyre, and Jamal Atif. Towards evading the limits of randomized smoothing: A theoretical analysis. *arXiv preprint arXiv:2206.01715*, 2022.

M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111, 2015.

Farzan Farnia, Jesse Zhang, and David Tse. Generalizable adversarial training via spectral normalization. In *International Conference on Learning Representations*, 2019.

Marc Fischer, Maximilian Baader, and Martin Vechev. Scalable certified segmentation via randomized smoothing. In *International Conference on Machine Learning*, pages 3340–3351. PMLR, 2021.

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

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

Sven Gowal, Krishnamurthy Dvijotham, Robert Stanforth, Rudy Bunel, Chongli Qin, Jonathan Uesato, Relja Arandjelovic, Timothy Mann, and Pushmeet Kohli. On the effectiveness of interval bound propagation for training verifiably robust models. *arXiv preprint arXiv:1810.12715*, 2018.

Jamie Hayes. Extensions and limitations of randomized smoothing for robustness guarantees. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 786–787, 2020.

Xiang He, Sibei Yang, Guanbin Li, Haofeng Li, Huiyou Chang, and Yizhou Yu. Non-local context encoder: Robust biomedical image segmentation against adversarial attacks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019.

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

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70, 1979.

Yujia Huang, Huan Zhang, Yuanyuan Shi, J Zico Kolter, and Anima Anandkumar. Training certifiably robust neural networks with efficient local lipschitz bounds. *Advances in Neural Information Processing Systems*, 34:22745–22757, 2021.

Xu Kang, Bin Song, Xiaojiang Du, and Mohsen Guizani. Adversarial attacks for image segmentation on multiple lightweight models. *IEEE Access*, 8:31359–31370, 2020.

Aounon Kumar and Tom Goldstein. Center smoothing: Certified robustness for networks with structured outputs. *Advances in Neural Information Processing Systems*, 34: 5560–5575, 2021.

Aounon Kumar, Alexander Levine, Tom Goldstein, and Soheil Feizi. Curse of dimensionality on randomized smoothing for certifiable robustness. In *International Conference on Machine Learning*, pages 5458–5467. PMLR, 2020.

Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *Artificial intelligence safety and security*, pages 99–112. Chapman and Hall/CRC, 2018.

Mathias Lecuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. Certified robustness to adversarial examples with differential privacy. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 656–672. IEEE, 2019.

Bai Li, Changyou Chen, Wenlin Wang, and Lawrence Carin. Certified adversarial robustness with additive noise. *Advances in neural information processing systems*, 32, 2019a.

Qiyang Li, Saminul Haque, Cem Anil, James Lucas, Roger B Grosse, and Jörn-Henrik Jacobsen. Preventing gradient attenuation in lipschitz constrained convolutional networks. *Advances in neural information processing systems*, 32, 2019b.

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

Xiaofeng Mao, Gege Qi, Yuefeng Chen, Xiaodan Li, Ranjie Duan, Shaokai Ye, Yuan He, and Hui Xue. Towards robust vision transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12042–12051, June 2022.

Laurent Meunier, Blaise J Delattre, Alexandre Araujo, and Alexandre Allauzen. A dynamical system perspective for lipschitz neural networks. In *International Conference on Machine Learning*, pages 15484–15500. PMLR, 2022.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018.

Jeet Mohapatra, Ching-Yun Ko, Tsui-Wei Weng, Pin-Yu Chen, Sijia Liu, and Luca Daniel. Higher-order certification for randomized smoothing. *Advances in Neural Information Processing Systems*, 33:4501–4511, 2020.

Jeet Mohapatra, Ching-Yun Ko, Lily Weng, Pin-Yu Chen, Sijia Liu, and Luca Daniel. Hidden cost of randomized smoothing. In *International Conference on Artificial Intelligence and Statistics*, pages 4033–4041. PMLR, 2021.

Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 891–898, 2014.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021.

Rafael Pinot, Laurent Meunier, Alexandre Araujo, Hisashi Kashima, Florian Yger, Cédric Gouy-Pailler, and Jamal Atif. Theoretical evidence for adversarial robustness through randomization. *Advances in neural information processing systems*, 32, 2019.

Bernd Prach and Christoph H Lampert. Almost-orthogonal layers for efficient general-purpose lipschitz networks. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXI*, pages 350–365. Springer, 2022.

Hadi Salman, Jerry Li, Ilya Razenshteyn, Pengchuan Zhang, Huan Zhang, Sebastien Bubeck, and Greg Yang. Provably robust deep learning via adversarially trained smoothed classifiers. *Advances in Neural Information Processing Systems*, 32, 2019.

Hadi Salman, Mingjie Sun, Greg Yang, Ashish Kapoor, and J Zico Kolter. Denoised smoothing: A provable defense for pretrained classifiers. *Advances in Neural Information Processing Systems*, 33:21945–21957, 2020.

Sahil Singla and Soheil Feizi. Skew orthogonal convolutions. In *International Conference on Machine Learning*, pages 9756–9766. PMLR, 2021.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Asher Trockman and J Zico Kolter. Orthogonalizing convolutional layers with the cayley transform. In *International Conference on Learning Representations*, 2021.

Yusuke Tsuzuku, Issei Sato, and Masashi Sugiyama. Lipschitz-margin training: Scalable certification of perturbation invariance for deep neural networks. *Advances in neural information processing systems*, 31, 2018.

Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3349–3364, 2020.

Yihan Wu, Aleksandar Bojchevski, Aleksei Kuvshinov, and Stephan Günnemann. Completing the picture: Randomized smoothing suffers from the curse of dimensionality for a large family of distributions. In *International Conference on Artificial Intelligence and Statistics*, pages 3763–3771. PMLR, 2021.

Chong Xiang, Charles R Qi, and Bo Li. Generating 3d adversarial point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9136–9144, 2019.

Cihang Xie, Jianyu Wang, Zhishuai Zhang, Yuyin Zhou, Lingxi Xie, and Alan Yuille. Adversarial examples for semantic segmentation and object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 1369–1378, 2017.

Xiaojun Xu, Linyi Li, and Bo Li. LOT: Layer-wise orthogonal training on improving l2 certified robustness. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

Greg Yang, Tony Duan, J Edward Hu, Hadi Salman, Ilya Razenshteyn, and Jerry Li. Randomized smoothing of all shapes and sizes. In *International Conference on Machine Learning*, pages 10693–10705. PMLR, 2020.

Maksym Yatsura, Kaspar Sakmann, N. Grace Hua, Matthias Hein, and Jan Hendrik Metzen. Certified defences against adversarial patch attacks on semantic segmentation. In *The Eleventh International Conference on Learning Representations*, 2023.

Tan Yu, Jun Li, Yunfeng Cai, and Ping Li. Constructing orthogonal convolutions in an explicit manner. In *International Conference on Learning Representations*, 2022.