Towards Robust Certified Defense via Improved Randomized Smoothing

Satyadwyoom Kumar
Dept. Electronics and Communication
Netaji Subhas Institute of Technology
New-Delhi, India
satyadwyoom.ec18@nsut.ac.in

Apurva Narayan
Dept. Computer Science
The University of British Columbia
Kelowna, BC, Canada
apurva.narayan@ubc.ca

2022 International Joint Conference on Neural Networks (IJCNN) | 978-1-7281-8671-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/IJCNN55064.2022.9892471

***Abstract*—Deep learning models change their prediction through a carefully optimized imperceptible change in the input termed as an adversarial perturbation. Researchers have been focusing on developing methods to counter such effects. Recently, randomized smoothing a highly scalable technique to develop a certified classifier was introduced. However, this technique involves training a neural network from scratch.**

In this paper we present an empirical insight into the technique of randomized smoothing and propose a framework for a generalizable defence that works on a novel way of adding gaussian noise to the randomized smoothing procedure and is applicable to black-box pre-trained classifiers. We perform extensive experimentation across a variety of classification models and multiple datasets such as Cifar10 and ImageNet. Our framework is model-agnostic and out-performs the recent state-of-the-art certified defence methods by a large margin without the requirement of any intensive training, thus achieving high certified as well as unperturbed performance.

***Index Terms*—Certified adversarial defence, Randomized smoothing**

I. Introduction

Deep learning methods are widely applied across many domains. Hence, there is an increasing concern related to the security and reliability of such methods under adversarial attacks [1]–[4]. Any attack on the model by an adversary not only affects the performance but also poses a serious concern related to the leakage of private data prompting researchers to identify these threats/issues and propose methodologies to counter and mitigate the effects of such threats. These concerns have led to the development of many adversarial defences broadly classified into empirical defence methods [5] and certified defence methods [6], [7]. Out of these methods certified defences provide a formal guarantee on model resilience to such attacks.

Earlier methods for adversarial defence [5]–[7] involve training the classification model against adversarial perturbations or random noise, expecting that the retrained version of the classifier has a higher robustness to adversarial attacks. Recently, Denoised Smoothing [8] utilized image denoisers to produce a model-agnostic certifiable defence that denoises the gaussian noise augmented input images and applies these denoised images directly to the underlying pretrained classifier.

Motivated by the work of [8], we propose a novel model-agnostic defence framework to produce certified classifiers out of pre-trained classifiers without requiring any prior knowledge about them during testing phase. Moreover, our framework does not require multiple surrogate (uses only a single surrogate) models during the training phase, opposite to the work of denoised-smoothing [8]. We also present an empirical study of the randomized smoothing procedure used by [7]. Our contributions can be summarised as follows:

- We argue how the traditional input noise augmentation in randomized smoothing reduces the important features required by the model to predict the correct label. Following this, we propose a novel strategy of adding noise in the randomized smoothing.
- We also propose an improved objective function to train our framework of certified transformation-based adversarial defence that results in high transferability across multiple pre-trained classification models without having any prior knowledge about them.
- We perform extensive experiments on our method using both low-dimensional (Cifar10) and high-dimensional (ImageNet) datasets.
- Our strategy provides a robust accuracy of 43% and standard accuracy of 81%, thereby achieving an increment of 30% on robust accuracy and an increment of 60% on standard accuracy of the previous state-of-the-art for certified transformation based defence on a black-box pre-trained classifiers for a l_2 perturbation of radius $\frac{255}{255}$.

II. Related Work

Certified defence methods focus on providing a certification bound in terms of perturbation radius or probability against any l_p norm-based attacks. We discuss a variety of such methods starting with [9] that present a differentiable robustness certification using an upper bound on the worst-case loss of a neural network using semidefinite programming. Their work focuses on l_∞ norm-based attacks. Their certificate of robustness is differentiable, one could simultaneously train a robust network by adding it as a regularisation term in the objective function.

Authors in [10] propose another approach that generates certified robust models against all norm-based attacks. The basic idea behind their approach is to minimize the worst-case loss over a region defined by the convex outer approximation

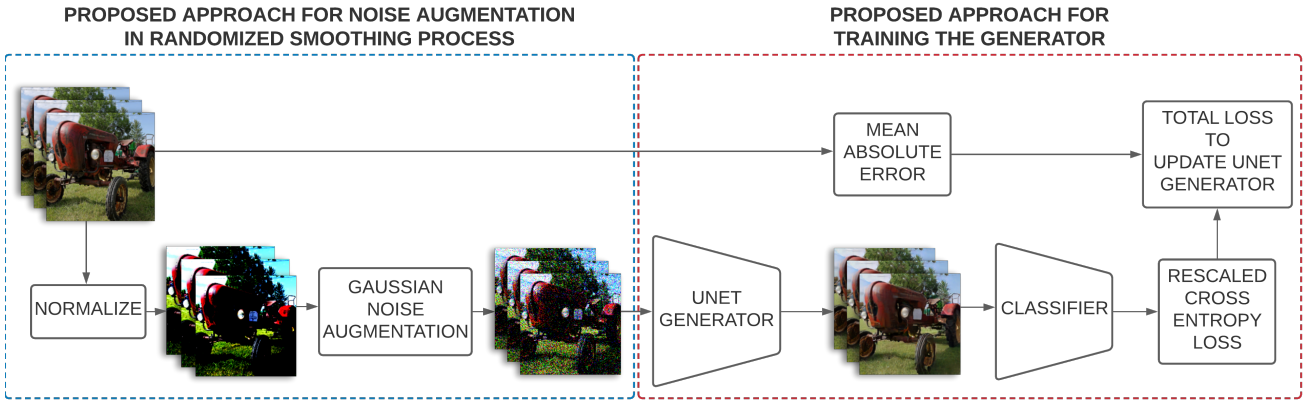


Fig. 1: IRS-Net: Proposed training framework. During test-time the input-test-image is passed to normalization module which normalizes the image using some pre-determined mean and standard deviation value. Then n-samples of gaussian noise having dimensions similar to input image are sampled and are added to the normalized images. These noise augmented normalized image are passed to a UNet denoiser that is learnt to remove gaussian noise from the augmented images and transforms it to pixel range expected by the classifier. Then a classifier uses these transformed images and predicts n-labels corresponding to each augmented n-normalized image. Then majority voting is used to determine the final class corresponding to the true input image.

of activation function values over norm bounded input perturbations. The only difference between [9] and [10] work is that the authors of [10] use a linear programming formulation to compute worst-case loss. However, both the methods are only scalable to networks containing few hidden layers used for small datasets like MNIST.

Authors in [6] took inspiration from the differential privacy literature and propose the technique of randomized smoothing that solved the problem of scalability faced by past works on certified defences [9], [10]. They propose to protect the privacy of each pixel using random noises that are applied directly at the image level or after hidden layers of a network during training and testing. They also propose an approach where an encoder-decoder model is trained to denoise the noise augmented images followed by re-training the underlying classification model to robustify both models, reducing the time requirement of the previously proposed random noise re-training involving only the classifier. The author uses gaussian noise for l_2 attacks and Laplacian noise for l_1 attacks.

Authors in [7] found that the methodology of [6] results in loose bounds for l_2 attacks on certification radius and proposed a new statistical formulation that provided tighter certification radius based on Neyman-Pearson lemma.

Authors in [8] presented the idea to robustify pre-trained classifier by using an off-the-shelf denoiser [11], [12] shifting from the traditional approach of training the classifiers [7]. Their [8] method was similar to what [6] used but with a different training strategy wherein the pre-trained classifier was not tweaked again using the robustified denoiser. The author also proposed multiple objective functions to train the denoisers. Out of all the objective functions, the certified performance in a white-box classifier was maximum with the stability objective (which involves the cross-entropy loss between logits value of the denoised-image and the predicted label of the ground-truth input image). However, all these objectives attain similar performance on a pre-trained black-box classifier which was low when compared to [7]. They also present the practical application of defending public vision APIs. To find the certificates of robustness, they used the formulation of [7].

[13]–[16] are some other recent works that try to scale randomized-smoothing using different proposed random noises to create provably robust classifier against different l_p norm-based attacks.

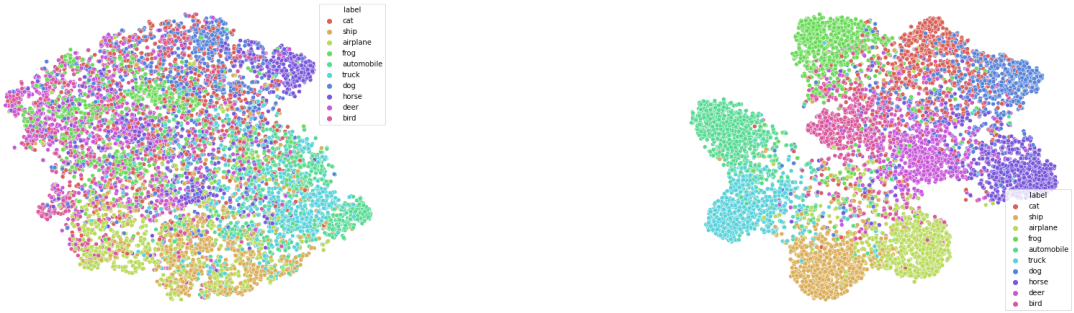
III. Technical Background

A. Randomized Smoothing

Randomized Smoothing is the procedure to develop a robust classifier from a standard classifier. The procedure takes into account a classification problem $\mathbf{R}^d \rightarrow y$, i.e., mapping d-dimensional inputs to true classes y. The idea is to create a smooth classifier g from a base classifier f, such that label given by the smoothed classifier g to an input x is equal to class c returned by the base classifier f to the random isotropic gaussian corruption of the input x, i.e.,

$$g(x) = argmax_{c \in y} \mathcal{P}(f(x+\delta) = c) \\ where, \delta \in \mathcal{N}(0, \sigma^2 I) \tag{1}$$

Here σ is the parameter that controls the robustness/accuracy trade-off: A large value of σ would mean that a classifier would not change the predicted label on a particular input for a large value of perturbation radius but, this results in a lower standard-accuracy. The author in [7] came up with a new statistical formulation to provide tighter bounds on the certified radius around an input x based on Neyman-Pearson



(a) (Image + Noise + Classifier)

(b) Our (Image + Normalization + Noise + Classifier)

Fig. 2: 2D t-SNE plots generated using the logit vectors obtained from a ResNet110 model using Cohen et al. (Fig a) and our proposed methodology for gaussian noise augmentation (Fig b) on the Cifar10 dataset and gaussian noise standard deviation(σ) of 1.00. This figure depicts how the decision boundary changes when a very high level of noise is added to input for different methods.

Lemma for l_2 norm, where p_a and p_b are the probability of predicted class c_a and runner class c_b and Φ^{-1} is the inverse gaussian CDF.

$$R = \frac{\sigma}{2}(\Phi^{-1}(p_a) - \Phi^{-1}(p_b)) \tag{2}$$

Since it is very hard to compute p_a and p_b if f is a neural network, authors in [7] proposed to use Monte Carlo sampling to replace p_a with a lower bound $\underline{p_a}$ and p_b with an upper bound $\overline{p_b}$ in Equation 2 to compute the certified radius R.

IV. Methodology

A. Overview

Our proposed framework works on the principle of randomized-smoothing [6] and use the statistical formulation proposed by [7] explained in section "Background" for calculating the certified radius. According to the definition provided by [7] random noise is added to the input used by f (a neural network) for inference purposes. However, in the experiments performed by the authors of [6]–[8] random noise is being added to the image $x \in (0, 1)$ which is then passed on to the classifier.

Since most modern deep learning classification pipelines normalize the input-image x using a mean and a standard deviation calculated from the input data before using them as the input. So, from a practical perspective, the classifier is using the normalized version of the image x as input and not the actual image $x \in$ (0,1). Therefore, adding noise when the image x is in the (0,1) domain may incur some loss in the informative features present in the normalized variant of input-image which is the cause of decreased standard accuracy of the classifier on noise-augmented inputs for [7] and [8].

This phenomenon of loss of informative features is being depicted using Figure 2 containing the 2D t-SNE [17] plot generated using logit-vectors obtained from the underlying classifier for the traditional way of adding noise in randomized smoothing procedure (image + noise + classifier) and our proposed way of adding noise to the randomized smoothing procedure (image + normalization + noise + classifier). This t-SNE plot maps the predicted logit-values of a classifier to a two-dimensional space to depict the learnt decision boundary on the input data. To generate these t-SNE values we train a ResNet110 model from scratch with different ways of gaussian noise augmentation. When the gaussian noise with a high standard-deviation value is added to the input image $x \in$ (0,1) of a classifier, the decision boundary worsens as can be seen in Figure 2(a). To counter this loss of informative features, we in our method add gaussian noise to the normalized input data as depicted in Figure 1, this change in the process of randomized smoothing helps the classifier improve upon its ability to correctly differentiate between inputs and assign them with a correct output-class even when a high value of gaussian noise standard deviation is added to the normalized input, this is visualized in Figure 2(b).

Since when a classifier is under an adversarial attack through an adversarial perturbation, each segmental input in the classifier internal structure also becomes adversarial. Hence, certifying any segmental input helps certify the downstream classifier components and classifier predictions. Moreover, since we only aim to remove the gaussian noise from the noise-augmented normalized image and not the adversarial perturbation that may exist in image $x \in$ (0,1).

Therefore, this use of normalized image instead of image $x \in$ (0,1) should not affect the theoretical assumptions taken by [7]. Hence, we convert the problem of creating a certified classifier through randomized-smoothing [7], to using to a problem of image-reconstruction, transforming noisy-normalized-images to a domain accepted by the underlying classifier while simultaneously converting an uncertified underlying pre-trained classification model without noisy-training [7] to a certified one. Our framework is different from the one proposed by [8] because we do not directly use the image $x \in$ (0,1) or use multiple surrogates for training in a black-box setting.

Since our motive is to defend a pre-trained classification model to which our defence mechanism has no prior knowl-

(a) Salman et al.(MSE+Stability)

(b) Our Method

Fig. 3: ImageNet dataset visualisations resulted from our methodology and Salman et al. (MSE+Stability) for a gaussian standard deviation of 1.00

TABLE I: Transferability in terms of standard accuracy across multiple model architectures when our and [8] methodology is transferred from ResNet18 to the listed models on Cifar10 dataset.

Models	Noise standard deviation(s)					
	0.25		0.50		1.00	
	Ours	Salman et al. [8]	Ours	Salman et al. [8]	Ours	Salman et al. [8]
VGG16	**90.27%**	49.24%	**84.87%**	25.10%	**72.96%**	24.82%
VGG19	**90.04%**	46.44%	**84.12%**	22.25%	**72.75%**	22.54%
MobileNetV2	**90.32%**	32.29%	**80.57%**	12.35%	**68.43%**	17.28%
GoogLeNet	**87.08%**	32.10%	**82.79%**	10.54%	**69.15%**	10.08%

edge, we use the UNet architecture [18] which is a simple and easy to train convolution neural network-based encoder-decoder model. We choose UNet architecture because we want to denoise the gaussian noise augmented normalized version of image $x \in (0,1)$ and transform the denoised normalized image to a pixel-range that is accepted by the underlying classifier. Therefore using UNet is an appropriate need of the framework, moreover as seen by [19] UNet, DnCNN, MemNet all have a similar denoising performance.

We also design a novel objective function which we call rescaled-reconstruction-loss (RRL) based on the reconstruction loss proposed by [8]. This proposed function is used to train the UNet network. This function takes into account the underlying classification task (cross-entropy loss) and the similarity calculated using the mean absolute error between original-ground truth required by the classifier and output of the encoder-decoder network, i.e., the ground truth input and the transformed image. The MAE loss ground truth input can be changed based upon what the underlying classifier expects as input which could be $x \in (0,1)$ or its normalized variant.

$$L = \lambda \cdot L_{xent}(\hat{X}, y) + L_{MAE}(X, \hat{X}) \tag{3}$$

The main improvement of equation 3 over the reconstruction loss of [8] is that we rescale the cross-entropy loss with the hyperparameter λ to bring the cross-entropy loss closer to the magnitude of the mean absolute error. Such that during reconstructor training the reconstructor does not give high preference to the underlying classifier rather than focusing on denoising/transforming the input. Another alternative to λ rescaling could be min-max normalization that brings the two losses to a similar scale. This re-scaling helps us achieve a high level of transferability (applying an approach to different classification models) across different unknown architectures for both Cifar10 and ImageNet datasets. As can be seen in Table I, our method achieves a high level of transferability across multiple black-box classifiers when compared to [8] (MSE + Stability) for a similar objective function. We also do not observe the visual artifacts reported by [8] in there denoised images for the similar objective function (MSE + stability [8]) which can be verified from Figure 3.

We see that the artifacts/textures within the denoised images generated by [8] are a cause of naively summing the stability-loss (or classification-loss) to the MSE-loss during model optimisation, since anytime during training the magnitude of stability-loss (or classification-loss) is very high compared to MSE-loss. This makes the denoiser model bias towards the classification model which was accessible during training. This is driving force to make [8] apply ensemble training to improve the applicability of their defence in a black-box setting. This phenomenon is similar to the concept of attacks based on adversarial patches [20] where the patch extensively over-tunes to the classification model to which the patch optimization process had access, leading to reduced performance on black-box classification models in a targeted situation.

B. Algorithm

To train our defence mechanism we use Algorithm 1 in which we begin by first initializing the encoder-decoder model $U_\theta(\cdot)$ with parameter θ, a classification model f, standard deviation σ for sampling gaussian noise vector $\delta \in N(0, \sigma^2 I)$ and the total number of training epochs T. We then sample a randomly chosen data-batch $\{x_i, y_i\}_1^n \in S$ followed by the

TABLE II: Certified accuracy on Cifar10 dataset on ResNet110 for various l_2 radii with standard accuracy in parenthesis

Defence Methods	ℓ_2 Perturbation Radius					
	0.25	0.5	0.75	1.0	1.25	1.5
Cohen et al. (Whitebox)	$^{(0.77)}$0.59	$^{(0.77)}$0.45	$^{(0.65)}$0.31	$^{(0.65)}$0.21	$^{(0.45)}$0.18	$^{(0.45)}$0.13
Salman et al. (Whitebox)	$^{(0.72)}$0.56	$^{(0.62)}$0.41	$^{(0.62)}$0.28	$^{(0.44)}$0.19	$^{(0.42)}$0.16	$^{(0.44)}$0.13
Salman et al. (Blackbox)	$^{(0.81)}$0.45	$^{(0.68)}$0.20	$^{(0.21)}$0.15	$^{(0.21)}$0.13	$^{(0.16)}$0.11	$^{(0.16)}$0.10
Ours(Blackbox)	$^{\mathbf{(0.86)}}$**0.77**	$^{\mathbf{(0.86)}}$**0.61**	$^{\mathbf{(0.81)}}$**0.52**	$^{\mathbf{(0.81)}}$**0.43**	$^{\mathbf{(0.67)}}$**0.35**	$^{\mathbf{(0.67)}}$**0.30**

normalization procedure where for each image x_i we find the normalized version $x_i^{'}$ using mean μ_o and standard-deviation σ_o calculated from dataset S, We then sample noise mask $\delta \in N(0, \sigma^2 I)$ using predefined value of σ and add it to $x_i^{'}$. Now, this noisy normalized image is passed onto to the encoder-decoder model U_θ that try to generate an image $\hat{x_i}$ similar to the original image x_i and when passed on to the classifier f leads to the correct label y. The loss value as given in Equation 3 is calculated using the above variable, Then this loss value is used perform a gradient descent step using the learning rate γ to update encoder-decoder model parameter θ.

ALGORITHM 1: Universal Certified Defence

Initialise: dataset S, training epochs T, batch size n, learning rate γ, normalization mean μ, normalization standard deviation σ_n, gaussian noise standard deviation σ, rescaling factor λ;

for t = *1 to* T **do**

  for *random batch* $\{x_i, y_i\}_1^n \in S$ **do**

    $x_i^{'} \leftarrow \dfrac{x_i - \mu_0}{\sigma_o}$;

    Sample: $\delta \in N(0, \sigma^2 I)$;

    $x_i^{'} \leftarrow x_i^{'} + \delta$;

    $\hat{x_i} \leftarrow U_\theta(x_i^{'})$;

    $L \leftarrow \lambda \cdot L_{xent}(f(\hat{x_i}), y_i) + L_{MAE}(x_i, \hat{x_i})$;

    $\theta \leftarrow \theta$ - $\gamma \cdot \dfrac{1}{n} \sum \{\nabla_\theta L\}$;

V. Experimentation

A. Datasets

We test our methodology across both small datasets like Cifar10 containing 10-classes with 50000 training and 10000 testing images of (32,32) dimensions and large datasets like ImageNet containing 1.2 million training samples, 50000 validation samples, and 100000 test sample for 1000 different object classes, Since the ground truth labels of testset of ImageNet are not publicly available we use its validation set as the testset.

B. Classification models

For Cifar10 we use only ResNet18 to train our defence strategy and ResNet110 for testing purpose. Similarly for ImageNet we use only ResNet18 to train our defence strategy and ResNet34 , ResNet50 for testing purpose.

C. Traintime Hyperparamter Setting

We use a pretrained ResNet18 as the underlying classifier, an SGD optimizer with a learning rate of 0.1 and algorithm 1 with $\lambda = 0.1$ for training the UNet architecture for a total of 20 epochs. Same hyperparameter setting is followed for both Cifar10 and ImageNet. For image normalization we use traditional ImageNet values, i.e., $\mu_0 = [0.485, 0.456, 0.406]$ and $\sigma_0 = [0.229, 0.224, 0.225]$.

D. Certification Details

To generate certifications for our methodology developed by putting an encoder-decoder based reconstructor on top of a pre-trained classifier, we too use the CERTIFY algorithm proposed by [7], where we put n = 10,000, n_0 = 100, α = 0.001, as was done by [8]. Such a hyperparameter setting resulted in easier comparisons. In all the comparisons of the certified radius of our method vs baselines, we use the best model/strategy of the baselines [7], [8]. We use the model weights provided by [8] available at [21].

E. Results

In this subsection, we discuss the results of various experiments performed on our defence methodology and compare them with our baselines, i.e., [7] and [8] which are two state-of-the-art methodologies in certified defences to verify our claims.

We begin our evaluation experiments with Cifar10 [22] dataset and use our encoder-decoder model on top of pre-trained ResNet110 [23] model to which our method had no access during training, i.e., a black-box setting. Based on the results achieved in Table II our methods outperform the baselines in terms of both the certified and standard accuracy. Our methodology results in an improvement of 16% on standard accuracy, 22% on certified accuracy when compared to the best white-box method [7] and an improvement of 60% on standard accuracy, 30% on certified accuracy when compared to the black-box setting of [8] for a l_2 perturbation radius of $\frac{255}{255}$. Figure 4 plots the certified performance for different values of σ on Cifar10 dataset, comparing the performance of our methodology with the white-box approach of [7], [8] and black-box setting of [8] using 14 different surrogate models during training. As the σ value increases our strategy provides much better-certified accuracy compared to the baselines.

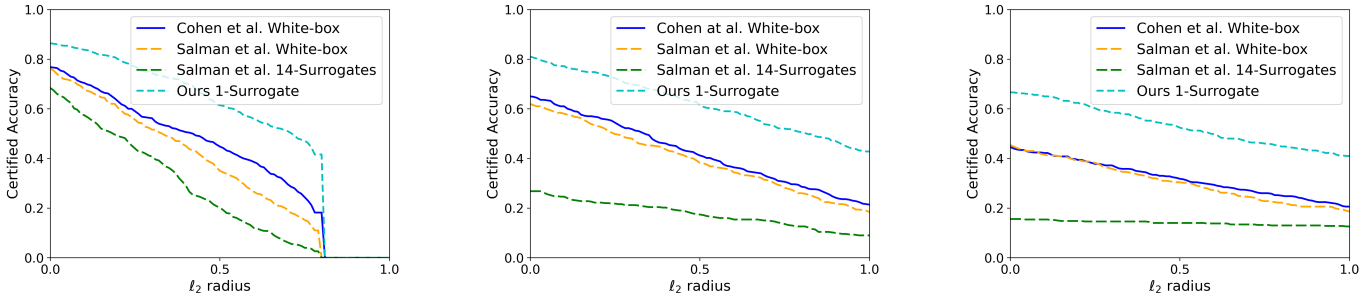
To verify the scalability as well as transferability of our method, we test our strategy on large datasets like Imagenet with two different models, namely ResNet34 and ResNet50 in a black-box setting. Through Table III, for ResNet34, our method provides a gain of 11% on standard accuracy, 16%

TABLE III: Certified accuracy on Imagenet dataset on ResNet34 for various l_2 radii with standard accuracy in parenthesis

Defence Methods	ℓ_2 Perturbation Radius					
	0.25	0.5	0.75	1.0	1.25	1.5
Cohen et al. (Whitebox)	$^{(0.60)}$0.50	$^{(0.53)}$0.44	$^{(0.53)}$0.39	$^{(0.53)}$0.33	$^{(0.53)}$0.28	$^{(0.42)}$0.22
Salman et al. (Whitebox)	$^{(0.64)}$0.47	$^{(0.55)}$0.32	$^{(0.55)}$0.19	$^{(0.35)}$0.12	$^{(0.35)}$0.08	$^{(0.16)}$0.04
Salman et al. (Blackbox)	$^{(0.65)}$0.47	$^{(0.53)}$0.32	$^{(0.53)}$0.18	$^{(0.34)}$0.12	$^{(0.34)}$0.08	$^{(0.34)}$0.03
Ours(Blackbox)	$^{(0.69)}$**0.62**	$^{(0.64)}$**0.58**	$^{(0.64)}$**0.53**	$^{(0.64)}$**0.49**	$^{(0.56)}$**0.44**	$^{(0.56)}$**0.41**

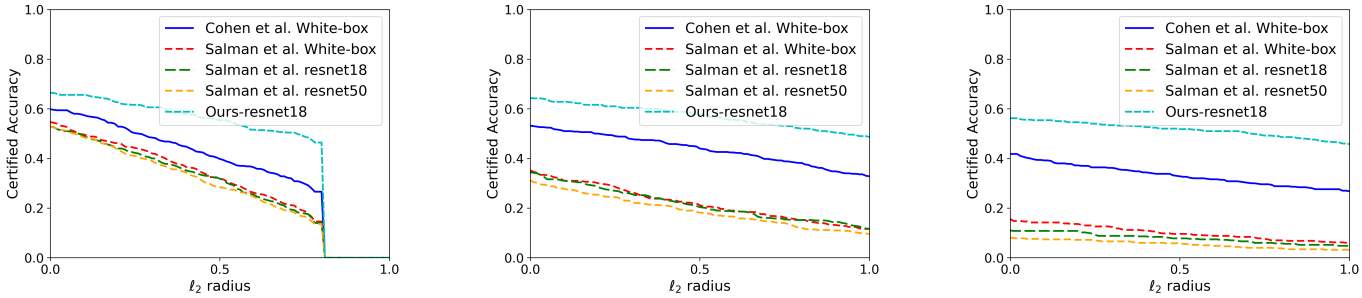
TABLE IV: Certified accuracy on Imagenet dataset on ResNet50 for various l_2 radii with standard accuracy in parenthesis

Defence Methods	ℓ_2 Perturbation Radius					
	0.25	0.5	0.75	1.0	1.25	1.5
Cohen et al. (Whitebox)	$^{(0.70)}$0.62	$^{(0.70)}$0.52	$^{(0.62)}$0.45	$^{(0.62)}$0.39	$^{(0.62)}$0.34	$^{(0.50)}$0.29
Salman et al. (Whitebox)	$^{(0.67)}$0.50	$^{(0.60)}$0.33	$^{(0.60)}$0.20	$^{(0.38)}$0.14	$^{(0.38)}$0.11	$^{(0.38)}$0.06
Salman et al. (Blackbox)	$^{(0.69)}$0.48	$^{(0.56)}$0.31	$^{(0.56)}$0.19	$^{(0.34)}$0.12	$^{(0.34)}$0.07	$^{(0.30)}$0.04
Ours(Blackbox)	$^{(0.74)}$**0.69**	$^{(0.72)}$**0.64**	$^{(0.72)}$**0.58**	$^{(0.72)}$**0.54**	$^{(0.64)}$**0.49**	$^{(0.64)}$**0.46**



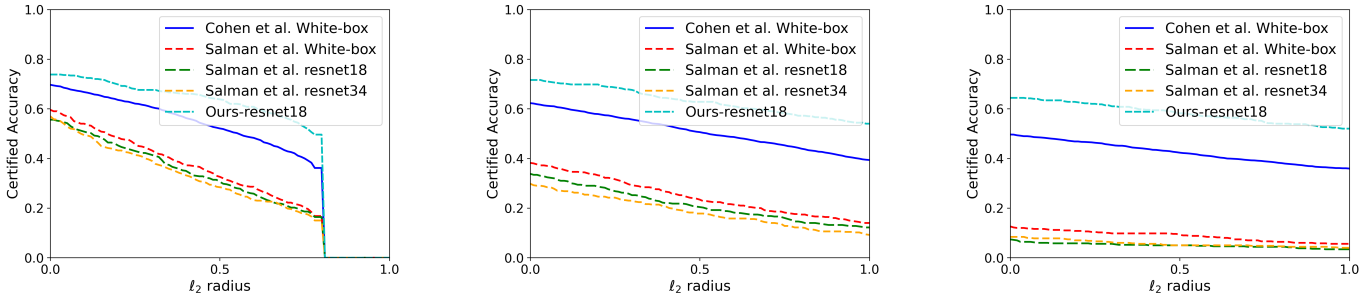
(a) $\sigma = 0.25$ (b) $\sigma = 0.50$ (c) $\sigma = 1.00$

Fig. 4: Certifying blackbox ResNet110 on Cifar10 for different values of σ



(a) $\sigma = 0.25$ (b) $\sigma = 0.50$ (c) $\sigma = 1.00$

Fig. 5: Certifying blackbox ResNet34 on ImageNet for different values of σ



(a) $\sigma = 0.25$ (b) $\sigma = 0.50$ (c) $\sigma = 1.00$

Fig. 6: Certifying blackbox ResNet50 on ImageNet for different values of σ

TABLE V: Accuracy against an adapted 20-step BPDA [24] based PGD-l2 attack applied on complete methodology modules

Method	Noise standard deviation (σ)		
	0.25	0.50	1.00
Ours (blackbox) (UNet) (with rescaled loss)	**73.6%**	**75.2%**	**66.8%**
Ours (blackbox) (UNet) (without rescaled loss)	72.5%	74.7%	40.2%
Salman et al. (blackbox) (DnCNN)	8.6%	9.6%	8.8%
Salman et al. (blackbox) (MemNet)	10.1%	10.6%	10.2%
Salman et al. (blackbox) (DnCNN-Wide)	2.0%	6.9%	6.6%

on certified accuracy when compared to the best white-box setting of [7] and a gain of 30% on standard accuracy, 37% on certified accuracy when compared to the black-box setting of [8] using only ResNet18 as a surrogate for a l_2 perturbation radius of $\frac{255}{255}$. Figure 5 compares the performance of different strategies on ResNet34 and visualises a plot between certified accuracy vs l_2 perturbation radius. Similar trends can be seen for ResNet50 using Table IV and Figure 6 where our method significantly outperforms the baselines.

To verify the real-life performance of different denoiser (or transformation) based strategies against adversaries, we test our and [8] methodologies on an adapted 20-step BPDA [24] based PGD-l_2 attack with a l_2-norm equal to 0.5 and a α equal to 0.01 for Cifar10-dataset in a black box-setting. Here the adversarial attack is performed on combined (denoiser + classifier) architecture. Note: For each methodology the adversarial perturbation is added to the original input image $\in$ (0,1). Table V shows the accuracies results for different strategies and using different proposed denoiser models. Similar to the certified results, we also achieve an increment in the adversarial accuracy. For [8] we use their pretrained denoisers trained using stability loss as they claimed such models achieve the best accuracy in a black box-setting.

Interestingly, when we apply this attack on our methodology trained without rescaling the losses, we see a huge decrement in blackbox performance at higher noise levels. This decrease in the blackbox performance signifies the overfitting that occur during the training of denoiser networks with pretrained surrogate model. A possible reason for this decreased performance is giving higher magnitude of classification loss (eg: cross-entropy loss) in comparison to reconstruction (eg: MSE-Loss/L1-loss etc), thus when these losses are added classification loss has a higher dominance. Hence, using the rescaled loss is an added advantage to our methodology.

VI. Conclusion

Through the extensive experiments performed within this study, we were able to address a few shortcomings faced by the past state-of-the-art on developing certified defences from scratch or by using the pre-trained models using the techniques of randomized-smoothing and denoised-smoothing. In particular, we set up an empirical insight into how the process of adding noise to image $x \in (0, 1)$ is the cause of loss of informative features from that image and how normalization-based noise-augmentation of input is a solution to it.

We use this insight to develop a scalable (in terms of dimensionality of input) approach that shows high transferability to pre-trained classification models in a black-box setting. We propose a novel objective function for the training framework that gives equal preference to the reconstruction and the classification loss. Our method achieves exceptionally better performance both in certified as well as standard accuracy at a higher level of σ when compared to previous state-of-the-art approaches proposed by [7] and [8] without any model-intensive training like ensemble training.

References

[1] R. Jia and P. Liang, "Adversarial examples for evaluating reading comprehension systems," *arXiv preprint arXiv:1707.07328*, 2017.

[2] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[3] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song, "Robust physical-world attacks on deep learning visual classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1625–1634.

[4] A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, and A. Madry, "Adversarial examples are not bugs, they are features," *arXiv preprint arXiv:1905.02175*, 2019.

[5] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.

[6] M. Lecuyer, V. Atlidakis, R. Geambasu, D. Hsu, and S. Jana, "Certified robustness to adversarial examples with differential privacy," in *2019 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2019, pp. 656–672.

[7] J. Cohen, E. Rosenfeld, and Z. Kolter, "Certified adversarial robustness via randomized smoothing," in *International Conference on Machine Learning*. PMLR, 2019, pp. 1310–1320.

[8] H. Salman, M. Sun, G. Yang, A. Kapoor, and J. Z. Kolter, "Denoised smoothing: A provable defense for pretrained classifiers," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[9] A. Raghunathan, J. Steinhardt, and P. Liang, "Certified defenses against adversarial examples," *arXiv preprint arXiv:1801.09344*, 2018.

[10] E. Wong and Z. Kolter, "Provable defenses against adversarial examples via the convex outer adversarial polytope," in *International Conference on Machine Learning*. PMLR, 2018, pp. 5286–5295.

[11] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising," *IEEE transactions on image processing*, vol. 26, no. 7, pp. 3142–3155, 2017.

[12] Y. Tai, J. Yang, X. Liu, and C. Xu, "Memnet: A persistent memory network for image restoration," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 4539–4547.

[13] B. Li, C. Chen, W. Wang, and L. Carin, "Certified adversarial robustness with additive noise," *arXiv preprint arXiv:1809.03113*, 2018.

[14] G.-H. Lee, Y. Yuan, S. Chang, and T. S. Jaakkola, "Tight certificates of adversarial robustness for randomly smoothed classifiers," *arXiv preprint arXiv:1906.04948*, 2019.

[15] K. D. Dvijotham, J. Hayes, B. Balle, Z. Kolter, C. Qin, A. György, K. Xiao, S. Gowal, and P. Kohli, "A framework for robustness certification of smoothed classifiers using f-divergences." in *ICLR*, 2020.

[16] G. Yang, T. Duan, J. E. Hu, H. Salman, I. Razenshteyn, and J. Li, "Randomized smoothing of all shapes and sizes," in *International Conference on Machine Learning*. PMLR, 2020, pp. 10 693–10 705.

[17] L. van der Maaten and G. Hinton, "Visualizing data using t-sne. journal of machine learning research 9," *Nov (2008)*, 2008.
[18] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.
[19] J. Gurrola-Ramos, O. Dalmau, and T. E. Alarcón, "A residual dense u-net neural network for image denoising," *IEEE Access*, vol. 9, pp. 31 742–31 754, 2021.
[20] T. B. Brown, D. Mané, A. Roy, M. Abadi, and J. Gilmer, "Adversarial patch," *arXiv preprint arXiv:1712.09665*, 2017.
[21] H. Salman, "Denoised-smoothing," https://github.com/microsoft/denoised-smoothing.git, 2020.
[22] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2009.
[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
[24] A. Athalye, N. Carlini, and D. Wagner, "Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples," in *International Conference on Machine Learning*. PMLR, 2018, pp. 274–283.