RANDOMIZED FEATURE SQUEEZING AGAINST UNSEEN l_p ATTACKS WITHOUT ADVERSARIAL TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep learning has made tremendous progress in the last decades; however, it is not robust to adversarial attacks. The most effective approach is perhaps adversarial training, although it is impractical because it requires prior knowledge about the attackers and incurs high computational costs. In this paper, we propose a novel approach that can train a robust network only through standard training with clean images without awareness of the attacker's strategy. We add a specially designed network input layer, which accomplishes a randomized feature squeezing to reduce the malicious perturbation. It achieves excellent robustness against unseen l_0, l_1, l_2 and l_∞ attacks at one time in terms of the computational cost of the attacker versus the defender through just 100 and 50 epochs of standard training with clean images in CIFAR-10 and ImageNet respectively. The thorough experimental results validate the high performance. Moreover, it can also defend against unlearnable examples generated by One-Pixel Shortcut which breaks down the adversarial training approach.

1 INTRODUCTION

Since the seminal work of Szegedy et al. (2014), the vulnerability of neural networks has been widely acknowledged by the deep learning community. A lot of solutions have been proposed to solve these problems. They can be categorized into three classes.

The first is preprocessing-based approaches that include bit-depth reduction (Xu et al., 2018), JPEG compression, total variance minimization, image quilting (Guo et al., 2018), and Defense-GAN (Samangouei et al., 2018). With this preprocessing, the hope is to reduce adversarial effect; however, it neglects the fact that the adversary can still take this operation into account and craft an effective attack through Backward Pass Differentiable Approximation (BPDA) (Athalye et al., 2018).

Secondly, perhaps the most effective method is adversarial training. In the training phase, the attack is launched through the backward gradient propagation concerning the current network state. A large volume of work falls into this class differing in ways to generate extra training samples. Madry et al. (2018) used a classical 7-step PGD attack. Other approaches are also possible, such as Mixup inference (Pang et al., 2020), feature scattering (Zhang & Wang, 2019), feature denoising (Xie et al., 2019), geometry-aware instance reweighting (Zhang et al., 2021), and channel-wise activation suppressing (Bai et al., 2021). External (Gowal et al., 2020) or generated data (Gowal et al., 2021; Rebuffi et al., 2021) are also beneficial for robustness. The inherent drawbacks are the large computation cost and the need for prior knowledge about attacks. This is certainly not realistic in practice. Also, there is a possibility of robust overfitting (Rice et al., 2020).

The last is adaptive test-time defenses. They try to purify the input iteratively as in Mao et al. (2021); Shi et al. (2021); Yoon et al. (2021) or adapt the model parameters or network structures to reverse the attack effect. For example, closed-loop control is applied in Chen et al. (2021), while a neural Ordinary Differential Equation (ODE) layer in Kang et al. (2021). Unfortunately, Croce et al. (2022) proved that most of them are not effective as claimed.

Overall, the progress is not optimistic, and marginal improvements in robust accuracy require huge computational costs while not valid for unseen attacks. So we ask a question: "can we design a novel network and loss function thereof that can drive the network to be robust on its own without awareness of adversarial attacks?" In other words, we do not intend to generate extra adversarial

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.

samples as most other approaches do, and standard training with clean images is enough. Indeed, there should be no prior knowledge of attacks needed at all. The disparity between clean accuracy and robust accuracy likely arises from the vastness of the input space $\mathcal{X}$. More specifically, the region in which the network generalizes well, denoted as $S_g \subseteq \mathcal{X}$, may not align with human perceptual space, denoted as $S_a \subseteq \mathcal{X}$. Both S_g and S_a occupy certain portions of the input space, which is so expansive that S_g and S_a may have little or no overlap, i.e., $S_g \cap S_a = \emptyset$. Consequently, we posit that a severe constraint on the input space (see Figure 1) forces a well-trained network—with its intrinsic generalization capability—to adjust its region of good generalization S_g into closer alignment with the human perceptual space S_a. This alignment, enforced within a reduced-dimensional space, is anticipated to yield a direct improvement in adversarial robustness.

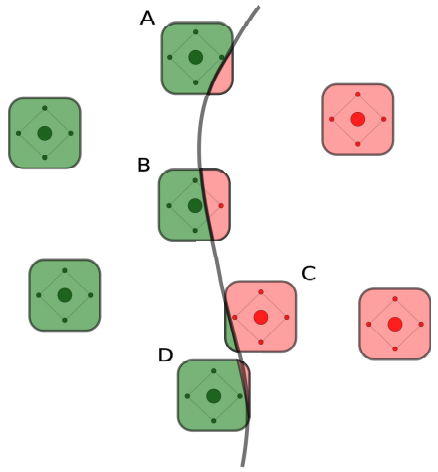

Figure 1: A conceptual illustration of feature squeezing for adversarial robustness. The rectangle around the center of a test sample represents the possible perturbation space, while only four vertices in a diamond can be allowed through feature squeezing. As a result, with feature squeezing, only B out of A, B, C, and D fails for adversarial defense, rather than all of them without feature squeezing.

In order to achieve better generalization with the reduced input space, our training and test framework is designed as shown in Figure 2. Unlike previous preprocessing methods Xu et al. (2018), our approach randomly squeezes the input features using parameters learned during training. Furthermore, during testing, this layer is simplified, which greatly facilitates evaluation. Experiments on CIFAR-10 and ImageNet demonstrate that the proposed method can enhance the robustness of networks.

Remarkably, although our primary motivation is adversarial defense against unseen attacks, it turns out that ours is much less influenced by the unlearnable examples, i.e., data intentionally manipulated for unauthorized usage for training DNNs. Recently, a One-Pixel Shortcut(OPS) has been proposed in Wu et al. (2023) and could effectively degrade model accuracy even to almost an untrained counterpart even equipped with adversarial training, while ours sustains around 60%. Image Shortcut Squeezing Liu et al. (2023b) can counter OPS, however it may not deal with adversarial attacks.

With all the source codes and pre-trained models online B, our work has the following contributions:

- We design a special input layer that uses reciprocal and multiplication to implement our randomized feature squeezing, which is very novel. Furthermore, it could be plunged simply into networks such as WideResNet and ConvNeXt with very different structures to boost performance.
- **Our work is the only one that does not require any prior knowledge about the attacks with standard training with clean images;** while achieves great robust accuracy.
- Our approach appears to be the only one that can effectively deal with both adversarial attacks and unlearnable examples generated by the state-of-the-art OPS without any data augmentation.

2 Related Works

Some works add extra preprocessing steps. For example, in Yang et al. (2019), pixels are randomly dropped and reconstructed using matrix estimation. Ours is not preprocessing. Ours only adds an extra layer inside the network, and the network is trained and tested as usual without explicit image

completion. Besides this, to get high robust accuracy, Yang et al. (2019) needs adversarial training, while we adopt standard training with clean images. Another related work is certified adversarial robustness via randomized smoothing (Cohen et al., 2019). The base classifier needs Gaussian data augmentation for training, and inference is the most likely output class of the input perturbed by isotropic Gaussian noise. Ours only uses standard training and testing, without perturbation-based training data augmentation involved at all.

Stochastic Neural Networks(SNNs) (Eustratiadis et al., 2021; Däubener & Fischer, 2022; Lee et al., 2023) achieve robustness by intentionally injecting noise into the intermediate layers of the preexisting networks, which is very different from ours. Motivated by the inherent weakness of the current network, we are trying to modify it so that adversarial defense can be achievable for both CIFAR-10 and ImagetNet for which SNNs are not available due to large image size. Unlike the usual SNN, Bart Raff et al. (2019) adopts a barrage of random transformations. However, unfortunately, its robustness is likely over-estimated as presented in Sitawarin et al. (2022). The key-based defense such as Rusu et al. (2022); AprilPyone & Kiya (2021) is related but completely different from ours. The secure key is private and not open to the attacker, so the evaluation with a complete white-box attack is impossible. In our case, the attacker has full access to the defender's source code to launch a white-box attack.

Recently, some works have addressed the robustness from the perspective of the network's architecture. Wu et al. (2021) investigates the impact of the network width on the model robustness and proposes Width Adjusted Regularization. Similarly, Huang et al. (2021) explores architectural ingredients of adversarially robust deep neural networks thoroughly. Liu et al. (2023a) established that the higher weight sparsity benefits adversarially robust generalization via Rademacher complexity. Wang et al. (2022) proposes batch normalization removal, such that adversarial training can be improved. Singla et al. (2021) shows that using activation functions with low curvature values reduces both the standard and robust generalization gaps in adversarial training. We note that Hou et al. (2025) enhances baseline robustness via network reprogramming, yet it is only effective on MNIST without adversarial training. Similarly, Yang et al. also deals solely with MNIST, while Salamah et al. only reports results under weak attacks (e.g., 5-step PGD with maximum $\varepsilon = 4$ on CIFAR-100)—far below standard adversarial benchmarks—and in fact its robust accuracy at $\varepsilon = 4$ approaches zero. Therefore, despite relying on standard training, these methods cannot be considered a feasible defense.

There are some attempts (Tramèr & Boneh, 2019; Maini et al., 2020; Croce & Hein, 2022; Laidlaw et al., 2021; Dai et al., 2022) to deal with multiple attacks simultaneously. **Among them, the only relevant works for unseen attacks are Perceptual Adversarial Training (PAT)Laidlaw et al. (2021), and adversarial training with variation regularization (AT-VR) Dai et al. (2022), but they adopt costly adversarial training, and only for CIFAR-10.** Some benchmarks (Dai et al., 2023; Kang et al., 2019) extend beyond the l_p attacks.

Adversarial purification is another research line to defend against unseen attacks, but it is very slow in test. For example, in Table 14 of (Nie et al., 2022), the inference time is around (100-300)x of standard one. Also, they need pre-trained diffusion models, which are very expensive to get. Another disadvantage is that the thorough evaluation of the robustness of these methods is impossible due to very high memory consumption. Indeed, the robust accuracy is overly estimated as pointed out by Lee & Kim (2023). Moreover, adversarial purification cannot deal with OPS, as the underlying static model has very low clean accuracy.

3 BACKGROUND

A standard classification can be described as follows:

$$\min_{\vartheta} E_{(x,y)\sim D}\left[L\left(x, y, \vartheta\right)\right], \tag{1}$$

where data examples $x \in R^d$ and corresponding labels $y \in [k]$ are taken from the underlying distribution D, and $\vartheta \in R^p$ is the model parameters to be optimized with respect to an appropriate function L, for instance cross-entropy loss. When $x \in R^d$ can be maliciously manipulated within a set of allowed perturbations $S \subseteq R^d$, which is usually chosen as a l_p-ball ($p \in \{0, 1, 2, \infty\}$) of

radius ϵ around x, Equation 1 should be modified as:

$$\min_{\vartheta} E_{(x,y)\sim D}\left[\max_{\delta\in S} L\left(x+\delta, y, \vartheta\right)\right]. \tag{2}$$

An adversary implements the inner maximization via various white-box or black-box attack algorithms. On the other hand, outer minimization is the goal of a defender. Adversarial training is the most effective approach to achieve this outer minimization via augmenting the training data with crafted samples. In fact, all current approaches, including test-time adaptive defense as it needs a base classifier, aim to learn the parameters of a pre-existing model to improve the robustness. In this paper, we try to increase the robustness through a specially designed input layer such that standard training with clean images can be adopted.

4 METHOD

4.1 INPUT LAYER

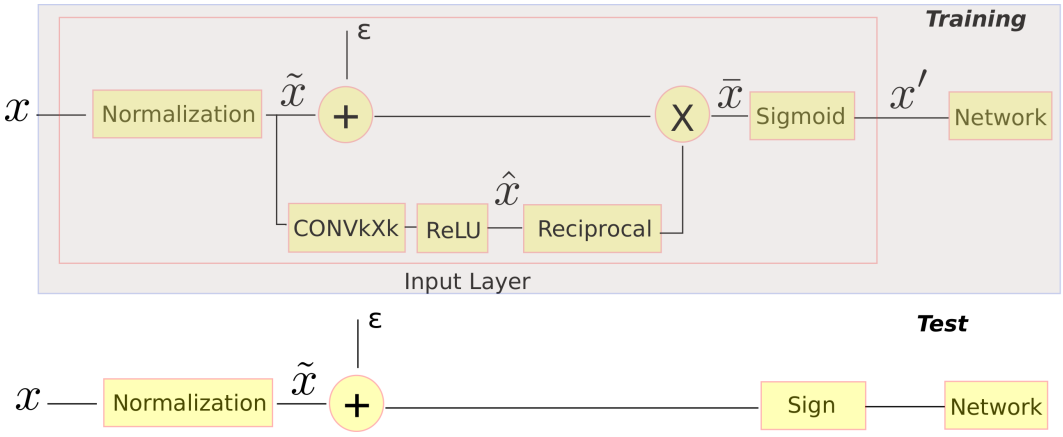

Figure 2: The shaded and non-shaded areas show the training and test framework respectively. In training, our specially designed input layer is inside the red rectangle. The input image x is first normalized, then undergoes two paths. On one path, independent Gaussian noise ϵ is added, and the other path includes $k \times k$ convolution and ReLU followed by element-wise reciprocal. Finally, these two terms are combined through element-wise multiplication and the result feeds to the Sigmoid. The final x' will be used as inputs to the classification network, the same as other training approaches. End-to-end training scheme is adopted to learn the parameters of $k \times k$ convolution. In test, the conv $k \times k$ path is removed wholly, and Sigmoid is replaced with Sign defined in Equation 4.

As we stated earlier, the goal of input layer is to squeeze the input feature in a random and controlled way. The whole procedure is depicted in Figure 2.

It consists of the following steps:

1. The input x with r, g, b channel will be normalized to a variable with a mean 0 and a standard deviation 1, through $\tilde{x} = \frac{x-mean}{std}$ in the input layer, where $mean$ and std are mean and standard deviation of training set. Then, it goes through top and bottom paths.
2. In the top path, each element of $\tilde{x}$ is corrupted independently by additive Gaussian noise ε, where $\varepsilon \sim N\left(0, \sigma^2\right)$.
3. In the bottom path, $\tilde{x}$ goes through a $k \times k$ 2D convolution and ReLU, and we get $\hat{x}$ with three channels, and then its element-wise reciprocal $\frac{1}{\hat{x}+\gamma}$, where γ is a small constant in order to make the denominator always positive, which is 1×10^{-5} in this paper.
4. The top and bottom paths are combined by element-wise multiplication, $\bar{x} = (\tilde{x} + \varepsilon) \times \frac{1}{\hat{x}+\gamma}$.

5. The final output x' is a Sigmoid of the $\bar{x}$, i.e., $x' = \frac{1}{1+\exp(-\bar{x})}$.

So essentially,

$$x' = \frac{1}{1 + \exp\left(-\frac{(\tilde{x}+\varepsilon)}{\hat{x}+\gamma}\right)}. \tag{3}$$

This formula can be interpreted this way. $\tilde{x} + \varepsilon$ is a polluted version of the input image,and $\frac{1}{\hat{x}+\gamma}$ tries to modulate the image based on the $\hat{x}$, named as sampling matrix having the same size as input x. The key motivation is that if we enforce $\hat{x}$ to be very small through some loss function, $\left|\frac{(\tilde{x}+\varepsilon)}{\hat{x}+\gamma}\right|$ will become big and the response of Sigmoid will be on the saturated region, i.e., most elements of x' will be either 0 or 1. In other words, the input feature will be squeezed in a random manner where the parameters of sampling matrix $\hat{x}$ are learned on the end-to-end training. Accordingly, ε plays the role of mimicking the attack that the adversary may launch. The appropriate value of σ should be chosen as the big one will degrade the clean accuracy while the network can not gain much robustness for the small one. Based on our analysis above, one may raise a big concern regarding the obfuscated gradients Athalye et al. (2018) which may be incurred by reciprocal and Sigmoid operator in robustness evaluation. On one hand, $\hat{x}$ is very small, so the gradient of reciprocal $\frac{1}{\hat{x}+\gamma}$ will be very big. On the other hand, $\bar{x} = (\tilde{x} + \varepsilon) \times \frac{1}{\hat{x}+\gamma}$ will reside on the saturated domain of Sigmoid, i.e., the gradients of x' with respect to $\bar{x}$ will be very small. Actually, this might also cause some trouble in training, as we need to learn the parameters of conv $k \times k$ for sampling matrix $\hat{x}$, although they might be canceled out by each other to some extent, as they are on the same path in backward pass gradient propagation. To resolve this, in training we adopt the BPDA-like optimization procedure. Namely, for the forward pass, we evaluate the reciprocal and Sigmoid as usual, however, in the backward pass, the gradient of the reciprocal is set to be -1, and 1 for Sigmoid. While in test, because Sigmoid often goes to two extreme values 0 and 1, dependent of the sign of $\tilde{x} + \varepsilon$, we just remove the bottom path wholly, and replace the Sigmoid with Sign which is defined as:

$$Sign(x) = \begin{cases} 0, & \text{if } x < 0, \\ 0.5, & \text{if } x = 0, \\ 1, & \text{if } x > 0. \end{cases} \tag{4}$$

This will greatly simplify our robust evaluation. Of course, it is also possible to train directly in the test framework. The detailed analysis is deferred to Section 5.4. In essence, the proposed training architecture introduces a refined optimization protocol for the target test framework. Rather than direct parameter tuning, this detour approach strategically incorporates auxiliary components—conv$k \times k$ layers, relu, reciprocal, and sigmoid—coupled with a customized loss function discussed in the next section. These transient modules are exclusively employed during the training and subsequently discarded during inference, yielding a streamlined test framework characterized by random noise injection and Sign.

4.2 Loss Function

As mentioned earlier, we have to design a loss function to implement our motivation to make the sampling matrix $\hat{x}$ small. For each $\hat{x}$, we get S, the average of all the elements of $\hat{x}$ that are greater than some threshold β. Formally,

$$S = \frac{\sum\limits_{i \in T} \hat{x}_i}{\#T}, \ \ where \ \ T = \{\, i | \, \hat{x}_i > \beta \}\,. \tag{5}$$

A small β means $\hat{x}$ will become sparse. The final loss function is:

$$L = \alpha \times L_{ce} + S, \tag{6}$$

where S is the sparse loss and L_{ce} is the cross-entropy loss with weight α. When α becomes large, the loss function falls back to standard cross-entropy. In summary, there are only four parameters, σ of noise, the size of convolution kernel k, weight α, and threshold $\beta = 0.2$ in this paper.

5 EXPERIMENTS

To verify the effectiveness of our approach, we conducted the experiments on both CIFAR-10 and ImageNet.

For CIFAR-10, we choose the wide residual network WideResNet-28-10 (Zagoruyko & Komodakis, 2016) as the base one, where we add our specially designed input layer as described in Section 4 with $\sigma = 0.65$, $\alpha = 0.1$, and conv 5×5. The initial learning rate of 0.1 is scheduled to drop at 30, 60, and 80 out of 100 epochs in total with a decay factor of 0.2. The weight decay factor is set to $5{\times}10^{-4}$, and the batch size is 200. To emphasize again, we only perform standard training through just 100 epochs. **Reemphasize that there is no work for unseen attacks with standard training, and if adversarial training is allowed, Laidlaw et al. (2021) and Dai et al. (2022) are the only two to this end with training costs 16 and 62 times as of ours with the same WideResNet-28-10, as shown in Table 1.**

ImageNet is the most challenging dataset for adversarial defense, and **there is no work dealing with unseen attacks even with adversarial training.** In this paper, ImageNet only refers to ImageNet-1k without explicit clarification, and robustness is only evaluated on the ImageNet validation set. For simplicity, we choose the architecture of ConvNeXt-T + ConvStem in Singh et al. (2023) with $\sigma = 1.4$, $\alpha = 0.5$, and conv 7×7. Our training scheme is very simple. All parameters are randomly initialized, followed by standard training for 50 epochs with heavy augmentations without CutMix (Yun et al., 2019) and MixUp (Zhang et al., 2018), as these will undermine the viability of our sampling matrix. While for the same ConvNeXt-T + ConvStem in Singh et al. (2023), although ConvStem is randomly initialized, the ConvNeXt-T part is from a strong pre-trained model which usually takes about 300 epochs. Thus the whole network needs extra standard training for 100 epochs to get good clean accuracy, followed by 300 epochs of adversarial training with 2-step APGD. So the total cost is up to $300 + 100 + 300 \times [2\,(\text{APGD steps}) + 1\,(\text{weights update})] = 1300$, which is around $1300/50 = 26$ times bigger than ours. As expected, our specially designed input layer changes the input x into x' that are extremely squeezed. On one hand, it poses a great challenge to the network; while on the other hand, it improves the robustness. Some of the example feature maps in our input layers are listed in Figure 3. Notably, thanks to the great capacity of deep network, our defense achieves reasonably good and stable clean accuracy on total images, i.e., 10k for CIFAR-10, and 5K for ImageNet; and due to resource constraints, we will demonstrate the robustness performance on only 1K images unless explicitly specified, which is enough to make a fair comparison, against l_0, l_1, l_2 and l_∞ attacks in the follow sections with l_1-ϵ=12, l_2-ϵ=1, l_∞-ϵ=$8/255$ for CIFAR-10; and l_1-ϵ=75, l_2-ϵ=2, l_∞-ϵ=$4/255$ for ImageNet.

Table 1: Clean and training cost comparison. For CIFAR10, the cost is defined as: #Epochs $\times$ [#PGD + 1 (weights update)] with respect to ours, which is denoted by 1. For ImageNet, please refer to the above main text. Since ours is random, we report mean and standard deviation for five runs.

Defense	Clean	#Epochs	#PGD	#Cost
CIFAR-10				
Ours	80.23±0.30	100	0	1
PATLaidlaw et al. (2021)	82.40	100	15	16
AT-VRDai et al. (2022)	72.73	200	30	62
ImageNet				
Ours	67.60±0.55	50	0	1
Singh et al. (2023)	72.74	400+300	2	26

The attack for a determined network only accepts correctly classified clean images and stops further operation once the network gets fooled. Since ours is random, we run out of the maximum allowed number of iterations for all input samples to ensure that the generated adversarial samples have a high probability of fooling the network. This principle goes throughout all experiments for scientific rigor.

5.1 BLACK-BOX ATTACKS

For l_0 we use Pomponi et al. (2022), which is based on rearranging the pixels inside a random selected patch without limits of the number of perturbed pixels, i.e., l_0 norm. The configuration for CIFAR-10 is 25 restarts with 10 max iterations per restart with patch dimension of 3, so the total budget is 250 iterations; while for ImageNet, 100 restarts with 50 max iterations per restart with the same dimension 3, accordingly 5000 in total. For CIFAR-10, the average iterations of competitors are only half of ours while the accuracies drop to random guess with l_0 less than ours. please refer to Table 5 for more details. For l_1 we use Square Andriushchenko et al. (2020), which is commonly adopted in adversarial defense evaluation. We use the default iterations 5000. For l_2, l_∞, we use both Square Andriushchenko et al. (2020) and SignHunter Al-Dujaili & O'Reilly (2020), which is a divide-and-conquer, adaptive, memory-efficient algorithm. As in Table 2, l_0, l_1, l_2, and l_∞ black-box attacks are almost impotent to ours, and in general our robust accuracy significantly outperforms others by a large margin except in l_2 for ImageNet. Our high performance comes from the very effective adversarial perturbation simulation during training operated by random noise, which also plays a role in misleading the attack in the test.

Table 2: Robustness comparison against $l_0, l_1, l_2,$ and l_∞black-box attacks. The iterations column shows the average number of iterations by the attack. Two columns in l_2 and l_∞ are respecively for Square and SignHunter (in bold). Adv-Trained refers to Singh et al. (2023).

Defense	Robust				Iterations			
CIFAR-10	l_0	l_1	l_2	l_∞	l_0	l_1	l_2	l_∞
Ours	61.10	79.00	77.60 **78.70**	76.70 **77.80**	250	5000	5000 **5000**	5000 **5000**
PAT	12.00	52.90	62.20 **61.20**	46.00 **42.10**	122	3480	4026 **3939**	3230 **2961**
AT-VR	11.40	27.70	53.20 **60.80**	53.10 **51.50**	109	2390	3915 **4263**	3807 **3657**
ImageNet	l_0	l_1	l_2	l_∞	l_0	l_1	l_2	l_∞
Ours	67.70	67.00	66.90 **68.40**	65.40 **67.40**	5000	5000	5000 **10000**	5000 **10000**
Adv-Trained	33.40	50.70	69.00 **71.00**	64.30 **64.30**	3227	3668	4791 **9752**	4521 **9001**

5.2 WHITE-BOX ATTACKS

Since Sign is non-differentiable, the backward pass differentiation should be approximated with some function to evaluate the robust accuracy on white-box attacks. We have tried different options for 1K samples: identity; $\frac{\text{Softsign}(x)+1}{2}$, where $\text{Softsign}(x) = \frac{x}{1+|x|}$; and Sigmoid($ax$), $a \in \{1, 3, 5, 7, 9\}$. Rusu et al. (2022) also used Sign, and only tested $\text{Softsign}(x)$ and Sigmoid(x). Of course, more options can convince us more of the robustness of our defense. We have tested against attacks of APGD-l_2 and APGD-l_∞, but their robust accuracies are higher than the plain version. Indeed, the similar observation is also reported in Lee & Kim (2023). The APGD-l_1 is adopted since it is stronger than Sparse l_1 Tramèr & Boneh (2019). The results are shown in Table 7. According to the worst accuracies among all BPDA in Table 7, to ensure the legitimate robust evaluation in the following sections, for CIFAR-10, we select $\frac{\text{Softsign}(x)+1}{2}$ for both APGD-l_1 and PGD-l_2, Sigmoid($9x$) for DDN-l_2, Sigmoid($3x$) for PGD-l_∞; while for ImageNet, x for APGD-l_1, $\frac{\text{Softsign}(x)+1}{2}$ for PGD-l_2, Sigmoid($5x$) for DDN-l_2, and Sigmoid($3x$) for l_∞. Based on the comparison in Table 3, ours converges on EOT-20. Interestingly, DDN (Rony et al., 2019) is ineffective against our method: EOT-20 performs almost identically to the standard attack, whereas all other defenses almost completely fail. This strongly highlights an inherent drawback of defenses that rely on adversarial training against a single attack type. DDN is a heuristic approach and may not be well-suited to our defense; consequently, EOT-20 can be weaker than standard one.

5.3 ONE-PIXEL SHORTCUT

Although our approach is motivated for adversarial defense, it turns out ours is much less impacted by OPS without any data augmentation. Ours exceeds others by 40+ in Table 6. Again, it is due to

Table 3: Robustness against l_1, l_2, and l_∞ white-box attacks on total images except the rows of EOT only for 1k images. than then st

Defense	Clean	APGD-l_1	PGD-l_2	DDN-l_2	PGD-l_∞
CIFAR-10		$\frac{\text{Softsign}(x)+1}{2}$	$\frac{\text{Softsign}(x)+1}{2}$	Sigmoid($9x$)	Sigmoid($3x$)
Ours	80.23	57.16	37.97	52.95	42.38
Ours-EOT20	81.02	38.50	33.90	58.20	34.00
Ours-EOT50	81.02	38.00	34.50	-	32.50
PATLaidlaw et al. (2021)	82.40	33.22	40.96	0.00	36.38
AT-VRDai et al. (2022)	72.73	9.06	31.21	0.10	52.73
ImageNet		x	$\frac{\text{Softsign}(x)+1}{2}$	Sigmoid($5x$)	Sigmoid($3x$)
Ours	67.60	57.06	40.46	53.96	24.18
Ours-EOT20	68.80	46.90	37.90	51.80	18.80
Ours-EOT50	68.80	46.60	38.10	-	17.40
Singh et al. (2023)	72.74	30.63	53.56	2.90	53.28

Table 4: Comparison between the normal training(N.T.) and training with test(T.T.) framework. The Clean is for total images; while the robust accuracy is on 1k.

Defense	Clean	APGD-l_1	PGD-l_2	PGD-l_∞
CIFAR-10				
N.T.	80.23±0.30	65.00	39.90	41.70
T.T.	77.92±0.18	58.10	31.50	31.80
ImageNet				
N.T.	67.60±0.55	65.80	42.40	25.40
T.T.	66.90±0.31	64.70	40.80	23.90

the random featured squeezing. Since we transform all pixels to 1 or 0, the pixel chosen by the OPS can not stand out from its neighbors.

5.4 Ablation Studies

Training in the configuration of the test can be regarded as special case of our detour training framework where the parameters of all the conv $k \times k$ for sampling matrix $\hat{x}$ are manually set to be zero, then the reciprocal will become very big, and the Sigmoid is approximately equal to Sign. Table 4 shows that this training scheme can achieves robust accuracy since it also implements random feature squeezing. However, it is always inferior to normal ones, especially for CIFAR-10; while performance gap is small on ImageNet. This could be due to the following reasons. With normal training, at the early stages, since the sparse loss S in Equation 6 is relatively big, the feeds to the network, x', is not highly squeezed to be 0 or 1 at the early stages, which seems to benefit robustness. For ImageNet, compared with CIFAR-10, due to the massive size of the datasets and the unique ConvNeXt-T+ConvStem structure, S drops quickly and x' goes to the extreme values much faster, thus only having marginal improvement. Please refer to Section G for more analysis. For additional ablation studies on parameter settings, please refer to Section H. Our method remains robust across different parameter configurations. Another possible concern may be related with transfer attack, i.e., using the training framework as the targeted net to generate adversarial examples. This is not effective as shown in Section F, since training network contains the sigmoid and reciprocal plus small conv $k \times k$, which is driven by our sparse loss term. This will cause some problems in gradient backpropagation, even though BPDA is adopted as done in training.

6 ROBUST ACCURACY FOR DETERMINISTIC MODEL

To further solidify our robust evaluation, we remove randomness and transform the test framework with random noise into a deterministic one. To our knowledge, no existing work comprising a random component has been evaluated with that component fixed. More specifically, we feed the N copies of the same test image to the test framework, each with a different but fixed seed of noise, and then the average logits of N outputs are used to get the final classification. It might be possible that our training scheme implements implicit adversarial training due to the added random noise; however, it is unclear how it relates to test robust accuracy. Interestingly, as expected, non-trivial training robust accuracy is achieved, and thanks to the feature squeezing, test robust accuracy also keeps up, with maximum discrepancy less than 10 for most cases shown in Table 11. Indeed, a deep connection exists between feature squeezing and Rademacher complexity for adversarially robust generalization in Yin et al. (2019). Let $x \in X \subseteq R^d$ and $y \in Y \subseteq \{-1, 1\}$ be the feature and label spaces, and the training set $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$. The experimental Rademacher complexity for the function class F is defined as : $R_s(F) = \frac{1}{n} E_\sigma \left[\sup_{f \in F} \sum_{i=1}^{n} \sigma_i f(x_i) \right]$, where $\sigma_1, \sigma_2, ..., \sigma_n$ are independent random variables uniformly chosen from $\{-1, 1\}$, and E_σ for expectation. The main result is as follows:

Theorem 6.1. *Let* $F := \left\{ \langle w, x \rangle : \|w\|_p \leq W \right\}$ *and* $\tilde{F} := \left\{ \min_{x' \in \mathrm{B}_x^\infty(\varepsilon)} y \langle w, x' \rangle : \|w\|_p \leq W \right\}$. *Suppose that* $\frac{1}{p} + \frac{1}{q} = 1$. *Then there exists a universal constant* $c \in (0, 1)$ *such that*

$$\frac{c}{2} \left(R_s(F) + \varepsilon W \frac{d^{\frac{1}{q}}}{\sqrt{n}} \right) \leq R_s(\tilde{F}) \leq R_s(F) + \varepsilon W \frac{d^{\frac{1}{q}}}{\sqrt{n}}.$$

In our setting, $F := \{\langle w, \mathrm{sign}(x) \rangle : \|w\|_p \leq W\}$. If for most of the d components in x_i, it holds that $|x_i^{(k)}| > \varepsilon$ (where k indexes the k-th component)—a condition satisfied in typical adversarial attacks where ε is small, then the sign of $x_i^{(k)}$ cannot be flipped under an ℓ_∞-perturbation of radius ε; i.e., $\langle w, \mathrm{sign}(x_i) \rangle \approx \min_{x_i' \in B_{x_i}^\infty(\varepsilon)} \langle w, \mathrm{sign}(x_i') \rangle$. This finding suggests that the empirical Rademacher complexity $R_s(\widetilde{F})$ in the adversarial setting is effectively independent of ε and closely approximates the standard empirical Rademacher complexity $R_s(F)$. Consequently, $R_s(\widetilde{F})$ is significantly reduced, which in turn keeps the test adversarial accuracy close to the training adversarial accuracy—a behavior consistent with our experimental results. Additionally, the dramatic compression of the input space (e.g., from $2^{224 \times 224 \times 3 \times 8}$ to $2^{224 \times 224 \times 3}$ for ImageNet, a compression ratio of $2^{224 \times 224 \times 3 \times 7}$) inherently restricts the feasible adversarial perturbations while preserving semantic information, thereby explaining the strong clean accuracy of our method. The random component in our design facilitates exploration within this compressed space, further enhancing model robustness without resorting to gradient obfuscation. This improvement generalizes from training to test data, as corroborated by the loss landscapes shown in Figure 5. Please refer to Section K for more analysis.

7 FUTURE WORK

Several promising directions remain for future work. First, the clean accuracy of the model could be further improved to maintain strong performance on benign examples. Second, more efficient noise injection schemes should be explored to enhance both l_2 and l_∞ robust accuracy. Third, providing a stronger theoretical robustness guarantee would be valuable for establishing the method's reliability under diverse threat models.

AI USE STATEMENT

In this work, we used generative AI tools solely to improve the English writing of the manuscript, including grammar correction, phrasing, and general language polishing. We have not used generative AI tools for generating synthetic data sets, developing theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims,

assisting in the writing of proofs, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, implementing methods, assisting with translation, cleaning and reformatting datasets, supporting qualitative and thematic data analysis, or interpreting results. These tasks are not applicable to this work.

We have reviewed all AI-assisted work. Specifically, all AI-assisted text was checked and edited by the authors to ensure that the meaning, claims, and technical content remained accurate and appropriate. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Abdullah Al-Dujaili and Una-May O'Reilly. Sign bits are all you need for black-box attacks. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=SygW0TEFwH.

Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square attack: A query-efficient black-box adversarial attack via random search. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (eds.), *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXIII*, volume 12368 of *Lecture Notes in Computer Science*, pp. 484–501. Springer, 2020. doi: 10.1007/978-3-030-58592-1_29. URL https://doi.org/10.1007/978-3-030-58592-1_29.

MaungMaung AprilPyone and Hitoshi Kiya. Block-wise image transformation with secret key for adversarially robust defense. *IEEE Trans. Inf. Forensics Secur.*, 16:2709–2723, 2021. doi: 10.1109/TIFS.2021.3062977.

Anish Athalye, Nicholas Carlini, and David A. Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In Jennifer G. Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pp. 274–283. PMLR, 2018. URL http://proceedings.mlr.press/v80/athalye18a.html.

Yang Bai, Yuyuan Zeng, Yong Jiang, Shu-Tao Xia, Xingjun Ma, and Yisen Wang. Improving adversarial robustness via channel-wise activation suppressing. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=zQTezqCCtNx.

Zhuotong Chen, Qianxiao Li, and Zheng Zhang. Towards robust neural networks via close-loop control. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=2AL06y9cDE-.

Jeremy M. Cohen, Elan Rosenfeld, and J. Zico Kolter. Certified adversarial robustness via randomized smoothing. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1310–1320. PMLR, 2019. URL http://proceedings.mlr.press/v97/cohen19c.html.

Francesco Croce and Matthias Hein. Adversarial robustness against multiple and single l_p-threat models via quick fine-tuning of robust classifiers. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 4436–4454. PMLR, 2022. URL https://proceedings.mlr.press/v162/croce22b.html.

Francesco Croce, Sven Gowal, Thomas Brunner, Evan Shelhamer, Matthias Hein, and A. Taylan Cemgil. Evaluating the adversarial robustness of adaptive test-time defenses. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 4421–4435. PMLR, 2022. URL https://proceedings.mlr.press/v162/croce22a.html.

Sihui Dai, Saeed Mahloujifar, and Prateek Mittal. Formulating robustness against unforeseen attacks. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/392ac56724c133c37d5ea746e52f921f-Abstract-Conference.html.

Sihui Dai, Saeed Mahloujifar, Chong Xiang, Vikash Sehwag, Pin-Yu Chen, and Prateek Mittal. Multirobustbench: Benchmarking robustness against multiple attacks. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 6760–6785. PMLR, 2023. URL https://proceedings.mlr.press/v202/dai23c.html.

Sina Däubener and Asja Fischer. How sampling impacts the robustness of stochastic neural networks. In *NeurIPS*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/429d69979c22b06d6baa65caf3ab1e10-Abstract-Conference.html.

Panagiotis Eustratiadis, Henry Gouk, Da Li, and Timothy M. Hospedales. Weight-covariance alignment for adversarially robust neural networks. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 3047–3056. PMLR, 2021. URL http://proceedings.mlr.press/v139/eustratiadis21a.html.

Panagiotis Eustratiadis, Henry Gouk, Da Li, and Timothy M. Hospedales. Attacking adversarial defences by smoothing the loss landscape. *CoRR*, abs/2208.00862, 2022. doi: 10.48550/ARXIV.2208.00862.

Sven Gowal, Chongli Qin, Jonathan Uesato, Timothy A. Mann, and Pushmeet Kohli. Uncovering the limits of adversarial training against norm-bounded adversarial examples. *CoRR*, abs/2010.03593, 2020. URL https://arxiv.org/abs/2010.03593.

Sven Gowal, Sylvestre-Alvise Rebuffi, Olivia Wiles, Florian Stimberg, Dan Andrei Calian, and Timothy A. Mann. Improving robustness using generated data. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 4218–4233, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/21ca6d0cf2f25c4dbb35d8dc0b679c3f-Abstract.html.

Chuan Guo, Mayank Rana, Moustapha Cissé, and Laurens van der Maaten. Countering adversarial images using input transformations. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=SyJ7ClWCb.

Zhichao Hou, MohamadAli Torkamani, Hamid Krim, and Xiaorui Liu. Robustness reprogramming for representation learning. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=SuH5SdOXpe.

Hanxun Huang, Yisen Wang, Sarah M. Erfani, Quanquan Gu, James Bailey, and Xingjun Ma. Exploring architectural ingredients of adversarially robust deep neural networks. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 5545–5559, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/2bd7f907b7f5b6bbd91822c0c7b835f6-Abstract.html.

Daniel Kang, Yi Sun, Dan Hendrycks, Tom Brown, and Jacob Steinhardt. Testing robustness against unforeseen adversaries. *CoRR*, abs/1908.08016, 2019. URL http://arxiv.org/abs/1908.08016.

Qiyu Kang, Yang Song, Qinxu Ding, and Wee Peng Tay. Stable neural ODE with lyapunov-stable equilibrium points for defending against adversarial attacks. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 14925–14937, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/7d5430cf85f78c4b7aa09813b14bce0d-Abstract.html.

Cassidy Laidlaw, Sahil Singla, and Soheil Feizi. Perceptual adversarial robustness: Defense against unseen threat models. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=dFwBosAcJkN.

Minjong Lee and Dongwoo Kim. Robust evaluation of diffusion-based adversarial purification. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 134–144. IEEE, 2023. doi: 10.1109/ICCV51070.2023.00019.

Sungyoon Lee, Hoki Kim, and Jaewook Lee. Graddiv: Adversarial robustness of randomized neural networks via gradient diversity regularization. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(2): 2645–2651, 2023. doi: 10.1109/TPAMI.2022.3169217.

Yandong Li, Lijun Li, Liqiang Wang, Tong Zhang, and Boqing Gong. NATTACK: learning the distributions of adversarial examples for an improved black-box attack on deep neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3866–3876. PMLR, 2019. URL http://proceedings.mlr.press/v97/li19g.html.

Aishan Liu, Shiyu Tang, Siyuan Liang, Ruihao Gong, Boxi Wu, Xianglong Liu, and Dacheng Tao. Exploring the relationship between architectural design and adversarially robust generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4096–4107, June 2023a.

Zhuoran Liu, Zhengyu Zhao, and Martha A. Larson. Image shortcut squeezing: Countering perturbative availability poisons with compression. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 22473–22487. PMLR, 2023b. URL https://proceedings.mlr.press/v202/liu23bb.html.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=rJzIBfZAb.

Pratyush Maini, Eric Wong, and J. Zico Kolter. Adversarial robustness against the union of multiple perturbation models. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6640–6650. PMLR, 2020. URL http://proceedings.mlr.press/v119/maini20a.html.

Chengzhi Mao, Mia Chiquier, Hao Wang, Junfeng Yang, and Carl Vondrick. Adversarial attacks are reversible with natural supervision. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pp. 641–651. IEEE, 2021. doi: 10.1109/ICCV48922.2021.00070.

Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Animashree Anandkumar. Diffusion models for adversarial purification. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 16805–16827. PMLR, 2022. URL https://proceedings.mlr.press/v162/nie22a.html.

Tianyu Pang, Kun Xu, and Jun Zhu. Mixup inference: Better exploiting mixup to defend adversarial attacks. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=ByxtC2VtPB.

Jary Pomponi, Simone Scardapane, and Aurelio Uncini. Pixle: a fast and effective black-box attack based on rearranging pixels. In *International Joint Conference on Neural Networks, IJCNN 2022, Padua, Italy, July 18-23, 2022*, pp. 1–7. IEEE, 2022. doi: 10.1109/IJCNN55064.2022.9892966.

Edward Raff, Jared Sylvester, Steven Forsyth, and Mark McLean. Barrage of random transforms for adversarially robust defense. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 6528–6537. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00669. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Raff_Barrage_of_Random_Transforms_for_Adversarially_Robust_Defense_CVPR_2019_paper.html.

Sylvestre-Alvise Rebuffi, Sven Gowal, Dan A. Calian, Florian Stimberg, Olivia Wiles, and Timothy A. Mann. Fixing data augmentation to improve adversarial robustness. *CoRR*, abs/2103.01946, 2021. URL https://arxiv.org/abs/2103.01946.

Leslie Rice, Eric Wong, and J. Zico Kolter. Overfitting in adversarially robust deep learning. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 8093–8104. PMLR, 2020. URL http://proceedings.mlr.press/v119/rice20a.html.

Jérôme Rony, Luiz G. Hafemann, Luiz S. Oliveira, Ismail Ben Ayed, Robert Sabourin, and Eric Granger. Decoupling direction and norm for efficient gradient-based L2 adversarial attacks and defenses. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 4322–4330. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00445. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Rony_Decoupling_Direction_and_Norm_for_Efficient_Gradient-Based_L2_Adversarial_Attacks_CVPR_2019_paper.html.

Andrei A. Rusu, Dan Andrei Calian, Sven Gowal, and Raia Hadsell. Hindering adversarial attacks with implicit neural representations. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 18910–18934. PMLR, 2022. URL https://proceedings.mlr.press/v162/rusu22a.html.

Ahmed H. Salamah, Kaixiang Zheng, Yiwen Liu, and En-Hui Yang. JPEG inspired deep learning. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net. URL https://openreview.net/forum?id=te2IdORabL.

Pouya Samangouei, Maya Kabkab, and Rama Chellappa. Defense-gan: Protecting classifiers against adversarial attacks using generative models. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=BkJ3ibb0-.

Changhao Shi, Chester Holtz, and Gal Mishne. Online adversarial purification based on self-supervised learning. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=_i3ASPp12WS.

Naman D. Singh, Francesco Croce, and Matthias Hein. Revisiting adversarial training for imagenet: Architectures, training and generalization across threat models. *CoRR*, abs/2303.01870, 2023. doi: 10.48550/arXiv.2303.01870.

Vasu Singla, Sahil Singla, Soheil Feizi, and David Jacobs. Low curvature activations reduce overfitting in adversarial training. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pp. 16403–16413. IEEE, 2021. doi: 10.1109/ICCV48922.2021.01611.

Chawin Sitawarin, Zachary J. Golan-Strieb, and David A. Wagner. Demystifying the adversarial robustness of random transformation defenses. In Kamalika Chaudhuri, Stefanie Jegelka,

Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 20232–20252. PMLR, 2022. URL https://proceedings.mlr.press/v162/sitawarin22a.html.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In Yoshua Bengio and Yann LeCun (eds.), *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. URL http://arxiv.org/abs/1312.6199.

Florian Tramèr and Dan Boneh. Adversarial training and robustness for multiple perturbations. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 5858–5868, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/5d4ae76f053f8f2516ad12961ef7fe97-Abstract.html.

Haotao Wang, Aston Zhang, Shuai Zheng, Xingjian Shi, Mu Li, and Zhangyang Wang. Removing batch normalization boosts adversarial training. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 23433–23445. PMLR, 2022. URL https://proceedings.mlr.press/v162/wang22ap.html.

Boxi Wu, Jinghui Chen, Deng Cai, Xiaofei He, and Quanquan Gu. Do wider neural networks really help adversarial robustness? In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 7054–7067, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/3937230de3c8041e4da6ac3246a888e8-Abstract.html.

Shutong Wu, Sizhe Chen, Cihang Xie, and Xiaolin Huang. One-pixel shortcut: On the learning preference of deep neural networks. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/pdf?id=p7G8t5FVn2h.

Cihang Xie, Yuxin Wu, Laurens van der Maaten, Alan L. Yuille, and Kaiming He. Feature denoising for improving adversarial robustness. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 501–509. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00059. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Xie_Feature_Denoising_for_Improving_Adversarial_Robustness_CVPR_2019_paper.html.

Weilin Xu, David Evans, and Yanjun Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. In *25th Annual Network and Distributed System Security Symposium, NDSS 2018, San Diego, California, USA, February 18-21, 2018*. The Internet Society, 2018. URL http://wp.internetsociety.org/ndss/wp-content/uploads/sites/25/2018/02/ndss2018_03A-4_Xu_paper.pdf.

En-Hui Yang, Shayan Mohajer Hamidi, Linfeng Ye, Renhao Tan, and Beverly Yang. Conditional mutual information constrained deep learning for classification. 36(8):15436–15448. doi: 10.1109/TNNLS.2025.3540014.

Yuzhe Yang, Guo Zhang, Zhi Xu, and Dina Katabi. Me-net: Towards effective adversarial robustness with matrix estimation. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 7025–7034. PMLR, 2019. URL http://proceedings.mlr.press/v97/yang19e.html.

Dong Yin, Kannan Ramchandran, and Peter L. Bartlett. Rademacher complexity for adversarially robust generalization. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 7085–7094. PMLR, 2019. URL http://proceedings.mlr.press/v97/yin19b.html.

Jongmin Yoon, Sung Ju Hwang, and Juho Lee. Adversarial purification with score-based generative models. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12062–12072. PMLR, 2021. URL http://proceedings.mlr.press/v139/yoon21a.html.

Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Seong Joon Oh, Youngjoon Yoo, and Junsuk Choe. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pp. 6022–6031. IEEE, 2019. doi: 10.1109/ICCV.2019.00612.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In Richard C. Wilson, Edwin R. Hancock, and William A. P. Smith (eds.), *Proceedings of the British Machine Vision Conference 2016, BMVC 2016, York, UK, September 19-22, 2016*. BMVA Press, 2016. URL http://www.bmva.org/bmvc/2016/papers/paper087/index.html.

Haichao Zhang and Jianyu Wang. Defense against adversarial attacks using feature scattering-based adversarial training. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 1829–1839, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/d8700cbd38cc9f30cecb34f0c195b137-Abstract.html.

Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=r1Ddp1-Rb.

Jingfeng Zhang, Jianing Zhu, Gang Niu, Bo Han, Masashi Sugiyama, and Mohan S. Kankanhalli. Geometry-aware instance-reweighted adversarial training. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=iAX0l6Cz8ub.

A FEATURE MAP

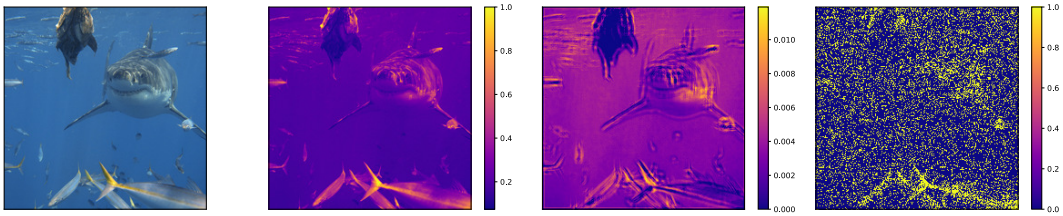

Figure 3: From the left to right are the great-white-shark x; the red channel; and the corresponding sampling matrix $\hat{x}$ and the final output x' where the continuous patterns are highly squeezed into two extreme values, 0 and 1, due to very small $\hat{x}$. Blue and green channels share a similar situation.

B SOURCE CODES AND PRE-TRAINED MODELS

1. CIFAR-10
   `https://rfsq.obs.cn-north-4.myhuaweicloud.com/cifar.zip`
2. ImageNet
   `https://rfsq.obs.cn-north-4.myhuaweicloud.com/imagenet.zip`

C l_0 ATTACK

Table 5: l_0 Robustness comparison. There is no specific constrains for l_0, and last column shows the average of l_0 of perturbed samples; while the iterations column shows the average number of iterations by the attack.

Defense	Clean	Robust	Iterations	l_0
CIFAR-10		l_0	l_0	
Ours	81.02	61.10	250	122
PATLaidlaw et al. (2021)	82.30	12.00	122	101
AT-VRDai et al. (2022)	73.00	11.40	109	92
ImageNet		l_0	l_0	
Ours	68.80	67.70	5000	1468
Singh et al. (2023)	73.40	33.40	3227	1706

D OPS

Table 6: Performance under One-Pixel Shortcut on ResNet-18 for different training strategies. The first two rows are excerpted from Wu et al. (2023). l_∞ AT stands for adversarial training with l_∞=8/255.

Training Strategy	Clean	OPS
Standard	94.01	15.56
l_∞ AT	82.72	11.08
Ours	82.03	56.25

E BPDA FOR RANDOM MODEL

Table 7: Comparison between the different BPDA. We choose the worst ones (in bold) as BPDA for tests of our defense. $\mathcal{S}$ stands for Sigmoid. The attacks of APGD including DLR loss are much weaker than PGD ones. Please note that without an explicit statement, we adopt cross-entropy loss in this paper.

DataSet	Attack	x	$\frac{\text{softsign}(x)+1}{2}$	$\mathcal{S}(x)$	$\mathcal{S}(3x)$	$\mathcal{S}(5x)$	$\mathcal{S}(7x)$	$\mathcal{S}(9x)$
CIFAR-10	APGD-l_1	63.50	**57.00**	57.60	60.00	65.00	66.50	69.00
	PGD-l_2	43.10	**36.40**	39.40	37.60	39.90	38.00	38.90
	APGD-l_2	51.50	52.90	52.20	51.80	51.20	52.20	49.70
	DDN-l_2	58.90	57.10	58.00	56.60	57.40	55.80	**54.30**
	APGD^{DLR}-l_2	56.20	58.00	55.70	56.20	57.80	55.50	55.30
	PGD-l_∞	41.90	42.50	41.20	**41.10**	41.70	42.60	42.10
	APGD-l_∞	53.70	55.40	54.40	53.40	54.70	54.60	57.40
	APGD^{DLR}-l_∞	60.20	59.80	59.60	59.00	58.40	59.50	57.80
ImageNet	APGD-l_1	**58.20**	61.90	59.20	64.10	65.80	66.50	67.10
	PGD-l_2	41.70	**38.90**	40.20	39.40	42.40	43.60	44.60
	APGD-l_2	49.20	50.60	47.90	51.80	55.10	55.50	55.60
	DDN-l_2	53.40	54.70	53.40	54.10	**52.50**	53.20	54.80
	APGD^{DLR}-l_2	50.40	53.80	50.90	55.20	58.20	58.90	60.50
	PGD-l_∞	23.70	24.70	24.50	**23.50**	25.40	25.30	23.60
	APGD-l_∞	34.10	33.10	33.40	34.80	34.80	32.70	35.50
	APGD^{DLR}-l_∞	37.30	40.40	39.30	37.20	39.10	39.40	38.20

F TRANSFER ATTACK

The transfer attack is launched using the training framework, which is weaker than a direct attack on the test framework, except for its closeness with PGD-l_∞.

Table 8: Robustness against EOT-20 transfer attack.

Defense	APGD-l_1	PGD-l_2	PGD-l_∞
CIFAR-10			
Transfer-EOT20	55.70	51.90	34.30
Normal-EOT20	38.50	33.90	34.00
ImageNet			
Transfer-EOT20	52.00	47.50	18.70
Normal-EOT20	46.90	37.90	18.80

G DETOUR TRAINING

The key distinction between the standard test framework training and detour learning lies in the initialization strategy and early-phase optimization dynamics. In training with the test framework where parameters are randomly initialized, x' saturates to binary values (0/1) at the start. This premature squeezing limits the model's exploration capacity, potentially trapping it in suboptimal local minima. By contrast, detour learning introduces a warm-up phase where x' is not highly squeezed to be 0 or 1, allowing parameters to discover better initialization regions. After that point, x' is highly squeezed to 0/1, so it does the training with the test framework to the same effect.

The following Figure 4 shows that although the cross-entropy loss shows similar trends, the sparse loss is very different, slowly for CIFAR-10 drops while quick for ImagetNet. This sparse loss evolution indicates that detour training for ImageNet is closer to direct training possibly due to massive image size and ConvNeXt-T+ConvStem structure, which leads to only marginal improvement.

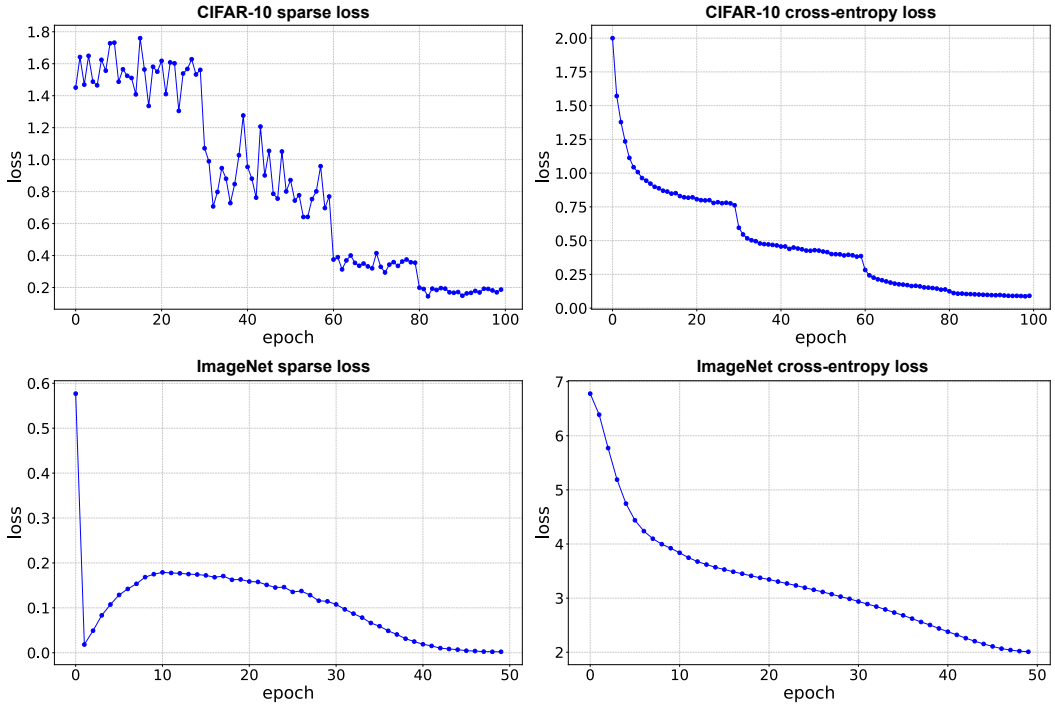

Figure 4: From the left to right are the sparse loss and cross-entropy loss.

H ADDITIONAL ABLATION STUDIES

Table 9: Additional ablation studies for ImageNet. The ones in bold correspond to the models used in our paper.

kernel_size	α	σ	**Clean**	**PGD-l_∞**
7×7	**0.5**	**1.4**	**67.80**	**23.74**
5×5	0.5	1.4	67.30	25.20
9×9	0.5	1.4	66.94	24.06
3×3	0.5	1.2	68.04	22.24
5×5	0.5	1.2	68.82	22.84

Table 10: Additional ablation studies for CIFAR-10. The ones in bold correspond to the models used in our paper. Our initial experiments indicated that a 9x9 kernel yielded the highest robust accuracy for CIFAR-10 under a PGD-l_∞ white-box attack. However, a more comprehensive evaluation revealed that these models did not maintain their superiority under the gradient-free NAttack for the deterministic models explained in Section 6. Consequently, we adopt a rigorously conservative evaluation framework to ensure the sustainability of our results.

kernel_size	α	σ	**Clean**	**PGD-l_∞**	**Nattack Li et al. (2019) Determinstic N=5/20/30**
5×5	**0.1**	**0.65**	**80.14**	**43.30**	**19.00/28.00/28.00**
5×5	0.1	0.6	80.92	40.40	–
5×5	0.1	0.7	79.50	42.20	20.00/25.00/26.00
5×5	0.2	0.7	79.69	46.80	–
5×5	0.3	0.7	78.99	45.20	–
5×5	0.4	0.7	68.97	42.00	–
5×5	0.5	0.7	60.34	47.20	–
7×7	0.2	0.7	79.71	51.90	–
9×9	0.2	0.7	79.94	55.80	19.00/24.00/25.00
11×11	0.2	0.7	79.83	53.30	–

I BPDA FOR DETERMINISTIC MODEL

Table 11: Accuracy for APGD attack on 1k images across training and test sets (in bold), with varying N. $\mathcal{S}(.)$ denotes Sigmoid for different BPDA configurations. As the model is deterministic, APGD can be safely applied for $l_{1,2,\infty}$ attacks. Further details are provided in Table 12. Notably, on ImageNet, our method achieves higher l_1 accuracy than Singh et al. (2023).

N	Clean	APGD-l_1	APGD-l_2	APGD-l_∞
CIFAR-10		$\mathcal{S}(9x)$	$\mathcal{S}(19x)$	$\mathcal{S}(19x)$
5	99.80 **83.80**	10.20 **6.90**	4.60 **2.80**	26.00 **20.30**
20	99.90 **84.90**	22.40 **17.50**	15.00 **12.20**	30.80 **24.80**
30	99.90 **85.60**	24.20 **19.20**	20.30 **16.40**	32.50 **25.80**
ImageNet		$\mathcal{S}(25x)$	$\mathcal{S}(25x)$	$\mathcal{S}(25x)$
5	84.70 **70.20**	7.30 **6.70**	0.40 **0.10**	3.50 **3.10**
20	85.40 **70.30**	37.20 **27.40**	7.80 **6.70**	5.90 **5.30**
30	85.50 **70.10**	45.60 **32.70**	13.70 **10.90**	7.00 **6.90**

Table 12: Comparison between the different BPDA on 1k images for deterministic model with $N = 5$. We choose the worst ones (in bold) as BPDA for tests of our defense. $\mathcal{S}$ stands for Sigmoid.

DataSet	Attack	x	$\frac{\text{softsign}(x)+1}{2}$	$\mathcal{S}(x)$	$\mathcal{S}(3x)$	$\mathcal{S}(5x)$	$\mathcal{S}(7x)$
CIFAR-10	APGD-l_1	32.70	10.70	22.20	11.00	9.10	7.70
	APGD-l_2	34.60	14.00	28.50	17.20	9.80	6.60
	APGD-l_∞	27.60	23.70	26.70	24.10	22.60	22.20
		$\mathcal{S}(9x)$	$\mathcal{S}(13x)$	$\mathcal{S}(17x)$	$\mathcal{S}(19x)$	$\mathcal{S}(21x)$	
CIFAR-10	APGD-l_1	**6.90**	7.20	7.30	7.40	8.20	
	APGD-l_2	5.00	3.30	2.90	**2.80**	3.40	
	APGD-l_∞	21.50	20.70	20.80	**20.30**	20.90	
		x	$\frac{\text{softsign}(x)+1}{2}$	$\mathcal{S}(x)$	$\mathcal{S}(3x)$	$\mathcal{S}(5x)$	$\mathcal{S}(7x)$
ImagNet	APGD-l_1	35.60	18.00	27.60	18.70	15.30	12.10
	APGD-l_2	32.60	11.30	26.10	14.00	8.60	5.90
	APGD-l_∞	10.10	5.40	8.20	5.50	4.60	4.00
		$\mathcal{S}(9x)$	$\mathcal{S}(15x)$	$\mathcal{S}(20x)$	$\mathcal{S}(25x)$	$\mathcal{S}(30x)$	
ImagNet	APGD-l_1	10.40	7.40	6.90	**6.70**	7.30	
	APGD-l_2	3.70	0.60	0.30	**0.10**	0.10	
	APGD-l_∞	3.90	3.40	3.20	**3.10**	3.10	

J LOSS LANDSCAPES

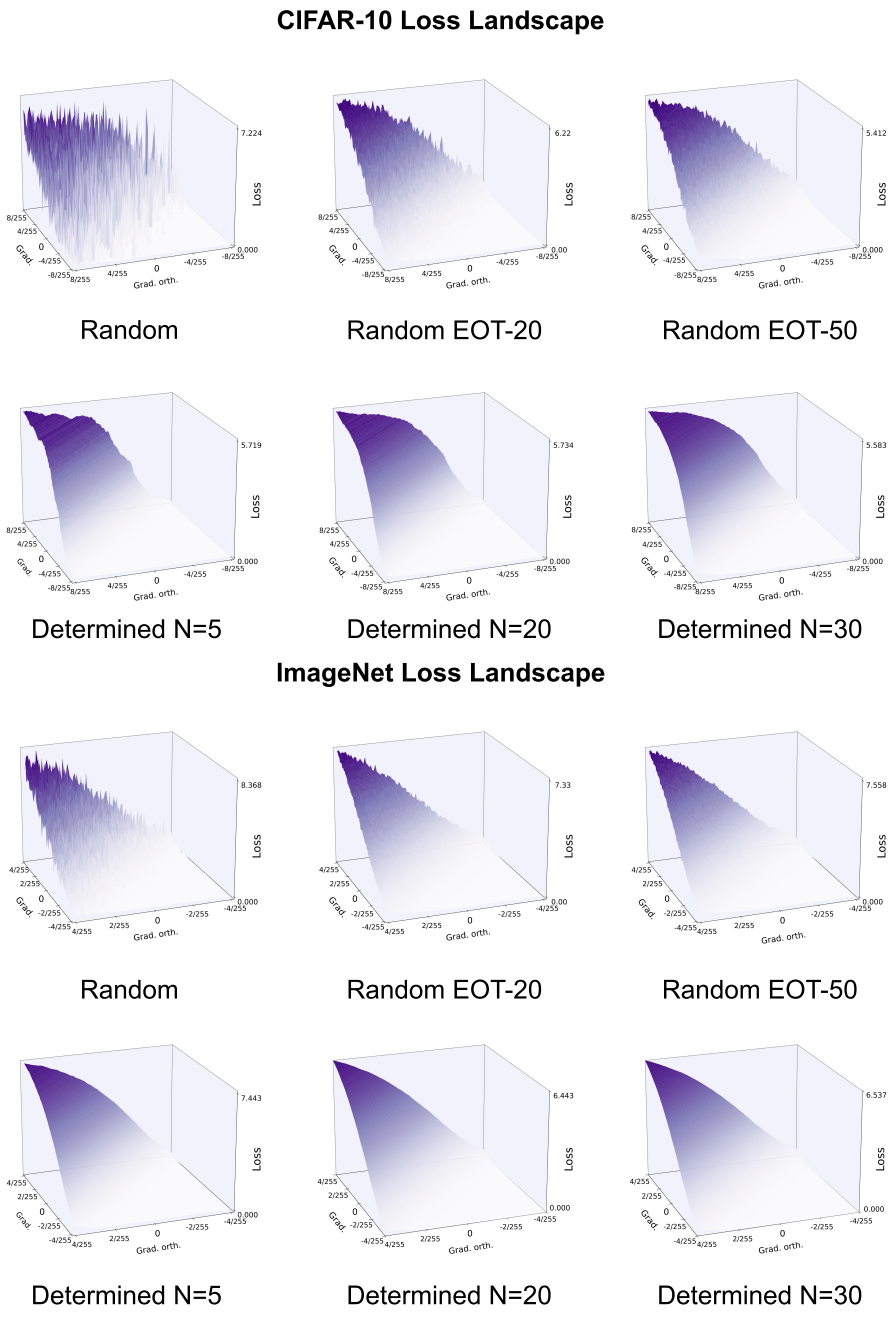

Figure 5: The loss landscapes generated using the code adapted from Eustratiadis et al. (2022) show that although a random version of our network exhibits a rough surface, it becomes smoother as the iterations of EOT increase. The determined versions are smooth. It suggests that EOT+BPDA is enough to give a robust evaluation without the risk of overestimation.

K MORE ANALYSIS REGARDING OBFUSCTED GRADIENTS

Our test-time model is:

$$f_{\text{test}}(x) = F_{\text{net}}(\text{Sign}(x + \epsilon))$$

where the underlying classifier F_{net} is defined on the discrete set $\{0,1\}^d$. The Sign function maps any $x \in \mathbb{R}^d$ into this discrete set, so f_{test} can be evaluated on the entire $\mathbb{R}^d$. However, the effective decision structure of f_{test} is entirely determined by its values on $\{0,1\}^d$—the continuous space between these discrete points contains no additional decision information.

The key observation is:

> ***Any continuous function g that agrees with F_{net} on the discrete set $\{0,1\}^d$ is a valid continuous surrogate for f_{test}.***

This means:

- The function f_{test} is piecewise constant on $\mathbb{R}^d$: it only changes value when the Sign output flips.
- The gradient of f_{test} is zero almost everywhere—not because it is "masked," but because the function is flat in those regions.
- Different BPDA surrogates simply choose different ways to interpolate between the discrete points. Their variation in attack success (Table 8) reflects different interpolation choices, not gradient obfuscation.

In the stochastic version of our defense, we do not fix the noise ϵ; instead, each forward pass draws a single fresh sample from $\mathcal{N}(0, \sigma^2)$.

From the discrete-space perspective, this means:

> ***For each input, the random noise effectively causes the discrete point*** $\text{Sign}(x + \epsilon)$ ***to "hop" among neighboring discrete points with certain probabilities.***

Crucially, the output is not an expectation over noise—it is a single sample from this distribution:

$$f_{\text{test}}(x; \epsilon) = F_{\text{net}}(\text{Sign}(x + \epsilon)), \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

Each forward pass yields a different discrete point, and therefore potentially a different classification output.

This has two important consequences for the attacker:

1. The attack must now cope with stochasticity: When the attacker queries the model, they receive a different output each time. This makes gradient estimation noisy and optimization unstable.
2. The surrogate must approximate both the discrete function and the randomness: Any continuous proxy must account for the fact that the discrete point is sampled stochastically. This is a strictly harder approximation problem than the deterministic case.

This explains the accuracy drop from the stochastic model (Table 3) to the deterministic model (Table 11): when we fix the noise (deterministic), we remove this "hopping" behavior, giving the attacker a simpler, more stable optimization target. The drop is therefore expected and does not indicate gradient masking—it is a direct consequence of the discrete decision structure and the role of randomness in complicating the attacker's optimization.