Randomized Smoothing under Attack: How Good is it in Pratice?

Thibault Maho, Teddy Furon, Erwan Le Merrer

RANDOMIZED SMOOTHING UNDER ATTACK: HOW GOOD IS IT IN PRACTICE?

Thibault Maho, Teddy Furon, Erwan Le Merrer*

Univ. Rennes, Inria, CNRS
IRISA, Rennes, France

ABSTRACT

Randomized smoothing is a recent and celebrated solution to certify the robustness of any classifier. While it indeed provides a theoretical robustness against adversarial attacks, the dimensionality of current classifiers necessarily imposes Monte Carlo approaches for its application in practice.

This paper questions the effectiveness of randomized smoothing as a *defense*, against state of the art black-box attacks. This is a novel perspective, as previous research works considered the certification as an unquestionable guarantee. We first formally highlight the mismatch between a theoretical certification and the practice of attacks on classifiers. We then perform attacks on randomized smoothing as a defense. Our main observation is that there is a major mismatch in the settings of the RS for obtaining high certified robustness or when defeating black box attacks while preserving the classifier accuracy.

***Index Terms*—** Classifiers, black-box attacks, randomized smoothing, randomized adversarial examples.

1. INTRODUCTION

The adoption of neural network-based classifiers has been crucial for performance in multiple security-sensitive fields: self-driving cars, face recognition, or alert detection to name a few. This success is unfortunately hampered by their vulnerability [1] to a wide list of so called adversarial attacks. This is particularly critical in decision-based black-box attacks [2, 3, 4, 5, 6] where the adversary still manages to lead powerful attacks although a minimal set of assumptions is granted. As far as image classification is concerned, the adversarial perturbations are close to invisible to humans.

Defenses have been proposed to increase the robustness of these classifiers against such attacks. Adversarial Training [7, 8] for instance involves retraining with adversarial inputs; such a defense is effective but costly. Other approaches reform the inputs to remove the perturbation [9]. Some others are adapted to specific domains such as time-series analysis, and frame the problem as an instance of outlier detection [10, 11].

*Thanks to ANR and AID french agencies for funding Chaire SAIDA.

A novel proposal is to *certify* the robustness of classifiers, with in particular randomized smoothing [12, 13, 14, 15]. Certification is a general and model-agnostic paradigm, which can be applied without additional retraining. Its advantage is to theoretically certify a level of robustness to attacks, with a correctness guarantee for the elected label in some radius around inputs sent to the classifier.

Randomized smoothing is with no doubt an important advance to approach the robustness of classifiers. Nevertheless, its application as a defense (and not only as a theoretical guarantee) comes with blind spots: *i)* The exact certified robustness is impossible to compute due to the dimensionality of the input space handled by current classifiers. Monte Carlo methods are used to estimate this certified robustness. There is a lack of understanding of the interplay between the theoretical certification for a radius that is fully spanned, and the practice where a limited amount of samples is key to tractability. In addition, since this defense is randomized in essence, the classic definition of an adversarial [1] is not applicable anymore. The defender lacks a definition of an adversary in the case of her randomized defense (since no attack trials is 100% adversarial). *ii)* The amont of samples required by this sampling approach is unclear and varying in the papers: between 100 [16, 17] and $100\,000$ [15, 18]. No results to date have shown the importance in this quantity on the effectiveness of attacks. *iii)* Finally, although this defense is in principle applicable without retraining, it is yet recommended [15] to mitigate the accuracy drop involved. Indeed, the larger the noise radius certified, the more robust the classifier, at the price of an important accuracy drop that can be limited by retraining on noisy data. The relation of the radius with the final accuracy and the effectiveness of the attacks is also unclear.

This paper makes the contribution to tackle these three issues, in a dedicated attempt to consider randomized smoothing as a practical defense. We first confront theory and practice for certification and defense in the context of randomized smoothing. We then evaluate the practical robustness of this defense with regards to the impact of the Monte Carlo sample sizes and the noise variance parameter. This study highlights the effectiveness of randomized smoothing in defeating state of the art black-box attacks with much smaller parameters that are suggested in the papers limited to considering it as a mere theoretical certification only.

2. RELATED WORK

2.1. Black-box attacks

Decision-based black-box attacks were first studied by Brendel *et al*. in [19] as the ultimate attack because it only relies on the decision (*i.e*. the top-1 label) that is returned by the model. The attack `HSJA` [2] significantly improves the efficiency by building surrogates of gradient: It estimates the gradient at a point on the boundary by bombarding the model with noisy version of this sensitive point. Papers [3, 4] improve the results by working in the frequency domain. `SurFree` [5] does not rely on gradient estimation but on a geometric approximation of the boundary. These are decision-based ℓ_2-attacks in the sense that their goal is to find adversarial examples with minimum ℓ_2 norm perturbation. Attacks for the ℓ_∞ norm also exist like `RayS` [6] which is also one the best for ℓ_2. This paper leverages these three attacks to test the practical robustness of randomized smoothing.

2.2. Randomized smoothing (RS)

Introduced by Lecuyer *et al*. [13], RS is a model-agnostic method to obtain a certified local robustness of a model. It guarantees a correct prediction within a certain radius around a given input. In other words, it certifies that no adversarial example lies at a distance smaller than this radius. The beauty of this literature is that the adversarial attacks are no longer considered since the robustness is formally guaranteed. In practice, RS is merely a Monte Carlo simulation requiring a large number of calls to the model. It produces a lower bound of the robustness and spoils the accuracy. Papers [14, 15, 20] found stronger bounds while [13, 16] improved the noise tolerance of classifiers. None of them considers attacking RS.

3. RANDOMIZED SMOOTHING: FROM THEORY TO PRACTICE

This section summarizes RS certification for ℓ_2 norm robustness, focusing on the differences between theory and practice.

3.1. A primer on random smoothing

For the sake of simplicity, consider a trained binary classifier $f : \mathbb{R}^d \to \{0, 1\}$. RS defines a new classifier g_σ as follows:

$$g_\sigma(\mathbf{x}) = \arg\max_{y\in\{0,1\}} \mathbb{P}[f(\mathbf{x}+\sigma\mathbf{N}) = y],\ \mathbf{N} \sim \mathcal{N}(0, I). \quad (1)$$

The main advantage of g_σ is that its robustness is certified. Assume that a genie reveals the value of the two probabilities $\pi_0(\mathbf{x}) := \mathbb{P}[f(\mathbf{x} + \sigma\mathbf{N}) = 0]$ and $\pi_1(\mathbf{x}) := 1 - \pi_0(\mathbf{x})$, then $\mathbf{x}$ is classified according to (1) with a certified robustness:

$$R(\mathbf{x}, \sigma) = \sigma\Phi^{-1}(\pi_{g_\sigma(\mathbf{x})}(\mathbf{x})). \quad (2)$$

where Φ is cumulative distribution function of the standard normal distribution $\mathcal{N}(0, I)$. All points at a distance from $\mathbf{x}$ lower than $R(\mathbf{x}, \sigma)$ are classified in the same way. Note that despite the term 'randomized smoothing', g_σ is indeed a deterministic classifier. Its frontier ∂g_σ is the locus of the points s.t. $\pi_0(\mathbf{x}) = 1/2$. For instance, if the base classifier f is linear, then $g_\sigma = f$.

In practice, there is no genie and the defender uses a Monte Carlo simulation over n random i.i.d. samples $\{\mathbf{n}_i\}_{i=1}^n$ distributed as $\mathbf{N}$ yielding n decisions $\{y_i\}_{i=1}^n$. The final predicted class is an aggregation of these n 'micro'-decisions s.a. the majority vote. They also give a confidence interval $\underline{\pi}_0(\mathbf{x}) < \pi_0(\mathbf{x})$ up to a given confidence level. This defines the classifier $g_{\sigma,n}$, a practical implementation of ideal g_σ function. The robustness is assessed up to the confidence level using (2) with $\underline{\pi}_0$, which yields $\underline{R}(\mathbf{x}, \sigma) < R(\mathbf{x}, \sigma)$. Maximizing the certified robustness around a given point $\mathbf{x}$ with a high confidence level requires large n and σ [13, 15, 16].

3.2. A critical point of view

The main argument of RS is the following: Leading adversarial attacks to gauge the security of a classifier is no longer needed since its robustness is certified. This has to be mitigated: $\underline{R}(\mathbf{x}, \sigma)$ certifies the robustness of the theoretical classifier g_σ which does not exist in practice. The practical classifier $g_{\sigma,n}$ behaves as g_σ only when $n \to \infty$.

More importantly, $g_{\sigma,n}$ *is not a deterministic function*. For $\mathbf{x} \in \partial g_\sigma$, $g_{\sigma,n}(\mathbf{x})$ acts as a random variable from one call to another since $\pi_0(\mathbf{x}) = \pi_1(\mathbf{x}) = 1/2$ even for large n. This challenges the concept of frontiers whence the definition of adversarial examples.

3.3. Pushing the frontiers

Consider a point $\mathbf{x}$ s.t. $f(\mathbf{x}) = 1$ and at a distance $\delta = \beta\sigma$ from the frontier ∂f of the base classifier. The so-called SORM in statistical reliability engineering approximates

$$\pi_0(\mathbf{x}) \approx \Phi(-\beta) \prod_{i=1}^{d-1} \frac{1}{\sqrt{1 + \beta\kappa_i}}, \quad (3)$$

where $\{\kappa_i\}$ are the signed principal curvatures of the surface ∂f. If flat, all the curvatures equal 0, and $\mathbf{x}$ lies on the boundary ∂g_σ of the ideal RS classifier if $\pi_0(\mathbf{x}) = 1/2$ implying $\delta = 0$. If ∂f is convex onward $\mathbf{x}$, the curvatures are all negatives, the second term gets larger and compensates $\Phi(-\beta)$ so that $\pi_0(\mathbf{x}) = 1/2$ for some $\beta > 0$. This shows that the frontier ∂g_σ is closer than ∂f when lying in a convexe region, and thus further away when sitting in a concave region. If the original images lie in concave regions, then RS pushes the frontier and thus increases the norm of the adversarial perturbation. In Fig. 1, a white-box attack against model f first finds an adversarial $\mathbf{x}_a \in \partial f$. We see that going along the direction $\mathbf{x}_a - \mathbf{x}_o$, we cross the frontier ∂g_σ (*i.e*. $\pi_0(\mathbf{x}) = 0.5$)

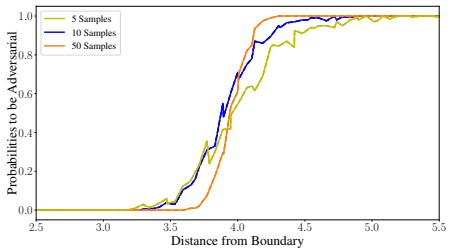
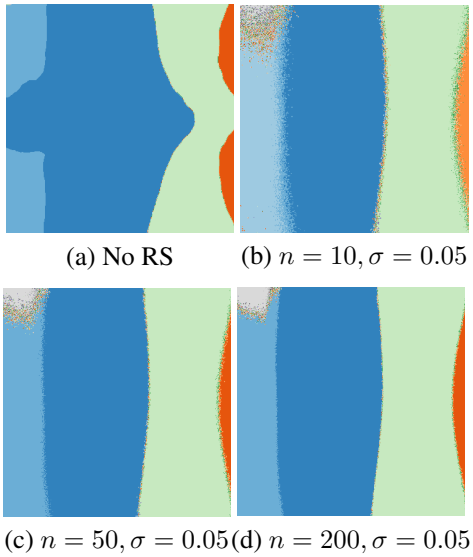


Fig. 1. Probability of being adversarial by following the direction $\mathbf{x}_a - \mathbf{x}_o$, Image x_o attacked with BP [21] to get the best adversarial x_a on the boundary of `ResNet50`.



(a) No RS (b) $n = 10, \sigma = 0.05$

(c) $n = 50, \sigma = 0.05$ (d) $n = 200, \sigma = 0.05$

Fig. 2. 2D slice in the image space of `ResNet50` with and without RS. Each point is an image, his color represents the elected label.

after $\mathbf{x}_a \in \partial f$ for a $\mathbf{x}$ s.t. $\|\mathbf{x} - \mathbf{x}_a\| \approx 4.0$. This is also illustrated in Fig. 2 on a 2D cut of $\mathbb{R}^d$.

3.4. Adversarial example with confidence level

We propose a new definition for untargeted attack: an adversarial example of $\mathbf{x}_o$ of level $P_a \in [0, 1]$ is a point $\mathbf{x}_a$ s.t.

$$\mathbb{P}[g_{\sigma,n}(\mathbf{x}_a) \neq g_\sigma(\mathbf{x}_o)] \geq P_a. \tag{4}$$

If the attacker is satisfied with a level $P_a = 1/2$, then the closest adversarial example lies on the frontier of g_σ at a distance bigger than $R(\mathbf{x}, \sigma) > \underline{R}(\mathbf{x}, \sigma)$. We believe that attackers are requiring stronger guarantee $P_a > 1/2$, hence the closest adversarial example is at an even bigger distance. Eq. (4) requires that $\sum y_i \sim \mathcal{B}(n, 1 - \pi_0(\mathbf{x}_a))$ takes a value greater than $n/2$ (due to the majority vote) with a probability larger than P_a. This holds for:

$$\pi_0(\mathbf{x}_a) < 1 - I^{-1}_{P_a}(\tilde{n}, \tilde{n}) < 1/2, \tag{5}$$

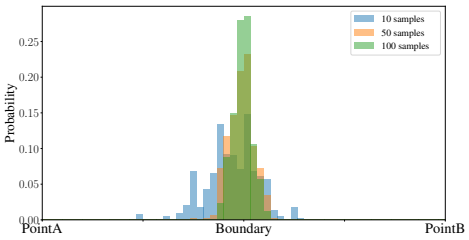


Fig. 3. Distribution of the output of a binary search with RS.

with $\tilde{n} = 1 + \lfloor n/2 \rfloor$ and $I_p^{-1}(a, b)$ is the inverse incomplete beta function. Applying (2) onto $\mathbf{x}_o$ and $\mathbf{x}_a$, it comes that

$$\begin{aligned} \|\mathbf{x}_o - \mathbf{x}_a\| &= \|\mathbf{x}_o - \mathbf{x}_b\| + \|\mathbf{x}_b - \mathbf{x}_a\| \\ &\geq R(\mathbf{x}_o, \sigma) + \sigma\Phi^{-1}(I^{-1}_{P_a}(\tilde{n}, \tilde{n})), \end{aligned} \tag{6}$$

where $\mathbf{x}_b \in [\mathbf{x}_o, \mathbf{x}_a] \cap \partial g_\sigma$. To conclude, the robustness $\underline{R}(\mathbf{x}_o, \sigma)$ certified by the practical implementation of RS is even less tight in practice.

3.5. Jeopardizing black box attacks

Black box attacks usually make two assumptions. First, from a point outside the class region, a binary search can find a point $\mathbf{x}_b$ right on the boundary within a controlled accuracy. However, RS classifier $g_{\sigma,n}$ is random in practice especially when n is small and this jeopardizes the binary search. Fig. 3 shows the distribution of the result of the binary search. It concentrates around ∂g_σ only when n is large.

Second, the boundary is smooth so that it is possible to estimate the normal vector of the tangent hyperplan locally around $\mathbf{x}_b$ on the boundary. This is usually done by bombarding the classifier with noisy versions of $\mathbf{x}_b$ and observing its outputs. Yet, RS randomizes the immediate neighborhood of boundaries, as seen in Fig. 2. This indeed does not spoil the estimation. We notice that normal vector estimations for $\mathbf{x}_b \in \partial g_\sigma$ with and without RS correlates very well, provided that the noise variance used for the estimate is larger than the variance σ^2 of RS. A large σ^2 may spoil the estimation but it is detrimental for the natural accuracy of $g_{\sigma,n}$. Yet, the estimation is indeed of poor quality due to the violation of the first assumption: the binary search may yield a point $\mathbf{x}_b$ not exactly on the boundary and this biaises the estimate. For instance, we notice that `HSJA` [2] sometimes crashes because all the noisy versions of $\mathbf{x}_b$ give the same output.

4. BLACK BOX ATTACKS VS. RS

4.1. Experimental Setup

We attack the classifier models with 200 random images from the ILSVRC2012's validation set with size $d = 3 \times 224 \times 224$.

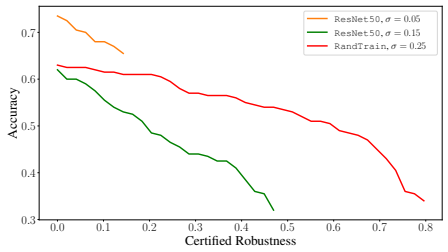


Fig. 4. Certification for `ResNet50` and `RandTrain`

Classifiers. The base classifier is `ResNet50` [22]. RS is performed with two noise standard deviations: $\sigma = 0.05$ gives an acceptable drop of accuracy of 3%, whereas $\sigma = 0.15$ yields larger certified robustness value but with a loss of 12% of accuracy (see Fig. 4).

Paper [15] proposes to re-train the model with noisy data in order to use a bigger σ without sacrificing too much accuracy. This new model is called `RandTrain`. With $\sigma = 0.25$, the accuracy loss is also around 12% but it delivers larger certified robustness (see Fig. 4).

We compare RS to the adversarially trained `ResNet50` from [7] that we denote `AdvTrain`.

Black-box attacks. Sect. 2 mentions three state of the art attacks. `RayS` [6], `SurFree` [5], and `HSJA` [2] achieve good results within $1,000$ calls to the classifier, but we use up to $2,000$ queries to be sure they reach their full potential.

Protocol. The distortion is measured as the Euclidean norm of the adversarial perturbation in the domain $[0,1]^d$. To assess that a point $\mathbf{x}_a$ complies with (4), the attacker needs to query $\ell = O(1/P_a)$ times the classifier $g_{\sigma,n}$. We speed up the simulation by considering that (4) holds if $\lceil nP_a \rceil$ micro-decisions are not correct.

4.2. Evaluation Results

Certified robustness vs. practice. The gap between theory and practice is salient when considering Fig. 4 and Fig. 5: Fig. 5 reports distortions at least 30 times larger than the certified robustness in Fig. 4.

New definition of adversarial needed. Attacks are not disturbed by level P_a (4). Being 80% adversarial forces to move away a little (Fig. 1) especially for small n. The robustness is only slightly better.

A small amount of noise is enough. Regardless of the attack, a large σ does not robustify the network whereas it spoils its accuracy. This is true even if the network has learned to handle noise: `RandTrain` has the same robustness and the same accuracy as the vanilla `ResNet50` with RS $\sigma = 0.15$. The situation is even worse against `RayS` [6]: noticeably `RandTrain` is less robust than `ResNet50` without RS.

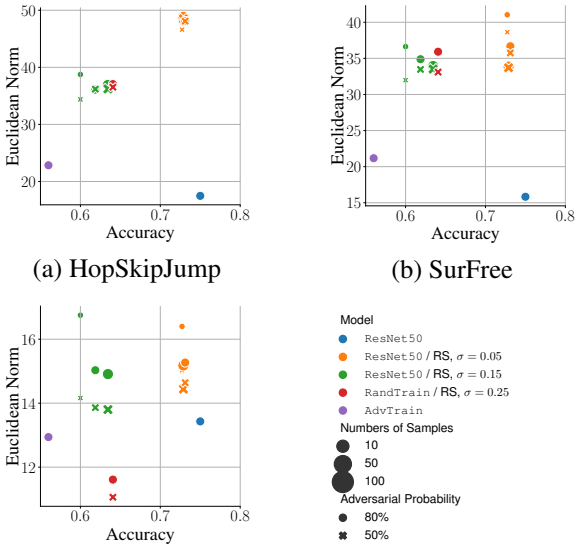


Fig. 5. Average adversarial ℓ_2 distortion vs. Accuracy

A small number of samples is enough. A big number of samples is key to get 'large' certified robustness. Fig. 5 shows another reality. The robustness against all 3 attacks is better with fewer samples. This confirms explanations in Sect. 3.5. Fewer samples makes the prediction at the boundary more random which jeopardizes more black box attacks.

Binary search is the only tool common to the 3 attacks. A point exactly on the boundary is crucial for `HSJA` [2] since it estimates the gradient. `SurFree` [5] does not do that but rely on the smoothness of the boundary. As for `RayS` [6], the binary search improves the distortion but it's not crucial for the convergence. It explains why `RayS` [6] is not as impacted as `SurFree` [5] and `HSJA` [2].

5. CONCLUSION

Certification with randomized smoothing is an important advance to apprehend the robustness of classifiers. Yet it was not considered as a practical defense; this paper chose this angle to reveal its real robustness facing state of the art black-box attacks. We *i)* illustrated formally the gap between a theoretical certification and a practical defense, and redefined what is an adversarial that faces a randomized defense. We found that the recommendations made in order to have larger certified bounds are often antagonistic with the concrete actions for obtaining a robust and accurate classifier in practice: *ii)* a low amount of samples is enough to fuzzy the frontiers; this is key to bother black box attacks, and *iii)* a high noise variance does not robustify the classier much, while it make accuracy drop; a small variance is enough.

6. REFERENCES

[1] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus, "Intriguing properties of neural networks," in *ICLR*, 2014.

[2] J. Chen, M. I. Jordan, and M. J. Wainwright, "HopSkipJumpAttack: A query-efficient decision-based attack," in *IEEE S&P*, 2020.

[3] A. Rahmati, S.-M. Moosavi-Dezfooli, P. Frossard, and H. Dai, "Geoda: a geometric framework for black-box adversarial attacks," in *CVPR*, 2020.

[4] H. Li, X. Xu, X. Zhang, S. Yang, and B. Li, "Qeba: Query-efficient boundary-based blackbox attack," in *CVPR*, 2020.

[5] Thibault Maho, Teddy Furon, and Erwan Le Merrer, "Surfree: a fast surrogate-free black-box attack," in *CVPR*, 2021.

[6] J. Chen and Q. Gu, "Rays: A ray searching method for hard-label adversarial attack," in *SIGKDD*, 2020.

[7] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *ICLR*, 2018.

[8] Monisankha Pal, Arindam Jati, Raghuveer Peri, Chin-Cheng Hsu, Wael AbdAlmageed, and Shrikanth Narayanan, "Adversarial defense for deep speaker recognition using hybrid adversarial training," in *ICASSP*, 2021.

[9] Haibin Wu, Xu Li, Andy T. Liu, Zhiyong Wu, H. Meng, and Hung yi Lee, "Adversarial defense for automatic speaker verification by cascaded self-supervised learning models," *ICASSP*, 2021.

[10] Mubarak G. Abdu-Aguye, Walid Gomaa, Yasushi Makihara, and Yasushi Yagi, "Detecting adversarial attacks in time-series data," in *ICASSP*, 2020.

[11] Kejiang Chen, Yuefeng Chen, Hang Zhou, Chuan Qin, Xiaofeng Mao, Weiming Zhang, and Nenghai Yu, "Adversarial examples detection beyond image space," in *ICASSP*, 2021.

[12] R. Olivier, B. Raj, and M. Shah, "High-frequency adversarial defense for speech and audio," in *ICASSP*, 2021.

[13] Mathias Lecuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana, "Certified robustness to adversarial examples with differential privacy," 2019.

[14] Bai Li, Changyou Chen, Wenlin Wang, and Lawrence Carin, "Second-order adversarial attack and certifiable robustness," 2019.

[15] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter, "Certified adversarial robustness via randomized smoothing," in *ICML*, 2019.

[16] Hadi Salman, Mingjie Sun, Greg Yang, Ashish Kapoor, and J. Zico Kolter, "Denoised smoothing: A provable defense for pretrained classifiers," in *NIPS*, 2020.

[17] Bai Li, Changyou Chen, Wenlin Wang, and Lawrence Carin, "Certified adversarial robustness with additive noise," in *NIPS*, 2019.

[18] Jinyuan Jia, Xiaoyu Cao, Binghui Wang, and Neil Zhenqiang Gong, "Certified robustness for top-k predictions against adversarial perturbations via randomized smoothing," *ICLR*, 2020.

[19] W. Brendel, J. Rauber, and M. Bethge, "Decision-based adversarial attacks: Reliable attacks against black-box machine learning models," in *ICLR*, 2018.

[20] Jamie Hayes, "Extensions and limitations of randomized smoothing for robustness guarantees," in *CVPR*, 2020.

[21] Hanwei Zhang, Yannis Avrithis, Teddy Furon, and Laurent Amsaleg, "Walking on the Edge: Fast, Low-Distortion Adversarial Examples," *IEEE TIFS*, 2020.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.