Certifiable Black-Box Attacks with Randomized Adversarial Examples: Breaking Defenses with Provable Confidence

Hanbin Hong
University of Connecticut
Storrs, Connecticut, USA

Xinyu Zhang
Zhejiang University
Hangzhou, Zhejiang, China

Binghui Wang
Illinois Institute of Technology
Chicago, Illinois, USA

Zhongjie Ba
Zhejiang University
Hangzhou, Zhejiang, China

Yuan Hong
University of Connecticut
Storrs, Connecticut, USA

ABSTRACT

Black-box adversarial attacks have demonstrated strong potential to compromise machine learning models by iteratively querying the target model or leveraging transferability from a local surrogate model. Recently, such attacks can be effectively mitigated by state-of-the-art (SOTA) defenses, e.g., detection via the pattern of sequential queries, or injecting noise into the model. To our best knowledge, we take the first step to study a new paradigm of black-box attacks with provable guarantees – certifiable black-box attacks that can guarantee the attack success probability (ASP) of adversarial examples before querying over the target model. This new black-box attack unveils significant vulnerabilities of machine learning models, compared to traditional empirical black-box attacks, e.g., breaking strong SOTA defenses with provable confidence, constructing a space of (infinite) adversarial examples with high ASP, and the ASP of the generated adversarial examples is theoretically guaranteed without verification/queries over the target model. Specifically, we establish a novel theoretical foundation for ensuring the ASP of the black-box attack with randomized adversarial examples (AEs). Then, we propose several novel techniques to craft the randomized AEs while reducing the perturbation size for better imperceptibility. Finally, we have comprehensively evaluated the certifiable black-box attacks on the CIFAR10/100, ImageNet, and LibriSpeech datasets, while benchmarking with 16 SOTA black-box attacks, against various SOTA defenses in the domains of computer vision and speech recognition. Both theoretical and experimental results have validated the significance of the proposed attack.[1]

CCS CONCEPTS

• **Security and privacy** → *Formal security models.*

KEYWORDS

Adversarial Attack, Black-box Attack, Certifiable Robustness

[1]The code and all the benchmarks are available at https://github.com/datasec-lab/CertifiedAttack, and the full version can be accessed at https://arxiv.org/abs/2304.04343.

CCS '24, October 14–18, 2024, Salt Lake City, UT, USA

ACM ISBN 979-8-4007-0636-3/24/10
https://doi.org/10.1145/3658644.3690343

ACM Reference Format:
Hanbin Hong, Xinyu Zhang, Binghui Wang, Zhongjie Ba, and Yuan Hong. 2024. Certifiable Black-Box Attacks with Randomized Adversarial Examples: Breaking Defenses with Provable Confidence. In *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security (CCS '24).* ACM, Salt Lake City, UT, USA, 15 pages. https://doi.org/10.1145/3658644.3690343

1 INTRODUCTION

Machine learning (ML) models have achieved unprecedented success and have been widely integrated into many practical applications. However, it is well known that minor perturbations injected into the input data are sufficient to induce model misclassification [52]. Many state-of-the-art (SOTA) adversarial attacks [2, 7, 11, 13, 15, 38, 44, 52, 55, 85–87] have been proposed to explore the vulnerabilities of a variety of ML models. Wherein, the stringent *black-box* attack is believed to be closer to real-world security practice [13, 58].

In black-box attacks, the adversary only has access to the target ML model's outputs (either prediction scores or hard labels). Through iteratively querying the target model, the adversary progressively updates the perturbation until convergence. Existing black-box attack methods primarily utilize gradient estimation [5, 15, 19, 25, 38], surrogate models [24, 57, 58, 69], or heuristic algorithms [2, 7, 8, 30, 47] to generate adversarial perturbations. Although these attack algorithms can empirically achieve relatively high attack success rates (e.g., on CIFAR-10 [43]), their query process is shown to be easy to detect or interrupt due to the minor perturbation changes and high reliance on the previous perturbation [12, 16, 46, 60]. For example, "Blacklight" [46] can achieve 100% detection rate on most of the existing black-box attacks by checking the similarity of queries; some "randomized defense" methods [12, 16, 33, 50, 60] inject random noise to the inputs, outputs, intermediate features or model parameters such that the performance of existing black-box attacks can be significantly degraded (since the query results are obfuscated to be unpredictable).

To break such types of SOTA defenses [16, 33, 46, 50, 60], it is challenging to design an effective attack equipped with both *high degree of randomness to bypass the strong detection* (e.g., Blacklight [46]) and *high robustness to resist randomized defense.* A feasible solution is to add random noise to the adversarial example by the adversary, but it will make the query intractable. Therefore, an innovative method is desirable to carefully craft the adversarial example based on feedback from queries using randomly generated inputs.

To this end, we propose a novel attack paradigm, termed *Certifiable Attack*, that ensures a provable attack success probability (ASP) on the randomized adversarial examples against the equipped defenses (or no defense). Specifically, our attack strategy integrates random noise into the queries while preserving the adversarial efficacy of these queries. In particular, we model the adversarial examples as a random variable in the input space following an underlying noise distribution φ, namely "*Adversarial Distribution*". Then, we design a novel query strategy and establish the theoretical foundation to guarantee the ASP of the distribution throughout the crafting process. A novel framework is also developed to find the initial *Adversarial Distribution*, optimize it, and use it to sample the adversarial examples.

1.1 Certifiable Attacks vs. Empirical Attacks

Compared with existing empirical black-box attacks, the Certifiable Attack demonstrates multi-faceted advantages (also see Figure 1):

(a) **Strong attack to break SOTA defenses.** The randomness in the certifiable attack allows it to effectively bypass detection methods that rely on the similarity between the attacker's sequence of queries (e.g., Blacklight [46]), while traditional empirical attacks often create a suspicious trajectory of highly similar perturbations. The certifiable attack also provides a provable guarantee of success for attacks using randomized inputs, by taking into account the equipped defense and target model, enhancing its resistance to randomized defense [12, 60].

(b) **Adversarial space vs. Adversarial example (AE).** Distinct from traditional empirical adversarial attacks, which uncover model vulnerabilities with sample-wise inputs, the Certifiable Attack seeks to explore an adversarial input space constructed by an *Adversarial Distribution*. This continuous space facilitates the generation of numerous (potentially infinite) adversarial examples with a high ASP, thus revealing a more consistent and severe vulnerability of the target model.

(c) **Adversarial Examples (AEs) sampled from the adversarial distribution are verification-free.** Empirical attacks search AEs by iteratively querying the target model and verifying the query outputs (*the final successful AE is also used to query over the target model; then it will be verified and recorded by the defender/target model*). Instead, the certifiable attack crafts the *adversarial distribution* with a guaranteed lower bound of the ASP. Due to the highly dimensional and continuous input space, AEs sampled from the adversarial distribution can be considered unique (with noise in all the dimensions) and have a negligible probability of being recorded by the defender/target model after verification. The ASP of such AEs are theoretically guaranteed (verification-free), and they are new to the defender, posing more challenges for mitigation.

1.2 Randomization for Certifiable Attacks

To pursue certifiable attacks, we theoretically bound the ASP of *Adversarial Distribution* based on a novel way of utilizing randomized smoothing [22], a technique achieving great success in the certified defenses with probabilistic guarantees.

The design for the randomization-based certifiable attack follows an intuitive goal, i.e., *ensuring that the classification results are consistently wrong over the distribution*. However, many new significant challenges should be addressed. First, existing theories (randomization for certified defenses, e.g., [22]) cannot be directly adapted to certifiable attacks since they have completely different goals and settings. Second, how to efficiently craft the *Adversarial Distribution* that can ensure the ASP is challenging since it requires maintaining the wrong prediction over a large number of randomized samples drawing from the distribution. Third, how to make the *Adversarial Distribution* as imperceptible as possible is also challenging due to their randomness. By addressing these new challenges, in this paper, we make the following significant contributions:

1) To our best knowledge, we introduce the first *certifiable attack theory* based on randomization for the black-box setting, which universally guarantees the attack success probability of AEs drawn from different noise distributions, e.g., Gaussian, Laplace, and Cauthy distributions, enabling a novel transition from deterministic to probabilistic adversarial attacks.
2) We propose a novel *certifiable attack framework* that can efficiently craft certifiable *Adversarial Distribution* with provable ASP and imperceptibility. Specifically, we design a novel *randomized parallel query method* to efficiently collect probabilistic query results from any target model, which supports the certifiable attack theory. We propose a novel *self-supervised localization* method as well as a binary-search localization method to efficiently generate certifiable *Adversarial Distribution*. We design a novel *geometric shifting* method to reduce the perturbation size for better imperceptibility while ensuring the ASP. Finally, we have validated that diffusion models [34] can be used to further denoise the randomized AEs with guaranteed ASP.
3) We comprehensively evaluate the performance of the certifiable attack with different settings on 4 datasets, while benchmarking with 16 SOTA empirical black-box attacks, against various defenses. Experimental results consistently demonstrate that our certifiable attack effectively breaks the SOTA defenses, including adversarial detection, randomized pre-processing and post-processing defenses, as well as adversarial training defenses (Also, Table 1 shows a summary of the certifiable attack vs. SOTA black-box attacks).

2 PROBLEM DEFINITION

Threat Model: We consider designing a certifiable attack where the target model may or may not be protected by a defense mechanism.

- **Adversary**: We focus on the hard-label black-box attack, where the adversary only knows the predicted label by querying the target ML model. The adversary's objective is to craft adversarial examples to fool the model based on the query results.
- **Model Owner**: The model owner pursues the model utility. We consider three different levels of the model owner's knowledge and capability: 1) The model owner has no awareness of the adversarial attacks and is not equipped with any defense; 2) The model owner is aware of the adversarial attack but has no knowledge of the attack method. The model owner can deploy general defense methods such as adversarial training [52]; 3) The model owner is aware of the adversarial attack and has

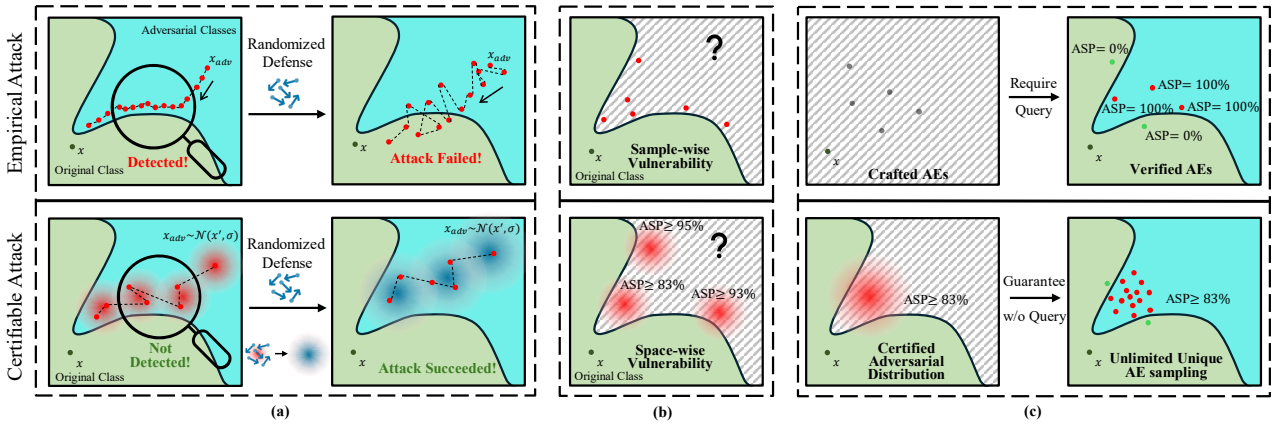


Figure 1: Empirical attacks vs. Certifiable attacks. (a) Certifiable attack can break the SOTA AE detection and randomized defenses. (b) Certifiable attack uncovers space-wise vulnerability rather than sample-wise vulnerability. (c) Once certified, Certifiable Attack can generate unlimited unique AEs with a guaranteed minimum ASP without querying the model for verification, while the empirical attack requires verifying the attack result of crafted AE by query.

Table 1: Comparison of state-of-the-art empirical black-box attacks with certifiable attack

Black-box Attacks	Query Type	Perturbation Type	ASP Guarantee	vs. Detection on Attacker's Queries	vs. Randomized Pre-process. Defense	vs. Randomized Post-process. Defense
Bandit [39], NES [38], Parsimonious [54], Sign [1], Square [2], ZOSignSGD [49]	Score-based	ℓ_∞-bounded	×	×	×	✓
GeoDA [62], HSJ [13], Opt [19], RayS [14], SignFlip [18], SignOPT [20]	Label-based	ℓ_∞-bounded	×	×	×	×
Bandit [39], NES [38], Simple [30], Square [2], ZOSignSGD [49]	Score-based	ℓ_2-bounded	×	×	×	✓
Boundary [7], GeoDA [62], HSJ [13], Opt [19], SignOPT [20]	Label-based	ℓ_2-bounded	×	×	×	×
PointWise [67], SparseEvo [78]	Label-based	Optimized	×	×	×	×
Certifiable Attack (ours)	Label-based	Optimized	✓	✓	✓	✓

knowledge about the attack method. The model owner can deploy adaptive defenses that are specifically designed for the attack.

Problem Formulation: We first briefly review adversarial examples, and then formally define our problem. Given an ML classifier f and a testing data $x \in \mathbb{R}^d$ with label y from a label set $\mathcal{Y} = [1, \cdots C]$ (where C is the number of classes). An adversary carefully crafts a perturbation on the data x such that the classifier f misclassifies the perturbed data x_{adv}, i.e., $f(x_{adv}) \neq y$ under $x_{adv} \in [\Pi_a, \Pi_b]^d$, where $[\Pi_a, \Pi_b]^d$ is the valid input space. The perturbed data x_{adv} is called *adversarial example*. Imperceptibility is usually achieved by restricting the ℓ_2 or ℓ_∞ norm of the perturbation $x_{adv} - x$, or by minimizing the magnitude of this perturbation.

In the black-box setting, an adversary can use *empirical* black-box attack techniques (details in Section 6) to iteratively query the classifier f and progressively update the perturbation until finding a successful adversarial example for a testing example. However, such attack strategies have key limitations: 1) query inefficient, usually > 100 queries per adversarial example; 2) easy to be detected by observing the query trajectory [12, 16, 46, 60]; and 3) lack of guaranteed attack performance, i.e., cannot provably guarantee a (un)successful adversarial example under a given budget.

We aim to address all these limitations and design an efficient and effective certifiable black-box attack in the paper. Particularly, instead of inefficiently searching adversarial *examples* one-by-one, we want to certifiably find the underlying adversarial *distribution* that the adversarial examples lie on.

Definition 2.1 (Certifiable black-box attack). Given a classifier $f : \mathbb{R}^d \to \mathcal{Y}$, a clean input $x \in \mathbb{R}^d$ with label $y \in \mathcal{Y}$, and an Attack Success Probability Threshold p, the certifiable attack is to find an *Adversarial Distribution* $\varphi(x', \boldsymbol{\kappa})$ with mean x' and parameters $\boldsymbol{\kappa}$[2], such that data sampled from φ have at least p probability of being misclassified (i.e., adversarial examples). That is,

$$\mathbb{P}_{x_{adv} \sim \varphi(x', \boldsymbol{\kappa})}[f(x_{adv}) \neq y] \geq p \tag{1}$$

$$\text{s.t. } x_{adv} \in [\Pi_a, \Pi_b]^d. \tag{2}$$

Design Goals: We expect our attack to achieve the below goals.

1) **Certifiable:** It can provide provable guarantees on the minimum attack success probability of the crafted adversarial examples.
2) **Verification free:** It can not only verify examples to be adversarial *after* querying the model, but also verify examples *before* the query by giving its ASP. This significantly boosts the effectiveness of adversarial examples generation.
3) **Query efficient:** It needs as few number of queries as possible. Fewer queries can definitely save the adversary's cost.
4) **Bypass defenses:** It can generate imperceptible adversarial perturbations that can bypass the existing detection and pre/post-processing based defenses [12, 16, 46, 60].

[2]If φ is a Gaussian distribution, κ is the standard deviation of φ. If φ is a Generalized normal distribution, $\kappa = (a, b)$, with a and b the scale and shape parameters of φ, respectively. Notice that, the distribution will be applied to all the dimensions in the input, and *Adversarial Distribution* is a noise distribution over the input space.

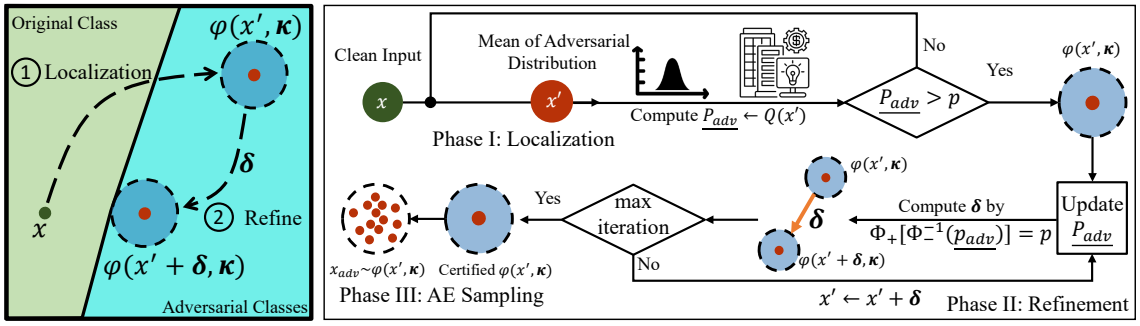


Figure 2: Overview of our certifiable black-box attack to generate certified adversarial distribution.

3 ATTACK OVERVIEW

At a high level, our certifiable black-box attack can be divided into three phases. The overview of our attack is depicted in Figure 2.

Phase I: Adversarial Distribution Localization. This phrase initially locates a feasible *Adversarial Distribution* φ with guarantees on the lower bound of attack success probabilities (i.e., satisfying Eq. (1)). There are a few challenges. First, computing the exact probability $\mathbb{P}[f(x_{adv}) \neq y]$ is intractable due to the high-dimensional continuous input space. Second, due to the black-box nature, there exists no gradient information that can be used. To address the first challenge and ensure query efficiency, we propose a Randomized Parallel Query (RPQ) strategy that can approximate the probability and ensure multiple queries are implemented in parallel. To address the second challenge, we design two localization strategies to enable learning a feasible adversarial distribution. The first strategy adapts the existing self-supervised perturbation (SSP) technique [57], which facilitates designing a classifier-unknown loss on a pretrained feature extractor such that the adversarial examples/perturbations can be optimized. The second one is based on binary search. It first randomly initializes a qualified *Adversarial Distribution* , and then reduces the perturbation size using the binary search algorithm. See Section 4.1 for more details.

Phase II: Adversarial Distribution Refinement. While successfully generating the adversarial distribution, the adversarial examples from it often induce relatively large perturbation sizes. This phrase further refines the adversarial distribution by reducing the perturbation size and maintains the guarantee of attack success probability as well. Particularly, we propose to shift the adversarial distribution close to the decision boundary of the classifier. This problem can be solved by two steps: *the first step finds the shifting direction, and the second step derives the shifting distance and maintains the guarantee.* We design a novel shifting method to find the local-optimal *Adversarial Distribution* by considering the geometric relationship between the decision boundary and *Adversarial Distribution*. Deciding the shifting distance can then be converted to an optimization problem. We then propose a binary search algorithm to achieve the goal. See Section 4.2 for more details.

Phase III: Adversarial Example Sampling. Phases I and II craft an *Adversarial Distribution* with guaranteed attack success probability, called "certifiable attack". To transform the *Adversarial Distribution* into concrete AEs, we need to sample the AE from the *Adversarial Distribution*. The sampled AEs naturally maintain the certified ASP without the need for additional model queries. Optionally, the adversary can verify the success of these sampled AEs to ensure a successful attack, turning the certifiable attack into an empirical attack. Specifically, the adversary can sequentially sample the adversarial examples from *Adversarial Distribution* and query the target model until finding the successful adversarial example(s).

4 CERTIFIABLE BLACK-BOX ATTACK

In this section, we present our certifiable black-box attack in detail. We first introduce the Randomized Parallel Query strategy that estimates the lower bound probability of being the adversarial example (Section 4.1.1). We then develop two algorithms to locate the feasible *Adversarial Distribution* (Section 4.1.2). Next, we propose our refinement method to reduce the perturbation size, while maintaining the guarantees of attack success probability (Section 4.2). We also provide the theoretical analysis of the convergence and confidence bound of the Shifting method.

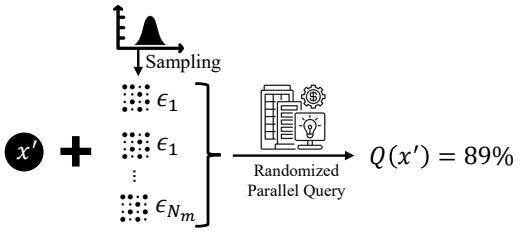


Figure 3: Illustration of randomized parallel query (returning the probability $Q(x')$ that $x' + \epsilon$ is an adversarial example).

4.1 Adversarial Distribution Localization

4.1.1 Randomized Parallel Query. As stated, computing the exact probability $\mathbb{P}[f(x_{adv}) \neq y]$ with $x_{adv} \sim \varphi(x', \kappa)$ is intractable. Here, we propose to estimate its low bound probability by the Monte Carlo method. This requires the adversary to query the classifier with random instances sampled from an *Adversarial Distribution*. By noting that random instances can be queried efficiently in parallel, we propose the Randomized Parallel Query (RPQ) to compute the

Algorithm 1 Lower Bound of Attack Success Probability

Input: Mean x' of the *Adversarial Distribution* φ, classifier f, confidence level α, Monte Carlo samples N_m, ground truth label y.
Output: The lower bound of attack success probability $\underline{p_{adv}}$

$\epsilon_1, \epsilon_2, ..., \epsilon_{N_m} \sim \varphi(0, \boldsymbol{\kappa})$
Incorrect prediction count $k \leftarrow \sum_{N_m}^{i=1} \mathbf{1}[f(x' + \epsilon_i) \neq y]$
return $\underline{p_{adv}} \leftarrow \text{LowerConfBound}(k, N_m, 1 - \alpha)$

Algorithm 2 Smoothed Self-Supervised Perturbation (SSSP)

Input: Clean input x, feature extractor $\mathcal{F}$, noise distribution $\varphi(0, \boldsymbol{\kappa})$, maximum iterations n_{max}, perturbation budget π, step size η, and noise sampling number N_s.
Output: Updated mean x' of *Adversarial Distribution*

$x' = x$
for $n = 1$ to n_{max} **do**
  $\mathcal{L}(x') \leftarrow \frac{1}{N_s} \sum_i^{N_s} [||\mathcal{F}(x' + \epsilon_i) - \mathcal{F}(x + \epsilon_i)||_2],\ \epsilon_i \sim \varphi$
  $x' \leftarrow x' + \eta\ sgn(\nabla_{x'} \mathcal{L})$
  $x' \leftarrow Clip(x',\ x - \pi,\ x + \pi)$
  $x' \leftarrow Clip(x',\ 0.0,\ 1.0)$ (if x is an image)
return x'

Algorithm 3 Smoothed SSP for Certifiable Attack Localization

Input: Clean input x, feature extractor $\mathcal{F}(\cdot)$, RPQ function $Q(\cdot)$, smoothed SSP algorithm $SSSP(\cdot)$ (Algorithm 2), initial perturbation budget π_{init}, step size γ, ASP Threshold p, maximum iterations N_{max}.
Output: Mean x' of *Adversarial Distribution* φ, number of RPQs q.

$x' = x, \pi = \pi_{init}, N = 0, q = 0$
while $Q(x') < p$ and $N < N_{max}$ **do**
  $N \leftarrow N + 1, q \leftarrow q + 1, \pi \leftarrow \pi + \gamma$
  $x' \leftarrow SSSP(x',\ \mathcal{F},\ \pi)$
if $Q(x') < p$ **then**
  return Abstain
else
  return x' and q

lower bound of the attack success probability as below:

$$Q(x') = \underline{p_{adv}} \leq \mathbb{P}_{x_{adv} \sim \varphi(x', \boldsymbol{\kappa})}[f(x_{adv}) \neq y] = \mathbb{P}_{\epsilon \sim \varphi(0, \boldsymbol{\kappa})}[f(x' + \epsilon) \neq y]. \quad (3)$$

With a given x', the lower bound probability $\underline{p_{adv}}$ can be estimated via the Binomial testing on a zero-mean distribution $\varphi(0, \boldsymbol{\kappa})$ using Clopper-Pearson confidence interval [45] following the Algorithm 1, where the $\text{LowerConfBound}(k, N_m, 1-\alpha)$ returns the one-sided $(1 - \alpha)$ lower confidence interval.

Now we can estimate $\underline{p_{adv}}$ given an *Adversarial Distribution* with known mean/location x'. The next question is how to decide x' to satisfy Eq. (1), i.e., locating the adversarial distribution that includes certifiable adversarial examples (with probability at least p).

The simplest way is random localization, where the input x is uniformly sampled from the input space $[\Pi_a, \Pi_b]^d$, e.g., $[0,\ 1]^d$, followed by the RPQ to check if $\underline{p_{adv}}$ is larger than p. However, random localization could not generate a good initial adversarial distribution due to the high-dimensional input space. Below we propose two practical localization methods to mitigate the issue.

Algorithm 4 Binary Search for Certifiable Attack Localization

Input: Clean input x, RPQ function $Q(\cdot)$, ASP Threshold p, random search iterations N_r, and binary search iteration N_b, error tolerance Ω.
Output: Mean of initial *Adversarial Distribution* x', number of RPQs q.

$n = 0, m = 0, q = 0, x^* = x$
while $Q(x') < p$ and $n \leq N_r$ **do**
  $x' \sim \text{Uniform}([0, 1]^d)$
  $q \leftarrow q + 1,$
if $n > N_r$ **then return Abstain**
while $m < N_b$ and $\|x' - x^*\|_2 \leq \Omega$ **do**
  if $Q(\frac{x^* + x'}{2}) \geq p$ **then**
    $x' = \frac{x^* + x'}{2}$
  else
    $x^* = \frac{x^* + x'}{2}$
return x'

4.1.2 Proposed Localization Algorithms. We notice the adversarial distribution localization is similar to empirical black-box attacks on generating adversarial examples. Here, we propose to adapt these empirical attack algorithms and design two localization algorithms.

Smoothed Self-Supervised Localization: To better locate the *Adversarial Distribution*, we propose to adapt the self-supervised perturbation (SSP) technique [57]. Specifically, SSP generates generic adversarial examples by distorting the features extracted by a pre-trained feature extractor on a large-scale dataset in a self-supervised manner. The rationale is that the extracted (adversarial) features can be transferred to other classifiers as well.

As our attack uses RPQ, we compute the feature distortion over *a set of random samples* from the *Adversarial Distribution*. Formally,

$$x' = \arg\max_{x'} \mathbb{E}_{\epsilon \sim \varphi(0, \boldsymbol{\kappa})}[\|\mathcal{F}(x' + \epsilon) - \mathcal{F}(x + \epsilon)\|_2] \quad \text{s.t. } \|x' - x\|_\infty \leq \pi \quad (4)$$

where $\mathcal{F}$ is a pre-trained feature extractor. The perturbation budget π is initially set to a small value and later increased in multiple attempts of localization, ensuring that smaller perturbations are identified first. This optimization problem can be solved via the Projected Gradient Ascent method [52]. Let the adversarial loss be $\mathcal{L}(x') \equiv \mathbb{E}_{\epsilon \sim \varphi}[\|\mathcal{F}(x' + \epsilon) - \mathcal{F}(x + \epsilon)\|_2]$. Then we can locate the *Adversarial Distribution* via iteratively update x' with $x' = x' + \eta\ sgn(\nabla_{x'} \mathcal{L})$, where $sgn(\cdot)$ is the sign function, and η denotes the step size. The details for localizing the *Adversarial Distribution* are summarized in Algorithms 2 and 3.

Binary Search Localization: Another method is to randomly initialize the location of *Adversarial Distribution* such that $\underline{p_{adv}} \geq p$, and then reduce the gap between $\underline{p_{adv}}$ and p, as well as the perturbation via binary search. The algorithm is presented in Algorithm 4. This method is efficient in reducing the perturbation size once the feasible *Adversarial Distribution* is found by random search. Figure 6 and 7 visualize some x_{adv} during the crafting process for both Binary Search Localization and SSSP Localization.

4.2 Adversarial Distribution Refinement

Though our localization algorithms can find an effective *Adversarial Distribution*, our empirical results found the perturbation size can be large (See Table 5.4.3). This occurs possibly because the pretrained

feature extractor is too generic and the generated adversarial perturbation is suboptimal for our target classifier. To mitigate the issue, we propose to reduce the perturbation by refining the *Adversarial Distribution* while still maintaining the condition Eq. (1).

Our key observation is that the optimal perturbation is achieved when the adversarial example is close to the decision boundary of the target classifier. Hence, we propose to shift the *Adversarial Distribution* until intersecting the decision boundary, thereby locating the locally optimal point on that boundary.

4.2.1 Certification for Adversarial Distribution *Shifting.* We propose a theory on shifting the *Adversarial Distribution* while maintaining the attack success probability. We denote $\varphi(x'+\delta, \boldsymbol{\kappa})$ as a shifted distribution for the *Adversarial Distribution* $\varphi(x', \boldsymbol{\kappa})$ by a shifting vector δ. Then, the shifted *Adversarial Distribution* ensures the ASP if δ satisfies the condition presented in Theorem 1.

Theorem 1. ***(Certifiable Adversarial Distribution Shifting)*** *Let f be a classifier, ϵ be the noise drawn from any continuous probability density function $\varphi(0, \boldsymbol{\kappa})$. Let p be the predefined attack success possibility threshold. Denote $\underline{p_{adv}}$ as the lower bound of the attack success probability. For any x' satisfies*

$$\mathbb{P}[f(x'+\epsilon) \neq y] \geq \underline{p_{adv}} = Q(x') \geq p, \tag{5}$$

$\mathbb{P}[f(x'+\delta+\epsilon) \neq y] \geq p$ is guaranteed for any shifting vector δ when

$$\Phi_+[\Phi_-^{-1}(\underline{p_{adv}})] \geq p \tag{6}$$

where Φ_-^{-1} is the inverse cumulative density function (CDF) of the random variable $\frac{\varphi(\epsilon-\delta,\boldsymbol{\kappa})}{\varphi(\epsilon,\boldsymbol{\kappa})}$, and Φ_+ the CDF of random variable $\frac{\varphi(\epsilon,\boldsymbol{\kappa})}{\varphi(\epsilon+\delta,\boldsymbol{\kappa})}$.

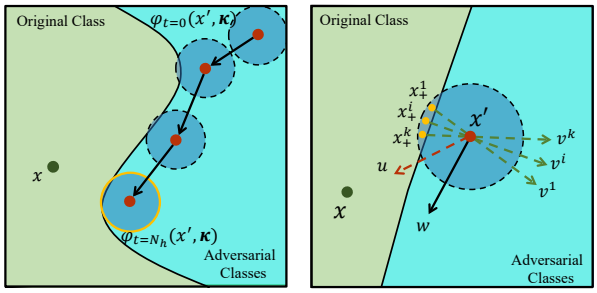
Proof. See detailed proof in Appendix A.1 [3]. □

Theorem 1 ensures the minimum attack success probability if Eq. (5) and Eq. (6) hold while without querying $\varphi(x'+\delta, \boldsymbol{\kappa})$. Eq. (5) requires finding a x' such that the RPQ on samples of $\varphi(x', \boldsymbol{\kappa})$ returns a $\underline{p_{adv}} \geq p$, and Eq. (6) ensures any δ meeting this condition will not reduce the attack success probability of the shifted *Adversarial Distribution* below p. Further, Theorem 1 works for any continuous noise distributions, e.g., Gaussian, Laplace, Exponential, and mixture PDFs. We also present the case when the noise is Gaussian in Corollary 2.1 in Appendix A.2. It shows the shifting perturbation δ should satisfy $||\delta||_2 \leq \sigma[\Phi^{-1}(\underline{p_{adv}}) - \Phi^{-1}(p)]$, where Φ^{-1} is the inverse of Gaussian CDF.

4.2.2 Obtaining Refined Adversarial Distribution. Since the *Adversarial Distribution* can be shifted by any δ satisfying Eq. (6), we propose to shift it toward the clean input with the maximum δ that does not break the guarantee. By iteratively executing the RPQ and applying the Theorem 1, the *Adversarial Distribution* can be repeatedly shifted with a guarantee until approaching the decision boundary (where $\underline{p_{adv}} = p$).

The problem of shifting the *Adversarial Distribution* to reduce the perturbation can be solved by two steps: *first finding the shifting direction, and then deriving the shifting distance while maintaining the guarantee.* Here, we design a novel shifting method to find the locally optimal *Adversarial Distribution* by considering the geometric relationship between the decision boundary and the *Adversarial Distribution*, which is called "Geometrical Shifting". Specifically, through using the noisy samples of *Adversarial Distribution* to "probe" the decision boundary, we shift the RandAE along the decision boundary and towards the clean input until finding the local optimal point on the decision boundary (see Figure 4 for the illustration). If none of the noisy samples can approach the decision boundary, we simply shift the *Adversarial Distribution* directly toward the clean input without considering the decision boundary.

[3]Detailed proofs are presented in the full version at https://arxiv.org/abs/2304.04343



Geometrically Shifting Geometric Relationship

Figure 4: Illustration of geometrically shifting.

Algorithm 5 Shifting Direction

Input: Mean of the *Adversarial Distribution* x', clean input x, vectors $\{v^i\}$, a vector u, maximum iteration M, updating step size η'.
Output: The shifting direction w

Initialize w with random noise
if $\{v^i\}$ is empty **then**
  $w = x - x'$
else
  for $j = 1$ to M **do**
    $w \leftarrow w + \eta'\ sgn[\nabla_w(\sum_{i=1}^{k} \sin(v^i, w) + \cos(u, w))]$
    $w \leftarrow \frac{w}{||w||_2}$
return w

Algorithm 6 Shifting Distance

Input: Mean of *Adversarial Distribution* x', noise distribution φ, randomized query function $Q(\cdot)$, the shifting direction algorithm $SD(\cdot)$ (Algorithm 5), error threshold e, ASP Threshold p.
Output: The shifting perturbation δ

$w \leftarrow SD(x')$, $\underline{p_{adv}} \leftarrow Q(x')$
find a scalar a such that $\delta = aw$ and $\Phi_+[\Phi_-^{-1}(\underline{p_{adv}})] > p$
find a scalar b such that $\delta = bw$ and $\Phi_+[\Phi_-^{-1}(\underline{p_{adv}})] < p$
while $\Phi_+[\Phi_-^{-1}(\underline{p_{adv}})] < p$ or $> p + e$ and $n \leq N_k$ **do**
  if $\Phi_+[\Phi_-^{-1}(\underline{p_{adv}})] > p$ **then**
    $a \leftarrow \frac{(a+b)}{2}$
  else
    $b \leftarrow \frac{(a+b)}{2}$
  $\delta \leftarrow \frac{(a+b)}{2} w$, $n \leftarrow n + 1$
return δ

Finding the Shifting Direction: The geometrical relationship is presented on the right-hand side of Figure 4. Denote x' as the mean of the current *Adversarial Distribution*. When sampling the adversarial examples from the *Adversarial Distribution*, we mark

Algorithm 7 Certifiable Attack Shifting

Input: Mean of *Adversarial Distribution* x', noise distribution φ, randomized query function $Q(\cdot)$, shifting distance algorithm $\textsc{Shift}(\cdot)$ (Algorithm 6), distance threshold e_s, ASP Threshold p, max iteration N_h.
Output: The shifted mean x'

$\underline{p_{adv}} \leftarrow Q(x')$, $\delta \leftarrow \textsc{Shift}(x')$, $n = 0$
while $\underline{p_{adv}} > p$ and $\|\delta\|_2 \geq e_s$ and $n \leq N_h$ **do**
  $x' \leftarrow x' + \delta$, $\underline{p_{adv}} \leftarrow Q(x')$, $\delta \leftarrow \textsc{Shift}(x')$, $n \leftarrow n + 1$
  if $\|x' - x\|_2 \leq \|\delta\|_2$ **then return** x
return x'

the failed adversarial examples as $x_+^1, x_+^2, ..., x_+^i, ..., x_+^k$, aka., "samples fell into the original class". The normalized vector from x_+^i to x' is denoted as v^i. The normalized vector from x' to x is denoted as u. If the *Adversarial Distribution* has no samples crossing the decision boundary, then we can shift the *Adversarial Distribution* straight toward the clean input (along the direction of u) until it intersects the decision boundary, otherwise, the shifting should be along the decision boundary but not cross it (without changing the certifiable attack guarantee). Note that the input space is high-dimensional, thus there could be many directions along the decision boundary. To reduce the perturbation, the direction should be similar to the vector u as much as possible. Based on these geometric analyses, the goals of the geometrical shifting can be summarized as: *The shifting direction should lie relatively parallel to the direction of u; and be relatively vertical to the vectors v^i.*

Formally, denoting the shifting direction as w, then the goal of finding the shifting direction can be formulated as:

$$w = \arg\max \sum_{i=1}^{k} \sin(v^i, w) + \cos(u, w) \tag{7}$$

where $\sin(\cdot)$ and $\cos(\cdot)$ denote the sine and cosine function, and Eq. (7) can be solved via the gradient ascent algorithm.

Calculating the Shifting Distance: The shifting distance can be determined by maximizing $||\delta||_2$ that satisfies the constraint of Eq. (6), i.e., when the equality holds. We use binary search to approach the equality and the Monte Carlo method to estimate the CDF of random variable $\frac{\varphi(\epsilon-\delta,\boldsymbol{\kappa})}{\varphi(\epsilon,\boldsymbol{\kappa})}$ and $\frac{\varphi(\epsilon,\boldsymbol{\kappa})}{\varphi(\epsilon+\delta,\boldsymbol{\kappa})}$, similar to [36]. Algorithm 5, 6, and 7 show the details of finding the shifting direction, shifting distance, and the shifting process, respectively.

Convergence Guarantee and Confidence Bound: Any δ computed by Algorithm 6 will satisfy the certifiable attack guarantee since it strictly ensures $\Phi_+[\Phi_-^{-1}(\underline{p_{adv}})] \geq p$. Further, with a centralized noise distribution, the shifting algorithm is guaranteed to converge once the located *Adversarial Distribution* is feasible.

Theorem 2. *If the PDF of noise distribution $\varphi(x)$ decreases as the $|x|$ increases, with the satisfaction of Eq. (5), given any direction vector w, the Shifting Distance algorithm guarantees to find δ such that $\Phi_+[\Phi_-^{-1}(\underline{p_{adv}})] = p$ with confidence $(1-\alpha)(1-2e^{-2N_m\Delta^2})^2$, where $(1-\alpha)$ is the confidence for of estimating $\underline{p_{adv}}$, N_m is the Monte Carlo samples, and Δ is the error bound for the CDF estimation.*

Proof. See detailed proof in Appendix A.3. □

4.3 Discussions on Our Attack

Realizing Our Certifiable Attack: Our certifiable attack does not have extra requirements on realization compared to empirical black-box attacks. To implement our attack, we only need to predefine a continuous noise distribution and a threshold of certified attack success probability. The adversary then adds the noise sampled from the distribution to the inputs and queries the target model. Then, *Adversarial Distribution* can be crafted by RPQ and our theory.

Randomized Query vs. Deterministic Query: The proposed randomized query returns a *probability* over a batch of inputs with injected random noises, while the traditional query returns a *deterministic* output (score or hard label) from the target model. This probability return may provide more information that better guides the attack. In addition, the randomized queries can be executed in parallel for query acceleration. See results in Section 5.3.

Imperceptibility with Diffusion Denoiser: The certifiable adversarial examples sampled from *Adversarial Distribution* are noise-injected inputs that still might be perceptible when the noise is large. We can further leverage the recent innovation for image synthesis, i.e., diffusion model [34], to denoise the adversarial examples for better imperceptibility. The key idea is to consider the noise-perturbed adversarial examples as the middle sample in the forward process of the diffusion model [10, 91]. This is shown to improve the imperceptibility and the diversity of the adversarial examples. More technical details are shown in Appendix B and results in Table 14 in Appendix C.4.

Extension to Certifiable White-Box Attack: Our certifiable attack can be readily extended to the white-box setting by adapting/designing a white-box localization method. Specifically, the Smoothed SSP localization method can directly compute the gradients of the noise-perturbed examples rather than leveraging the feature extractor, which may significantly improve the certified accuracy of the certifiable attack. In our experiments, when leveraging the PGD-like white-box attacks as the localization method, the certified accuracy can be increased to 100% for ResNet and CIFAR10, compared to the 92.54% certified accuracy in the black-box setting.

Extension to Targeted Certifiable Attack: Our attack design focuses on the untargeted certifiable attack. It can also be generalized to the targeted attack setting, where we require the majority of the noise-perturbed inputs to be *certifiably* misclassified to a specific *target* label. However, we admit it would be more challenging to find a successful *Adversarial Distribution* in this scenario.

Attacks under Adaptive Blacklight: The defender might design an adaptive countermeasure, such as an adaptive blacklight defense, to mitigate certified attacks. For instance, the defender could attempt to eliminate randomness by assuming the noise distribution is known. However, this approach presents several challenges: 1) The defender would need detailed knowledge about the attack's design, including the noise distribution, which is often an impractical assumption. 2) Even if the noise distribution were known, the sampled adversarial examples would remain random, making it difficult to accurately estimate the center of the noise distribution.

Table 2: Summary of Experiments

Experiments	Dataset	Model	Reference
Comparison with empirical attacks against Blacklight detection	CIFAR10	VGG16	Table 15
	CIFAR10	ResNet110	Table 16
	CIFAR10	ResNext29	Table 17
	CIFAR10	WRN28	Table 18
	CIFAR100	VGG16	Table 19
	CIFAR100	ResNet110	Table 20
	CIFAR100	ResNext29	Table 21
	CIFAR100	WRN28	Table 22
	ImageNet	ResNet18	Table 3
Comparison with empirical attacks against RAND pre-processing defense	CIFAR10	VGG16	Table 23
	CIFAR10	ResNet110	Table 24
	CIFAR10	ResNext29	Table 25
	CIFAR10	WRN28	Table 26
	CIFAR100	VGG16	Table 27
	CIFAR100	ResNet110	Table 28
	CIFAR100	ResNext29	Table 29
	CIFAR100	WRN28	Table 30
	ImageNet	ResNet18	Table 4
Comparison with empirical attacks against RAND post-processing defense	CIFAR10	VGG16	Table 31
	CIFAR10	ResNet110	Table 32
	CIFAR10	ResNext29	Table 33
	CIFAR10	WRN28	Table 34
	CIFAR100	VGG16	Table 35
	CIFAR100	ResNet110	Table 36
	CIFAR100	ResNext29	Table 37
	CIFAR100	WRN28	Table 38
	ImageNet	ResNet18	Table 5
Comparison with empirical attack against adversarial training	CIFAR10	ResNet110 (ℓ_2)	Table 6
	CIFAR10	ResNet110 (ℓ_∞)	
Ablation: CA vs. different noise variance	CIFAR10	ResNet110	Table 7
	ImageNet	ResNet50	
	LibriSpeech	ECAPA-TDNN	Table 39
Ablation: CA vs. different p	CIFAR10	ResNet110	Table 8
	ImageNet	ResNet50	
	LibriSpeech	ECAPA-TDNN	Table 40
Ablation: CA vs. different Localization/Shifting	CIFAR10	ResNet110	Table 9
	ImageNet	ResNet50	
	LibriSpeech	ECAPA-TDNN	Table 41
Ablation: CA vs. different noise PDF	CIFAR10	ResNet110	Table 10
	ImageNet	ResNet50	
Ablation: CA w/ and w/o Diffusion Denoise	CIFAR10	ResNet110	Table 14
	ImageNet	ResNet50	
CA vs. Feature Squeezing	CIFAR10	ResNet110	Figure 8
CA vs. Adaptive Denoiser	CIFAR10	ResNet110	Table 11
CA vs. Rand. Smoothing	CIFAR10	ResNet110	Table 13

5 EVALUATIONS

We comprehensively evaluate our certifiable black-box attack in various experimental settings. Particularly, we would like to study the following research questions:

- **RQ1:** How effective is the learnt *Adversarial Distribution*? Particularly, how large is the probability of samples from it being successful adversarial examples?
- **RQ2:** Can our certifiable attack outperform empirical attacks in terms of attack effectiveness and query efficiency?
- **RQ3:** How effective is our attack to break SOTA defenses?
- **RQ4:** What is the impact of the design components and their hyperparameters on our attack?

Accordingly, we first assess the empirical attack success possibility of the *Adversarial Distribution* in Section 5.2. Then, we evaluate our certifiable attack on various models with defenses while benchmarking with empirical black-box attacks in Section 5.3. In Section 5.4, we conduct ablation studies to explore in-depth our certifiable attack. *All sets of experiments are summarized in Table 2 for reference.*[4]

5.1 Experimental Setup

Datasets and Models. We use three benchmark datasets for image classification: CIFAR10/CIFAR100 [43] and ImageNet [65]. CIFAR10 and CIFAR100, both consisting of 60,000 $32x32$ color images split into 10 and 100 classes, respectively. ImageNet is a large-scale dataset with 1,000 classes. The training set contains 1,281,167 images and the validation set contains 50,000 images (resized to $3 \times 224 \times 224$). we use VGG [70], ResNet [32], ResNext [84], and WRN [89] as the target model. We use a pre-trained ResNet34 on ImageNet as the feature extractor (in the Smoothed SSP localization). We also test our attacks on the audio dataset LibriSpeech [42] for the speaker verification task, and results are shown in Appendix C.5.

Baseline Attacks. We compare our certifiable (hard label-based) black-box attack with SOTA black-box attacks including 7 *hard label-based* black-box attacks: GeoDA [62], HSJ [13], Opt [19], RayS [14], SignFlip [18], SignOPT [20], and Boundary [7]; and 7 *score-based* black-box attacks: Bandit [39], NES [38], Parsimonious [54], Sign [1], Square [2], ZOSignSGD [49], Simple attack [30]. As our method does not constrain the perturbation budget but minimizing the perturbation, we also compare with two similar attacks: SparseEvo [78] and PointWise [67]. We evaluate our attack with both SSSP localization and binary-search localization. For optimized-based attacks, we limit the AEs in the valid image space. For a fair comparison with optimization-based attacks, the perturbation budget for ℓ_p-bounded attacks are set to 0.1 for ℓ_∞ and 5 for ℓ_2 on CIFAR10 and CIFAR100, while on ImageNet, they are $\ell_\infty = 0.1$ and $\ell_2 = 40$. The maximum query limits are 10,000 for CIFAR10 and CIFAR100 and 1,000 for ImageNet. We evaluate 1,000 randomly selected images for each dataset.

Defenses. We select 4 SOTA defenses against black-box attacks for evaluation: Blacklight detection [46], Randomized pre-processing defense (RAND-Pre) [60], Randomized post-processing defense (RAND-Post) [12], and Adversarial Training based TRADES [90]. Blacklight has recently proposed to mitigate query-based black-box attacks by utilizing the similarity among queries. It has been shown to detect 100% adversarial examples generated in multiple attacks. RAND-Pre and RAND-Post respectively add noise to the inputs and prediction logits to obfuscate the gradient estimation or local search. TRADES has demonstrated SOTA robustness performance against adversarial attacks by training on adversarial examples.

Metrics. We use the below metrics to evaluate all compared attacks.

- **Model Accuracy:** the model accuracy under attack and defense.
- **Number of RPQ (# RPQ):** the number of the randomized parallel query for certifiable attack.
- **Number of Query (# Q):** the total number of queries for empirical attack. For our method, it is equal to Monte Carlo Sampling Number × # RPQ + additional queries for sampling from the *Adversarial Distribution*.
- **Certified Accuracy@p:** the certified accuracy at the ASP Threshold p. It is the percentage of the testing samples that have the

[4]Additional results including Tables 12-41 and Figures 8-9 are presented in Appendix in the full version at https://arxiv.org/abs/2304.04343

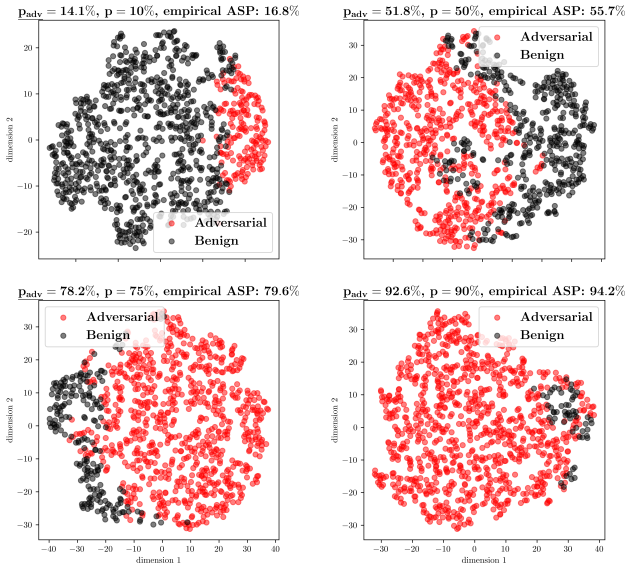


Figure 5: t-SNE visualization of adversarial example sampling from the adversarial distribution.

certified ASP at least p, e.g., a 95% certified accuracy with ASP Threshold $p = 90\%$ means the adversary can guarantee to have 90% probability to attack successfully for 95% testing samples.

- ℓ_2 **Perturbation Size (Dist. ℓ_2):** ℓ_2 distance between the adversarial example x_{adv} and the clean input x, i.e., $\|x_{adv} - x\|_2$.
- ℓ_2 **Mean Distance (Mean Dist. ℓ_2):** ℓ_2 distance between the mean x' of *Adversarial Distribution* and clean input x, i.e., $\|x' - x\|_2$.
- **Detection Success Rate (Det. Rate):** the detection success rate of Blacklight detection.
- **Average # Queries for Detection (# Q to Det.):** the average number of queries before Blacklight detects an AE.
- **Detection Coverage (Det. Cov.):** the percent of queries in an attack's query sequence that Blacklight identified as attack queries.

Parameters Settings. There exist many parameters that may affect the performance of our certifiable attack. For instance, the Monte Carlo sampling number, the attack success probability p, and the family of the adversarial distribution and its parameters. If not specified, we set Monte Carlo sampling number to be 50, $p = 10\%$, and use Gaussian distribution with variance $\sigma = 0.025$. We will also study the impact of these parameters in Section 5.3. All the parameter details are summarized in Table 12 in Appendix C.1.

Experimental Environment. We implemented a PyTorch library[5] including 16 black-box attacks, 4 defenses, 6 datasets, and 9 models by integrating several open-source libraries[6]. The experiments were run on a server with AMD EPYC Genoa 9354 CPUs (32 Core, 3.3GHz), and NVIDIA H100 Hopper GPUs (80GB each).

5.2 Verifying the Adversarial Distribution

We first assess the ASP of the crafted *Adversarial Distribution*. Specifically, given an ASP Threshold p and the certified *Adversarial Distribution* $\varphi(x', \kappa)$, we randomly sample 1,000 examples $x_{adv} \sim \varphi(x', \kappa)$, and query the model. We visualize the query results for 4 certified *Adversarial Distributions* with different p using 2D t-SNE[7] [77]. We also report the provable lower bound of ASP $\underline{p_{adv}}$ and the empirical ASP in Figure 5. It validates that the sampled AEs ensure the minimum ASP via the *Adversarial Distribution*, and the *Adversarial Distribution* lies on the decision boundary.

Table 3: Attack performance under Blacklight detection on ResNet and ImageNet (Clean Accuracy: 67.9%)

Attack	Query Type	Pert. Type	Det. Rate %	# Q to Det.	Det. Cov. %	Model Acc.	# Q	Dist. ℓ_2
Bandit	Score	ℓ_∞	100.0	1.0	64.2	1.9	25	25.42
NES	Score	ℓ_∞	100.0	10.3	17.3	7.0	337	8.28
Parsimonious	Score	ℓ_∞	100.0	2.0	96.7	3.8	282	25.24
Sign	Score	ℓ_∞	100.0	2.0	91.5	0.5	126	25.50
Square	Score	ℓ_∞	100.0	2.0	66.9	0.0	14	25.54
ZOSignSGD	Score	ℓ_∞	100.0	2.0	50.2	12.5	322	8.53
GeoDA	Label	ℓ_∞	100.0	1.0	88.9	5.1	151	17.99
HSJ	Label	ℓ_∞	100.0	7.3	94.9	35.6	212	9.82
Opt	Label	ℓ_∞	99.9	8.4	81.4	61.2	646	0.98
RayS	Label	ℓ_∞	100.0	4.4	83.5	4.2	260	29.63
SignFlip	Label	ℓ_∞	100.0	8.5	70.3	4.4	148	27.64
SignOPT	Label	ℓ_∞	99.9	8.4	69.8	55.9	570	1.32
Bandit	Score	ℓ_2	100.0	1.0	99.5	1.7	431	9.60
NES	Score	ℓ_2	100.0	10.2	32.8	61.2	571	0.45
Simple	Score	ℓ_2	100.0	1.0	99.9	53.6	883	0.88
Square	Score	ℓ_2	100.0	2.0	68.8	0.0	16	26.30
ZOSignSGD	Score	ℓ_2	100.0	2.0	52.4	65.1	531	0.30
Boundary	Label	ℓ_2	100.0	7.2	76.3	37.9	60	11.63
GeoDA	Label	ℓ_2	100.0	1.0	89.3	3.9	181	19.14
HSJ	Label	ℓ_2	100.0	7.3	93.4	11.4	255	22.21
Opt	Label	ℓ_2	100.0	8.5	67.9	41.2	610	16.71
SignOPT	Label	ℓ_2	99.9	8.4	62.9	36.7	485	17.54
PointWise	Label	Opt.	100.0	1.0	99.8	0.0	920	13.53
SparseEvo	Label	Opt.	100.0	1.0	99.9	0.0	1000	7.68
CA (sssp)	Label	Opt.	0.0	∞	0.0	1.4	148	13.74
CA (bin search)	Label	Opt.	0.0	∞	0.0	0.0	603	33.14

5.3 Attack Performance against SOTA Defenses

In this section, we evaluate our certifiable attack and empirical attacks against the 4 studied SOTA defenses.

5.3.1 Attack Performance under Blacklight Detection [46]. We use the default setting from [46] with a threshold of 25. The results are presented in Table 3 and Tables 15-22 in Appendix C. We have the following key observations: 1) Our certifiable attack consistently circumvents Blacklight with 0% detection success rate, and 0% detection coverage on all settings. This indicates that none of the queries from our attack are detected. In contrast, existing black-box attacks are highly susceptible to Blacklight, with most achieving a 100% detection success rate on various datasets and models. Even the most resilient attack, as shown in Appendix C, Table 19, attains an 86.5% detection success rate on the CIFAR100 dataset using the VGG16 model. 2) With the strong ability to bypass the detection, our certifiable attack still maintains top attack performance on all the datasets and models such that the model accuracy can be attacked to 0% with moderate ℓ_2 perturbation size and few queries. The high attack accuracy and low detection rate of certifiable attacks stem from the randomness of *Adversarial Distribution* and the guarantee of the attack success probability.

5.3.2 Attack Performance under RAND-Pre [60]. We follow [60] to inject the Gaussian noise with standard deviation 0.02 to the

[5]The codes are available at https://github.com/datasec-lab/CertifiedAttack
[6]BlackboxBench, pytorch image classification, Blacklight, SparseEvo, and TRADES
[7]t-SNE reduces the prediction logits of the random samples to 2-dimension.

Table 4: Attack performance under RAND Pre-processing Defense on ResNet and ImageNet (Clean Accuracy: 67.0%)

Attack	Query Type	Perturbation Type	# Query	Model Acc.	Dist. ℓ_2
Bandit	Score	ℓ_∞	10	6.7	25.26
NES	Score	ℓ_∞	428	49.8	10.26
Parsimonious	Score	ℓ_∞	243	62.7	25.12
Sign	Score	ℓ_∞	116	40.6	25.20
Square	Score	ℓ_∞	27	10.4	24.96
ZOSignSGD	Score	ℓ_∞	428	49.4	10.36
GeoDA	Label	ℓ_∞	150	40.0	18.08
HSJ	Label	ℓ_∞	232	58.5	8.76
Opt	Label	ℓ_∞	905	69.4	0.44
RayS	Label	ℓ_∞	235	47.9	28.14
SignFlip	Label	ℓ_∞	46	52.8	13.06
SignOPT	Label	ℓ_∞	394	59.1	0.39
Bandit	Score	ℓ_2	583	58.2	12.99
NES	Score	ℓ_2	341	66.8	0.43
Simple	Score	ℓ_2	258	67.2	0.10
Square	Score	ℓ_2	18	13.6	25.96
ZOSignSGD	Score	ℓ_2	249	67.3	0.28
Boundary	Label	ℓ_2	38	49.2	15.12
GeoDA	Label	ℓ_2	149	47.4	17.70
HSJ	Label	ℓ_2	225	55.7	14.30
Opt	Label	ℓ_2	1000	58.2	12.28
SignOPT	Label	ℓ_2	406	52.2	15.41
PointWise	Label	Optimized	942	54.6	16.90
SparseEvo	Label	Optimized	1000	61.7	11.33
CA (sssp)	Label	Optimized	154	1.7	13.98
CA (bin search)	Label	Optimized	603	0.0	32.16

Table 5: Attack performance under RAND Post-processing Defense on ResNet and ImageNet (Clean Accuracy: 68.0%)

Attack	Query Type	Perturbation Type	# Query	Model Acc.	Dist. ℓ_2
Bandit	Score	ℓ_∞	17	2.7	25.51
NES	Score	ℓ_∞	378	18.6	9.53
Parsimonious	Score	ℓ_∞	253	47.9	25.46
Sign	Score	ℓ_∞	124	8.1	25.81
Square	Score	ℓ_∞	18	0.8	25.44
ZOSignSGD	Score	ℓ_∞	376	21.4	9.71
GeoDA	Label	ℓ_∞	143	38.6	17.62
HSJ	Label	ℓ_∞	212	52.7	8.82
Opt	Label	ℓ_∞	1000	65.3	0.67
RayS	Label	ℓ_∞	243	43.9	28.09
SignFlip	Label	ℓ_∞	86	47.2	15.44
SignOPT	Label	ℓ_∞	412	63.6	0.64
Bandit	Score	ℓ_2	596	6.0	13.96
NES	Score	ℓ_2	344	59.7	0.44
Simple	Score	ℓ_2	241	58.9	0.10
Square	Score	ℓ_2	23	0.4	26.46
ZOSignSGD	Score	ℓ_2	275	61.4	0.29
Boundary	Label	ℓ_2	24	48.0	12.64
GeoDA	Label	ℓ_2	146	40.6	16.89
HSJ	Label	ℓ_2	238	49.5	14.59
Opt	Label	ℓ_2	1000	53.9	12.62
SignOPT	Label	ℓ_2	411	46.3	15.96
PointWise	Label	Optimized	969	55.1	16.01
SparseEvo	Label	Optimized	1000	66.7	9.10
CA (sssp)	Label	Optimized	147	1.4	13.70
CA (bin search)	Label	Optimized	603	0.0	32.67

query (in the input space). The experimental results are presented in Table 4 and Tables 23-30 in Appendix C. Based on a comprehensive analysis of all results, it is evident that the RAND-Pre consistently reduces the attack success rate of existing black-box attacks. Specifically, the defense reduces the average attack success rate of empirical black-box attacks from 92% to 30% on CIFAR10, from 95% to 29% on CIFAR100, and from 69% to 25% on ImageNet, respectively. However, our attack still achieves the average attack success rate of 93%, 99%, and 99% respectively on the three datasets under RAND-Pre. Further, we highlight that, with RAND-Pre applied across all datasets and models, the average ℓ_2 perturbation size and number of queries in our certifiable attack *decrease by* 4.2% *and* 2.1%, respectively. This intriguing observation matches our findings in Section 5.4.1 where a larger variance leads to smaller ℓ_2 mean distance and # RPQ in the certifiable attack. This is because the Gaussian noise injected by the defense (e.g., $\epsilon_1 \sim \mathcal{N}(0, v)$) is added to the adversary's noise (e.g., $\epsilon_2 \sim \mathcal{N}(0, u)$), leading to a larger variance $v + u$ and hence further enhancing our attack.

5.3.3 Attack Performance under RAND-Post [12]. We follow [12] to inject the Gaussian noise with standard deviation 0.2 to the output logits of each query (applied to both hard label-based and score-based attacks). The experimental results are presented in Table 5, and Tables 31-38. Similarly, we find that RAND-Post can strongly degrade the average attack success rate of hard label-based empirical attacks from 84% to 41% on CIFAR10, from 89% to 45% on CIFAR100, and from 60% to 24% on ImageNet, respectively. On the other hand, it moderately degrades the average attack success rate of score-based empirical attacks from 100% to 91% on CIFAR10, from 100% to 95% on CIFAR100, and from 72% to 62% on ImageNet. The discrepancy between label-based and score-based empirical attacks may stem from variations in the richness and smoothness of the query information. The loss value (score), providing a smoother evaluation, is less susceptible to noise interference and discloses finer-grained details. In contrast, labels are more likely to be impacted by injected noise, resulting in more randomized query outcomes. However, our hard-label certifiable attack shows strong resilience against RAND-Post, by maintaining the average attack success rate at 93%, 99%, and 99% on CIFAR10, CIFAR100, and ImageNet, respectively. This advantage over empirical attacks, particularly the label-based ones, originates from Randomized Parallel Querying—It precisely assesses query results with a lower bound of the ASP.

5.3.4 Attack Performance under TRADES [90]. We consider both ℓ_∞ and ℓ_2 perturbations to generate adversarial examples, and TRADES respectively uses ℓ_2 or ℓ_∞ adversarial examples for adversarial training. We set the perturbation size to be $\ell_\infty = 0.1$ and $\ell_2 = 5$, following [90]. We then evaluate all attacks against TRADES. The results on CIFAR10 are presented in Table 6[8]. We observe our attack requires much less query number than the empirical attacks. Also, our attack can achieve 100% attack success rate (with the binary search localization), but at the cost of a relatively larger perturbation size.

5.4 Ablation Study

In this section, we explore in-depth our certifiable attack—we study its performance with varying noise variances, ASP thresholds, localization and shifting methods, and noise PDFs. We mainly show results on the image datasets and defer results on the audio dataset to Appendix C, where similar performance can be observed.

[8]It is computationally intensive and time-consuming to train TRADES on CIFAR100 and ImageNet

Table 6: Attack performance under TRADES Adversarial Training on ResNet and CIFAR10

Defense	Attack	Query Type	Pert. Type	# Query	Model Acc.	Dist. ℓ_2
ℓ_∞ Adversarial Training (Clean Accuracy: 80.9%)	Bandit	Score	ℓ_∞	1601	10.2	4.32
	NES	Score	ℓ_∞	1474	27.4	2.27
	Parsimonious	Score	ℓ_∞	630	5.3	4.35
	Sign	Score	ℓ_∞	439	4.4	4.37
	Square	Score	ℓ_∞	854	5.7	4.39
	ZOSignSGD	Score	ℓ_∞	1196	37.0	2.21
	GeoDA	Label	ℓ_∞	1358	41.9	1.99
	HSJ	Label	ℓ_∞	2149	36.5	1.92
	Opt	Label	ℓ_∞	1871	73.0	0.19
	RayS	Label	ℓ_∞	721	7.3	4.29
	SignFlip	Label	ℓ_∞	2240	24.4	3.36
	SignOPT	Label	ℓ_∞	832	69.3	0.15
	PointWise	Label	Opt.	3460	9.3	4.38
	SparseEvo	Label	Opt.	8691	9.1	5.10
	CA (sssp)	Label	Opt.	548	21.2	4.29
	CA (bin search)	Label	Opt.	412	9.8	6.31
ℓ_2 Adversarial Training (Clean Accuracy: 59.2%)	Bandit	Score	ℓ_2	860	1.5	2.44
	NES	Score	ℓ_2	3535	9.5	0.99
	Simple	Score	ℓ_2	4062	2.1	1.29
	Square	Score	ℓ_2	991	4.6	2.95
	ZOSignSGD	Score	ℓ_2	3505	15.1	0.77
	Boundary	Label	ℓ_2	771	40.7	1.19
	GeoDA	Label	ℓ_2	1506	14.3	2.85
	HSJ	Label	ℓ_2	1332	5.1	3.53
	Opt	Label	ℓ_2	2890	41.6	2.39
	SignOPT	Label	ℓ_2	1766	33.6	2.76
	PointWise	Label	Opt.	4845	0.6	5.36
	SparseEvo	Label	Opt.	9697	0.4	6.03
	CA (sssp)	Label	Opt.	809	20.4	6.06
	CA (bin search)	Label	Opt.	461	0.0	8.18

5.4.1 Attack Performance on Different Noise Variances. Table 7 shows the performance of our attack with varying noise variances used in φ. We have the following key observations: 1) As the variance increases, the ℓ_2 perturbation size increases, since larger variance results in larger noise. 2) The ℓ_2 mean distance tends to decrease as the variance increases. This could be because that larger variance covers a larger decision space, and without moving the mean far away from the clean input, we can easily find a large portion of adversarial samples under the distribution with a large variance. 3) As the variance increases, the number of RPQ decreases. This is because the larger variance usually leads to a larger shifting step. It takes fewer iterations to move to the decision boundary when the variance increases. 4) Finally, a larger certified accuracy means that it is easier to determine the *Adversarial Distribution*. The results on CIFAR10 show that it is easier to find a small area of adversarial examples than a large area of adversarial examples. On ImageNet, we observe nearly 100% certified accuracy, which means it is relatively easy to find the adversarial examples on datasets with a large number of classes (since 999 out of 1,000 classes in ImageNet are all false classes) or with high feature dimension.

5.4.2 Attack Performance on Different ASP Thresholds. We study the relationship between the performance of our attack and the ASP threshold, and Table 8 shows the results. As p increases, so do the ℓ_2 perturbation size, the ℓ_2 mean distance, and the number of RPQ. On one hand, a larger p means it requires more adversarial examples to fall into the false classes. When the noise variance is fixed, the mean of the *Adversarial Distribution* should be further away from the decision boundary to allow more adversarial examples to fall into the false classes. On the other hand, the smaller p results in a larger shifting distance, which depends on the gap between p and $\underline{p_{adv}}$ (see the Gaussian-case of Theorem 1 in Appendix A.2). With a larger shifting distance, the required number of RPQ can be fewer. We also observe that a smaller p results in a higher certified accuracy on CIFAR10, since a smaller p allows more "failed" adversarial examples. On ImageNet, the certified accuracy is consistently $\sim$ 100%, no matter p's value. This might still because it is much easier to find adversarial examples with a much larger number of classes.

Table 7: Attack performance of our certifiable attack with varying Gaussian noise variances σ (p = 90%)

	σ	Dist. ℓ_2	Mean Dist. ℓ_2	# RPQ	Certified Acc.
CIFAR10	0.10	7.39	3.96	18.34	94.17%
	0.25	12.95	2.34	14.35	91.21%
	0.50	19.41	0.43	11.38	90.00%
ImageNet	0.10	41.80	16.78	32.55	99.80%
	0.25	87.47	16.78	17.02	99.60%
	0.50	135.47	2.27	8.31	100.00%

Table 8: Attack performance of our certifiable attack with varying p under the Gaussian variance σ = 0.25

	p	Dist. ℓ_2	Mean Dist. ℓ_2	# RPQ	Certified Acc.
CIFAR10	50%	12.65	1.63	9.34	97.17%
	60%	12.72	1.86	11.09	95.85%
	70%	12.80	2.05	11.94	94.72%
	80%	12.87	2.18	12.37	93.17%
	90%	12.95	2.34	14.35	91.21%
	95%	13.09	2.65	15.93	90.37%
ImageNet	50%	85.88	9.89	12.85	100.00%
	60%	86.20	11.30	13.63	100.00%
	70%	86.45	12.64	14.33	100.00%
	80%	87.03	14.64	16.02	100.00%
	90%	87.47	16.78	17.02	99.60%
	95%	88.42	19.98	19.81	100.00%

Table 9: Attack performance of our certifiable attack on different localization/refinement algorithms (σ = 0.25, p = 90%)

Localization	Refinement	Dist. ℓ_2	Mean Dist. ℓ_2	# RPQ	Cert. Acc.
sssp	none	11.46	1.35	2.30	92.54
binary search	none	11.29	0.34	9.07	92.54
random	geo.	11.80	1.73	67.53	92.54
sssp	geo.	11.20	0.49	3.70	91.54
binary search	geo.	11.28	0.27	10.08	92.53

5.4.3 Attack Performance on Different Localization/Refinement Algorithms. In this experiment, we compare our proposed Smoothed SSP and binary-search localization methods with the random localization baseline; and compare our proposed geometric shifting method with a no-shifting baseline. Results are shown in Table 9. We observe that the combination of the localization and refinement methods yields the smallest perturbation size, i.e., the smallest Dist. ℓ_2 and Mean Dist. ℓ_2. This demonstrates that they are both effective in improving the imperceptibility of adversarial examples.

Visualization. We also visualize the adversarial examples x_{adv} while crafting the Certifiable Attack for Binary-search Localization (Figure 6) and SSSP Localization (Figure 7). It shows that when $\sigma = 0.025$, both the Binary-search and SSSP-based certifiable attack can craft imperceptible perturbations. The difference is that the binary search method starts from a random x' and requires more # RPQ to update the *Adversarial Distribution*, while the SSSP can easily find an initial *Adversarial Distribution*with small perturbation and thus requires fewer # RPQ.

5.4.4 Attack Performance on Different Noise Distributions. Our attack can use any continuous noise distribution to craft the *Adversarial Distribution*. Besides the Gaussian noise distribution, in this experiment, we also evaluate the performance of our certifiable attack using other noise distributions including the Cauthy distribution, Hyperbolic Secant distribution, and general normal distributions. Note that we adjust the parameters in these distributions to ensure consistent variances for a fair comparison.

The results are presented in Table 10, and the noise distributions are plotted in Figure 9 in Appendix. On both datasets, we observe the ℓ_2 perturbation size is decreasing while the ℓ_2 mean distance is increasing as the PDF of the noise distribution is more centralized. This result may share a similar nature with results in Table 7—when the adversarial samples are more widely distributed, they tend to fall into an adversarial class (the majority of all classes). It is hard to determine which distribution is better since there is a trade-off between the perturbation size and the number of RPQ.

5.5 Defending against Our Certifiable Attack

In this subsection, we discuss potential defenses and mitigation strategies against our attacks.

Noise Detection based Defenses: Our certifiable attack injects noise into the adversarial examples. Here, we suppose the adversary is aware of the noise injection and designs a detection method by training a binary classifier to distinguish the noise-injected inputs and clean inputs. Specifically, the defender (i.e., model owner) uses ResNet110 (as powerful as the target model) to train a noise detector to distinguish the inputs with and without noise. The experimental results show the noise detection rate can be as high as 99% with the noise variance $\sigma = 0.5$, which means this detector can be used as a strong defense against our certifiable attacks with a larger noise. However, this defense does not work when the noise scale is smaller (i.e., $\sigma = 0.025$), where the detection rate is less than 1%. Especially, the adversary may design a novel method to hide this noise in the image texture, e.g., using the diffusion model for denoising, which may circumvent the detection.

White-Box Adaptive Defenses against Our Attack: We assume the model owner knows the noise distribution used by our attack and performs a "white-box" defense. Particularly, it applies a denoiser to eliminate the injected noise, so that the adversarial examples can be restored to clean inputs. The denoiser can be deployed as a pre-processing module and is pre-trained by the model owner. Specifically, we use a U-Net structure [63] as the denoiser and denote it as $\mathcal{D}$. Then, the loss function for the training is

$$\mathbb{E}_{\epsilon\sim\mathcal{N}(0,\sigma_d)}[||\mathcal{D}(x+\epsilon)-x||_2+||f(\mathcal{D}(x+\epsilon))-f(x)||_2] \quad (8)$$

Taking Gaussian noise as an example (e.g., the model owner knows the Gaussian variance $\sigma = 0.25$ used in the certifiable attack), we train the denoiser to eliminate Gaussian noise with $\sigma = 0.25$ while evaluating the certifiable attack with Gaussian noise generated by different σ. Table 11 shows the results. We can observe that this defense can significantly degrade the performance of a certifiable attack. Notably, by choosing the same variance σ as the adversary, the adaptive defense can increase the Mean Dist. ℓ_2 significantly. However, the certified accuracy is still near 90%.

6 RELATED WORK

Adversarial Attack. It aims to mislead learnt ML models by perturbing testing data with imperceptible perturbations. It can be divided into white-box attacks [11, 29, 52, 55, 82] and black-box attacks [2, 5, 7, 7, 8, 13, 15, 15, 19, 24, 25, 30, 37, 38, 47, 57, 58, 69], per the access that the adversary holds. White-box attacks have full access to the model parameters, and can leverage the gradient of the loss function w.r.t. the inputs to guide the adversarial example generation. Instead, black-box attacks only know the outputs (in the form of prediction scores or labels) of a target model via sending queries. It is widely believed that black-box attack is more practical in real-world scenarios [6, 13, 58]. Therefore, we focus on the black-box attacks in this paper.

Black-Box Attack. Existing black-box attack methods can be classified into three types: gradient estimation based [5, 15, 19, 25, 38, 61, 72, 79, 80], surrogate models based [24, 57, 58, 69], or local search based algorithms [2, 7, 8, 27, 30, 47, 56]. Gradient estimation based attack is mainly based on zero-order estimation since the true gradient is unknown [15]. Surrogate model-based methods first perform white-box attacks on an offline surrogate model to generate adversarial examples, and then use these generated adversarial examples to test the target model. The attack performance largely depends on the transferability of such generated adversarial examples. Local search-based methods craft adversarial examples by searching the effective perturbation direction, e.g., Boundary Attack [7] traverses the decision boundary to craft the *least imperceptible* perturbations.

All existing black-box attacks rely on querying the target model until finding a successful adversarial example or reaching the maximum number of queries. However, none of them can ensure the success rate of the adversarial examples that have not been queried. Further, they are shown to be easily detected/removed via adversarial detection and randomized pre/post-processing-based defenses.

Empirical Defense. It defends against adversarial attacks without guarantees. Empirical defenses against white-box attacks can be roughly categorized into four classes. Gradient-masking defenses [23, 59, 83] modify the model inference process to obstacle the gradient computation. Input-transformation defenses [9, 31, 35, 48, 66, 71] use pre-processing methods to transform the inputs so that the malicious effects caused by the perturbations can be reduced. Detection-based defenses [40, 51, 53, 64, 74] identify features that expect to separate adversarial examples and clean examples, and train a binary classifier to detect adversarial examples. Another branch of works [17, 21, 26, 46] detects the adversarial examples based on the similarity of the queries, demonstrating high detection accuracy in practice. Among these, Blacklight [46] has shown supreme detection performance without assumptions on the user

Figure 6: Visualization of successful adversarial examples crafting by certifiable attack with binary-search localization

Table 10: Attack performance of our certifiable attack with different noise distributions

	Distribution	Density	Parameter	$\sqrt{\|\|\epsilon\|\|^2/d}$	Dist. ℓ_2	Mean Dist. ℓ_2	# RPQ	Certified Acc.
CIFAR10	Gaussian	$\propto e^{-\|z/a\|^2}$	$a = 0.25$	0.25	12.95	2.34	14.35	91.21%
	Cauthy	$\propto \frac{a^2}{z^2+a^2}$	$a = 0.01969$	0.25	7.82	4.87	32.77	94.12%
	Hyperbolic Secant	$\propto sech(\|z/a\|)$	$a = 0.1592$	0.25	12.51	2.43	14.59	91.67%
	General Normal ($b = 1.5$)	$\propto e^{-\|z/a\|^b}$	$a = 0.2909$, $b = 1.5$	0.25	12.74	2.37	14.15	91.39%
	General Normal ($b = 3.0$)	$\propto e^{-\|z/a\|^b}$	$a = 0.4092$, $b = 3$	0.25	13.16	2.38	14.15	91.25%
ImageNet	Gaussian	$\propto e^{-\|z/a\|^2}$	$a = 0.25$	0.25	87.47	16.78	17.02	99.60%
	Cauthy	$\propto \frac{a^2}{z^2+a^2}$	$a = 0.01969$	0.25	46.18	23.94	59.94	99.60%
	Hyperbolic Secant	$\propto sech(\|z/a\|)$	$a = 0.1592$	0.25	85.57	21.29	20.89	99.80%
	General Normal ($b = 1.5$)	$\propto e^{-\|z/a\|^b}$	$a = 0.2909$, $b = 1.5$	0.25	86.69	19.05	17.58	99.80%
	General Normal ($b = 3.0$)	$\propto e^{-\|z/a\|^b}$	$a = 0.4092$, $b = 3$	0.25	88.51	15.58	14.99	100.00%

Figure 7: Visualization of successful adversarial examples crafting by certifiable attack with SSSP localization (SSSP requires fewer # RPQ)

Table 11: White-box adaptive defense against our attack ($\sigma = 0.25$, $p = 90\%$) on CIFAR10

Defense Para.	Dist. ℓ_2	Mean Dist. ℓ_2	# RPQ	Cert. Acc.
$\sigma_d = 0.10$	9.99	7.73	34.11	87.51%
$\sigma_d = 0.25$	15.40	10.21	29.80	88.31%
$\sigma_d = 0.50$	20.46	8.11	26.52	86.56%

accounts. These three types of defenses show certain effectiveness when they target specific known attacks, but can be broken by adaptive attacks [4]. Lastly, adversarial training-based defenses [52, 68, 75, 76] have achieved the SOTA performance against adaptive attacks. The main idea is to augment training data with "adversarial examples", but they are reassigned the correct label. As to defend against *black-box* attacks, RAND-Post [12], RAND-Pre [60], Adversarial Training based TRADES [52], and Blacklight [46] are the SOTA in each category. Thus, we evaluated our certifiable attack under these defenses.

Certified Defense. Certified defense [3, 28, 36, 41, 81, 92] was proposed to guarantee constant classification prediction on a set of adversarial examples. Recently, randomized smoothing (RS) [22] has achieved great success in the certified defense since it is the first method to certify arbitrary classifiers of any scale. Specifically, RS can guarantee the prediction if the perturbation is bounded by a distance in ℓ_p-norm, i.e., certified radius [22, 36, 73, 88]. RS adds noise from a distribution (e.g., Gaussian) to the inputs and uses hypothesis testing to quantify the prediction probability. Then the bound on the perturbations (usually a ℓ_p norm constraint) for ensuring the consistent prediction is derived. This method is widely used in certified defense to ensure consistent and correct prediction under attack. However, in this paper, we propose to use this method to ensure consistent and wrong prediction on the *Adversarial Distribution*, resulting in a reliable and strong certifiable attack.

7 CONCLUSION

Certifiable attack lays a novel direction for adversarial attacks, enabling the transition from deterministic to probabilistic adversarial attacks. Compared with empirical black-box attacks, certifiable attacks share significant benefits including breaking SOTA strong detection and randomized defense, revealing consistent and severe robustness vulnerability of models, and guaranteeing the minimum ASP for numerous unique AEs without verifying via the query.

ACKNOWLEDGMENTS

We sincerely thank the anonymous reviewers for their constructive comments and suggestions. This work is supported in part by the National Science Foundation (NSF) under Grants No. CNS-2308730, CNS-2302689, CNS-2319277, CMMI-2326341, ECCS-2216926, CNS-2241713, CNS-2331302 and CNS-2339686. It is also partially supported by the Cisco Research Award and the Synchrony Fellowship.

REFERENCES

[1] Abdullah Al-Dujaili and Una-May O'Reilly. Sign bits are all you need for black-box attacks. In *International Conference on Learning Representations*, 2020.
[2] Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square attack: A query-efficient black-box adversarial attack via random search. In *ECCV*, 2020.
[3] Cem Anil, James Lucas, and Roger B. Grosse. Sorting out lipschitz function approximation. In *ICML*, volume 97, pages 291–301. PMLR, 2019.
[4] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International conference on machine learning*, pages 274–283. PMLR, 2018.
[5] Arjun Nitin Bhagoji, Warren He, Bo Li, and Dawn Song. Practical black-box attacks on deep neural networks using efficient query mechanisms. In *ECCV*. Springer, 2018.
[6] Siddhant Bhambri, Sumanyu Muku, Avinash Tulasi, and Arun Balaji Buduru. A survey of black-box adversarial attacks on computer vision models. *arXiv preprint arXiv:1912.01667*, 2019.
[7] Wieland Brendel, Jonas Rauber, and Matthias Bethge. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. In *ICLR*. OpenReview.net, 2018.
[8] Thomas Brunner, Frederik Diehl, Michael Truong-Le, and Alois C. Knoll. Guessing smart: Biased sampling for efficient black-box adversarial attacks. In *ICCV*, pages 4957–4965. IEEE, 2019.
[9] Jacob Buckman, Aurko Roy, Colin Raffel, and Ian J. Goodfellow. Thermometer encoding: One hot way to resist adversarial examples. In *ICLR*, 2018.
[10] Nicholas Carlini, Florian Tramer, Krishnamurthy Dj Dvijotham, Leslie Rice, Mingjie Sun, and J Zico Kolter. (certified!!) adversarial robustness for free! *arXiv preprint arXiv:2206.10550*, 2022.
[11] Nicholas Carlini and David A. Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy*, 2017.
[12] Varun Chandrasekaran, Kamalika Chaudhuri, Irene Giacomelli, Somesh Jha, and Songbai Yan. Exploring connections between active learning and model extraction. In *USENIX Security*. USENIX Association, 2020.
[13] Jianbo Chen, Michael I. Jordan, and Martin J. Wainwright. Hopskipjumpattack: A query-efficient decision-based attack. In *IEEE Symposium on Security and Privacy*, pages 1277–1294. IEEE, 2020.
[14] Jinghui Chen and Quanquan Gu. Rays: A ray searching method for hard-label adversarial attack. In *KDD*, 2020.
[15] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. ZOO: zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *AISec@CCS*, 2017.
[16] Sizhe Chen, Zhehao Huang, Qinghua Tao, Yingwen Wu, Cihang Xie, and Xiaolin Huang. Adversarial attack on attackers: Post-process to mitigate black-box score-based query attacks. *NeurIPS*, 2022.
[17] Steven Chen, Nicholas Carlini, and David Wagner. Stateful detection of black-box adversarial attacks. In *Proceedings of the 1st ACM Workshop on Security and Privacy on Artificial Intelligence*, pages 30–39, 2020.
[18] Weilun Chen, Zhaoxiang Zhang, Xiaolin Hu, and Baoyuan Wu. Boosting decision-based black-box adversarial attacks with random sign flip. In *ECCV*, 2020.
[19] Minhao Cheng, Thong Le, Pin-Yu Chen, Huan Zhang, Jinfeng Yi, and Cho-Jui Hsieh. Query-efficient hard-label black-box attack: An optimization-based approach. In *ICLR*. OpenReview.net, 2019.
[20] Minhao Cheng, Simranjit Singh, Patrick H. Chen, Pin-Yu Chen, Sijia Liu, and Cho-Jui Hsieh. Sign-opt: A query-efficient hard-label adversarial attack. In *ICLR*. OpenReview.net, 2020.
[21] SeokHwan Choi, Jin-Myeong Shin, and Yoon-Ho Choi. PIHA: detection method using perceptual image hashing against query-based adversarial attacks. *Future Gener. Comput. Syst.*, 145:563–577, 2023.
[22] Jeremy M. Cohen, Elan Rosenfeld, and J. Zico Kolter. Certified adversarial robustness via randomized smoothing. In *ICML*, 2019.
[23] Guneet S. Dhillon, Kamyar Azizzadenesheli, Zachary C. Lipton, Jeremy Bernstein, Jean Kossaifi, Aran Khanna, and Animashree Anandkumar. Stochastic activation pruning for robust adversarial defense. In *ICLR*, 2018.
[24] Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *CVPR*, 2019.
[25] Yali Du, Meng Fang, Jinfeng Yi, Jun Cheng, and Dacheng Tao. Towards query efficient black-box attacks: An input-free perspective. In *ACM CCS*, 2018.
[26] Bardia Esmaeili, Amin Azmoodeh, Ali Dehghantanha, Hadis Karimipour, Behrouz Zolfaghari, and Mohammad Hammoudeh. Iiot deep malware threat hunting: From adversarial example detection to adversarial scenario detection. *IEEE Trans. Ind. Informatics*, 18(12):8477–8486, 2022.
[27] Houxiang Fan, Binghui Wang, Pan Zhou, Ang Li, Zichuan Xu, Cai Fu, Hai Li, and Yiran Chen. Reinforcement learning-based black-box evasion attacks to link prediction in dynamic graphs. In *2021 IEEE 23rd Int Conf on High Performance Computing & Communications*, 2021.
[28] Matteo Fischetti and Jason Jo. Deep neural networks and mixed integer linear optimization. *Constraints An Int. J.*, 23(3):296–309, 2018.
[29] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015*, 2015.
[30] Chuan Guo, Jacob R. Gardner, Yurong You, Andrew Gordon Wilson, and Kilian Q. Weinberger. Simple black-box adversarial attacks. In *ICML 2019*.
[31] Chuan Guo, Mayank Rana, Moustapha Cissé, and Laurens van der Maaten. Countering adversarial images using input transformations. In *ICLR*, 2018.
[32] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.
[33] Zhezhi He, Adnan Siraj Rakin, and Deliang Fan. Parametric noise injection: Trainable randomness to improve deep neural network robustness against adversarial attack. In *CVPR*, 2019.
[34] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.
[35] Hanbin Hong, Yuan Hong, and Yu Kong. An eye for an eye: Defending against gradient-based attacks with gradients. *arXiv preprint arXiv:2202.01117*, 2022.
[36] Hanbin Hong, Binghui Wang, and Yuan Hong. Unicr: Universally approximated certified robustness via randomized smoothing. In *European Conference on Computer Vision*, pages 86–103. Springer, 2022.
[37] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Query-efficient black-box adversarial examples. 2017.
[38] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks with limited queries and information. In *ICML*, 2018.
[39] Andrew Ilyas, Logan Engstrom, and Aleksander Madry. Prior convictions: Black-box adversarial attacks with bandits and priors. In *International Conference on Learning Representations*, 2019.
[40] Shubham Jain, Ana-Maria Crețu, and Yves-Alexandre de Montjoye. Adversarial detection avoidance attacks: Evaluating the robustness of perceptual hashing-based client-side scanning. In *USENIX Security Symposium*, 2022.
[41] Guy Katz, Clark W. Barrett, David L. Dill, Kyle Julian, and Mykel J. Kochenderfer. Reluplex: An efficient SMT solver for verifying deep neural networks. In *Computer Aided Verification - 29th International Conference, CAV 2017*, 2017.
[42] Matěj Korvas, Ondřej Plátek, Ondřej Dušek, Lukáš Žilka, and Filip Jurčíček. Free English and Czech telephone speech corpus shared under the CC-BY-SA 3.0 license. In *(LREC 2014)*, 2014.
[43] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.
[44] Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial machine learning at scale. In *International Conference on Learning Representations*, 2017.
[45] Mathias Lecuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. Certified robustness to adversarial examples with differential privacy. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 656–672. IEEE, 2019.
[46] Huiying Li, Shawn Shan, Emily Wenger, Jiayun Zhang, Haitao Zheng, and Ben Y. Zhao. Blacklight: Scalable defense for neural networks against query-based black-box attacks. In *USENIX Security*, 2022.
[47] 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 *ICML*, 2019.
[48] Fangzhou Liao, Ming Liang, Yinpeng Dong, Tianyu Pang, Xiaolin Hu, and Jun Zhu. Defense against adversarial attacks using high-level representation guided denoiser. In *CVPR*, 2018.
[49] Sijia Liu, Pin-Yu Chen, Xiangyi Chen, and Mingyi Hong. signsgd via zeroth-order oracle. In *International Conference on Learning Representations*, 2019.
[50] Xuanqing Liu, Minhao Cheng, Huan Zhang, and Cho-Jui Hsieh. Towards robust neural networks via random self-ensemble. In *Proceedings of the european*

conference on computer vision (ECCV), pages 369–385, 2018.

[51] Jiajun Lu, Theerasit Issaranon, and David A. Forsyth. Safetynet: Detecting and rejecting adversarial examples robustly. In *IEEE International Conference on Computer Vision, ICCV 2017*, pages 446–454. IEEE Computer Society, 2017.

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

[53] Dongyu Meng and Hao Chen. Magnet: A two-pronged defense against adversarial examples. In Bhavani Thuraisingham, David Evans, Tal Malkin, and Dongyan Xu, editors, *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, CCS 2017*, pages 135–147. ACM, 2017.

[54] Seungyong Moon, Gaon An, and Hyun Oh Song. Parsimonious black-box adversarial attacks via efficient combinatorial optimization. In *ICML*, 2019.

[55] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. *CoRR*, abs/1610.08401, 2016.

[56] Jiaming Mu, Binghui Wang, Qi Li, Kun Sun, Mingwei Xu, and Zhuotao Liu. A hard label black-box adversarial attack against graph neural networks. In *ACM SIGSAC Conference on Computer and Communications Security*, 2021.

[57] Muzammal Naseer, Salman H. Khan, Munawar Hayat, Fahad Shahbaz Khan, and Fatih Porikli. A self-supervised approach for adversarial robustness. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020*, pages 259–268. Computer Vision Foundation / IEEE, 2020.

[58] Nicolas Papernot, Patrick D. McDaniel, Ian J. Goodfellow, Somesh Jha, Z. Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In Ramesh Karri, Ozgur Sinanoglu, Ahmad-Reza Sadeghi, and Xun Yi, editors, *AsiaCCS 2017*.

[59] Nicolas Papernot, Patrick D. McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *IEEE Symposium on Security and Privacy, SP 2016*, pages 582–597. IEEE Computer Society, 2016.

[60] Zeyu Qin, Yanbo Fan, Hongyuan Zha, and Baoyuan Wu. Random noise defense against query-based black-box attacks. pages 7650–7663, 2021.

[61] Wenjie Qu, Youqi Li, and Binghui Wang. A certified radius-guided attack framework to image segmentation models. In *IEEE European Symposium on Security and Privacy*, 2023.

[62] Ali Rahmati, Seyed-Mohsen Moosavi-Dezfooli, Pascal Frossard, and Huaiyu Dai. Geoda: a geometric framework for black-box adversarial attacks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020.

[63] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, 2015.

[64] Kevin Roth, Yannic Kilcher, and Thomas Hofmann. The odds are odd: A statistical test for detecting adversarial examples. In *ICML*, 2019.

[65] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[66] 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*, 2018.

[67] Lukas Schott, Jonas Rauber, Matthias Bethge, and Wieland Brendel. Towards the first adversarially robust neural network model on MNIST. In *ICLR*, 2019.

[68] Ali Shafahi, Mahyar Najibi, Amin Ghiasi, Zheng Xu, John P. Dickerson, Christoph Studer, Larry S. Davis, Gavin Taylor, and Tom Goldstein. Adversarial training for free! In *NeurIPS*, 2019.

[69] Yucheng Shi, Siyu Wang, and Yahong Han. Curls & whey: Boosting black-box adversarial attacks. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019*, pages 6519–6527, 2019.

[70] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[71] Yang Song, Taesup Kim, Sebastian Nowozin, Stefano Ermon, and Nate Kushman. Pixeldefend: Leveraging generative models to understand and defend against adversarial examples. In *6th International Conference on Learning Representations, ICLR 2018*. OpenReview.net, 2018.

[72] Jingyu Sun, Bingyu Liu, and Yuan Hong. Logbug: Generating adversarial system logs in real time. In *CIKM*, pages 2229–2232. ACM, 2020.

[73] Jiaye Teng, Guang-He Lee, and Yang Yuan. $\backslash ell\_1$ adversarial robustness certificates: a randomized smoothing approach. 2019.

[74] Florian Tramèr. Detecting adversarial examples is (nearly) as hard as classifying them. In *ICML 2022*.

[75] 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, editors, *Advances in Neural Information Processing Systems, NeurIPS 2019*, pages 5858–5868, 2019.

[76] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian J. Goodfellow, Dan Boneh, and Patrick D. McDaniel. Ensemble adversarial training: Attacks and defenses. In *ICLR 2018*.

[77] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[78] Viet Quoc Vo, Ehsan Abbasnejad, and Damith Ranasinghe. Query efficient decision based sparse attacks against black-box deep learning models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.

[79] Binghui Wang, Youqi Li, and Pan Zhou. Bandits for structure perturbation-based black-box attacks to graph neural networks with theoretical guarantees. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[80] Binghui Wang, Meng Pang, and Yun Dong. Turning strengths into weaknesses: A certified robustness inspired attack framework against graph neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16394–16403, 2023.

[81] Eric Wong and J. Zico Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 5283–5292. PMLR, 2018.

[82] Eric Wong, Frank R. Schmidt, and J. Zico Kolter. Wasserstein adversarial examples via projected sinkhorn iterations. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019*, volume 97 of *Proceedings of Machine Learning Research*, pages 6808–6817. PMLR, 2019.

[83] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan L. Yuille. Mitigating adversarial effects through randomization. In *6th International Conference on Learning Representations, ICLR 2018*. OpenReview.net, 2018.

[84] Saining Xie, Ross B. Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 5987–5995. IEEE Computer Society, 2017.

[85] Shangyu Xie, Han Wang, Yu Kong, and Yuan Hong. Universal 3-dimensional perturbations for black-box attacks on video recognition systems. In *In Proceedings of the 43rd IEEE Symposium on Security and Privacy*, 2022.

[86] Shangyu Xie, Yan Yan, and Yuan Hong. Stealthy 3d poisoning attack on video recognition models. *IEEE Trans. Dependable Secur. Comput.*, 2023.

[87] Shenao Yan, Shen Wang, Yue Duan, Hanbin Hong, Kiho Lee, Doowon Kim, and Yuan Hong. An llm-assisted easy-to-trigger backdoor attack on code completion models: Injecting disguised vulnerabilities against strong detection. In *USENIX Security*, 2024.

[88] Greg Yang, Tony Duan, J. Edward Hu, Hadi Salman, Ilya P. Razenshteyn, and Jerry Li. Randomized smoothing of all shapes and sizes. In *ICML*, 2020.

[89] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In Richard C. Wilson, Edwin R. Hancock, and William A. P. Smith, editors, *Proceedings of the British Machine Vision Conference*, 2016.

[90] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, 2019.

[91] Jiawei Zhang, Zhongzhu Chen, Huan Zhang, Chaowei Xiao, and Bo Li. Diffsmooth: Certifiably robust learning via diffusion models and local smoothing. In *USENIX Security*, 2023.

[92] Xinyu Zhang, Hanbin Hong, Yuan Hong, Peng Huang, Binghui Wang, Zhongjie Ba, and Kui Ren. Text-crs: A generalized certified robustness framework against textual adversarial attacks. In *IEEE Symposium on Security and Privacy*, 2024.