Closing the Gap: Achieving Better Accuracy-Robustness Tradeoffs against Query-Based Attacks

Pascal Zimmer[1], **Sébastien Andreina**[2], **Giorgia Azzurra Marson**[2], **Ghassan Karame**[1]
[1]Ruhr-Universität Bochum, Germany
[2]NEC Labs Europe, Germany
{pascal.zimmer, ghassan.karame}@rub.de, {sebastien.andreina, giorgia.marson}@neclab.eu

Abstract

Although promising, existing defenses against query-based attacks share a common limitation: they offer increased robustness against attacks at the price of a considerable accuracy drop on clean samples. In this work, we show how to efficiently establish, at test-time, a solid tradeoff between robustness and accuracy when mitigating query-based attacks. Given that these attacks necessarily explore low-confidence regions, our insight is that activating dedicated defenses, such as random noise defense and random image transformations, only for low-confidence inputs is sufficient to prevent them. Our approach is independent of training and supported by theory. We verify the effectiveness of our approach for various existing defenses by conducting extensive experiments on CIFAR-10, CIFAR-100, and ImageNet. Our results confirm that our proposal can indeed enhance these defenses by providing better tradeoffs between robustness and accuracy when compared to state-of-the-art approaches while being completely training-free.

Introduction

Even though deep neural networks (DNNs) are currently enjoying broad applicability, they are unfortunately fragile to (easily realizable) manipulations of their inputs. Namely, *adversarial samples* pose a severe threat to the deployment of DNNs in safety-critical applications, e.g., autonomous driving and facial recognition. For instance, in the context of image classification, adversarial samples can deceive a classifier with carefully crafted and visually almost imperceptible perturbations applied to an input image. This results in (un-)targeted misclassification with identical semantic information to the human eye.

Initially, all attack strategies were designed in the so-called white-box model (Biggio et al. 2013; Goodfellow, Shlens, and Szegedy 2015), in which the attacker has full domain knowledge, e.g., model architecture, trained parameters, and training data. More recent attacks, such as query-based attacks, consider a better grounded and more realistic threat model, in which the attacker has no knowledge about the classifier's internals and training data and is only able to observe outputs to supplied inputs, i.e., through oracle access to the classifier. This black-box setting faithfully mimics many real-world applications, such as existing machine learning as a service (MLaaS) deployments. Query-based black-box attacks can be categorized as *score-based* or *decision-based* attacks (Chen, Jordan, and Wainwright 2020; Maho, Furon, and Le Merrer 2021). The former category assumes that the adversary can acquire information from the output score of the classifier, while the latter mimics a more realistic setting where the adversary has only access to the top-1 label of the classifier's prediction.

In an attempt to design defensive strategies to mitigate adversarial inputs, an arms race has been sparked in the community. A number of proposals have explored the use of randomization to improve adversarial robustness. While most randomization strategies are ineffective in the original white-box setting (Athalye et al. 2018; Athalye, Carlini, and Wagner 2018), recent findings suggest that embedding random noise within the input could effectively mitigate query-based black-box attacks (Byun, Go, and Kim 2022). As such, while randomization-based defenses could be effective in thwarting query-based attacks, they inevitably damage the classifier's clean accuracy (Tsipras et al. 2019). For example, the random noise defense (Qin et al. 2021) (using the noise level $\sigma = 0.07$ as a hyperparameter) can increase the robust accuracy against PopSkipJump (PSJA) (Simon-Gabriel, Sheikh, and Krause 2021), the strongest known black-box attack against randomized classifiers, by almost 13% at the cost of a significant drop of almost 30% in main task accuracy in the CIFAR-10 dataset.

In this work, we set forth to establish a stronger accuracy-robustness tradeoff against query-based black-box attacks by leveraging a different hyperparameter grounded on the confidence τ of classifying incoming inputs. Our approach relies on the insight that, while query-based attacks necessarily need to explore low-confidence regions, most genuine inputs are classified with relatively high confidence. By cleverly differentiating between low- and high-confidence regions, we aim to establish a strong tradeoff between adversarial robustness and the model's clean accuracy. To do so, we propose to obstruct the search for adversarial inputs by activating a defensive layer (e.g., based on input randomization) *only for low-confidence inputs*. That is, we propose to only activate (existing) randomization defenses on inputs x such that $\max_i f_i(x) < \tau$, for an appropriate threshold $0 \leq \tau \leq 1$, while high-confidence inputs (with

confidence at least τ) are processed normally.

We show that our approach can be instantiated with existing test-time defenses without the need for retraining and show that it can strike robust tradeoffs that could not be reached otherwise with existing off-the-shelf defenses. While our approach is generic, it is naturally apt to particularly thwart decision-based attacks, as it is harder for the adversary to avoid low-confidence regions when she does not have access to the scores output by the classifier.

We conduct an extensive robustness evaluation of our approach with lightweight off-the-shelf defenses and find across-the-board improvements in accuracy-robustness tradeoffs over all considered defenses. For instance, our experiments on CIFAR-10 and CIFAR-100 show that for PSJA, one of the most powerful state-of-the-art decision-based attacks, our method improves robust accuracy by up to 9% and 20%, with a negligible impact on clean task accuracy of at most 2%. For SurFree, a geometric decision-based attack, we report robustness improvements of up to 34% with almost no degradation of main task accuracy[1].

Related Work

Black-box attacks. In contrast to white-box attackers, who can easily generate adversarial samples using gradients of the model, a black-box attacker is unaware of the classifier's internals (Sharad et al. 2020). In a black-box attack, the adversary only accesses the target classifier as an oracle and uses its responses to generate adversarial examples.

Transfer-based attacks (Dong et al. 2018; Xie et al. 2019; Naseer et al. 2021) have access to (similar) training data that has been used to train the target classifier. Using such datasets, they train a local "surrogate" model and use it to craft adversarial samples with white-box strategies against the surrogate model. Due to the transferability property of DNNs, the generated samples often fool the original target classifier. To be effective, these attacks require knowledge of the target classifier's architecture and/or its training data. On the other hand, query-based attacks (Brendel, Rauber, and Bethge 2018; Andriushchenko et al. 2020; Chen, Jordan, and Wainwright 2020) do not require access to the training data itself and instead interact with the target classifier to obtain predictions on inputs of their choice. By adaptively generating a sequence of images based on the classifier's predictions, query-based attacks can derive adversarial examples with minimal distortion. Score-based attacks assume additional access to the individual probabilities of the possible classes, while *decision-based* attacks typically work with access to the top-1 label.

Decision-based attacks. As decision-based attacks can only obtain the top-1 label of a prediction, they cannot leverage the classifier's confidence.

Hence they attempt to locate the classifier's decision boundary, as seen in Figure 1, and to estimate the shape or the gradient near the boundary or by exploiting its geometric

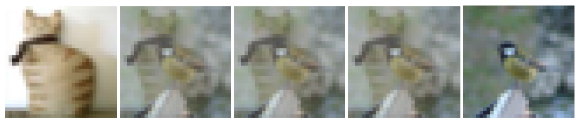


Figure 1: Selected iterations (or steps) from the binary search procedure typically used in a decision-based attack. Here, a source image, classified as 'cat', is blended with a target image, classified as 'bird'. The procedure crosses a low-confidence region before outputting a boundary sample with a slightly higher probability for the target class.

properties. The ultimate goal of such attacks is to undergo a series of steps with the aim of gradually minimizing the distortion of each candidate adversarial sample.

The first known attack (Brendel, Rauber, and Bethge 2018) is based on a naive rejection-sampling exploration of the decision boundary to converge to a low-distortion adversarial example, which requires a significant amount of queries. Recent proposals can primarily be divided into two broader categories: gradient-based methods, such as HopSkipJump (HSJA) (Chen, Jordan, and Wainwright 2020), build a surrogate gradient inspired by zeroth-order-optimization for faster convergence, and geometry-based attacks, such as SurFree (Maho, Furon, and Le Merrer 2021), exploit geometric properties of the decision boundary and perform careful query trials along it.

Since decision-based attacks are based on the assumption of a deterministic classifier, they are especially susceptible to input-randomization defenses (Chen, Jordan, and Wainwright 2020; Qin et al. 2021). By adapting HSJA to noisy environments, PSJA (Simon-Gabriel, Sheikh, and Krause 2021) overcomes this fragility.

Defenses. Existing mitigation strategies against adversarial examples can be categorized into two main groups: training- and inference-time approaches. The most popular training-time defense is adversarial training with various instantiations (Madry et al. 2018; Zhang et al. 2019), which essentially augments the training dataset with adversarial examples labeled with the semantically correct class. This process is especially costly as the adversarial images have to be generated at training time to allow their use during the training procedure. Other approaches, such as Parametric Noise Injection (PNI) (He, Rakin, and Fan 2019), inject Gaussian distributed noise upon weights in a layer-wise fashion or insert a noise layer before each convolution layer, as done in Random Self Ensemble (RSE) (Liu et al. 2018).

Inference-time defenses are generally more lightweight. Popular examples include randomized pre-processing strategies such as random noise defense (RND) (Qin et al. 2021), random resize and cropping (RCR) (Guo et al. 2018; Xie et al. 2018). These defenses promise increased robustness against decision-based attacks and come at a small additional inference cost. Deterministic inference-time defenses

[1]The full paper version is available at (Zimmer et al. 2023). The corresponding code is available at https://github.com/RUB-InfSec/closing_the_gap.

include applying the JPEG compression algorithm and have been shown to enhance the robustness of a classifier (Guo et al. 2018; Dziugaite, Ghahramani, and Roy 2016).

However, all these defenses share a common limitation: they affect benign samples similarly to adversarial examples, hence inevitably reducing main task accuracy. For instance, RND requires a noise level below $\sigma = 0.02$ in order to preserve main task accuracy, while $\sigma = 0.05$ is necessary to reach a reasonable level of robustness above 70% for CIFAR-10 and PSJA.

Accuracy-Robustness Tradeoff. There is already great progress in analyzing existing tradeoffs between accuracy and robustness (Zhang et al. 2019; Stutz, Hein, and Schiele 2019; Schmidt et al. 2018; Tsipras et al. 2019). For instance, recent findings show that robust and accurate classifiers are possible for certain (realistic) classification tasks, as long as different classes are sufficiently separated (Yang et al. 2020).

Most proposals that aim at improving the accuracy-robustness tradeoff are focused on adversarial-training defenses based on data-augmentation methods, hence leading to high computational overhead at training time (Raghunathan et al. 2020). Other approaches allow for a "free" adjustment with a hyperparameter, without requiring any retraining after initially augmenting the model with a new model-conditional training approach (Wang et al. 2020).

Methodology

Preliminaries and Notations

We define an adversarial sample x' as a genuine image x to which carefully crafted adversarial noise p is added, i.e., $x' = x + p$ for a small perturbation p such that x' and x are perceptually indistinguishable to the human eye and yet are classified differently.

Let $f\colon \mathbb{R}^d \to \Delta^n$ be a DNN model assigning d-dimensional inputs to n classes, where Δ^n is the probability vector of n classes, and let $C\colon \mathbb{R}^d \to [n]$ be the associated classifier defined as $C(x) := \arg\max_{i\in[n]} f_i(x)$. The highest prediction probability $\max_{i\in[n]} f_i(x)$ is called the classifier's *confidence*.

Given a genuine input $x_0 \in \mathbb{R}^d$ predicted as $C(x_0) = s$ (source class), x' is an *adversarial sample* of x_0 if $C(x') \neq s$ and $\|x' - x_0\|_p \leq \varepsilon$ for a given distortion bound $\varepsilon \in \mathbb{R}^+$ and l_p norm. Formally, we have:

$$\begin{cases} C(x') \neq s & \text{(untargeted attack)}, \\ C(x') = t & \text{(targeted attack)}. \end{cases} \tag{1}$$

The objective of the adversary can then be expressed as follows:

$$\mathcal{A}_{x_0}(x) := \begin{cases} \max\limits_{i\neq s} f_i(x) - f_s(x) & \text{(untargeted attack)}, \\ f_t(x) - \max\limits_{i\neq t} f_i(x) & \text{(targeted attack)}. \end{cases} \tag{2}$$

For an adversarial sample x' to be successful, it must have a low distortion and satisfy $\mathcal{A}_{x_0}(x') > 0$. The attacker searches for inputs x' solving the optimization problem,

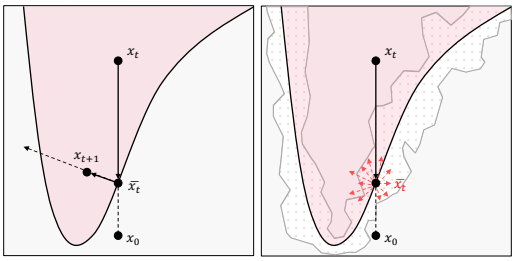


Figure 2: Iteration of a decision-based attack for an unprotected classifier (left) and in the presence of our approach (right). The attacker starts with a sample x_t in the target class, and in each iteration, she updates the current adversarial sample, from x_t to x_{t+1}, by locating the boundary sample $\bar{x}_t$. Intuitively, activating the defense only in low-confidence regions (right) can effectively obstruct the search for adversarial perturbations.

while she is restricted by the query budget Q in her number of permitted queries to the classifier:

$$\min_{x'} \|x' - x_0\|_p \quad \text{such that} \quad \mathcal{A}_{x_0}(x') > 0. \tag{3}$$

In contrast to score-based attacks—that receive the classifier's scores $(f_i(x))_{i\in[n]}$ for each queried input x and can thus evaluate the function $\mathcal{A}_{x_0}(x)$—decision-based attacks only obtain the final prediction $\arg\max_i f_i(x)$ and cannot compute $\mathcal{A}_{x_0}(x)$. Nevertheless, the prediction $C(x')$ is sufficient to determine the sign of $\mathcal{A}_{x_0}(x')$ as per Equation (2).

Main Intuition

We propose an approach that specifically aims to impede the search for adversarial perturbations in low-confidence regions while optimistically treating high-confidence samples as non-adversarial. A central component of decision-based attacks is a binary search procedure that is used to locate the decision boundary in a low-confidence region for further exploration.

More specifically, we propose to dynamically trigger the activation of a suitable defensive layer depending on the classifier's confidence on its input. Here, we consider lightweight defenses, such as RND, RCR, and JPEG.

Formally, our method is parameterized by a *confidence threshold* $\tau \in [0, 1]$ as a hyperparameter and can be generically combined with any defensive technique D applied at inference time.

Definition 1 (Low-confidence region) *We define a* low-confidence region *with respect to classifier f and confidence threshold τ as the space of samples for which f has confidence below τ, i.e., $\{x \in \mathcal{X} : \max_{i\in[n]} f_i(x) < \tau\}$.*

Namely, let f and f_{D} denote the (unprotected) model and its protected version with defense D, respectively. Our strategy consists of invoking the defense only when the classifier has confidence below the threshold, else the unprotected classifier is used, resulting in the following classifier:

$$F_{\mathsf{D},\tau}(x) := \begin{cases} f_{\mathsf{D}}(x) & \text{if } \max_{i\in[n]} f_i(x) < \tau, \\ f(x) & \text{otherwise.} \end{cases} \tag{4}$$

Due to the limited information retrieved in the decision-based model, attacks usually start with a highly distorted image (i.e, $x_1 = x + p$, $p >> \varepsilon$) such that $\mathcal{A}_{x_0}(x_1) > 0$ as a starting point. Afterward, it follows the iterative method displayed in Figure 2 (left) where it first locates the point $\bar{x}_t$ at the decision boundary through a binary search between x_0 and x_t, before pursuing further exploration strategies. An example is a gradient estimation to minimize the distortion and generate x_{t+1}.

Our intuition is that obstructing the generation of low-confidence adversarial inputs is sufficient to thwart the exact location of $\bar{x}_t$ and the further exploration of the decision boundary at that point as illustrated in Figure 2 (right).

As most genuine samples are classified correctly with high confidence, the defense has minimal impact on the classification accuracy of genuine samples while still being effective at preventing attacks, as attackers have to visit the low-confidence region where the defense is enabled in order to create adversarial samples.

Calibration. Predictions of recent model architectures are typically miscalibrated in the sense that they usually classify inputs with rather high confidence. To make our approach agnostic to the deployment instance, we calibrated our models by leveraging temperature scaling (Guo et al. 2017), i.e., using a single parameter $T > 0$ to scale the logits z before passing them through softmax σ and updating the confidence prediction p with $p = \sigma(z/T)$.

The parameter T is computed so as to minimize the difference between the reported confidence (i.e., the value p) and the actual accuracy. We group the predictions into M interval bins and denote the set of images falling in the *m*-th bin as B_m. Then, we derive T by solving the optimization problem below:

$$\min_T \sum_{m=1}^{M} \frac{|B_m|}{n} \left| \frac{1}{|B_m|} \sum_{i \in B_m} \mathbf{1}(\hat{y}_i = y_i) - \frac{1}{|B_m|} \sum_{i \in B_m} \hat{p}_i \right| \tag{5}$$

where n is the overall number of samples, y_i the true class label, $\hat{y}_i$ the predicted class label, and $\hat{p}_i$ the confidence of the model when classifying image i. The indicator function $\mathbf{1}$ is defined as:

$$\mathbf{1}(x) := \begin{cases} 1 \text{ if } x \text{ is true} \\ 0 \text{ otherwise.} \end{cases}$$

We stress that our approach is independent of model calibration (i.e., low-confidence regions must still be explored by attacks). Namely, we merely expect that the threshold τ in our approach to slightly vary among calibrated and uncalibrated models—with no impact on the resulting accuracy or robustness.

Theoretical Motivation. Recall that the supervised training procedure for a generic multi-class classifier C aims at finding the optimal set of parameters θ that minimize the aggregated loss over the entire training set $\mathcal{X}$: $\min_\theta \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(y_i, f(x_i, \theta))$, for $N = |\mathcal{X}|$, sample x, ground-truth y, and loss-function $\mathcal{L}$.

The optimization procedure itself can be instantiated with varying algorithms, such as SGD, which come with varying convergence guarantees. Depending on the use case, loss-function $\mathcal{L}$ is, for example, the cross-entropy loss.

In what follows, we show that *(1)* genuine (i.e., non-adversarial) samples are unlikely to bear low-confidence, and *(2)* existing query-based attacks necessarily have to investigate low-confidence regions to ensure convergence.

Proposition 1 *A set of parameters θ trained on $\mathcal{X}$ are likely to classify images following this distribution with high confidence.*

Consider the loss function $\mathcal{L}$ with the n-class cross-entropy loss:

$$\mathcal{L}(y, \mathbf{p}) := -\sum_{c=1}^{n} \mathbf{1}(y = c) \log(p_c) \tag{6}$$

with $\mathbf{p}$ being the probability vector output by the classifier.

Now, let us define $q := \mathbf{1}(y = c)$ as a fixed reference probability distribution, i.e., the one-hot encoding of y. Due to Gibbs' inequality (MacKay 2003) the cross-entropy function takes its minimum when $p = q$, i.e., the network is optimizing towards predictions of high confidence. Notice that this result corroborates findings in (Guo et al. 2017; DeVries and Taylor 2018) which also suggest that models tend to output high confidence predictions for samples following the training distribution $\mathcal{X}$. An empirical validation of Proposition 1 can be found in the full version of this paper (Zimmer et al. 2023).

Proposition 2 *During the binary-search procedure of a decision-based attack,* $\max_{i \in [n]} f_i(x) \leq 1/2$.

Decision-based attacks, such as PSJA (Simon-Gabriel, Sheikh, and Krause 2021) and SurFree (Maho, Furon, and Le Merrer 2021), use the classifier's feedback on input queries in order to locate the *boundary*:

$$\mathrm{bd}(\mathcal{A}_{x_0}) := \{x \in \mathbb{R}^d : \mathcal{A}_{x_0}(x) = 0\} \tag{7}$$

A binary-search computes an interpolation x between a source image x_0 and a target image x_t, satisfying $\mathcal{A}_{x_0}(x_t) > 0$ – more precisely:

$$x := \mathrm{bs}(x_t, x_0, k) = (1 - k)x_t + kx_0 \tag{8}$$

with interpolation factor $k \in [0, 1]$.

Figure 2 (left) illustrates one iteration of a binary search (and gradient sampling); here, each iteration reduces the distance from the candidate sample x_t and the original sample x_0. Therefore, the attacker needs to query the classifier on intermediate samples $\bar{x}_t$ that lie on the boundary, i.e., *where the classifier's confidence is low* in order to estimate the gradient or the geometric shape of the boundary.

The proof for Proposition 2 can be found in the full version of this paper (Zimmer et al. 2023).

Experiments

In this section, we empirically evaluate our approach on the CIFAR-10, CIFAR-100, and ImageNet datasets and compare its efficacy to related work in the area.

Datasets. In line with the literature, we evaluate our approach on datasets of varying input dimensions and number of classes, i.e., CIFAR-10, CIFAR-100 (Krizhevsky 2009), and ImageNet (Russakovsky et al. 2015) datasets. The former two contain $50,000$ train and $10,000$ test images of size 32×32 pixels, divided into 10 and 100 different classes, respectively. The latter contains 1.2 million training images and a validation set of $50,000$ images.

Attack choice. Our choice of attacks is mainly motivated by a study that provides an overview of decision-based attacks and outlines a comparison of their contributions. The study can be found in the full version of this paper (Zimmer et al. 2023). We selected PSJA (Simon-Gabriel, Sheikh, and Krause 2021) as it is the only attack that considers a probabilistic classifier. It is based on gradient estimation and is a direct improvement over HSJA (Chen, Jordan, and Wainwright 2020). We further selected the SurFree attack (Maho, Furon, and Le Merrer 2021) as it is the most recent attack, which trades the costly gradient estimation step and instead exploits geometric properties of the decision boundary. We argue that our selected attacks exhibit good coverage over attacks in the literature.

Defense choice. Our approach can be generically instantiated using any inference-time defense D. We selected three lightweight transformations to instantiate our approach: RND and RCR for the probabilistic setting and JPEG compression for the deterministic setting.

RND applies additive centered Gaussian noise to the input $f_{\mathsf{RND}}(\nu, x) = f(x + \nu r) \quad \text{for} \quad r \leftarrow \mathcal{N}(0, I)$, where the parameter $\nu \in \mathbb{R}$ controls the noise magnitude.

RCR is a function γ_ν that randomly crops and bilinearly interpolates the image back to its original size: $f_{\mathsf{RCR}}(\nu, x) = f(\gamma_\nu(x))$, where ν denotes the cropping size.

Finally, JPEG applies the JPEG compression algorithm ϕ with $f_{\mathsf{JPEG}}(\nu, x) = f(\phi_\nu(x))$,where $\nu \in [0, 100\%]$ is the quality parameter. We include a deterministic defense to evaluate attacks, such as SurFree, that rely on deterministic classification, as they are otherwise easily defeated by probabilistic defenses (Chen, Jordan, and Wainwright 2020).

Models. For CIFAR-10, we use a DenseNet-121 model with an accuracy of 95%; and a ResNet-50 for CIFAR-100 and ImageNet with an accuracy of 60% and 81%, respectively. All our models are calibrated (Guo et al. 2017).

Metrics

Attack success rate and robust accuracy. Let $S \subset \mathcal{X} \times \mathcal{Y}$ denote the set of (labeled) genuine samples provided to the attacker, let $n := |S|$, let Q denote the query budget, and let ε denote the distortion budget. To compute the attack success rate (ASR) in practice. we determine the number of successful adversarial samples generated by the attacker:

$$n_{\text{succ}} := |\{x \in \mathcal{A}(S) \mid \mathcal{A}_{x_0}(x) > 0 \wedge \|x - x_0\|_p \leq \varepsilon\}|, \quad (9)$$

where $\mathcal{A}(S)$ denotes the set of candidate adversarial samples output by $\mathcal{A}$ in a run of the attack on input S. The ASR is defined as $\mathsf{ASR} := {}^{n_{\text{succ}}}/_{n}$, i.e., the ratio of successful adversarial samples. The complement of the ASR is the robust accuracy (RA) of the classifier, i.e., $\mathsf{RA} = 1 - \mathsf{ASR}$.

CA-RA Pareto frontier. In multi-objective optimization problems, there is often no solution that maximizes all objective functions simultaneously. The Pareto frontier emerges as an effective tool to evaluate tradeoffs between the various objectives. To evaluate the accuracy-robustness tradeoffs of different defenses, we consider CA and RA as our objectives and empirically determine the Pareto frontier. Specifically, we consider the following optimization problem:

$$\max_{\omega \in \Omega} (\mathsf{CA}(\omega), \mathsf{RA}(\omega)), \quad (10)$$

where $\mathsf{CA}(\omega)$ and $\mathsf{RA}(\omega)$ denote the clean accuracy and robust accuracy of a given (protected) classifier as functions of the defense parameter ω – for a fixed attack and experiment setting.

A solution ω^* is *Pareto optimal* if there exists no other solution that improves all objectives simultaneously. Formally, given two solutions ω_1 and ω_2, we write $\omega_1 \succ \omega_2$ if ω_1 dominates ω_2, i.e., if $\mathsf{CA}(\omega_1) \geq \mathsf{CA}(\omega_2) \wedge \mathsf{RA}(\omega_1) \geq \mathsf{RA}(\omega_2)$. The *Pareto frontier* is the set of Pareto-optimal solutions:

$$PF(\Omega) = \{\omega^* \in \Omega \mid \nexists \omega \in \Omega \text{ s.t. } \omega \succ \omega^*\}. \quad (11)$$

Evaluation Setup

To evaluate the effectiveness of our approach, we instantiate our proposal with three existing defenses D, namely RND, RCR, and JPEG compression, and empirically measure the accuracy and robust accuracy of the resulting classifiers $F_{\mathsf{D},\tau}$ against state-of-the-art decision-based attacks PSJA and SurFree in the untargeted setting.

We generate adversarial examples under the l_2 norm constraint with maximum distortion $\varepsilon = 3$ for both CIFAR-10 and CIFAR-100, assuming a query budget of $Q = 20,000$. Due to the significantly larger input dimension of ImageNet, we allow for a higher query budget of $Q = 40,000$ in accordance with previous works and we select a comparable $\varepsilon = 21$. We evaluate our proposal with $\tau \in \{0.0, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.97, 0.99, 1.0\}$. In extreme cases, $\tau = 0$ means that the defense is never activated, while $\tau = 1$ triggers the defense on all inputs.

In combination with the aforementioned thresholds, we select a spectrum of noise levels to achieve a diverse set of CA and RA. More concretely, we select the noise parameter ν of the RND defense from the set $\{0.01, 0.02, 0.05, 0.07, 0.08, 0.1\}$, for RCR $\nu \in \{29, 27, 26, 25, 22, 18, 14\}$ for CIFAR-10/100 and $\nu \in \{200, 176, 152, 128, 104, 80, 56\}$ for ImageNet. Lastly, we use a quality setting of $\nu \in \{85, 75, 60, 50, 35, 25, 10\}$ for the JPEG defense.

Throughout our evaluation, we evaluate PSJA against randomized defenses, i.e., RND, RCR, and SurFree against the deterministic defense, i.e., JPEG. We always measure the clean accuracy (or CA) of the classifier over the entire test set of each respective dataset and the robust accuracy (RA) on $n = 100$ randomly selected images from the test set that are correctly classified by the undefended classifier.

Our experiments have been conducted on a server equipped with two AMD EPYC 7542 CPUs, two Nvidia A40 GPUs, and 256GB RAM. The system runs Ubuntu 22.04., Python 3.9, PyTorch 1.13.0, and CUDA 11.7.

	Dataset		CIFAR-10							CIFAR-100							
	ν / τ		0.5	0.6	0.7	0.8	0.97	0.99	1.0	0.5	0.6	0.7	0.8	0.9	0.97	0.99	1.0
PSJA with RND	0.02	CA	**0.95**	**0.95**	**0.94**	**0.94**	**0.94**	**0.94**	0.94	**0.59**	**0.59**	**0.59**	**0.59**	**0.59**	**0.58**	**0.58**	0.58
		RA	**0.51**	**0.54**	**0.57**	**0.57**	**0.62**	**0.55**	0.53	**0.26**	**0.29**	**0.28**	**0.33**	**0.31**	**0.34**	**0.33**	0.26
	0.05	CA	0.95	0.94	0.94	**0.94**	**0.87**	**0.83**	0.82	0.58	0.56	**0.55**	**0.53**	0.50	0.48	0.47	0.46
		RA	0.44	0.49	0.47	**0.59**	**0.64**	**0.71**	0.71	0.23	0.26	**0.31**	**0.38**	0.33	0.38	0.40	0.46
	0.07	CA	0.95	**0.94**	0.94	0.93	0.82	0.65	0.65	0.56	0.54	0.53	0.49	0.46	0.41	0.38	0.37
		RA	0.44	**0.54**	0.48	0.51	0.54	0.58	0.60	0.22	0.20	0.28	0.35	0.37	0.37	0.44	0.37
PSJA with RCR	29	CA	0.95	**0.95**	**0.94**	0.94	0.92	0.91	0.91	**0.59**	**0.59**	**0.58**	**0.56**	**0.55**	**0.54**	**0.54**	0.53
	200	RA	0.37	**0.47**	**0.44**	0.35	0.32	0.33	0.37	**0.24**	**0.29**	**0.31**	**0.37**	**0.42**	**0.36**	**0.44**	0.35
	27	CA	**0.95**	0.95	**0.94**	**0.94**	**0.90**	0.88	0.88	**0.59**	**0.58**	**0.57**	**0.56**	**0.53**	0.51	**0.50**	0.50
	176	RA	**0.43**	0.39	**0.41**	**0.41**	**0.42**	0.39	0.34	**0.25**	**0.37**	**0.40**	**0.42**	**0.43**	0.41	**0.49**	0.46
	26	CA	0.95	0.95	**0.94**	0.94	0.89	0.86	0.86	0.58	**0.58**	**0.56**	**0.54**	**0.52**	**0.49**	**0.47**	0.47
	152	RA	0.38	0.41	**0.45**	0.39	0.33	0.34	0.39	0.21	**0.37**	**0.36**	**0.39**	**0.49**	**0.47**	**0.50**	0.48
	25	CA	0.95	0.95	**0.94**	0.94	**0.88**	0.83	0.83	**0.58**	**0.57**	**0.56**	**0.54**	0.50	0.47	0.45	0.44
	128	RA	0.41	0.35	**0.43**	0.39	**0.46**	0.34	0.33	**0.27**	**0.28**	**0.34**	**0.44**	0.42	0.45	0.47	0.47
	22	CA	**0.95**	**0.94**	**0.94**	**0.94**	**0.81**	**0.66**	0.67	0.57	0.55	0.53	0.50	0.46	0.39	0.36	0.34
	104	RA	**0.42**	**0.42**	**0.45**	**0.43**	**0.43**	**0.51**	0.49	0.24	0.27	0.29	0.36	0.40	0.37	0.43	0.43
	18	CA	**0.95**	**0.94**	**0.94**	**0.93**	**0.72**	0.39	0.38	0.55	0.52	0.50	0.46	0.40	0.30	0.25	0.22
	80	RA	**0.48**	**0.49**	**0.46**	**0.46**	**0.48**	0.39	0.41	0.23	0.28	0.34	0.28	0.29	0.35	0.33	0.31
SurFree with JPEG	85	CA	**0.95**	**0.95**	**0.94**	**0.94**	**0.92**	0.92	0.92	**0.59**	0.58	**0.58**	0.57	**0.56**	0.56	0.56	0.56
		RA	**0.52**	**0.36**	**0.40**	**0.67**	**0.83**	0.65	0.73	**0.79**	0.47	**0.90**	0.51	**0.94**	0.85	0.49	0.85
	75	CA	**0.95**	**0.94**	**0.94**	**0.94**	**0.91**	0.90	0.90	0.59	**0.58**	**0.57**	0.56	0.56	0.55	**0.54**	0.54
		RA	**0.37**	**0.57**	**0.49**	**0.40**	**0.84**	0.71	0.67	0.45	**0.78**	**0.81**	0.50	0.51	0.60	**0.89**	0.82
	60	CA	**0.95**	**0.94**	0.94	**0.94**	**0.90**	0.87	0.87	0.58	0.57	**0.56**	0.55	0.53	0.52	0.52	0.52
		RA	**0.41**	**0.50**	0.31	**0.44**	**0.77**	0.60	0.60	0.65	0.52	**0.89**	0.60	0.50	0.50	0.55	0.83
	50	CA	**0.95**	**0.94**	0.94	0.94	**0.88**	0.85	0.85	0.58	**0.57**	0.56	**0.54**	0.52	0.51	0.50	0.50
		RA	**0.55**	**0.46**	0.32	0.38	**0.75**	0.60	0.64	0.49	**0.82**	0.56	**0.88**	0.52	0.56	0.60	0.83

Table 1: Selected results for RA and CA based on the defense parameter ν and the threshold τ in the CIFAR-10 and the CIFAR-100 datasets. Improvements beyond the Pareto frontier achieved in vanilla constructs (i.e., when $\tau = 1.0$) are bolded.

Evaluation Results

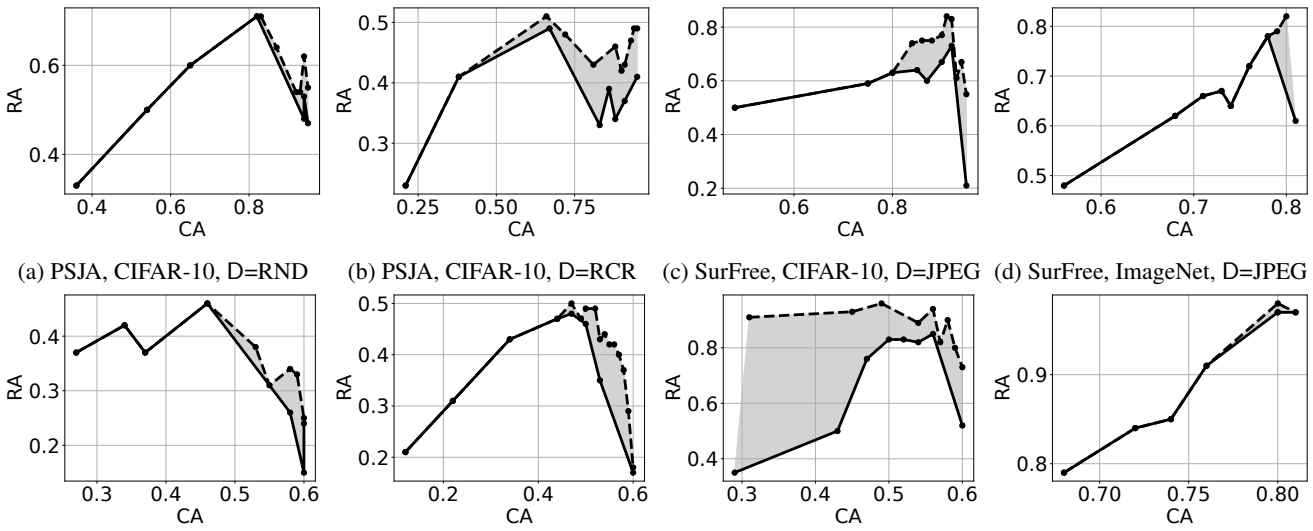
Accuracy-robustness tradeoffs. In Figure 3, we use the Pareto frontier to showcase the tradeoff improvements of our proposal compared to the baseline relying only on the defense parameter ν. We plot the points $(\mathsf{CA}(\omega^*), \mathsf{RA}(\omega^*))$ for empirically-determined Pareto-optimal solutions ω^* for the baseline defense ($\omega^* = \nu^*$, solid line) and for our approach ($\omega^* = (\nu^*, \tau^*)$, dotted line). The accuracy-robustness tradeoffs achieved by our proposal for selected threshold values are highlighted in grey. The specific combinations of ν, τ that outperform the baseline tradeoffs ($\tau = 1.0$) are highlighted in Table 1 and Table 2. Note that for RCR the crop size for CIFAR-10/100 and ImageNet are shown due to the differently sized input dimensions.

Tradeoffs in PSJA. When evaluated against PSJA on the CIFAR-10 dataset, our proposal combined with the RND defense can provide an improvement in RA while preserving the CA of the baseline. For example, it improves RA by 8% (out of a maximum of 9%) for $\mathsf{CA} = 0.95$ and RA by up to 9% for a tradeoff of at most 1% decrease in CA. Compared to the baseline ($\tau = 1.0, \nu = 0.07$), we improve CA by 29%, while decreasing RA by just 1% (cf. Figure 3a).

When evaluated with the RCR defense, we can observe several improvements in Figure 3b. For example, given a RA of 49%, our approach can increase CA from 67% to 94%. At a RA of 41%, our proposal achieves an 8% improvement in robustness compared to the baseline without compromising CA. This is in line with the results for RND.

Our results on CIFAR-100 shown in Figures 3e and 3f are consistent with our CIFAR-10 results. More precisely, we measure an improvement of RA of up to 8% for a CA of 58% when evaluating with the RND defense. Compared to the baseline with a CA/RA of 37%, our approach improves the CA by 16% and RA by 1%. When considering the RCR defense in Figure 3f, our approach can improve CA by 19% for the same RA of 43%. When $\mathsf{CA} = 0.53$, our approach yields an improvement in RA by 8%. For ImageNet, we see an improvement of up to 1% in RA (cf. Figure 3h)—we discuss these results in more detail in (Zimmer et al. 2023).

Tradeoffs in SurFree. Figure 3c, Figure 3g, and Figure 3d depict the evaluation results against the SurFree attack, for which our proposal can obtain consistent improvements compared to the baseline. At no decrease in CA, we observe significant improvements in RA. For CIFAR-10, it increases by 34% and for CIFAR-100 by 21%. This trend is also observable for ImageNet—we obtain an increase of



(a) PSJA, CIFAR-10, D=RND (b) PSJA, CIFAR-10, D=RCR (c) SurFree, CIFAR-10, D=JPEG (d) SurFree, ImageNet, D=JPEG

(e) PSJA, CIFAR-100, D=RND (f) PSJA, CIFAR-100, D=RCR (g) SurFree,CIFAR-100,D=JPEG (h) PSJA, ImageNet, D=RND

Figure 3: Pareto frontier when applying our approach in conjunction with RND, RCR, and JPEG on CIFAR-10, CIFAR-100, and ImageNet using the attacks PSJA and SurFree.

	Dataset		ImageNet					
	ν / τ		0.0	0.3	0.5	0.6	0.9	1.0
PSJA with RND	0.01	CA	0.81	0.81	**0.80**	0.80	**0.80**	0.80
		RA	0.97	0.97	**0.98**	0.97	**0.98**	0.97
	0.02	CA	0.81	0.81	**0.80**	**0.80**	0.80	0.80
		RA	0.97	0.97	**0.98**	**0.98**	0.97	0.97
	0.05	CA	0.81	0.80	0.79	0.78	0.77	0.76
		RA	0.97	0.96	0.94	0.94	0.91	0.91
SurFree with JPEG	85	CA	0.81	**0.80**	**0.79**	**0.79**	0.78	0.78
		RA	0.61	**0.77**	**0.79**	**0.79**	0.78	0.78
	75	CA	0.81	**0.80**	**0.79**	0.78	0.76	0.76
		RA	0.61	**0.82**	**0.79**	0.76	0.70	0.72
	60	CA	0.81	**0.80**	**0.79**	0.78	0.74	0.74
		RA	0.61	**0.80**	**0.76**	0.76	0.65	0.64

Table 2: Selected results for RA and CA based on the defense parameter ν and the threshold τ in the ImageNet dataset. Our complete results are included in (Zimmer et al. 2023).

		Training-time				Inference-time	
	$\sim$ CA	OAT	PNI	RSE	AT	Baseline ($\tau = 1.0$)	Ours
PSJA	0.91	0.67	–	0.20	**0.83**	0.37	0.43
	0.88	**0.77**	–	–	–	0.34	0.46
	0.84	–	**0.98**	–	–	0.33	0.43
SurFree	0.91	0.47	–	0.19	0.76	0.73	**0.84**
	0.88	**0.82**	–	–	–	0.60	0.75
	0.84	–	**0.97**	–	–	0.64	0.74

Table 3: RA for state-of-the-art defenses on CIFAR-10. The baseline instantiates PSJA with D = RCR and SurFree with D = JPEG.

RA of 21% with a negligible decrease of CA by 1%. When fixing RA to 67%, we obtain an improvement in CA of 4% for CIFAR-10. Moreover, we see an improvement of CA by 3% at a fixed RA $= 0.82$ for CIFAR-100.

Comparison to related work. We now compare our proposal to various training-time defenses, which require the costly retraining of a model and cannot be retrofitted to existing models. We compare against OAT (Wang et al. 2020) because it is the only reconfigurable, i.e., adjustment of accuracy-robustness tradeoff after training, adversarial-training solution with an open-source implementation. In addition, we include traditional, non-adjustable, training-time defenses, such as PNI (He, Rakin, and Fan 2019), RSE (Liu et al. 2018), and a state-of-the-art robust model (AT) (Gowal et al. 2020) from RobustBench for further comparison. We also include results from the previously considered vanilla inference-time defenses RCR and JPEG with $\tau = 1.0$, i.e., traditional deployment of defenses for this comparison.

In Table 3, we evaluate RA at those values of CA obtained with these defenses against PSJA and SurFree on CIFAR-10. Note that some values of CA cannot be achieved with some defenses. In these cases, we report the RA that results in the closest (by up to $1 - 2\%$) CA; when such a close estimate cannot be found, we do not report any value for RA.

Against SurFree, our proposal outperforms all training-/inference-time-based defenses, outperforming AT by 8%, while being completely training-free. For CA $= 0.88$, we observe a difference to the RA of OAT of just 7%. For CA $= 0.84$, a larger difference of 23% in RA to the one of PNI can be measured. We conclude that, although it is training-free, our proposal manages to outperform or closely perform similarly to existing defenses against SurFree. In

the case of PSJA, training-time defenses outperform our approach by a minimum of 31% in RA. Compared to OAT, this difference is reduced to 24%. Our approach, however, results in significant improvements compared to all training-free defenses.

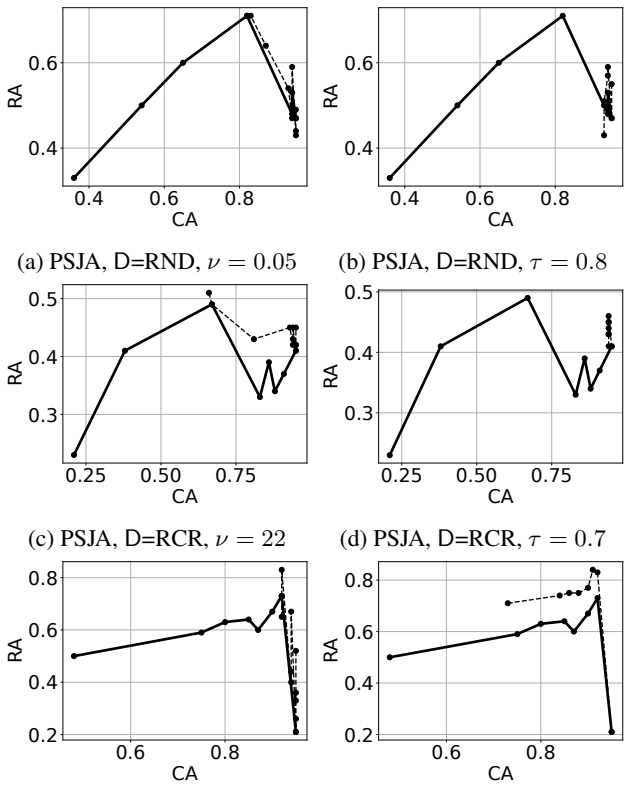
Ablation Study

Our proposal relies on two main parameters: the confidence threshold τ and the genuine defense's parameter ν. To evaluate the impact of each of these parameters on robustness, we present an ablation study. We detail our results in Table 1 and Table 2, where we show the achieved CA (top of each row) and RA (bottom of each row) for different values of τ and ν. We highlight the values that result in an improvement of the Pareto frontier in bold. When $\tau = 1.0$, our proposal instantiates the baseline since the defense is always enabled.

Impact of τ. We vary the value of $\tau \in \{0.0, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.97, 0.99, 1.0\}$. For the CIFAR-10 dataset, we first consider the RND defense with noise $\nu = 0.05$, which offers the highest RA $= 0.71$ at CA $= 0.82$ and the best tradeoffs. As we reduce the value of τ, we observe a significant improvement in CA, increasing from 82% when $\tau \geq 0.99$ to 87% when $\tau = 0.97$. By further reducing τ to 0.8, our proposal achieves 94% CA, fairly close to the original CA of 95%. On the other hand, RA starts at 71% for the baseline, remains consistent at 71% when $\tau = 0.99$, and decreases to 64% when $\tau = 0.97$. For $\tau \leq 0.8$, RA remains at 59% and then drops sharply to a RA as low as 43%. The Pareto frontier for the fixed ν and varying τ can be found in Figure 4a.

For RCR, we show the Pareto frontier for $\nu = 22$ in Figure 4c and for JPEG for $\nu = 85$ in Figure 4e. For both defenses, we observe consistent improvements in the baseline tradeoffs. We argue that such a noise level introduced by these two defenses is similar to $\nu = 0.05$ for the RND defense. These observations are not exclusive to CIFAR-10 but can also be clearly seen with CIFAR-100 with $\nu = 26$ for RCR, while it is observable for $\nu = 0.02$ and $\nu = 50$ for RND and JPEG, respectively.

Impact of ν. We now vary the defense-specific parameter ν for a fixed τ for CIFAR-10. For JPEG we identify a wide range of tradeoffs for $\tau = 0.97$, i.e., CA between 73% and 92% and RA between 71% and 83%. In line with previous results, we find an optimal $\tau = 0.97$ for which the Pareto frontiers outperform the baseline ($\tau = 1.0$) in terms of obtainable tradeoffs, which is highlighted in Figure 4f. For RND, we set τ to the identified value of 0.8, as it achieves the best tradeoff across varying ν (cf. Figure 4b). It has a negligible impact of $\pm 2\%$ on CA but allows us to control RA between 43% and 59%, with the most optimal point found at $\nu = 0.05$. For RCR, we notice a similar behavior at $\tau = 0.7$, for which we obtain a $3-5\%$ improvement over the baseline as seen in Figure 4d. For the CIFAR-100, we notice a similar trend for a threshold of $0.7/0.8$. For the ImageNet dataset, we consider JPEG with threshold $\tau = 0.3$ and vary the value of ν as before. When ν decreases from 85 to 75, we observe an improvement in both RA by 5% without hampering CA. Reducing ν further has a negligible impact on CA, while it decreases RA close to the initial value to 77%.



(a) PSJA, D=RND, $\nu = 0.05$ (b) PSJA, D=RND, $\tau = 0.8$

(c) PSJA, D=RCR, $\nu = 22$ (d) PSJA, D=RCR, $\tau = 0.7$

(e) SurFree, D=JPEG, $\nu = 85$ (f) SurFree, D=JPEG, $\tau = 0.97$

Figure 4: Pareto frontier for various τ and ν for CIFAR-10.

Overall, by appropriately setting the confidence threshold τ, our approach can improve the accuracy-robustness tradeoff compared to the baseline for all values of noise ν.

Conclusion

In this paper, we showed that limiting the invocation of an inference-time defense to low-confidence inputs might be sufficient to obstruct the search for adversarial samples in query-based attacks. Our approach can be applied generically to existing inference-time defenses and is training-free. We therefore hope to motivate further research in this area.

Acknowledgments

This work has been co-funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC 2092 CASA - 390781972, by the German Federal Ministry of Education and Research (BMBF) through the project TRAIN (01IS23027A), and by the European Commission through the HORIZON-JU-SNS-2022 ACROSS project (101097122). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

References

Andriushchenko, M.; Croce, F.; Flammarion, N.; and Hein, M. 2020. Square Attack: A Query-Efficient Black-Box Adversarial Attack via Random Search. In Vedaldi, A.; Bischof, H.; Brox, T.; and Frahm, J.-M., eds., *Computer Vision – ECCV 2020*, 484–501. Springer International Publishing. ISBN 978-3-030-58592-1.

Athalye, A.; Carlini, N.; and Wagner, D. A. 2018. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *ICML*, volume 80 of *Proceedings of machine learning research*, 274–283. PMLR.

Athalye, A.; Engstrom, L.; Ilyas, A.; and Kwok, K. 2018. Synthesizing Robust Adversarial Examples. In Dy, J.; and Krause, A., eds., *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, 284–293. PMLR.

Biggio, B.; Corona, I.; Maiorca, D.; Nelson, B.; Šrndić, N.; Laskov, P.; Giacinto, G.; and Roli, F. 2013. Evasion Attacks against Machine Learning at Test Time. In Blockeel, H.; Kersting, K.; Nijssen, S.; and Železný, F., eds., *Machine Learning and Knowledge Discovery in Databases*, 387–402. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN 978-3-642-40994-3.

Brendel, W.; Rauber, J.; and Bethge, M. 2018. Decision-based Adversarial Attacks: Reliable Attacks Against Black-Box Machine Learning Models. In *International Conference on Learning Representations*, 12.

Byun, J.; Go, H.; and Kim, C. 2022. On the effectiveness of small input noise for defending against query-based black-box attacks. In *WACV*, 3819–3828. IEEE.

Chen, J.; Jordan, M. I.; and Wainwright, M. J. 2020. HopSkipJumpAttack: A Query-Efficient Decision-Based Attack. In *2020 IEEE Symposium on Security and Privacy (SP)*, 1277–1294. IEEE. ISBN 978-1-72813-497-0.

DeVries, T.; and Taylor, G. W. 2018. Learning Confidence for Out-of-Distribution Detection in Neural Networks. arXiv:1802.04865.

Dong, Y.; Liao, F.; Pang, T.; Su, H.; Zhu, J.; Hu, X.; and Li, J. 2018. Boosting Adversarial Attacks With Momentum. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Dziugaite, G. K.; Ghahramani, Z.; and Roy, D. M. 2016. A study of the effect of JPG compression on adversarial images. arXiv:1608.00853.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2015. Explaining and Harnessing Adversarial Examples. arXiv:1412.6572.

Gowal, S.; Qin, C.; Uesato, J.; Mann, T.; and Kohli, P. 2020. Uncovering the limits of adversarial training against norm-bounded adversarial examples. *arXiv preprint arXiv:2010.03593*.

Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. Q. 2017. On Calibration of Modern Neural Networks. In Precup, D.; and Teh, Y. W., eds., *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, 1321–1330. PMLR.

Guo, C.; Rana, M.; Cisse, M.; and van der Maaten, L. 2018. Countering adversarial images using input transformations. In *International conference on learning representations*.

He, Z.; Rakin, A. S.; and Fan, D. 2019. Parametric noise injection: Trainable randomness to improve deep neural network robustness against adversarial attack. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 588–597.

Krizhevsky, A. 2009. Learning multiple layers of features from tiny images.

Liu, X.; Cheng, M.; Zhang, H.; and Hsieh, C.-J. 2018. Towards robust neural networks via random self-ensemble. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 369–385.

MacKay, D. 2003. *Information Theory, Inference, and Learning Algorithms*, volume 50. ISBN 978-0-521-64298-9.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks. In *International Conference on Learning Representations*.

Maho, T.; Furon, T.; and Le Merrer, E. 2021. SurFree: a fast surrogate-free black-box attack. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10425–10434. IEEE. ISBN 978-1-66544-509-2.

Naseer, M.; Khan, S.; Hayat, M.; Khan, F. S.; and Porikli, F. 2021. On Generating Transferable Targeted Perturbations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 7708–7717.

Qin, Z.; Fan, Y.; Zha, H.; and Wu, B. 2021. Random Noise Defense Against Query-Based Black-Box Attacks. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*, volume 34, 7650–7663. Curran Associates, Inc.

Raghunathan, A.; Xie, S. M.; Yang, F.; Duchi, J. C.; and Liang, P. 2020. Understanding and mitigating the tradeoff between robustness and accuracy. In *ICML*, volume 119 of *Proceedings of machine learning research*, 7909–7919. PMLR.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; Berg, A. C.; and Fei-Fei, L. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3): 211–252.

Schmidt, L.; Santurkar, S.; Tsipras, D.; Talwar, K.; and Madry, A. 2018. Adversarially Robust Generalization Requires More Data. In Bengio, S.; Wallach, H.; Larochelle, H.; Grauman, K.; Cesa-Bianchi, N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Sharad, K.; Marson, G. A.; Truong, H. T. T.; and Karame, G. 2020. On the Security of Randomized Defenses Against Adversarial Samples. In Sun, H.; Shieh, S.; Gu, G.; and Ateniese, G., eds., *ASIA CCS '20: The 15th ACM Asia Conference on Computer and Communications Security, Taipei, Taiwan, October 5-9, 2020*, 381–393. ACM.

Simon-Gabriel, C.-J.; Sheikh, N. A.; and Krause, A. 2021. PopSkipJump: Decision-Based Attack for Probabilistic Classifiers. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning*, Proceedings of Machine Learning Research, 9712–9721. PMLR.

Stutz, D.; Hein, M.; and Schiele, B. 2019. Disentangling Adversarial Robustness and Generalization. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6969–6980. IEEE. ISBN 978-1-72813-293-8.

Tsipras, D.; Santurkar, S.; Engstrom, L.; Turner, A.; and Madry, A. 2019. Robustness May Be at Odds with Accuracy. In *International Conference on Learning Representations*.

Wang, H.; Chen, T.; Gui, S.; Hu, T.; Liu, J.; and Wang, Z. 2020. Once-for-All Adversarial Training: In-Situ Tradeoff between Robustness and Accuracy for Free. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 7449–7461. Curran Associates, Inc.

Xie, C.; Wang, J.; Zhang, Z.; Ren, Z.; and Yuille, A. 2018. Mitigating Adversarial Effects Through Randomization. In *International Conference on Learning Representations*.

Xie, C.; Zhang, Z.; Zhou, Y.; Bai, S.; Wang, J.; Ren, Z.; and Yuille, A. L. 2019. Improving Transferability of Adversarial Examples With Input Diversity. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, 2730–2739. Computer Vision Foundation / IEEE.

Yang, Y.-Y.; Rashtchian, C.; Zhang, H.; Salakhutdinov, R. R.; and Chaudhuri, K. 2020. A Closer Look at Accuracy vs. Robustness. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 8588–8601. Curran Associates, Inc.

Zhang, H.; Yu, Y.; Jiao, J.; Xing, E.; Ghaoui, L. E.; and Jordan, M. 2019. Theoretically Principled Trade-off between Robustness and Accuracy. In Chaudhuri, K.; and Salakhutdinov, R., eds., *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, 7472–7482. PMLR.

Zimmer, P.; Andreina, S.; Marson, G. A.; and Karame, G. 2023. Closing the Gap: Achieving Better Accuracy-Robustness Tradeoffs Against Query-Based Attacks. arXiv:2312.10132.