BENCHMARKING ADVERSARIAL ROBUSTNESS

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep neural networks are vulnerable to adversarial examples, which becomes one of the most important problems in the development of deep learning. While a lot of efforts have been made in recent years, it is of great significance to perform correct and complete evaluations of the adversarial attack and defense algorithms. In this paper, we establish a comprehensive, rigorous, and coherent benchmark to evaluate adversarial robustness on image classification tasks. After briefly reviewing plenty of representative attack and defense methods, we perform large-scale experiments with two robustness curves as the fair-minded evaluation criteria to fully understand the performance of these methods. Based on the evaluation results, we draw several important findings and provide insights for future research.

1 INTRODUCTION

Recent progress in deep learning (DL) has led to substantial improvements in a wide range of domains, such as image understanding (Krizhevsky et al., 2012; He et al., 2016), speech recognition (Graves et al., 2013), and natural language processing (Devlin et al., 2019). However, the existing DL models are highly vulnerable to adversarial examples (Szegedy et al., 2014; Goodfellow et al., 2015), which are maliciously generated by an adversary to make a model produce erroneous predictions. As DL models have been integrated into various security-sensitive applications (e.g., autonomous driving, healthcare, and finance), the study of the adversarial robustness issue has attracted increasing attention with an enormous number of adversarial attack and defense methods proposed. Therefore, it is crucial to conduct correct and rigorous evaluations of these methods for understanding their pros and cons, comparing their performance, and providing insights for building new methods (Carlini et al., 2019).

The research on adversarial robustness is faced with an "arms race" between attacks and defenses: a defense method proposed to prevent existing attacks was soon evaded by new attacks, and vice versa (Carlini & Wagner, 2017a;b; He et al., 2018; Athalye et al., 2018a; Uesato et al., 2018; Zhang et al., 2019b). For instance, defensive distillation (Papernot et al., 2016c) was proposed to improve the robustness, but was later shown to be ineffective against a strong attack (Carlini & Wagner, 2017b). Many methods were introduced to build robust models by causing obfuscated gradients, which can be defeated by the adaptive ones (Athalye et al., 2018a; Uesato et al., 2018). As a result, it is particularly challenging to understand their effects, identify the real progress, and advance the field.

Moreover, the current attacks and defenses are often evaluated incompletely. First, most defenses are only tested against a small set of attacks under limited threat models, and many attacks are evaluated on a few models or defenses. Second, the robustness evaluation metrics are too simple to show the performance of these methods. The accuracy of a defense against an attack for a given perturbation budget (Kurakin et al., 2018) and the minimum distance of the adversarial perturbation (Brendel et al., 2018b) are used as the primary evaluation metrics, which are often insufficient to characterize the behavior of the attacks and defenses totally. Consequently, the incomplete evaluation cannot provide a comprehensive understanding of the strengths and limitations of the attack and defense methods.

In this paper, we establish a comprehensive, rigorous, and coherent benchmark to evaluate adversarial robustness, which can provide a comprehensive understanding of the effects of existing methods under different scenarios, with a hope to facilitate the future research. In particular, we focus on the robustness of image classifiers under the ℓ_p norm threat models, since the adversarial robustness issue has been extensively studied on image classification tasks with the ℓ_p additive noises. We incorporate a lot of typical and state-of-the-art attack and defense methods for robustness evaluation, including 15 attack methods and 16 defense models—8 on CIFAR-10 (Krizhevsky & Hinton, 2009) and 8 on

ImageNet (Russakovsky et al., 2015). To fully demonstrate the performance of these methods, we adopt two complementary robustness curves as the major evaluation metrics to present the results. Then, we carry out large-scale experiments on the cross evaluation of the attack and defense methods under complete threat models[1], including 1) untargeted and targeted attacks; 2) ℓ_∞ and ℓ_2 attacks; 3) white-box, transfer-based, score-based, and decision-based attacks.

By analyzing the quantitative results, we have some important findings. First, the relative robustness between defenses against an attack could be different under varying perturbation budgets or attack iterations. So it is hard to conclude that a defense is more robust than another against an attack by using a specific configuration. However, this is common in previous works. Second, although various defense techniques have been proposed, the most robust defenses are still the adversarially trained models. The robustness of these defenses can also generalize to other threat models, under which they are not trained to be robust. Third, defenses based on randomization are generally more robust to black-box attacks based on the query feedback. More detailed discussions can be found in Sec. 5.3.

All evaluation experiments are conducted on a new adversarial robustness platform[2] developed by us, since the existing platforms (e.g., CleverHans (Papernot et al., 2016a), Foolbox (Rauber et al., 2017), etc) cannot fully support our comprehensive evaluations (details in Appendix A). We hope that our platform could continuously incorporate and evaluate more methods, and be helpful for future works.

2 Threat Models

Precisely defining threat models is fundamental to perform adversarial robustness evaluations. According to Carlini et al. (2019), a threat model specifies the adversary's goals, capabilities, and knowledge under which an attack is performed and a defense is built to be robust. We first define the notations and then illustrate the three aspects of a threat model, respectively.

A classifier can be denoted as $C(\boldsymbol{x}) : \mathcal{X} \to \mathcal{Y}$, where $\boldsymbol{x} \in \mathcal{X} \subset \mathbb{R}^d$ is the input, and $\mathcal{Y} = \{1, 2, ..., L\}$ with L being the number of classes. Let y denote the ground-truth label of $\boldsymbol{x}$, and $\boldsymbol{x}^{adv}$ denote an adversarial example corresponding to $\boldsymbol{x}$.

2.1 Adversary's Goals

An adversary can have different goals of generating adversarial examples. We consider the *untargeted* and *targeted* adversarial examples in this paper. An untargeted adversarial example aims to cause misclassification of the classifier, as $C(\boldsymbol{x}^{adv}) \neq y$. A targeted one is crafted to be misclassified as the adversary-desired target class by the classifier, as $C(\boldsymbol{x}^{adv}) = y^*$, where y^* is the target class.

2.2 Adversary's Capabilities

As adversarial examples are usually assumed to be indistinguishable from the corresponding original ones to human eyes (Szegedy et al., 2014; Goodfellow et al., 2015), the adversary can only make small changes to the inputs. In this paper, we study the well-defined and widely used ℓ_p norm threat models, although there also exist other threat models (Xiao et al., 2018; Song et al., 2018b; Engstrom et al., 2019). Under the ℓ_p norm threat models, the adversary is allowed to add a small perturbation measured by the ℓ_p norm to the original input. Specifically, we consider the ℓ_∞ and ℓ_2 norms.

To achieve the adversary's goal, two strategies could be adopted to craft adversarial examples with small perturbations. The first seeks to craft an adversarial example $\boldsymbol{x}^{adv}$ that satisfies $\|\boldsymbol{x}^{adv} - \boldsymbol{x}\|_p \leq \epsilon$, where ϵ is the perturbation budget, while misleads the model. This could be achieved by solving a constrained optimization problem. For instance, the adversary can get an untargeted adversarial example by maximizing a loss function $\mathcal{J}$ (e.g., the cross-entropy loss) in the restricted region as

$$\boldsymbol{x}^{adv} = \operatorname*{arg\,max}_{\boldsymbol{x}': \|\boldsymbol{x}' - \boldsymbol{x}\|_p \leq \epsilon} \mathcal{J}(\boldsymbol{x}', y). \tag{1}$$

We call it the adversarial example with a *constrained* perturbation.

[1] Note that the purpose of this paper is not to contradict the previous results of some methods, but to provide a comprehensive evaluation of the existing methods. So we consider complete threat models even if a defense does not claim to be robust under some threat models.

[2] Our code is available at an anonymous link: `https://git.io/JeGsU`

The second strategy is generating an adversarial example by finding the minimum perturbation as

$$\boldsymbol{x}^{adv} = \operatorname*{arg\,min}_{\boldsymbol{x}':\boldsymbol{x}' \text{is adversarial}} \|\boldsymbol{x}' - \boldsymbol{x}\|_p. \tag{2}$$

We call $\boldsymbol{x}^{adv}$ generated by solving Eq. (2) the adversarial example with an *optimized* perturbation. However, it is usually intractable to solve Eq. (1) or Eq. (2) exactly, and thus various attack methods have been proposed to get an approximate solution.

2.3 Adversary's Knowledge

An adversary can have different levels of knowledge of the target model, from white-box access to the model architectures and parameters, to black-box access to the training data or model predictions. Based on the different knowledge of the model, we consider four attack scenarios, including *white-box attacks*, *transfer-based*, *score-based*, and *decision-based black-box attacks*.

White-box attacks rely on detailed information of the target model, including architecture, parameters, and gradient of the loss w.r.t. the input. For the defense models, the adversary also has access to the specific defense mechanisms, and designs adaptive attacks to evade them. Transfer-based black-box attacks are based on the transferability of adversarial examples (Papernot et al., 2016b). These attacks do not rely on model information but assume the availability of the training data. It is used to train a substitute model from which the adversarial examples are generated. Score-based black-box attacks can only acquire the output probabilities by querying the target model. And decision-based black-box attacks solely rely on the predicted classes of the queries. Score-based and decision-based attacks are also restricted by a limited number of queries to the target model.

3 Attacks and Defenses

In this section, we summarize the typical adversarial attack and defense methods.

3.1 Attack Methods

White-box Attacks: Most white-box attacks craft adversarial examples based on the input gradient. To solve Eq. (1), the fast gradient sign method (**FGSM**) (Goodfellow et al., 2015) linearizes the loss function in the input space and generates an adversarial example by an one-step update. The basic iterative method (**BIM**) (Kurakin et al., 2017) extends FGSM by iteratively taking multiple small gradient steps. Similar to BIM, Madry et al. (2018) apply the projected gradient descent (**PGD**) with random starts as a universal first-order adversary. To solve Eq. (2), **DeepFool** (Moosavi-Dezfooli et al., 2016) has been proposed to generate an adversarial example with the minimum perturbation. The Carlini & Wagner (2017b)'s method (**C&W**) takes a Lagrangian form and adopts Adam (Kingma & Ba, 2015) for optimization. However, some defenses can be robust against these gradient-based attacks by causing obfuscated gradients (Athalye et al., 2018a). To circumvent them, the adversary can use **BPDA** (Athalye et al., 2018a) to provide an approximate gradient when the true gradient is unavailable or useless, or **EOT** (Athalye et al., 2018b) when the gradient is random.

Transfer-based Black-box Attacks: Transfer-based attacks generate adversarial examples against a substitute model, which have a probability to fool black-box models due to the transferability. Besides the above methods, some others have been proposed to improve the transferability. The momentum iterative method (**MIM**) (Dong et al., 2018) integrates a momentum term into BIM to stabilize the update direction during the attack iterations. The diverse inputs method (**DIM**) (Xie et al., 2019b) applies the gradient of the randomly resized and padded input for adversarial example generation.

Score-based Black-box Attacks: In this setting, although the white-box access to the model gradient is unavailable, it can be estimated by some gradient-free methods through queries. **ZOO** (Chen et al., 2017) estimates the gradient at each coordinate through finite differences and adopts C&W for attacks based on the estimated gradient. **NES** (Ilyas et al., 2018) and **SPSA** (Uesato et al., 2018) can give the full gradient estimation based on drawing random samples and acquiring the corresponding loss values. $\mathcal{N}$**ATTACK** (Li et al., 2019) does not estimate the gradient but learns a Gaussian distribution centered around the input such that a sample drawn from it is likely an adversarial example.

Decision-based Black-box Attacks: This setting is more challenging since the model only provides discrete hard-label predictions. Brendel et al. (2018a) propose the first method in this setting, called the **Boundary** attack, which is based on random walk on the decision boundary. An **optimization-based** method (Cheng et al., 2019) formulates this problem as a continuous optimization problem and estimates the gradient to solve it. An **evolutionary** attack method (Dong et al., 2019) is further proposed to improve the query efficiency based on the evolution strategy.

3.2 DEFENSES

Due to the threat of adversarial examples, extensive research has been conducted on building robust models to defend against adversarial attacks. In this paper, we roughly classify the defense techniques into five categories, including *robust training*, *input transformation*, *randomization*, *model ensemble*, and *certified defenses*. Note that these defense categories are not exclusive, i.e., a defense can belong to many categories. Below we introduce each category.

Robust Training: The basic principle of robust training is to make the classifier robust against small perturbations internally. One line of work is based on adversarial training (Goodfellow et al., 2015; Tramèr et al., 2018; Madry et al., 2018; Kannan et al., 2018; Zhang et al., 2019a), which augments the training data by adversarially generated examples. Another line of work trains robust models by regularizations, including those on the Lipschitz constant (Cisse et al., 2017), input gradients (Hein & Andriushchenko, 2017; Ross & Doshi-Velez, 2018), or perturbation norm (Yan et al., 2018).

Input Transformation: Several defenses transform the inputs before feeding them to the classifier, including JPEG compression (Dziugaite et al., 2016), bit-depth reduction (Xu et al., 2018), total variance minimization (Guo et al., 2018), autoencoder-based denoising (Liao et al., 2018), and projecting adversarial examples onto the data distribution through generative models (Samangouei et al., 2018; Song et al., 2018a). However, these defenses can cause shattered gradients or vanishing/exploding gradients (Athalye et al., 2018a), which can be evaded by adaptive attacks.

Randomization: The classifiers can be made random to mitigate adversarial effects. The randomness can be added to either the input (Xie et al., 2018) or the model (Dhillon et al., 2018; Liu et al., 2018). The randomness can also be modeled by Bayesian neural networks (Liu et al., 2019). These methods partially rely on random gradients to prevent adversarial attacks, and can be defeated by attacks that take the expectation over the random gradients (He et al., 2018; Athalye et al., 2018a).

Model Ensemble: An effective defense strategy in practice is to construct an ensemble of individual models (Kurakin et al., 2018). Besides aggregating the output of each model in the ensemble, some different ensemble strategies have been proposed. Random self-ensemble (Liu et al., 2018) averages the predictions over random noises injected to the model, which is equivalent to ensemble an infinite number of noisy models. Pang et al. (2019) propose to promote the diversity among the predictions of different models, and introduce an adaptive diversity promoting regularizer to achieve this.

Certified Defenses: There are a lot of works (Raghunathan et al., 2018a; Sinha et al., 2018; Wong & Kolter, 2018; Wong et al., 2018; Raghunathan et al., 2018b; Xiao et al., 2019) on training certified defenses, which are provably guaranteed to be robust to adversarial perturbations under some threat models. Recently, certified defenses (Zhang & Liang, 2019; Cohen et al., 2019) can apply to ImageNet (Russakovsky et al., 2015), showing the scalability of this type of defenses.

4 EVALUATION METHODOLOGY

With the growing number of adversarial attacks and defenses being proposed, the correct and rigorous evaluation of these methods becomes increasingly important to help us better understand the strengths and limitations of these methods. However, there still lacks a comprehensive understanding of the effects of these methods due to incorrect or incomplete evaluations. To address this issue and further advance the field, we establish a comprehensive, rigorous, and coherent benchmark to evaluate adversarial robustness empirically. We incorporate 15 attack methods and 16 defense models on two image datasets in our benchmark for robustness evaluation. We also adopt two complementary robustness curves as the fair-minded evaluation metrics to better show the results.

4.1 EVALUATION METRICS

We first introduce the evaluation metrics that are adopted in our benchmark.

Table 1: We show the defense models that are incorporated into our benchmark for adversarial robustness evaluation. We also show the defense type and accuracy (%) on clean data of each method. The accuracy is re-calculated by ourselves. More details about their model architectures are shown in Appendix B.

CIFAR-10 (Krizhevsky & Hinton, 2009)			ImageNet (Russakovsky et al., 2015)		
Defense Model	Category	Acc.	Defense Model	Category	Acc.
Res-56 (He et al., 2016)	natural training	92.6	Inc-v3 (Szegedy et al., 2016)	natural training	78.0
PGD-AT (Madry et al., 2018)	robust training	87.3	Ens-AT (Tramèr et al., 2018)	robust training	73.5
DeepDefense (Yan et al., 2018)	robust training	79.7	ALP (Kannan et al., 2018)	robust training	49.0
TRADES (Zhang et al., 2019a)	robust training	84.9	FD (Xie et al., 2019a)	robust training	64.3
Convex (Wong et al., 2018)	(certified) robust training	66.3	JPEG (Dziugaite et al., 2016)	input transformation	77.3
JPEG (Dziugaite et al., 2016)	input transformation	80.9	Bit-Red (Xu et al., 2018)	input transformation	61.8
RSE (Liu et al., 2018)	randomization & ensemble	86.1	R&P (Xie et al., 2018)	(input) randomization	77.0
ADP (Pang et al., 2019)	ensemble	94.1	RandMix (Zhang & Liang, 2019)	(certified input) randomization	52.4

Given an attack method $\mathcal{A}_{\epsilon,p}$ that generates an adversarial example $\boldsymbol{x}^{adv} = \mathcal{A}_{\epsilon,p}(\boldsymbol{x})$ for an input $\boldsymbol{x}$ with perturbation budget ϵ under the ℓ_p norm[3], and a (defense) classifier C defined in Sec. 2, the accuracy of the classifier against the attack is defined as

$$\mathrm{Acc}(C, \mathcal{A}_{\epsilon,p}) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\big(C(\mathcal{A}_{\epsilon,p}(\boldsymbol{x}_i)) = y_i\big), \tag{3}$$

where $\{\boldsymbol{x}_i, y_i\}_{i=1}^N$ is the test set, $\mathbf{1}(\cdot)$ is the indicator function; while the attack success rate of the attack on the classifier is defined as

$$\mathrm{Asr}(\mathcal{A}_{\epsilon,p}, C) = \frac{1}{M}\sum_{i=1}^{N} \mathbf{1}\big(C(\boldsymbol{x}_i) = y_i \wedge C(\mathcal{A}_{\epsilon,p}(\boldsymbol{x}_i)) \neq y_i\big) \text{ or } \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\big(C(\mathcal{A}_{\epsilon,p}(\boldsymbol{x}_i)) = y_i^*\big) \tag{4}$$

for the untargeted and targeted cases, respectively, where $M = \sum_{i=1}^N \mathbf{1}\big(C(\boldsymbol{x}_i) = y_i\big)$.

The previous methods usually report the accuracy or the attack success rate for some chosen perturbation budgets ϵ, which may not reflect their behavior totally. In this paper, we adopt two complementary robustness curves to clearly and thoroughly show the robustness and resistance of the classifier against the attack, as well as the effectiveness and efficiency of the attack on the classifier.

The first one is the *accuracy (attack success rate) vs. perturbation budget* curve, which can give a global understanding of the robustness of the classifier and the effectiveness of the attack. To generate such a curve, we need to calculate the accuracy (attack success rate) for all values of ϵ. This can be efficiently done for attacks that find the minimum perturbations, by counting the number of the adversarial examples, the ℓ_p norm of whose perturbations is smaller than each ϵ. For attacks that craft adversarial examples with constrained perturbations, we perform a binary search on ϵ to find its minimum value that enables the generated adversarial example to fulfill the adversary's goal.

The second curve is the *accuracy (attack success rate) vs. attack strength* curve, where the attack strength is defined as the number of iterations or model queries based on different attack methods. This curve can show the efficiency of the attack, as well as the resistance of the classifier to the attack, e.g., a defense whose accuracy drops to zero against an attack with 100 iterations is considered to be more resistant to this attack than another defense that is totally broken by the same attack with 10 iterations, although the worst-case accuracy of both models is zero.

4.2 Evaluated Datasets and Algorithms

Datasets: We use the CIFAR-10 (Krizhevsky & Hinton, 2009) and ImageNet (Russakovsky et al., 2015) datasets to perform adversarial robustness evaluation in this paper. We use the test set containing $10,000$ images of CIFAR-10, and randomly choose $1,000$ images from the ImageNet validation set for evaluation. For each image, we select a target class uniformly over all other classes except its true class at random, which is used for targeted attacks.

Defense Models: For fair evaluation, we test 16 representative defense models whose original source codes and pre-trained models are publicly available. The selected models cover all defense categories and the state-of-the-art models in each category are included in them. On CIFAR-10, we choose 8 models—naturally trained ResNet-56 (Res-56) (He et al., 2016), PGD-based adversarial training

[3] For attacks that find minimum perturbations, e.g., DeepFool and C&W, we let $\mathcal{A}_{\epsilon,p}(\boldsymbol{x}) = \boldsymbol{x}$ if the ℓ_p norm of the minimum perturbation is larger than ϵ.

Table 2: We show the attack methods that are implemented in our benchmark for adversarial robustness evaluation. We also show the adversary's knowledge, goals, capability, and distance metrics of each attack method.

Attack Method	Knowledge	Goals	Capability	Distance
FGSM (Goodfellow et al., 2015)	white-box & transfer-based	untargeted & targeted	constrained	ℓ_∞, ℓ_2
BIM (Kurakin et al., 2017)	white-box & transfer-based	untargeted & targeted	constrained	ℓ_∞, ℓ_2
MIM (Dong et al., 2018)	white-box & transfer-based	untargeted & targeted	constrained	ℓ_∞, ℓ_2
DeepFool (Moosavi-Dezfooli et al., 2016)	white-box	untargeted	optimized	ℓ_∞, ℓ_2
C&W (Carlini & Wagner, 2017b)	white-box	untargeted & targeted	optimized	ℓ_2
DIM (Xie et al., 2019b)	transfer-based	untargeted & targeted	constrained	ℓ_∞, ℓ_2
ZOO (Chen et al., 2017)	score-based	untargeted & targeted	optimized	ℓ_2
NES (Ilyas et al., 2018)	score-based	untargeted & targeted	constrained	ℓ_∞, ℓ_2
SPSA (Uesato et al., 2018)	score-based	untargeted & targeted	constrained	ℓ_∞, ℓ_2
$\mathcal{N}$ATTACK (Li et al., 2019)	score-based	untargeted & targeted	constrained	ℓ_∞, ℓ_2
Boundary (Brendel et al., 2018a)	decision-based	untargeted & targeted	optimized	ℓ_2
Evolutionary (Dong et al., 2019)	decision-based	untargeted & targeted	optimized	ℓ_2

(PGD-AT) (Madry et al., 2018), DeepDefense (Yan et al., 2018), TRADES (Zhang et al., 2019a), convex outer polytope (Convex) (Wong et al., 2018), JPEG compression (Dziugaite et al., 2016), random self-ensemble (RSE) (Liu et al., 2018), and adaptive diversity promoting (ADP) (Pang et al., 2019). On ImageNet, we also choose 8 models—naturally trained Inception v3 (Inc-v3) (Szegedy et al., 2016), ensemble adversarial training (Ens-AT) (Tramèr et al., 2018), adversarial logit pairing (ALP) (Kannan et al., 2018), feature denoising (FD) (Xie et al., 2019a), JPEG compression (Dziugaite et al., 2016), bit-depth reduction (Bit-Red) (Xu et al., 2018), random resizing and padding (R&P) (Xie et al., 2018), and RandMix (Zhang & Liang, 2019). We use the natural models as the backbone classifiers for defenses based on input transformation (e.g., JPEG). Table 1 shows the defense details.

Attacks: We implement 15 typical and widely used attack methods in our benchmark, including 5 white-box attacks—FGSM, BIM, MIM, DeepFool, and C&W, 4 transfer-based attacks—FGSM, BIM, MIM, and DIM, 4 score-based attacks—ZOO, NES, SPSA, and $\mathcal{N}$ATTACK, and 2 decision-based attacks—Boundary and Evolutionary. More details of these attacks are outlined in Table 2. Note that 1) we do not evaluate PGD since PGD and BIM are very similar; 2) for transfer-based attacks, we craft adversarial examples by those white-box methods on a substitute model; 3) for defenses that rely on obfuscated gradients, we implement the white-box attacks adaptively by replacing the true gradient with an approximate one when it is unavailable or an expected one when it is random.

Platform: All of the attacks and defenses are implemented on a new adversarial robustness platform developed by us. We also conduct the experiments based on the platform. The comparisons between the existing platforms and ours are detailed in Appendix A. We acknowledge that many good works are not included in our current benchmark. We hope that our platform could continuously incorporate and evaluate more methods, and be helpful for future works.

5 Evaluation Results

We present the evaluation results on CIFAR-10 in Sec. 5.1, and ImageNet in Sec. 5.2. Due to the space limitation, we mainly provide the accuracy vs. perturbation budget and attack strength curves of the defense models against untargeted attacks under the ℓ_∞ norm in the section, and leave the full experimental results (including targeted attacks under the ℓ_∞ norm, untargeted and targeted attacks under the ℓ_2 norm, and attack success rate curves) in Appendix C. We also report some key findings in Sec. 5.3. Most experiments are conducted on a NVIDIA DGX-1 server with 8 Tesla P100 GPUs.

5.1 Evaluation Results on CIFAR-10

We show the accuracy of the 8 models on CIFAR-10 against white-box, transfer-based, score-based, and decision-based attacks in this section. To get the *accuracy vs. perturbation budget* curves, we fix the attack strength (i.e., attack iterations or queries) for different budgets. To generate the *accuracy vs. attack strength* curves, we use a fixed perturbation budget as $\epsilon = \frac{8.0}{255.0}$ for ℓ_∞ attacks and $\epsilon = 1.0$ for ℓ_2 attacks, with images in $[0, 1]$. The detailed parameters of each attack are provided in Appendix B.

White-box Attacks: We show the *accuracy vs. perturbation budget* and *accuracy vs. attack strength* curves of the 8 models against untargeted FGSM, BIM, MIM, and DeepFool attacks under the ℓ_∞ norm in Fig. 1 and Fig. 2. Note that there is no *accuracy vs. attack strength* curve for FGSM since it is an one-step attack. The accuracy of the models drops to zero against iterative attacks with the increasing perturbation budget. For a given perturbation budget, the accuracy decreases along with

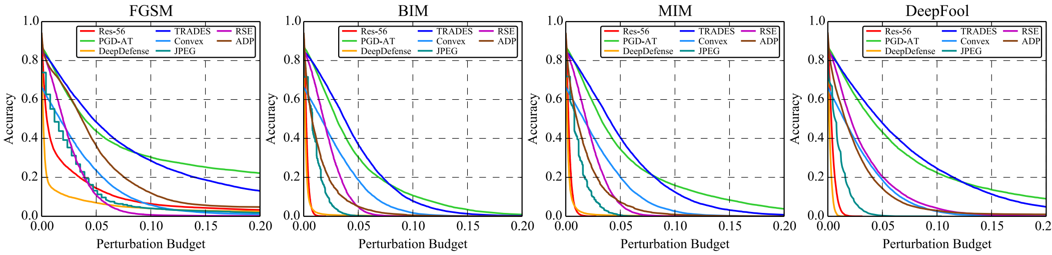

Figure 1: The *accuracy vs. perturbation budget* curves of the 8 models on CIFAR-10 against untargeted white-box attacks under the ℓ_∞ norm.

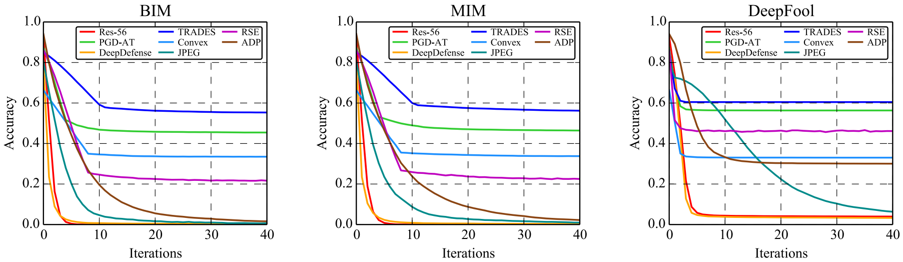

Figure 2: The *accuracy vs. attack strength* curves of the 8 models on CIFAR-10 against untargeted white-box attacks under the ℓ_∞ norm.

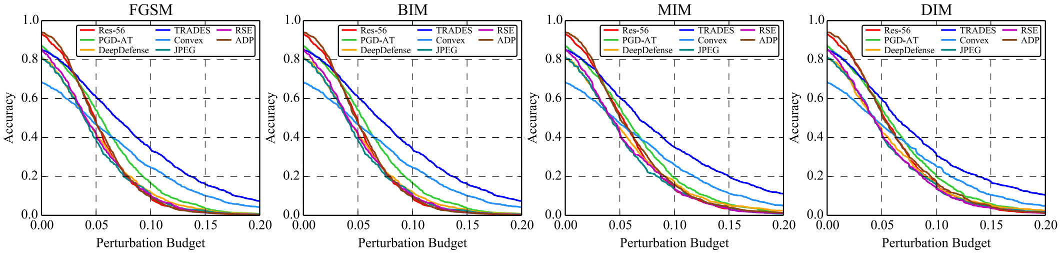

Figure 3: The *accuracy vs. perturbation budget* curves of the 8 models on CIFAR-10 against untargeted transfer-based attacks under the ℓ_∞ norm.

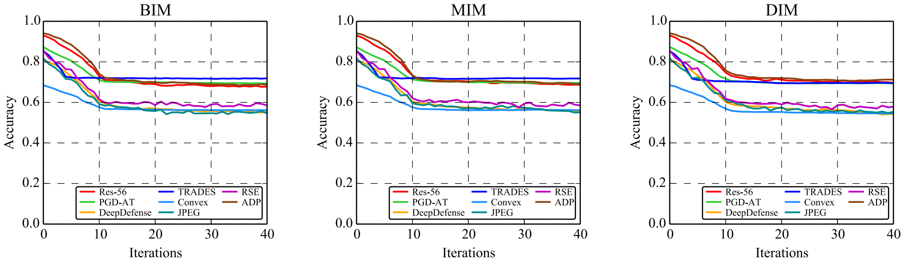

Figure 4: The *accuracy vs. attack strength* curves of the 8 models on CIFAR-10 against untargeted transfer-based attacks under the ℓ_∞ norm.

the attack iterations and finally converges after a few or dozens of iterations. Based on the results, we observe that under white-box attacks, the adversarially trained models (i.e., PGD-AT, TRADES) are more robust than other models, because they are trained on the worst-case adversarial examples. We also observe that the relative robustness between two models against an attack could be different under different perturbation budgets or attack iterations. For instance, the accuracy of TRADES is higher than that of PGD-AT against white-box attacks when the perturbation budget is small (e.g., $\epsilon = 0.05$), but is lower when it is large (e.g., $\epsilon = 0.15$). This finding implies that the comparison between the defense models at a chosen perturbation budget or attack iteration, which is common in previous works, cannot fully demonstrate the performance of a model. But the robustness curves adopted in this paper can better show the global behavior of the attack and defense methods.

Transfer-based Black-box Attacks: We show the *accuracy vs. perturbation budget* and *accuracy vs. attack strength* curves of the 8 models against untargeted transfer-based FGSM, BIM, MIM, and DIM attacks under the ℓ_∞ norm in Fig. 3 and Fig. 4. In this experiment, we choose TRADES as the substitute model to attack the others, and use PGD-AT to attack TRADES, since these two models demonstrate superior white-box robustness compared with other models, and thus the adversarial examples generated on other models can rarely transfer to TRADES and PGD-AT. From the results, the accuracy of the defenses also drops with the increasing perturbation budget, and it converges after a few attack iterations. We also observe that the recent attacks (e.g., MIM, DIM) for improving the transferability do not actually perform better than the baseline BIM method.

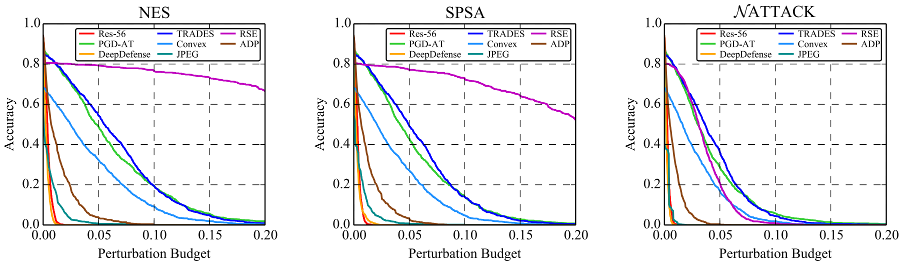

Figure 5: The *accuracy vs. perturbation budget* curves of the 8 models on CIFAR-10 against untargeted score-based attacks under the ℓ_∞ norm.

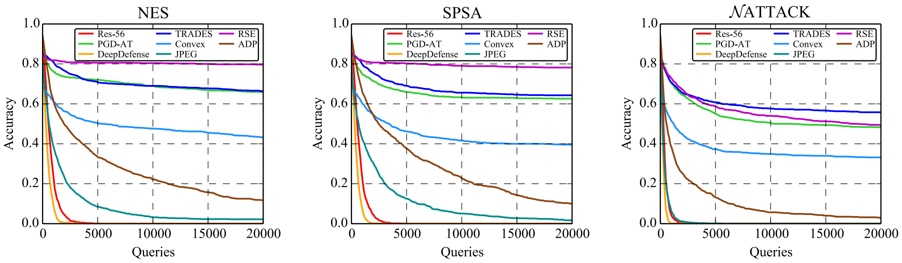

Figure 6: The *accuracy vs. attack strength* curves of the 8 models on CIFAR-10 against untargeted score-based attacks under the ℓ_∞ norm.

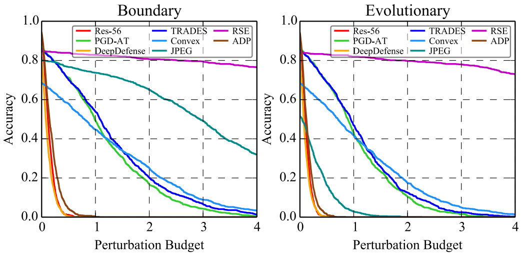

Figure 7: The *accuracy vs. perturbation budget* curves of the 8 models on CIFAR-10 against untargeted decision-based attacks under the ℓ_2 norm.

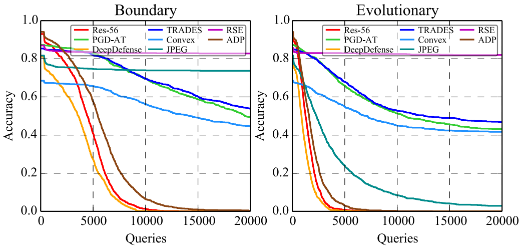

Figure 8: The *accuracy vs. attack strength* curves of the 8 models on CIFAR-10 against untargeted decision-based attacks under the ℓ_2 norm.

Score-based Black-box Attacks: We show the two accuracy curves against untargeted score-based NES, SPSA, and $\mathcal{N}$ATTACK under the ℓ_∞ norm in Fig. 5 and Fig. 6. We set the maximum number of queries as $20,000$ in these attack methods. The accuracy of the defenses also decreases along with the increasing perturbation budget or the number of queries. $\mathcal{N}$ATTACK is a more effective attack method, as can be seen from the figures. From the results, we notice that RSE is quite resistant to score-based attacks, especially NES and SPSA. We think that the randomness of the predictions given by RSE makes the estimated gradients of NES and SPSA useless for attacks.

Decision-based Black-box Attacks: Since the decision-based Boundary and Evolutionary methods can only be used for ℓ_2 attacks, we present the accuracy curves of the 8 models against untargeted Boundary and Evolutionary attacks under the ℓ_2 norm in Fig. 7 and Fig. 8. The behavior of the defenses is similar to that of score-based attacks. It can be observed that RSE is also resistant to decision-based attacks compared with other defenses due to the randomness of the predictions.

5.2 Evaluation Results on ImageNet

We present the experimental results on ImageNet in this section. We use the same settings with those on CIFAR-10 to get the evaluation curves. Since the input image size is different for the defenses, we adopt the normalized ℓ_2 distance defined as $\bar{\ell}_2(\boldsymbol{a}) = \|\boldsymbol{a}\|_2/\sqrt{d}$ as the measurement for ℓ_2 attacks, where d is the dimension of a vector $\boldsymbol{a}$. To get the *accuracy (attack success rate) vs. attack strength* curves, we fix the perturbation budget as $\epsilon = \frac{16.0}{255.0}$ for ℓ_∞ attacks and $\epsilon = \sqrt{0.001}$ for ℓ_2 attacks.

White-box Attacks: We show the accuracy curves of the 8 models on ImageNet against untargeted FGSM, BIM, and MIM under the ℓ_∞ norm in Fig. 9 and Fig. 10. We find that FD exhibits superior performance over all other models. FD is also trained by the adversarial training method in Madry et al. (2018), demonstrating the effectiveness of PGD-based adversarial training on ImageNet.

Transfer-based Black-box Attacks: We use a ResNet-152 model (He et al., 2016) as the substitute model. The results of the defenses against untargeted transfer-based FGSM, BIM, MIM, and DIM

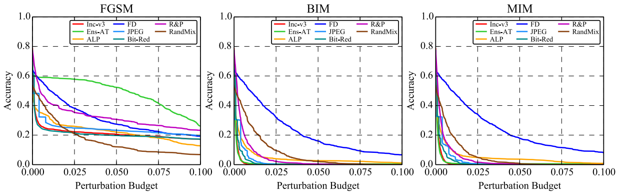

Figure 9: The *accuracy vs. perturbation budget* curves of the 8 models on ImageNet against untargeted white-box attacks under the ℓ_∞ norm.

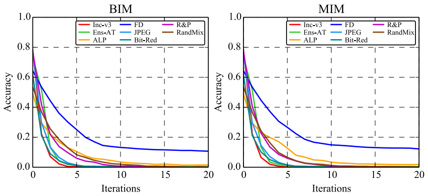

Figure 10: The *accuracy vs. attack strength* curves of the 8 models on ImageNet against untargeted white-box attacks under the ℓ_∞ norm.

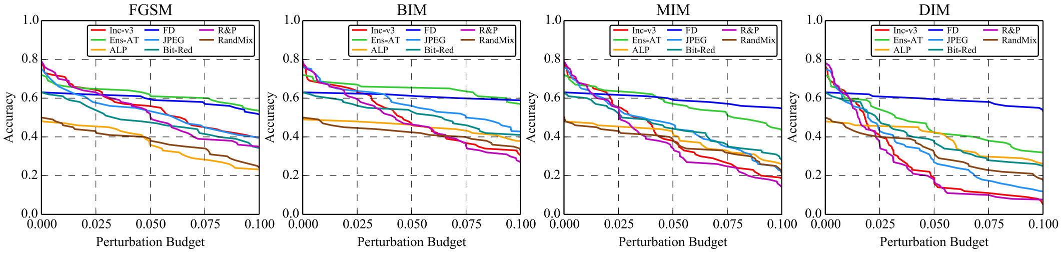

Figure 11: The *accuracy vs. perturbation budget* curves of the 8 models on ImageNet against untargeted transfer-based attacks under the ℓ_∞ norm.

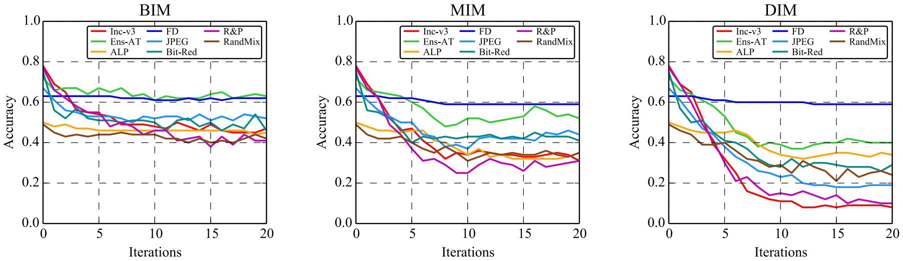

Figure 12: The *accuracy vs. attack strength* curves of the 8 models on ImageNet against untargeted transfer-based attacks under the ℓ_∞ norm.

under the ℓ_∞ norm are shown in Fig. 11 and Fig. 12. Different from the results on CIFAR-10, MIM and DIM improve the transferability of adversarial examples over FGSM and BIM, resulting in lower accuracy of the black-box models. A potential reason is that the image size of ImageNet is much larger, and the adversarial examples generated by BIM can "overfit" to the substitute model (Dong et al., 2018), making them difficult to transfer to other black-box models.

Score-based and Decision-based Attacks: Fig. 13 and Fig. 14 show the two accuracy curves of the defense models on ImageNet against untargeted score-based attacks under the ℓ_∞ norm, while Fig. 15 and Fig. 16 show the results for untargeted decision-based attacks under the ℓ_2 norm. Similar to the results on CIFAR-10, we find that the two defenses based on randomization, i.e., R&P and RandMix, have higher accuracy than other methods in most cases. JPEG and Bit-Red that are based on input transformations also improve the robustness over the baseline model.

5.3 Discussions

Based on the above results and more results in Appendix C, we highlight some key findings.

First, the relative robustness between defenses against the same attack could be different under varying attack parameters, such as the perturbation budget or the number of attack iterations. Not only the results of PGD-AT and TRADES in Fig. 1 can prove it, but also the results in many different scenarios show the similar phenomenon. Given this observation, the comparison between defenses at a specific attack configuration cannot fully demonstrate the superiority of a method upon another. We therefore strongly *advise the researchers to adopt the robustness curves as the major evaluation metrics to present the robustness results.*

Second, among the defenses studied in this paper, we find that the most robust models are obtained by PGD-based adversarial training. Their robustness not only is good for the threat model under which they are trained (i.e., the ℓ_∞ threat model), but can also generalize to other threat models (e.g., the ℓ_2 threat model). However, adversarial training usually leads to a reduction of natural accuracy and high training cost. A research direction is to develop new methods that maintain the natural accuracy or reduce the training cost. And we have seen several works (Shafahi et al., 2019) in this direction.

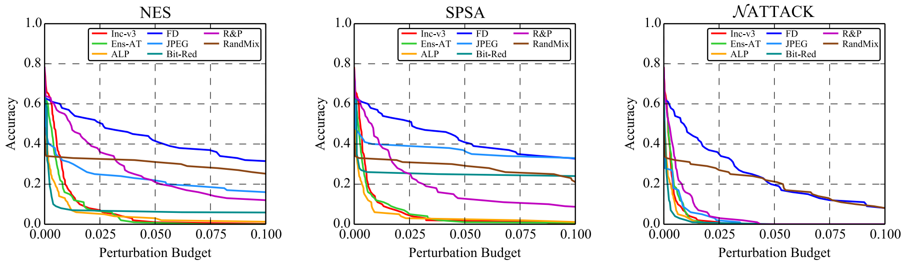

Figure 13: The *accuracy vs. perturbation budget* curves of the 8 models on ImageNet against untargeted score-based attacks under the ℓ_∞ norm.

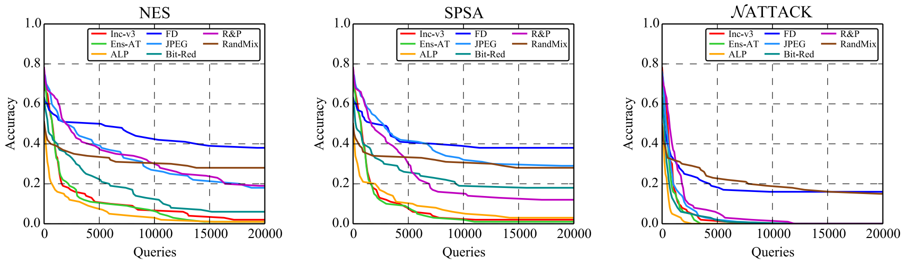

Figure 14: The *accuracy vs. attack strength* curves of the 8 models on ImageNet against untargeted score-based attacks under the ℓ_∞ norm.

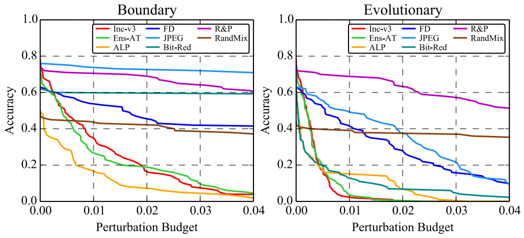

Figure 15: The *accuracy vs. perturbation budget* curves of the 8 models on ImageNet against untargeted decision-based attacks under the ℓ_2 norm.

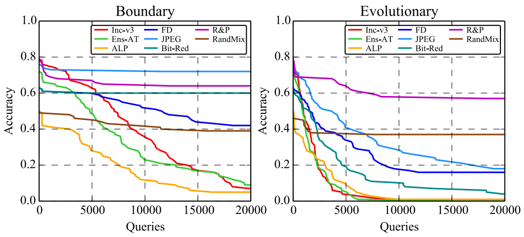

Figure 16: The *accuracy vs. attack strength* curves of the 8 models on ImageNet against untargeted decision-based attacks under the ℓ_2 norm.

Third, we observe that the defenses based on randomization are quite resistant to score-based and decision-based attacks, which rely on the query feedback of the black-box models. We argue that the robustness of the randomization-based defenses against these attacks is due to the random predictions given by the models, making the estimated gradients or search directions unreliable for attacks. A potential research direction is to develop more powerful score-based and decision-based attacks that can efficiently evade the randomization-based defenses.

Fourth, the defenses based on input transformations (e.g., JPEG, Bit-Red) can sightly improve the robustness over the undefended models, and sometimes get much higher accuracy against black-box attacks. Since these methods are quite simple, they may be combined with other types of defenses to build more powerful defenses.

Fifth, we find that different transfer-based attack methods exhibit similar performance on CIFAR-10, while the recent methods (e.g., MIM, DIM) can improve the transferability of adversarial examples over BIM on ImageNet. One potential reason is that the input dimension of the models on ImageNet is much higher than that on CIFAR-10, and thus the adversarial examples generated by BIM can easily "overfit" to the substitute model (Dong et al., 2018), resulting in poor transferability. And the recent methods proposed to solve this issue can generate more transferable adversarial examples.

Note that these findings are based on our current benchmark, which may be strengthened or falsified in the future if new results are given.

6 Conclusion

In this paper, we established a comprehensive, rigorous, and coherent benchmark to evaluate adversarial robustness of image classifiers. We performed large-scale experiments with two robustness curves as the fair-minded evaluation criteria to facilitate a better understanding of the representative and state-of-the-art adversarial attack and defense methods. We drew some key findings based on the evaluation results, which may be helpful for future research.

References

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 (ICML)*, 2018a.

Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. In *International Conference on Machine Learning (ICML)*, 2018b.

Wieland Brendel, Jonas Rauber, and Matthias Bethge. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. In *International Conference on Learning Representations (ICLR)*, 2018a.

Wieland Brendel, Jonas Rauber, Alexey Kurakin, Nicolas Papernot, Behar Veliqi, Marcel Salathé, Sharada P Mohanty, and Matthias Bethge. Adversarial vision challenge. *arXiv preprint arXiv:1808.01976*, 2018b.

Nicholas Carlini. A critique of the deepsec platform for security analysis of deep learning models. *arXiv preprint arXiv:1905.07112*, 2019.

Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *ACM Workshop on Artificial Intelligence and Security*, 2017a.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy*, 2017b.

Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, and Aleksander Madry. On evaluating adversarial robustness. *arXiv preprint arXiv:1902.06705*, 2019.

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 *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pp. 15–26. ACM, 2017.

Minhao Cheng, Thong Le, Pin-Yu Chen, Jinfeng Yi, Huan Zhang, and Cho-Jui Hsieh. Query-efficient hard-label black-box attack: An optimization-based approach. In *International Conference on Learning Representations (ICLR)*, 2019.

Moustapha Cisse, Piotr Bojanowski, Edouard Grave, Yann Dauphin, and Nicolas Usunier. Parseval networks: Improving robustness to adversarial examples. In *International Conference on Machine Learning (ICML)*, 2017.

Jeremy M Cohen, Elan Rosenfeld, and J Zico Kolter. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning (ICML)*, 2019.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2019.

Guneet S Dhillon, Kamyar Azizzadenesheli, Zachary C Lipton, Jeremy Bernstein, Jean Kossaifi, Aran Khanna, and Anima Anandkumar. Stochastic activation pruning for robust adversarial defense. In *International Conference on Learning Representations (ICLR)*, 2018.

Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Yinpeng Dong, Hang Su, Baoyuan Wu, Zhifeng Li, Wei Liu, Tong Zhang, and Jun Zhu. Efficient decision-based black-box adversarial attacks on face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Gintare Karolina Dziugaite, Zoubin Ghahramani, and Daniel M Roy. A study of the effect of jpg compression on adversarial images. *arXiv preprint arXiv:1608.00853*, 2016.

Logan Engstrom, Brandon Tran, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. Exploring the landscape of spatial robustness. In *International Conference on Machine Learning (ICML)*, 2019.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2015.

Alex Graves, Abdel Rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6645–6649, 2013.

Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens van der Maaten. Countering adversarial images using input transformations. In *International Conference on Learning Representations (ICLR)*, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Warren He, Bo Li, and Dawn Song. Decision boundary analysis of adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2018.

Matthias Hein and Maksym Andriushchenko. Formal guarantees on the robustness of a classifier against adversarial manipulation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks with limited queries and information. In *International Conference on Machine Learning (ICML)*, 2018.

Harini Kannan, Alexey Kurakin, and Ian Goodfellow. Adversarial logit pairing. *arXiv preprint arXiv:1803.06373*, 2018.

Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2012.

Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *International Conference on Learning Representations (ICLR) Workshops*, 2017.

Alexey Kurakin, Ian Goodfellow, Samy Bengio, Yinpeng Dong, Fangzhou Liao, Ming Liang, Tianyu Pang, Jun Zhu, Xiaolin Hu, Cihang Xie, et al. Adversarial attacks and defences competition. *arXiv preprint arXiv:1804.00097*, 2018.

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 *International Conference on Machine Learning (ICML)*, 2019.

Fangzhou Liao, Ming Liang, Yinpeng Dong, Tianyu Pang, Xiaolin Hu, and Jun Zhu. Defense against adversarial attacks using high-level representation guided denoiser. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Xiang Ling, Shouling Ji, Jiaxu Zou, Jiannan Wang, Chunming Wu, Bo Li, and Ting Wang. Deepsec: A uniform platform for security analysis of deep learning model. In *IEEE Symposium on Security and Privacy*, 2019.

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)*, 2018.

Xuanqing Liu, Yao Li, Chongruo Wu, and Cho-Jui Hsieh. Adv-bnn: Improved adversarial defense through robust bayesian neural network. In *International Conference on Learning Representations (ICLR)*, 2019.

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 (ICLR)*, 2018.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Maria-Irina Nicolae, Mathieu Sinn, Minh Ngoc Tran, Ambrish Rawat, Martin Wistuba, Valentina Zantedeschi, Nathalie Baracaldo, Bryant Chen, Heiko Ludwig, Ian M Molloy, et al. Adversarial robustness toolbox v0. 4.0. *arXiv preprint arXiv:1807.01069*, 2018.

Tianyu Pang, Kun Xu, Chao Du, Ning Chen, and Jun Zhu. Improving adversarial robustness via promoting ensemble diversity. In *International Conference on Machine Learning (ICML)*, 2019.

Nicolas Papernot, Fartash Faghri, Nicholas Carlini, Ian Goodfellow, Reuben Feinman, Alexey Kurakin, Cihang Xie, Yash Sharma, Tom Brown, Aurko Roy, et al. Technical report on the cleverhans v2. 1.0 adversarial examples library. *arXiv preprint arXiv:1610.00768*, 2016a.

Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against deep learning systems using adversarial examples. *arXiv preprint arXiv:1602.02697*, 2016b.

Nicolas Papernot, Patrick 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*, 2016c.

Aditi Raghunathan, Jacob Steinhardt, and Percy Liang. Certified defenses against adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2018a.

Aditi Raghunathan, Jacob Steinhardt, and Percy S Liang. Semidefinite relaxations for certifying robustness to adversarial examples. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018b.

Jonas Rauber, Wieland Brendel, and Matthias Bethge. Foolbox v0. 8.0: A python toolbox to benchmark the robustness of machine learning models. *arXiv preprint arXiv:1707.04131*, 2017.

Andrew Slavin Ross and Finale Doshi-Velez. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. In *Thirty-Second AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

Pouya Samangouei, Maya Kabkab, and Rama Chellappa. Defense-gan: Protecting classifiers against adversarial attacks using generative models. In *International Conference on Learning Representations (ICLR)*, 2018.

Ali Shafahi, Mahyar Najibi, Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. Adversarial training for free! In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Aman Sinha, Hongseok Namkoong, and John Duchi. Certifying some distributional robustness with principled adversarial training. In *International Conference on Learning Representations (ICLR)*, 2018.

Yang Song, Taesup Kim, Sebastian Nowozin, Stefano Ermon, and Nate Kushman. Pixeldefend: Leveraging generative models to understand and defend against adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2018a.

Yang Song, Rui Shu, Nate Kushman, and Stefano Ermon. Constructing unrestricted adversarial examples with generative models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018b.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations (ICLR)*, 2014.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. In *International Conference on Learning Representations (ICLR)*, 2018.

Jonathan Uesato, Brendan O'Donoghue, Aaron van den Oord, and Pushmeet Kohli. Adversarial risk and the dangers of evaluating against weak attacks. In *International Conference on Machine Learning (ICML)*, 2018.

Eric Wong and Zico Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *International Conference on Machine Learning (ICML)*, 2018.

Eric Wong, Frank Schmidt, Jan Hendrik Metzen, and J Zico Kolter. Scaling provable adversarial defenses. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Chaowei Xiao, Jun-Yan Zhu, Bo Li, Warren He, Mingyan Liu, and Dawn Song. Spatially transformed adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2018.

Kai Y Xiao, Vincent Tjeng, Nur Muhammad Shafiullah, and Aleksander Madry. Training for faster adversarial robustness verification via inducing relu stability. In *International Conference on Learning Representations (ICLR)*, 2019.

Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan Yuille. Mitigating adversarial effects through randomization. In *International Conference on Learning Representations (ICLR)*, 2018.

Cihang Xie, Yuxin Wu, Laurens van der Maaten, Alan L Yuille, and Kaiming He. Feature denoising for improving adversarial robustness. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019a.

Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019b.

Weilin Xu, David Evans, and Yanjun Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. In *Proceedings of the Network and Distributed System Security Symposium (NDSS)*, 2018.

Ziang Yan, Yiwen Guo, and Changshui Zhang. Deep defense: Training dnns with improved adversarial robustness. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P Xing, Laurent El Ghaoui, and Michael I Jordan. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning (ICML)*, 2019a.

Huan Zhang, Hongge Chen, Zhao Song, Duane Boning, Inderjit S Dhillon, and Cho-Jui Hsieh. The limitations of adversarial training and the blind-spot attack. In *International Conference on Learning Representations (ICLR)*, 2019b.

Y. Zhang and P. Liang. Defending against whitebox adversarial attacks via randomized discretization. In *Artificial Intelligence and Statistics (AISTATS)*, 2019.

A Adversarial Robustness Platforms

There are several public platforms for adversarial machine learning, such as CleverHans (Papernot et al., 2016a), Foolbox (Rauber et al., 2017), ART (Nicolae et al., 2018), etc. However, we observe that these platforms do not totally support our comprehensive evaluations in this paper. First, some attacks evaluated in this paper are not included in these platforms. There are less than 10 out of the 15 attacks adopted in this paper that are already implemented in each platform. And most of the available methods are white-box methods. Second, although these platforms incorporate a few defenses, they do not use the pre-trained models. But we use the original source codes and pre-trained models to perform unbiased evaluations. Third, the evaluation metrics defined by the two robustness curves in this paper are not provided in the existing platforms. Therefore, we develop a new adversarial robustness platform to satisfy our requirements.

Another similar work to ours is DeepSec (Ling et al., 2019), which also provides a uniform platform for adversarial robustness evaluation of DL models. However, as argued in Carlini (2019), DeepSec has several flaws, including 1) it evaluates the defenses by using the adversarial examples generated against undefended models; 2) it has some incorrect implementations; 3) it evaluates the robustness of the defenses as an average, etc. We try our best to avoid these issues in this paper. Our work differs from DeepSec in three main aspects: 1) we consider complete threat models and use various attack methods in different settings; 2) we use the original source codes and pre-trained models provided by the authors to prevent implementation errors; 3) we adopt two complementary robustness curves as the fair-minded evaluation metrics to present the results. We think that our evaluations can truly reflect the behavior of the attack and defense methods, and provide us with a detailed understanding of these methods.

B Evaluation Details

In this section, we provide additional evaluation details. Table 3 shows the network architecture of each defense model. Below we show the details of the attack methods as well as their parameters in our experiments. For clarity, we only introduce the untargeted attacks.

FGSM (Goodfellow et al., 2015) generates an untargeted adversarial example under the ℓ_∞ norm as

$$\boldsymbol{x}^{adv} = \boldsymbol{x} + \epsilon \cdot \mathrm{sign}(\nabla_{\boldsymbol{x}} \mathcal{J}(\boldsymbol{x}, y)), \tag{5}$$

where $\mathcal{J}$ is the cross-entropy loss. It can be extended to an ℓ_2 attack as

$$\boldsymbol{x}^{adv} = \boldsymbol{x} + \epsilon \cdot \frac{\nabla_{\boldsymbol{x}} \mathcal{J}(\boldsymbol{x}, y)}{\|\nabla_{\boldsymbol{x}} \mathcal{J}(\boldsymbol{x}, y)\|_2}. \tag{6}$$

To get the accuracy (attack success rate) vs. perturbation budget curves, we perform a line search followed by a binary search on ϵ to find its minimum value.

BIM (Kurakin et al., 2017) extends FGSM by iteratively taking multiple small gradient updates as

$$\boldsymbol{x}_{t+1}^{adv} = \mathrm{clip}_{\boldsymbol{x},\epsilon}\big(\boldsymbol{x}_t^{adv} + \alpha \cdot \mathrm{sign}(\nabla_{\boldsymbol{x}} \mathcal{J}(\boldsymbol{x}_t^{adv}, y))\big), \tag{7}$$

where $\mathrm{clip}_{\boldsymbol{x},\epsilon}$ projects the adversarial example to satisfy the ℓ_∞ constrain and α is the step size. It can also be extended to an ℓ_2 attack similar to FGSM. For most experiments, we set $\alpha = 0.15 \cdot \epsilon$. To get the accuracy (attack success rate) vs. perturbation budget curves, we also perform a binary search

Table 3: We show the network architecture of each defense model. Defenses based on input transformations use the baseline natural models as the backbone classifiers. DeepDefense uses a very simple 5-layer CNN. FD modifies a ResNet-152 architecture with the proposed denoising layers. ADP ensembles the predictions of 3 ResNet-110 models. Convex uses a ResNet model with architecture provided in Wong et al. (2018).

CIFAR-10 (Krizhevsky & Hinton, 2009)		ImageNet (Russakovsky et al., 2015)	
Defense Model	Architecture	Defense Model	Architecture
Res-56 (He et al., 2016)	ResNet-56	Inc-v3 (Szegedy et al., 2016)	Inception v3
PGD-AT (Madry et al., 2018)	Wide ResNet-34-10	Ens-AT (Tramèr et al., 2018)	Inception v3
DeepDefense (Yan et al., 2018)	5-layer CNN	ALP (Kannan et al., 2018)	ResNet-50
TRADES (Zhang et al., 2019a)	Wide ResNet-34-10	FD (Xie et al., 2019a)	ResNet-152 with denoising layers
Convex (Wong et al., 2018)	ResNet	JPEG (Dziugaite et al., 2016)	Inception v3
JPEG (Dziugaite et al., 2016)	ResNet-56	Bit-Red (Xu et al., 2018)	Inception v3
RSE (Liu et al., 2018)	VGG	R&P (Xie et al., 2018)	Inception v3
ADP (Pang et al., 2019)	ResNet-110 ×3	RandMix (Zhang & Liang, 2019)	Inception v3

on ϵ. For each ϵ during the binary search, we set the number of iterations as 20 in white-box attacks and 10 in transfer-based attacks.

MIM integrates a momentum term into BIM as

$$\boldsymbol{g}_{t+1} = \mu \cdot \boldsymbol{g}_t + \frac{\nabla_{\boldsymbol{x}} \mathcal{J}(\boldsymbol{x}_t^{adv}, y)}{\|\nabla_{\boldsymbol{x}} \mathcal{J}(\boldsymbol{x}_t^{adv}, y)\|_1} \quad \text{and} \quad \boldsymbol{x}_{t+1}^{adv} = \text{clip}_{\boldsymbol{x},\epsilon}(\boldsymbol{x}_t^{adv} + \alpha \cdot \text{sign}(\boldsymbol{g}_{t+1})). \tag{8}$$

MIM can similarly be extended to the ℓ_2 case. We set the step size α and the number of iterations identical to those in BIM. We set the decay factor as $\mu = 1.0$.

DeepFool (Moosavi-Dezfooli et al., 2016) is also an iterative attack method, which generates an adversarial example on the decision boundary of a classifier with the minimum perturbation. We set the maximum number of iterations as 100 in DeepFool, and it will early stop when the solution at an intermediate iteration is already adversarial.

C&W (Carlini & Wagner, 2017b) is a powerful optimization-based attack method, which generates an ℓ_2 adversarial example by solving

$$\boldsymbol{x}^{adv} = \arg\min_{\boldsymbol{x}'} \big\{ \|\boldsymbol{x}' - \boldsymbol{x}\|_2^2 + c \cdot \max(Z(\boldsymbol{x}')_y - \max_{i \neq y} Z(\boldsymbol{x}')_i, 0) \big\}, \tag{9}$$

where $Z(\boldsymbol{x}')$ is the logit output of the classifier and c is a constant. This optimization problem is solved by an Adam (Kingma & Ba, 2015) optimizer. c is found by binary search. To get the accuracy (attack success rate) vs. perturbation budget curves, we optimize Eq. (9) for 100 iterations. To get the accuracy (attack success rate) vs. attack strength curves, we optimize Eq. (9) for 10, 20, 30, 40 iterations on CIFAR-10, and 10, 20 iterations on ImageNet to show the results.

DIM (Xie et al., 2019b) randomly resizes and pads the input, and uses the transformed input for gradient calculation. It also adopts the momentum technique. In our experiments, we set the common parameters the same as those of MIM. For its own parameters, we set the input $\boldsymbol{x} \in \mathbb{R}^{s \times s \times 3}$ is first resized to a $rnd \times rnd \times 3$ image, with $rnd \in [0.9 * s, s]$, and then padded to the original size.

ZOO (Chen et al., 2017) has been proposed to optimize Eq. (9) in the black-box manner through queries. It estimates the gradient at each coordinate as

$$\hat{g}_i = \frac{\mathcal{L}(\boldsymbol{x} + \sigma \boldsymbol{e}_i, y) - \mathcal{L}(\boldsymbol{x} - \sigma \boldsymbol{e}_i, y)}{2\sigma} \approx \frac{\partial \mathcal{L}(\boldsymbol{x}, y)}{\partial x_i}, \tag{10}$$

where $\mathcal{L}$ is the objective in Eq. (9), σ is a small constant, and $\boldsymbol{e}_i$ is the i-th unit basis vector. In our experiments, we perform one update with $\hat{g}_i$ at one randomly sampled coordinate. We set $\sigma = 10^{-4}$.

NES (Ilyas et al., 2018) and **SPSA** (Uesato et al., 2018) adopt the update rule in Eq. (7) for adversarial example generation. Although the true gradient is unavailable, NES and SPSA give the full gradient estimation as

$$\hat{\boldsymbol{g}} = \frac{1}{q} \sum_{i=1}^{q} \frac{\mathcal{J}(\boldsymbol{x} + \sigma \boldsymbol{u}_i, y) - \mathcal{J}(\boldsymbol{x} - \sigma \boldsymbol{u}_i, y)}{2\sigma} \cdot \boldsymbol{u}_i, \tag{11}$$

where we use $\mathcal{J}(\boldsymbol{x}, y) = Z(\boldsymbol{x})_y - \max_{i \neq y} Z(\boldsymbol{x})_i$ instead of the cross-entropy loss, $\{\boldsymbol{u}_i\}_{i=1}^q$ are the random vectors sampled from a Gaussian distribution in NES, and a Rademacher distribution in SPSA. We set $\sigma = 0.001$ and $q = 100$ in experiments.

$\mathcal{N}$**ATTACK** (Li et al., 2019) does not estimate the gradient but learns a Gaussian distribution centered around the input such that a sample drawn from it is likely an adversarial example. We set the sampling variance as 0.1, the learning rate as 0.02, the number of samples per iteration as 100 in $\mathcal{N}$ATTACK.

The decision-based black-box attacks—**Boundary** (Brendel et al., 2018a) and **Evolutionary** (Dong et al., 2019) rely on heuristic search on the decision boundary. They need a starting point, which is already adversarial, to initialize an attack. For untargeted attacks, we sample each pixel of the initial image from a uniform distribution. For targeted attacks, we specify the starting point as a sample that is classified by the model as the target class. We use the default hyperparameters of these two attacks given by their authors.

C Full Evaluation Results

We provide the full evaluation results in this section.

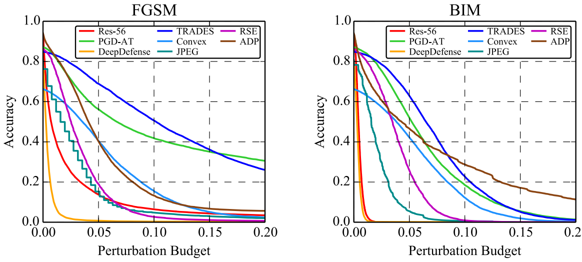

Figure 17: The *accuracy vs. perturbation budget* curves of the 8 models on CIFAR-10 against targeted white-box attacks under the ℓ_∞ norm.

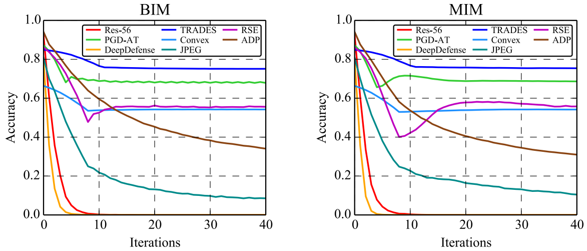

Figure 18: The *accuracy vs. attack strength* curves of the 8 models on CIFAR-10 against targeted white-box attacks under the ℓ_∞ norm.

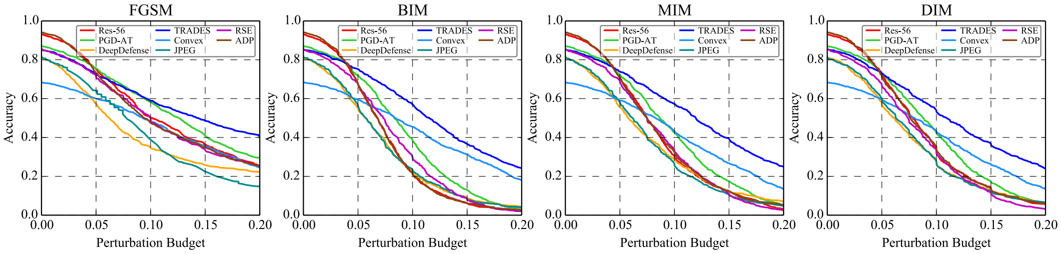

Figure 19: The *accuracy vs. perturbation budget* curves of the 8 models on CIFAR-10 against targeted transfer-based attacks under the ℓ_∞ norm.

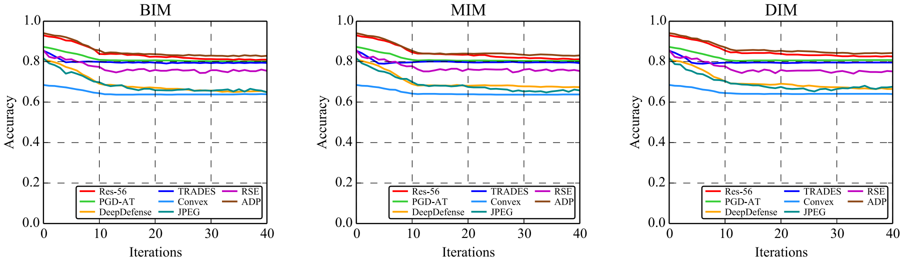

Figure 20: The *accuracy vs. attack strength* curves of the 8 models on CIFAR-10 against targeted transfer-based attacks under the ℓ_∞ norm.

C.1 Full Evaluation Results on CIFAR-10

Attacks under the ℓ_∞ norm: We have shown the accuracy curves of the defense models against untargeted attacks under the ℓ_∞ norm in Sec. 5.1. We next show the results of targeted attacks under the ℓ_∞ norm and the attack success rate curves. Fig. 17 and Fig. 18 show the accuracy curves of the defenses on CIFAR-10 against targeted white-box attacks under the ℓ_∞ norm. Fig. 19 and Fig. 20 show the accuracy curves of the defenses on CIFAR-10 against targeted transfer-based attacks under the ℓ_∞ norm. Fig. 21 and Fig. 22 show the accuracy curves of the defenses on CIFAR-10 against targeted score-based attacks under the ℓ_∞ norm. Fig. 23 to Fig. 28 show the *attack success rate vs. perturbation budget* and *attack success rate vs. attack strength* curves of white-box, transfer-based, and score-based attacks under the ℓ_∞ norm on the 8 models on CIFAR-10.

Attacks under the ℓ_2 norm: We show the accuracy curves of the defenses on CIFAR-10 against untargeted and targeted white-box attacks under the ℓ_2 norm in Fig. 29, Fig. 30, Fig. 31, and Fig. 32. We show the accuracy curves of the defenses on CIFAR-10 against untargeted and targeted transfer-based attacks under the ℓ_2 norm in Fig. 33, Fig. 34, Fig. 35, and Fig. 36. We show the accuracy curves of the defenses on CIFAR-10 against untargeted and targeted score-based attacks under the ℓ_2 norm

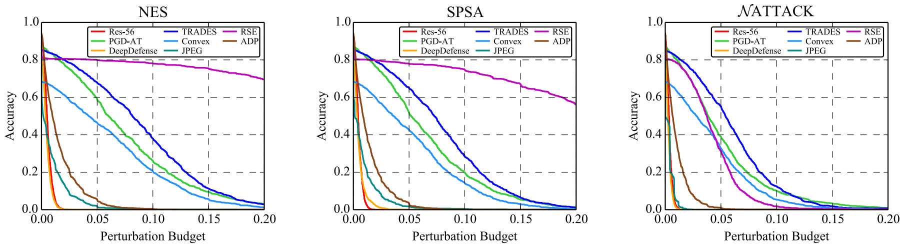

Figure 21: The *accuracy vs. perturbation budget* curves of the 8 models on CIFAR-10 against targeted score-based attacks under the ℓ_∞ norm.

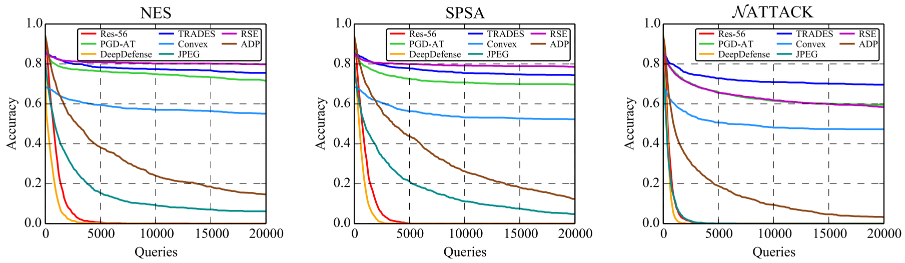

Figure 22: The *accuracy vs. attack strength* curves of the 8 models on CIFAR-10 against targeted score-based attacks under the ℓ_∞ norm.

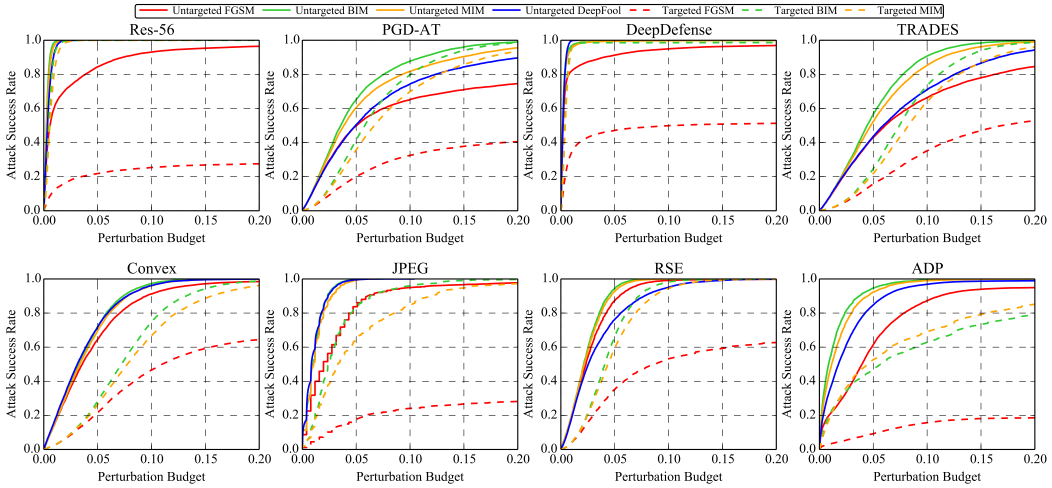

Figure 23: The *attack success rate vs. perturbation budget* curves of white-box attacks under the ℓ_∞ norm on the 8 models on CIFAR-10.

in Fig. 37, Fig. 38, Fig. 39, and Fig. 40. We show the accuracy curves of the defenses on CIFAR-10 against untargeted and targeted decision-based attacks under the ℓ_2 norm in Fig. 7, Fig. 8, Fig. 41, and Fig. 42. Fig. 43 to Fig. 50 show the *attack success rate vs. perturbation budget* and *attack success rate vs. attack strength* curves of white-box, transfer-based, score-based, and decision-based attacks under the ℓ_2 norm on the 8 models on CIFAR-10.

C.2 Full Evaluation Results on ImageNet

Attacks under the ℓ_∞ norm: Similar to CIFAR-10, we show the results of targeted attacks under the ℓ_∞ norm and the attacks success rate curves here. Fig. 51 and Fig. 52 show the accuracy curves of the defenses on ImageNet against targeted white-box attacks under the ℓ_∞ norm. Fig. 53 and Fig. 54 show the accuracy curves of the defenses on ImageNet against targeted transfer-based attacks under the ℓ_∞ norm. Fig. 55 and Fig. 56 show the accuracy curves of the defenses on ImageNet against targeted score-based attacks under the ℓ_∞ norm. Fig. 57 to Fig. 62 show the *attack success rate vs. perturbation budget* and *attack success rate vs. attack strength* curves of white-box, transfer-based, and score-based attacks under the ℓ_∞ norm on the 8 models on ImageNet.

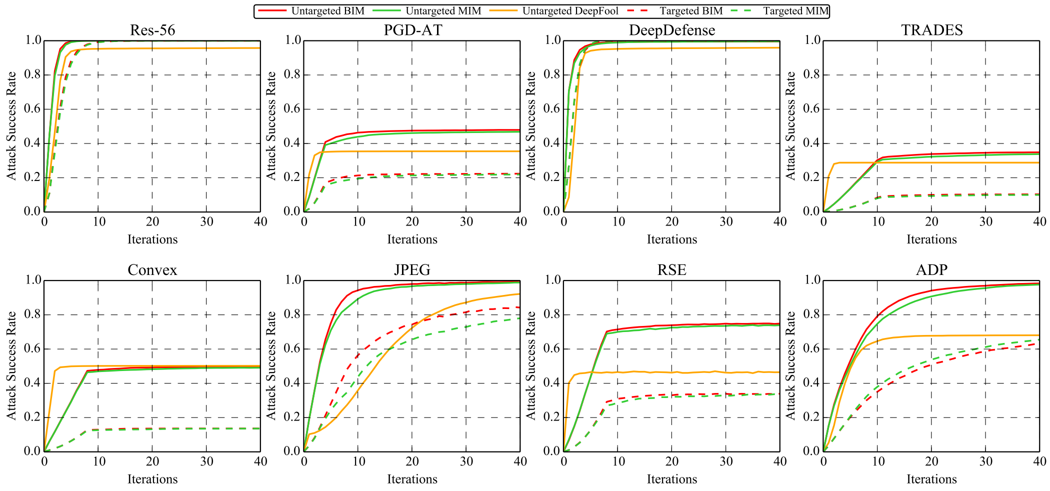

Figure 24: The *attack success rate vs. attack strength* curves of white-box attacks under the ℓ_∞ norm on the 8 models on CIFAR-10.

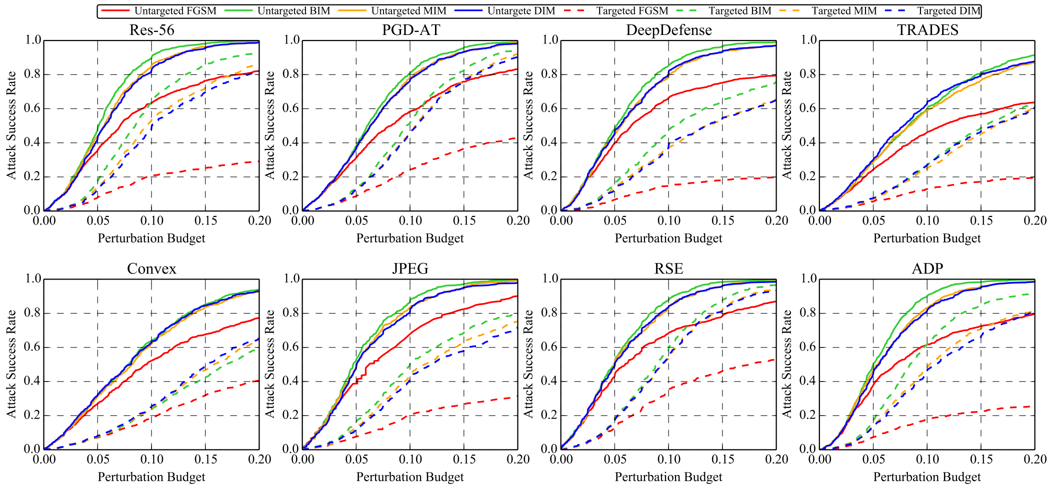

Figure 25: The *attack success rate vs. perturbation budget* curves of transfer-based attacks under the ℓ_∞ norm on the 8 models on CIFAR-10.

Attacks under the ℓ_2 norm: We show the accuracy curves of the defenses on ImageNet against untargeted and targeted white-box attacks under the ℓ_2 norm in Fig. 63, Fig. 64, Fig. 65, and Fig. 66. We show the accuracy curves of the defenses on ImageNet against untargeted and targeted transfer-based attacks under the ℓ_2 norm in Fig. 67, Fig. 68, Fig. 69, and Fig. 70. We show the accuracy curves of the defenses on ImageNet against untargeted and targeted score-based attacks under the ℓ_2 norm in Fig. 71, Fig. 72, Fig. 73, and Fig. 74. We show the accuracy curves of the defenses on ImageNet against untargeted and targeted decision-based attacks under the ℓ_2 norm in Fig. 15, Fig. 16, Fig. 75, and Fig. 76. Fig. 77 to Fig. 84 show the *attack success rate vs. perturbation budget* and *attack success rate vs. attack strength* curves of white-box, transfer-based, score-based, and decision-based attacks under the ℓ_2 norm on the 8 models on ImageNet.

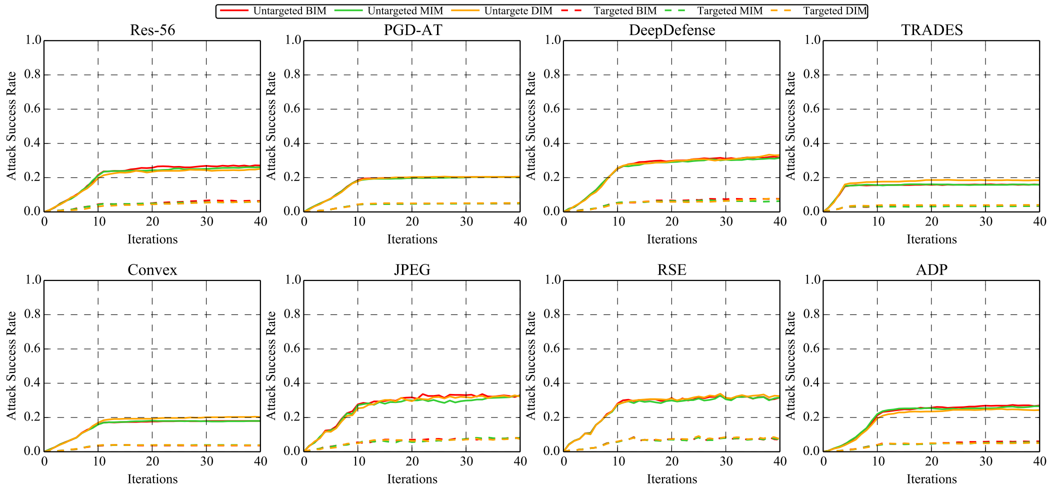

Figure 26: The *attack success rate vs. attack strength* curves of transfer-based attacks under the ℓ_∞ norm on the 8 models on CIFAR-10.

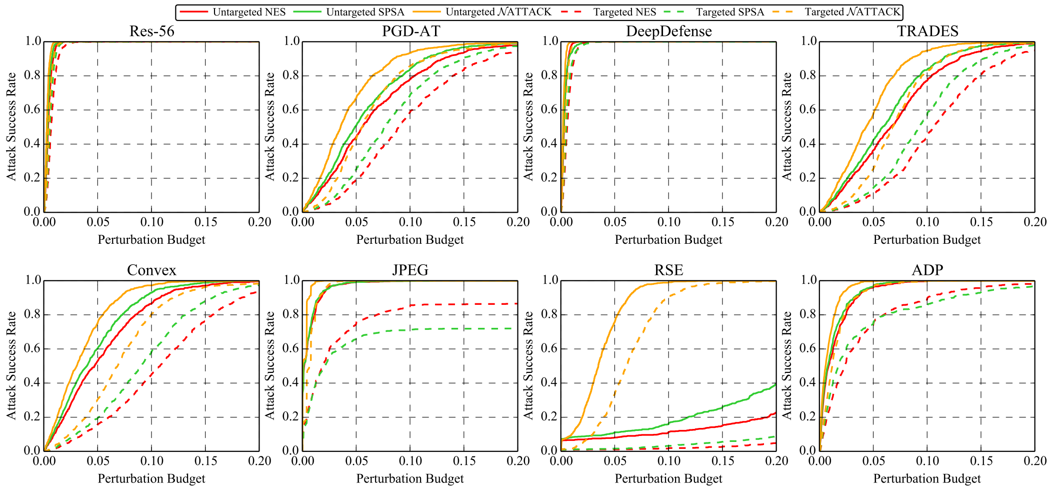

Figure 27: The *attack success rate vs. perturbation budget* curves of score-based attacks under the ℓ_∞ norm on the 8 models on CIFAR-10.

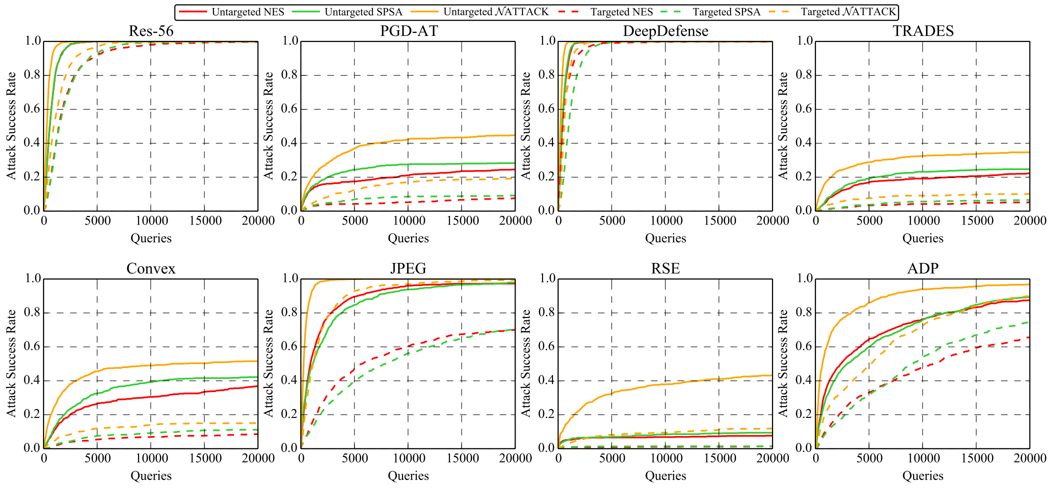

Figure 28: The *attack success rate vs. attack strength* curves of score-based attacks under the ℓ_∞ norm on the 8 models on CIFAR-10.

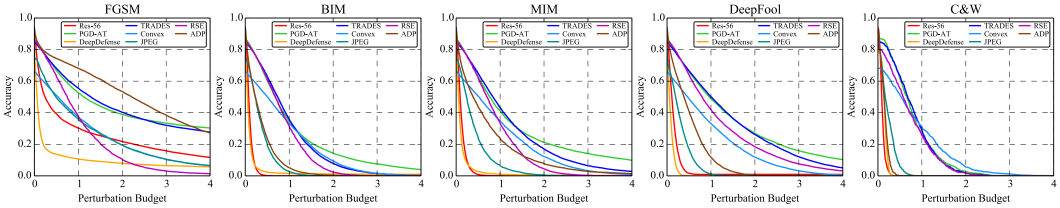

Figure 29: The *accuracy vs. perturbation budget* curves of the 8 models on CIFAR-10 against untargeted white-box attacks under the ℓ_2 norm.

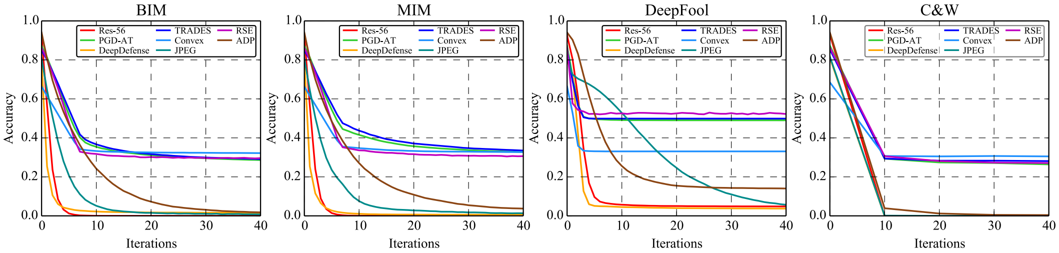

Figure 30: The *accuracy vs. attack strength* curves of the 8 models on CIFAR-10 against untargeted white-box attacks under the ℓ_2 norm.

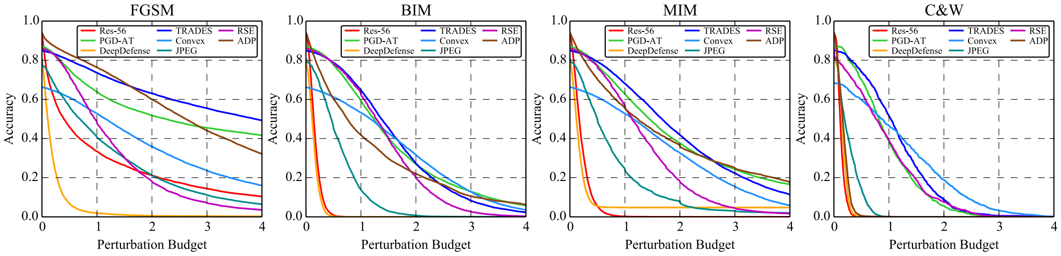

Figure 31: The *accuracy vs. perturbation budget* curves of the 8 models on CIFAR-10 against targeted white-box attacks under the ℓ_2 norm.

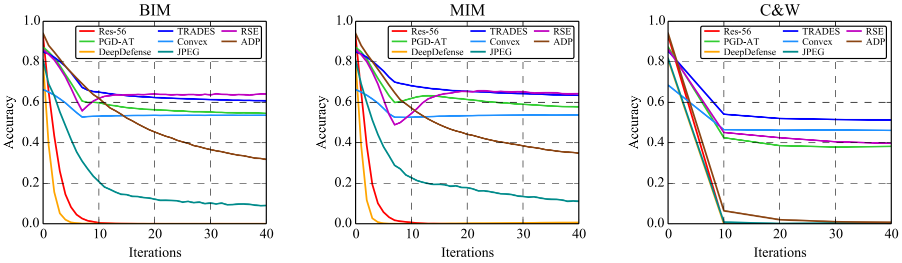

Figure 32: The *accuracy vs. attack strength* curves of the 8 models on CIFAR-10 against targeted white-box attacks under the ℓ_2 norm.

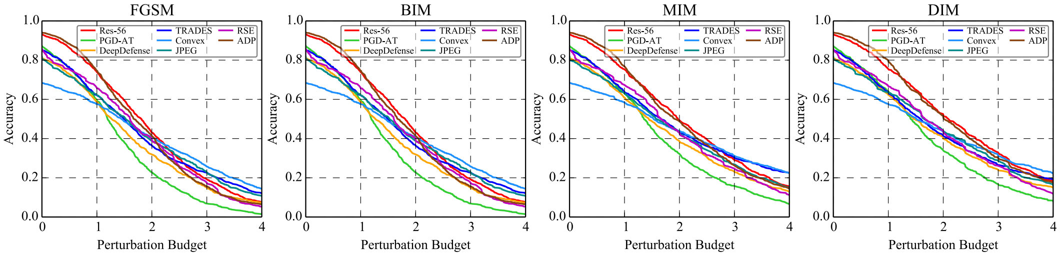

Figure 33: The *accuracy vs. perturbation budget* curves of the 8 models on CIFAR-10 against untargeted transfer-based attacks under the ℓ_2 norm.

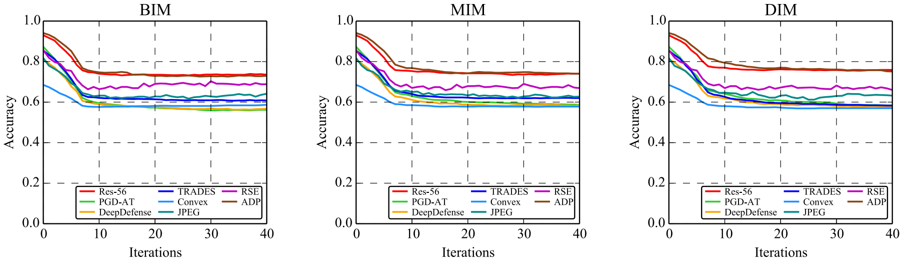

Figure 34: The *accuracy vs. attack strength* curves of the 8 models on CIFAR-10 against untargeted transfer-based attacks under the ℓ_2 norm.

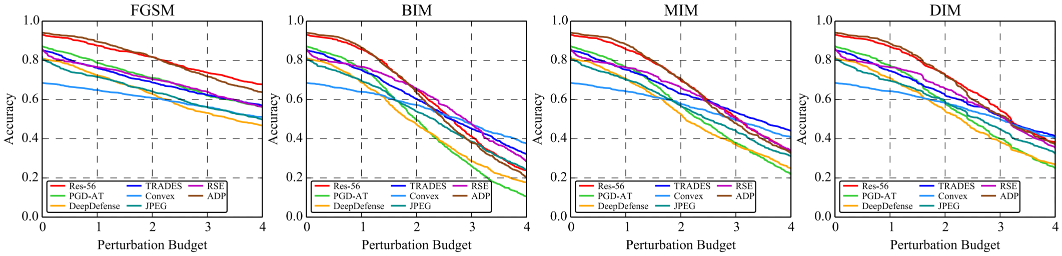

Figure 35: The *accuracy vs. perturbation budget* curves of the 8 models on CIFAR-10 against targeted transfer-based attacks under the ℓ_2 norm.

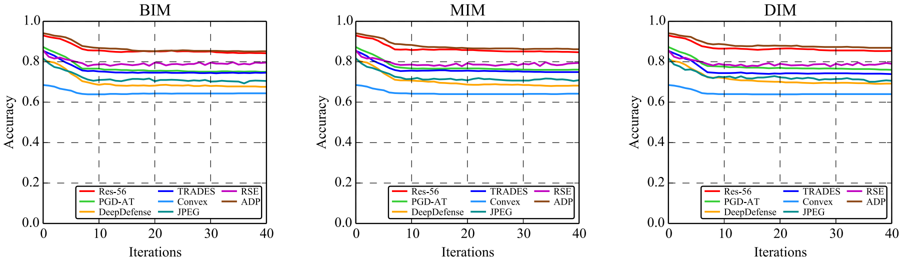

Figure 36: The *accuracy vs. attack strength* curves of the 8 models on CIFAR-10 against targeted transfer-based attacks under the ℓ_2 norm.

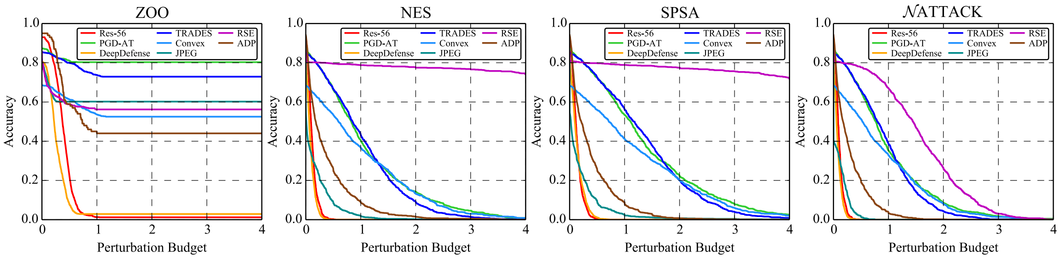

Figure 37: The *accuracy vs. perturbation budget* curves of the 8 models on CIFAR-10 against untargeted score-based attacks under the ℓ_2 norm.

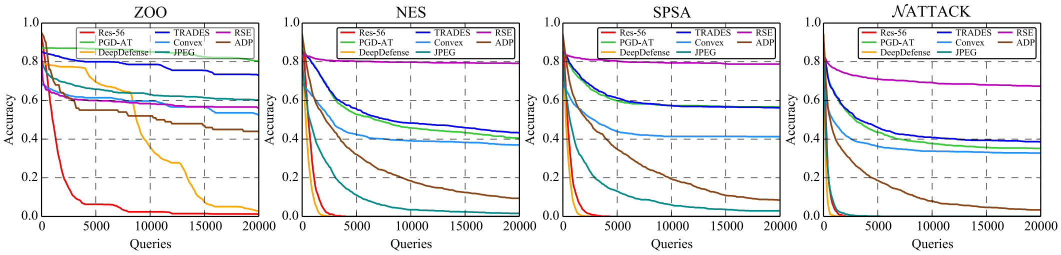

Figure 38: The *accuracy vs. attack strength* curves of the 8 models on CIFAR-10 against untargeted score-based attacks under the ℓ_2 norm.

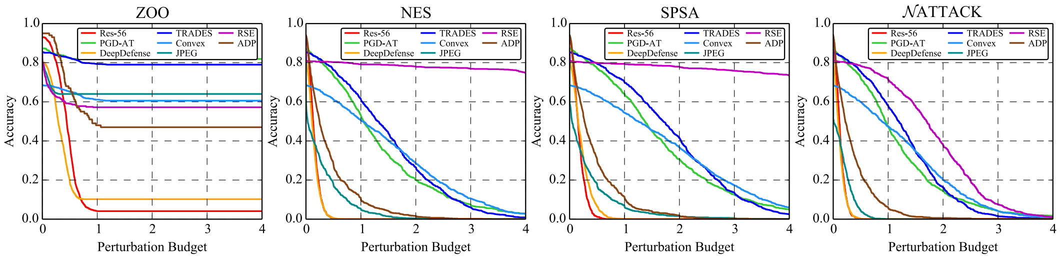

Figure 39: The *accuracy vs. perturbation budget* curves of the 8 models on CIFAR-10 against targeted score-based attacks under the ℓ_2 norm.

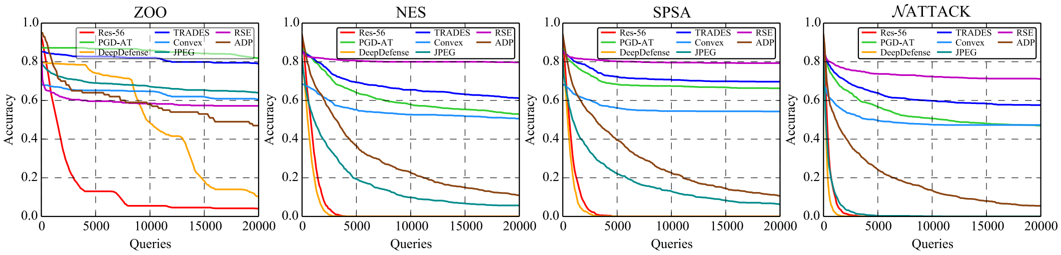

Figure 40: The *accuracy vs. attack strength* curves of the 8 models on CIFAR-10 against targeted score-based attacks under the ℓ_2 norm.

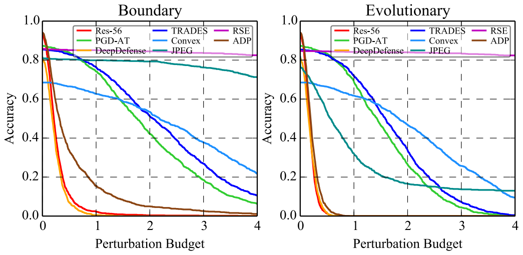

Figure 41: The *accuracy vs. perturbation budget* curves of the 8 models on CIFAR-10 against targeted decision-based attacks under the ℓ_2 norm.

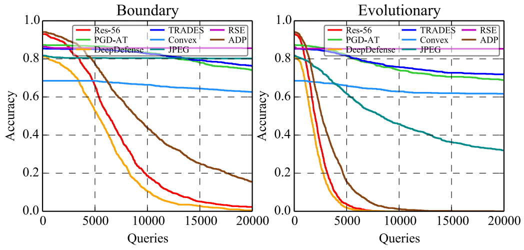

Figure 42: The *accuracy vs. attack strength* curves of the 8 models on CIFAR-10 against targeted decision-based attacks under the ℓ_2 norm.

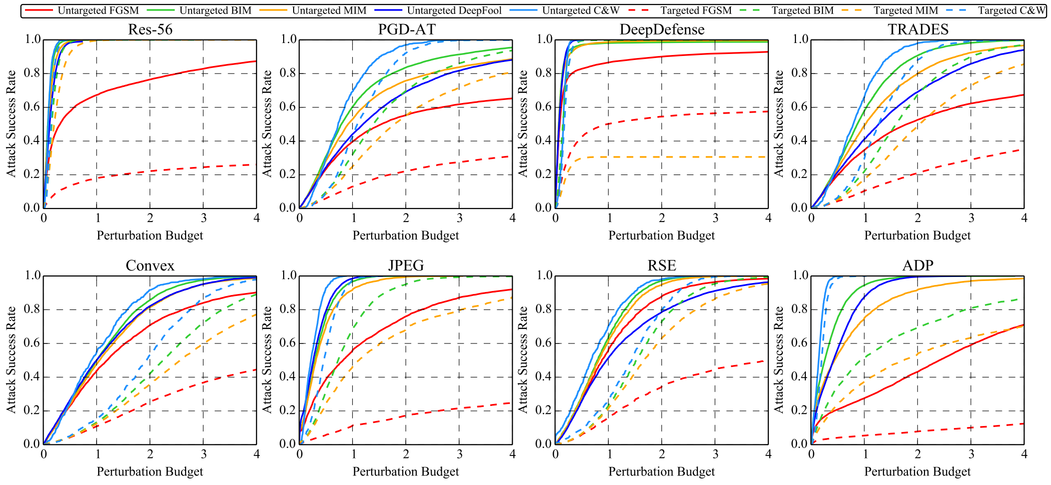

Figure 43: The *attack success rate vs. perturbation budget* curves of white-box attacks under the ℓ_2 norm on the 8 models on CIFAR-10.

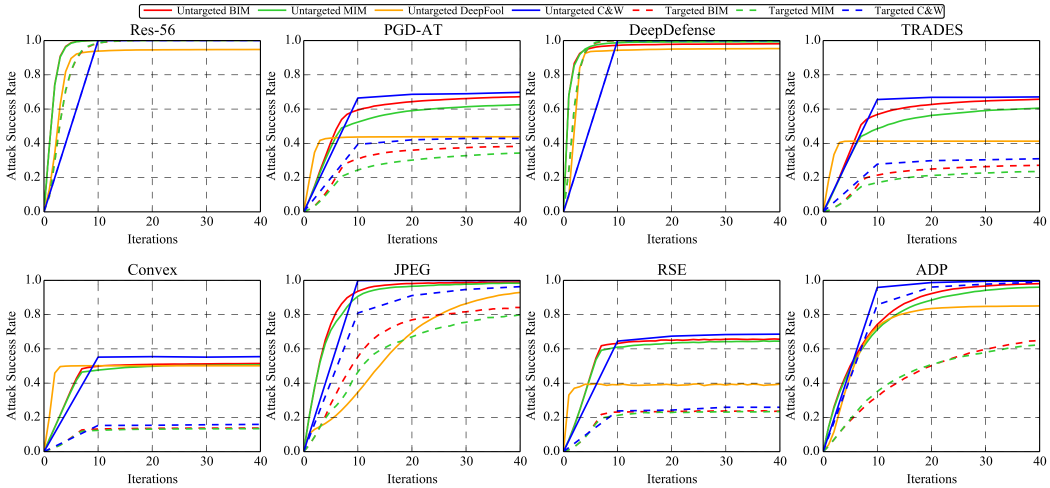

Figure 44: The *attack success rate vs. attack strength* curves of white-box attacks under the ℓ_2 norm on the 8 models on CIFAR-10.

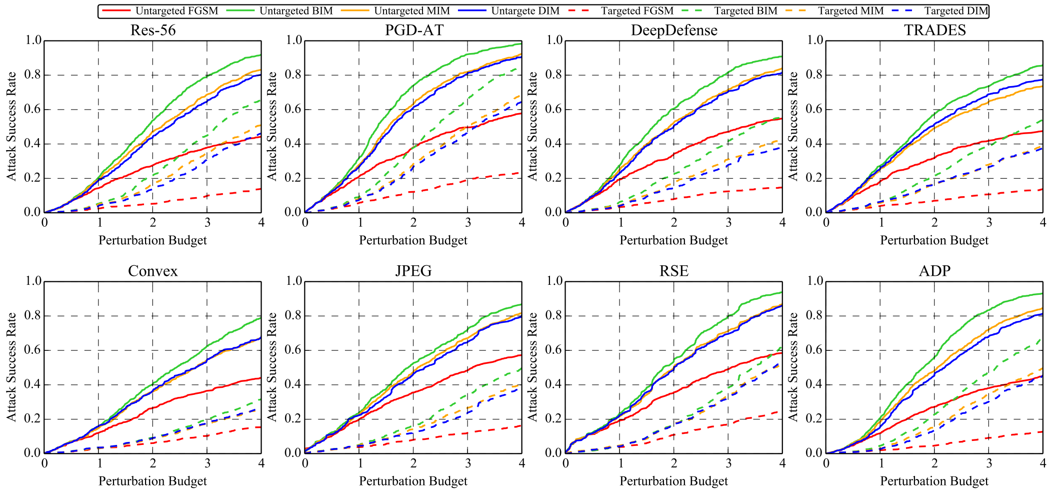

Figure 45: The *attack success rate vs. perturbation budget* curves of transfer-based attacks under the ℓ_2 norm on the 8 models on CIFAR-10.

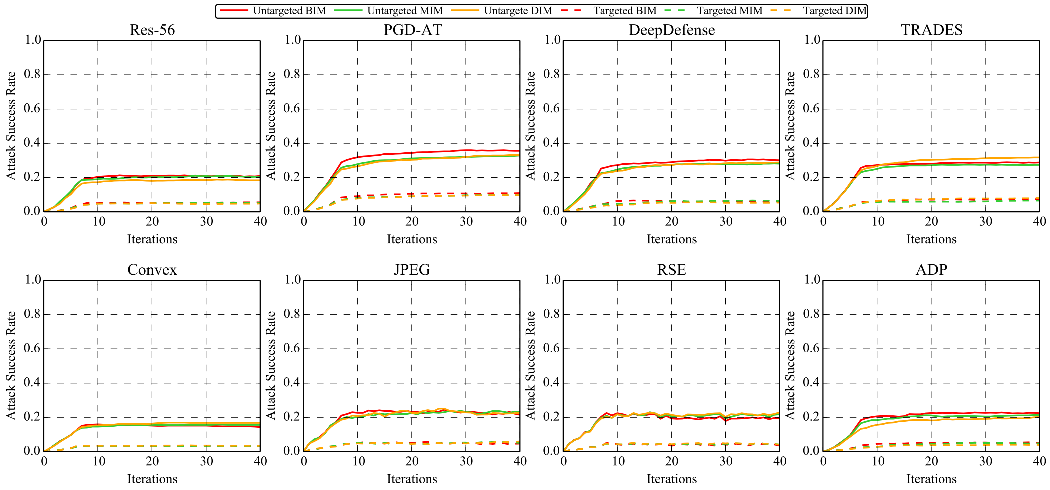

Figure 46: The *attack success rate vs. attack strength* curves of transfer-based attacks under the ℓ_2 norm on the 8 models on CIFAR-10.

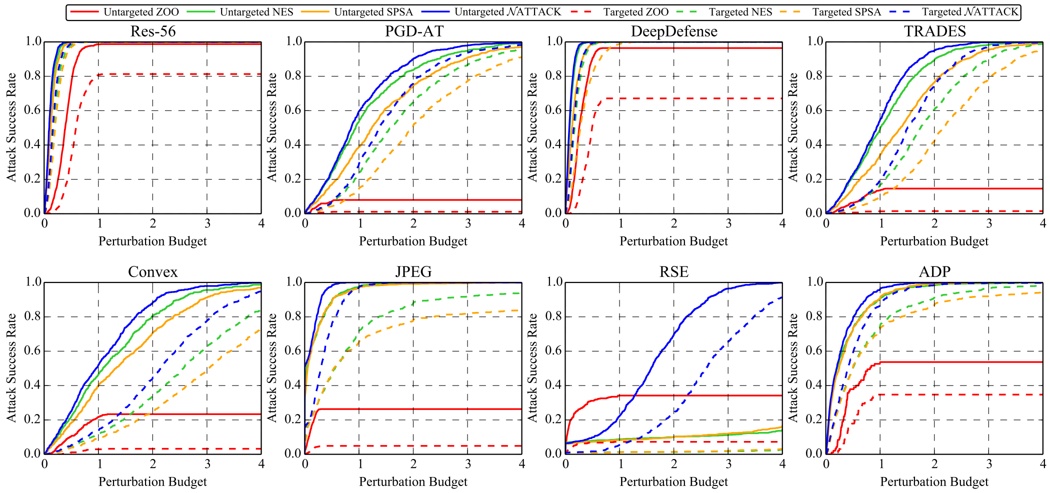

Figure 47: The *attack success rate vs. perturbation budget* curves of score-based attacks under the ℓ_2 norm on the 8 models on CIFAR-10.

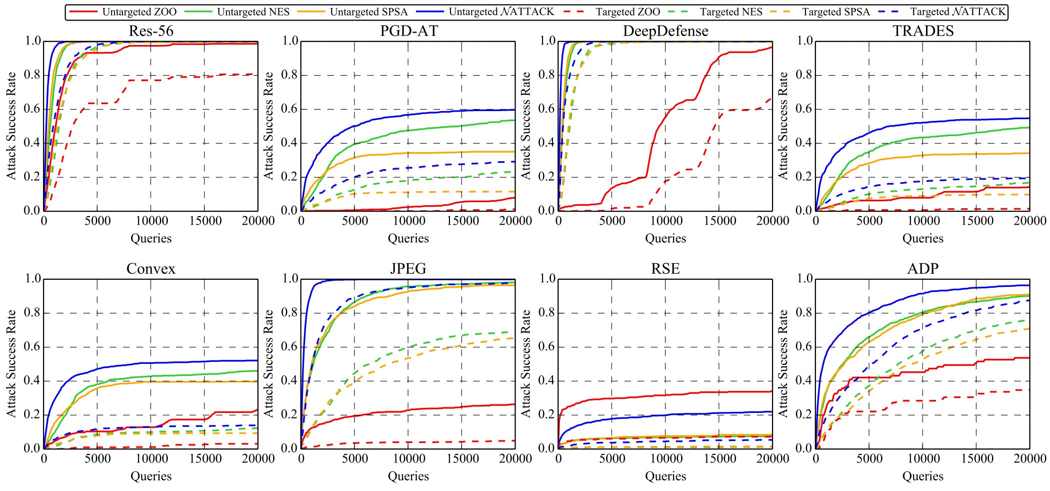

Figure 48: The *attack success rate vs. attack strength* curves of score-based attacks under the ℓ_2 norm on the 8 models on CIFAR-10.

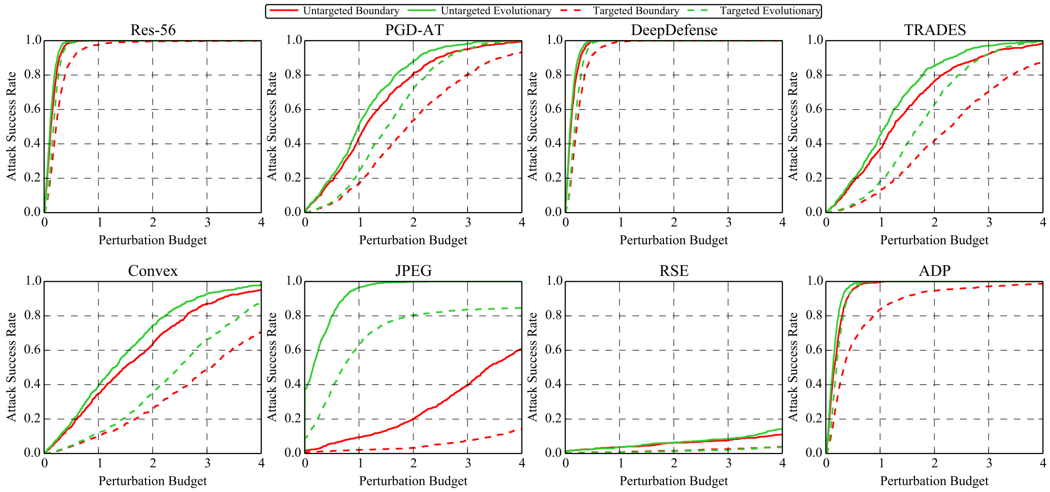

Figure 49: The *attack success rate vs. perturbation budget* curves of decision-based attacks under the ℓ_2 norm on the 8 models on CIFAR-10.

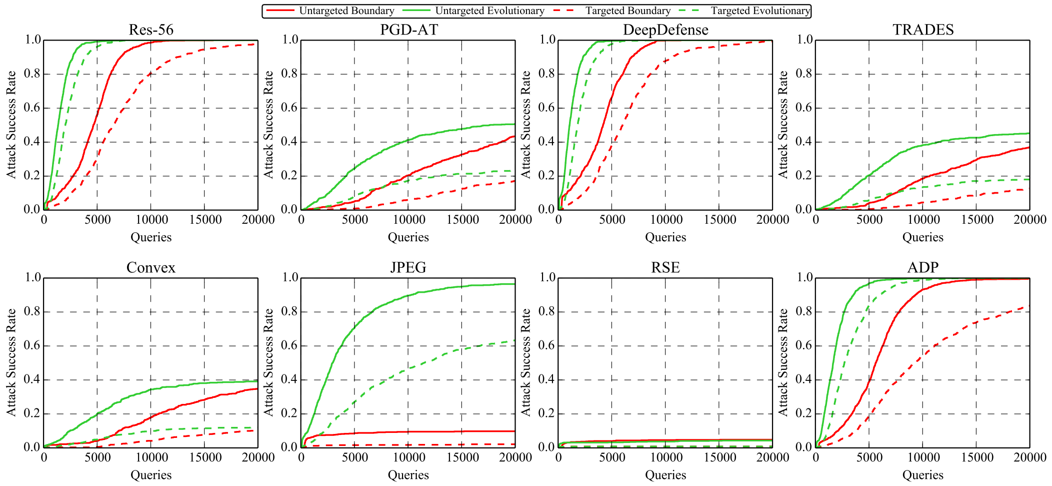

Figure 50: The *attack success rate vs. attack strength* curves of decision-based attacks under the ℓ_2 norm on the 8 models on CIFAR-10.

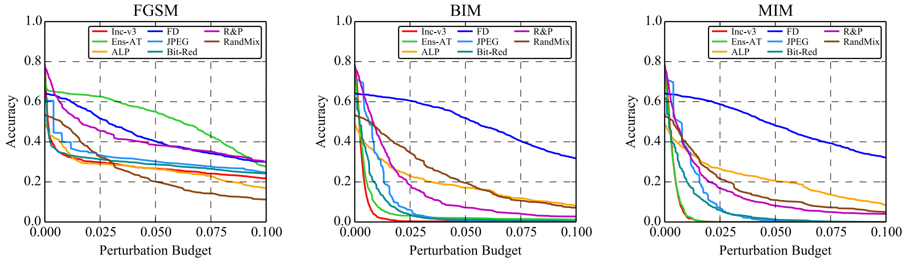

Figure 51: The *accuracy vs. perturbation budget* curves of the 8 models on ImageNet against targeted white-box attacks under the ℓ_∞ norm.

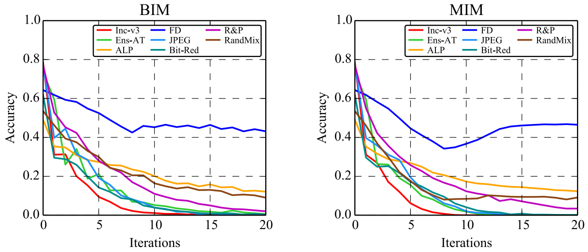

Figure 52: The *accuracy vs. attack strength* curves of the 8 models on ImageNet against targeted white-box attacks under the ℓ_∞ norm.

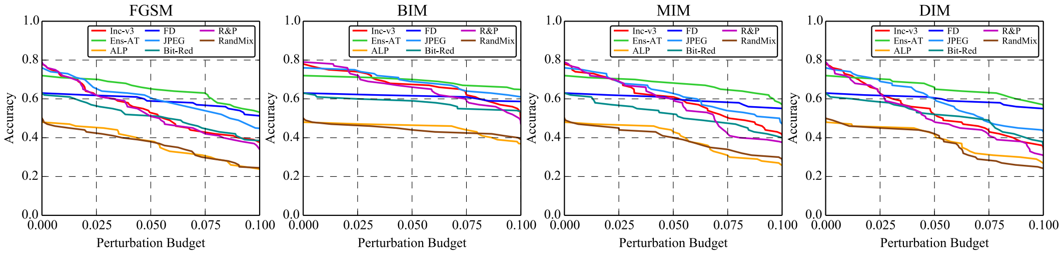

Figure 53: The *accuracy vs. perturbation budget* curves of the 8 models on ImageNet against targeted transfer-based attacks under the ℓ_∞ norm.

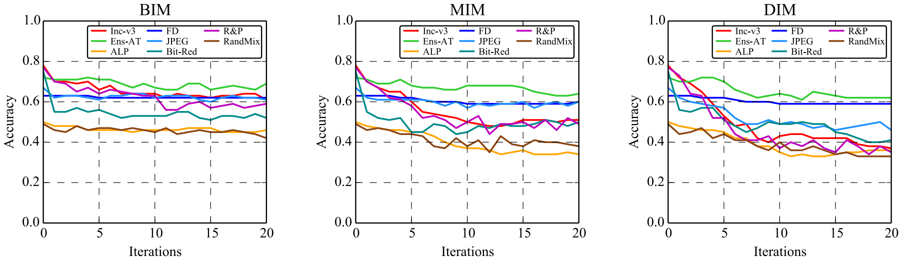

Figure 54: The *accuracy vs. attack strength* curves of the 8 models on ImageNet against targeted transfer-based attacks under the ℓ_∞ norm.

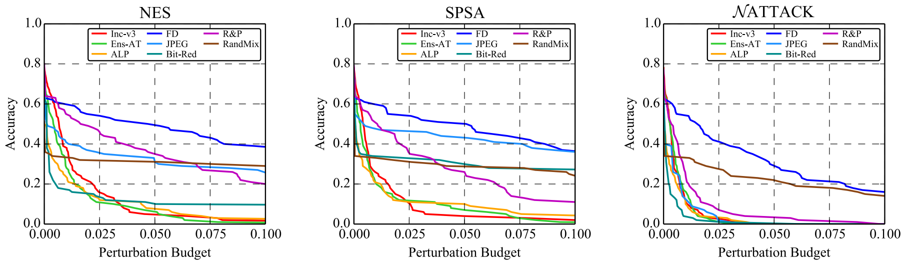

Figure 55: The *accuracy vs. perturbation budget* curves of the 8 models on ImageNet against targeted score-based attacks under the ℓ_∞ norm.

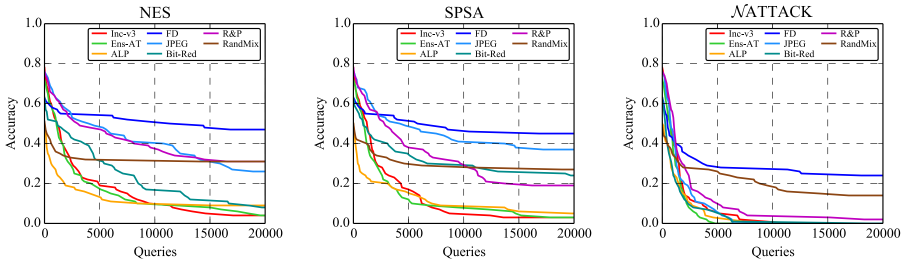

Figure 56: The *accuracy vs. attack strength* curves of the 8 models on ImageNet against targeted score-based attacks under the ℓ_∞ norm.

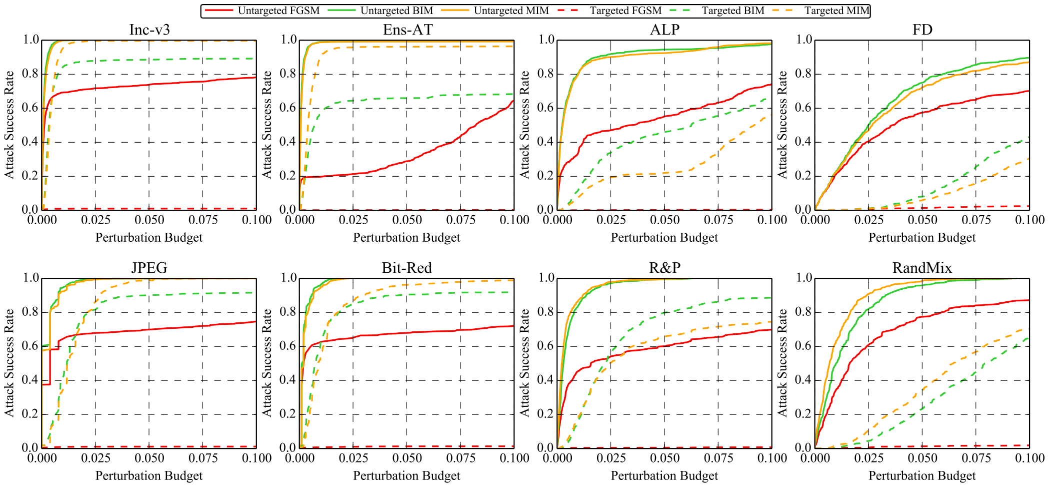

Figure 57: The *attack success rate vs. perturbation budget* curves of white-box attacks under the ℓ_∞ norm on the 8 models on ImageNet.

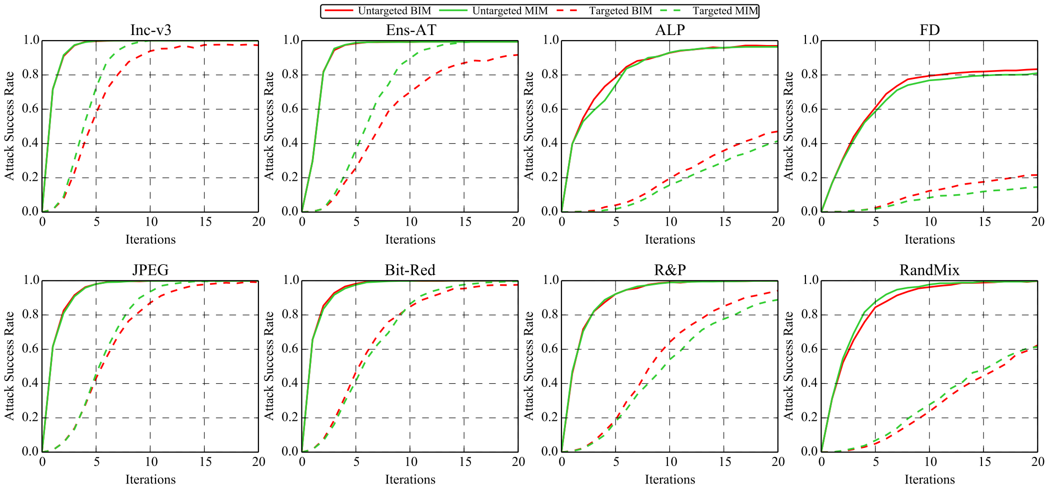

Figure 58: The *attack success rate vs. attack strength* curves of white-box attacks under the ℓ_∞ norm on the 8 models on ImageNet.

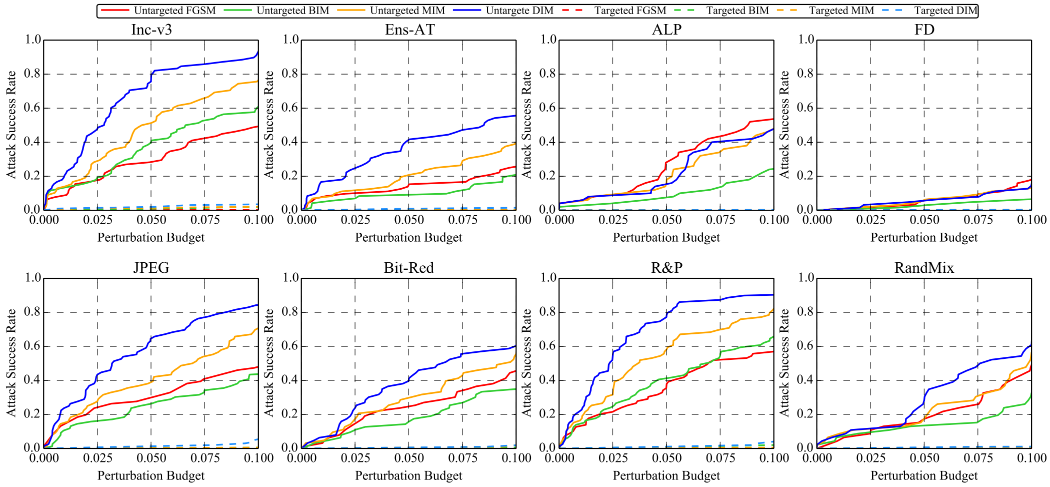

Figure 59: The *attack success rate vs. perturbation budget* curves of transfer-based attacks under the ℓ_∞ norm on the 8 models on ImageNet.

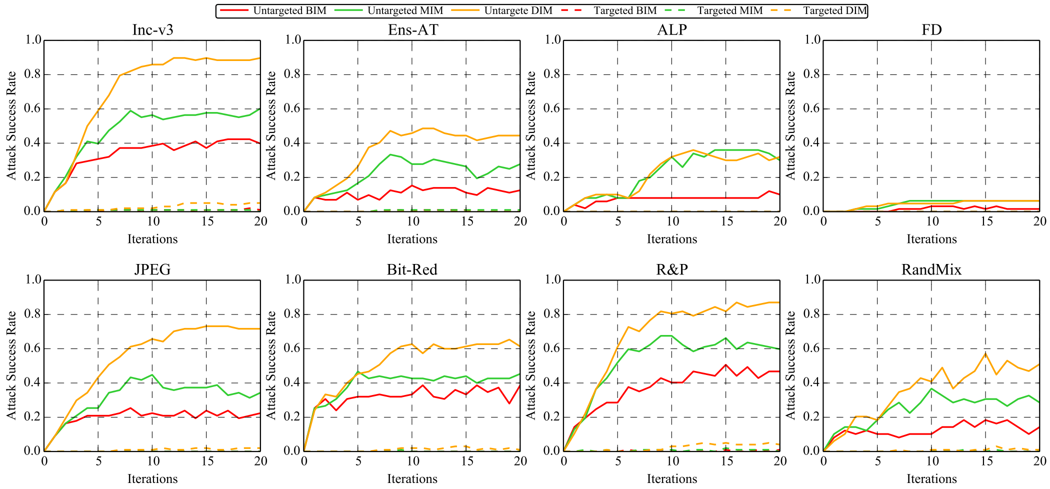

Figure 60: The *attack success rate vs. attack strength* curves of transfer-based attacks under the ℓ_∞ norm on the 8 models on ImageNet.

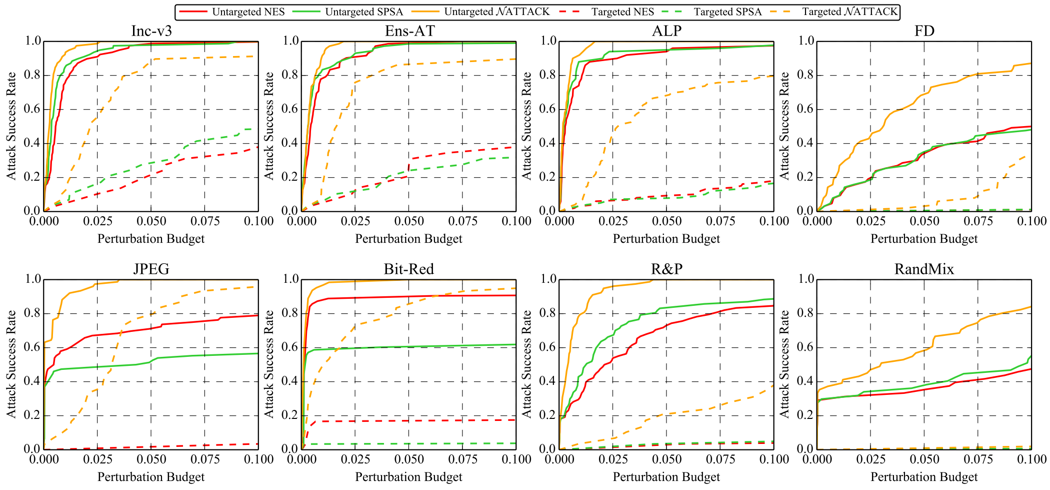

Figure 61: The *attack success rate vs. perturbation budget* curves of score-based attacks under the ℓ_∞ norm on the 8 models on ImageNet.

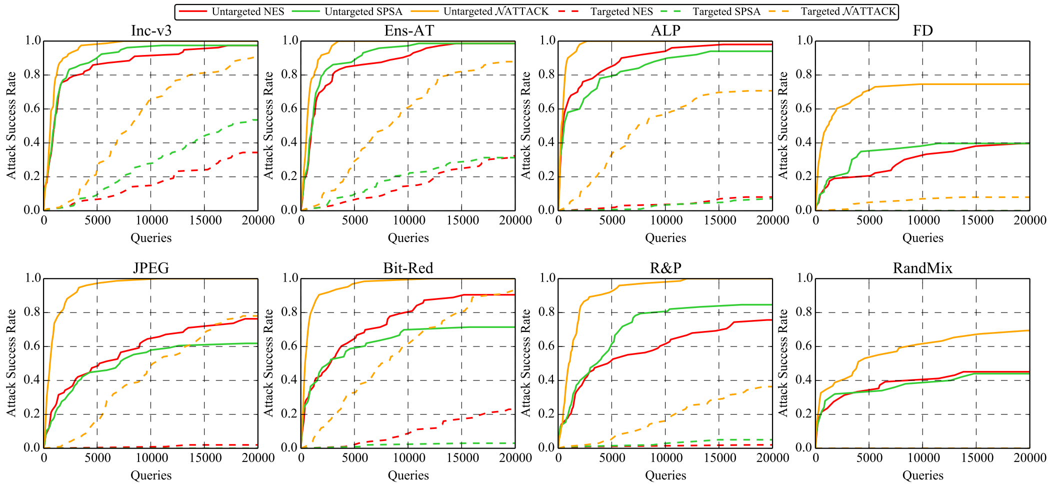

Figure 62: The *attack success rate vs. attack strength* curves of score-based attacks under the ℓ_∞ norm on the 8 models on ImageNet.

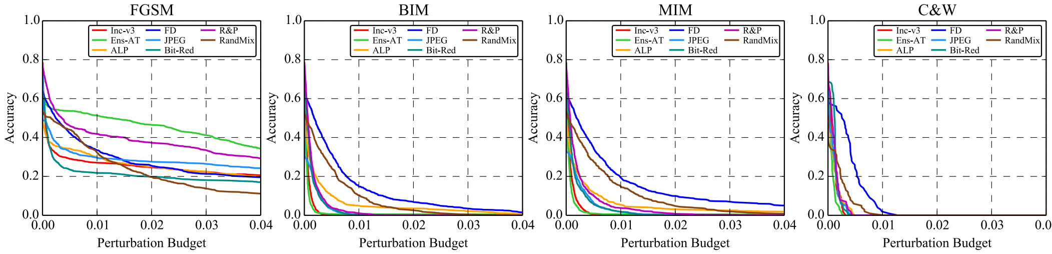

Figure 63: The *accuracy vs. perturbation budget* curves of the 8 models on ImageNet against untargeted white-box attacks under the ℓ_2 norm.

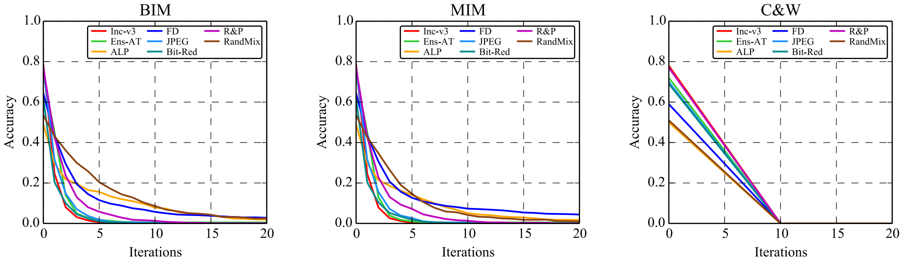

Figure 64: The *accuracy vs. attack strength* curves of the 8 models on ImageNet against untargeted white-box attacks under the ℓ_2 norm.

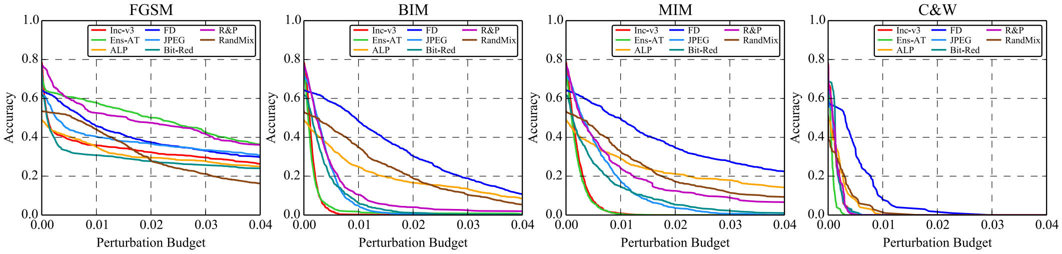

Figure 65: The *accuracy vs. perturbation budget* curves of the 8 models on ImageNet against targeted white-box attacks under the ℓ_2 norm.

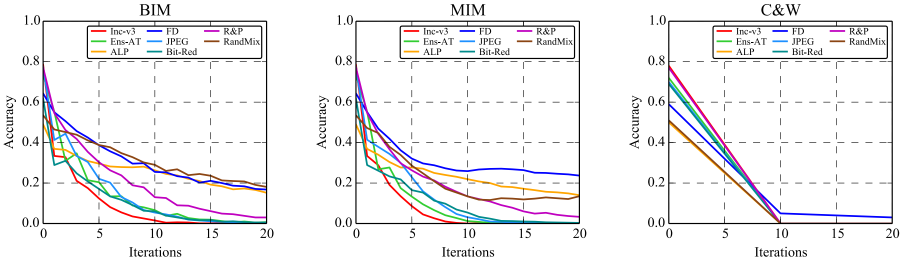

Figure 66: The *accuracy vs. attack strength* curves of the 8 models on ImageNet against targeted white-box attacks under the ℓ_2 norm.

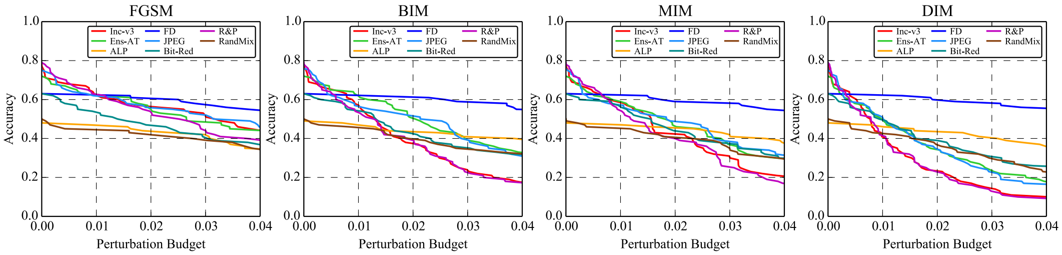

Figure 67: The *accuracy vs. perturbation budget* curves of the 8 models on ImageNet against untargeted transfer-based attacks under the ℓ_2 norm.

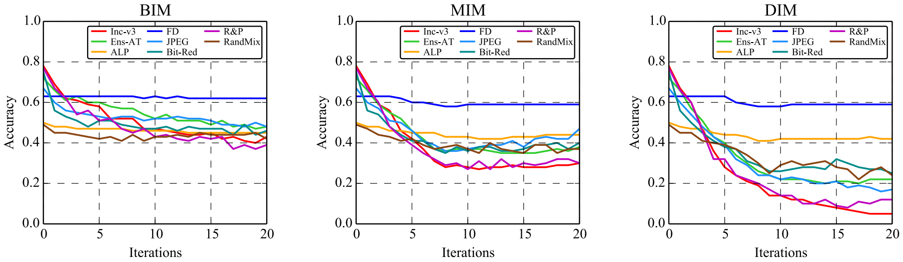

Figure 68: The *accuracy vs. attack strength* curves of the 8 models on ImageNet against untargeted transfer-based attacks under the ℓ_2 norm.

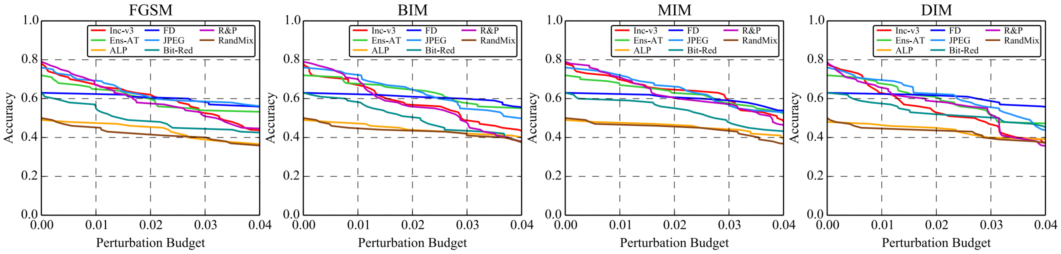

Figure 69: The *accuracy vs. perturbation budget* curves of the 8 models on ImageNet against targeted transfer-based attacks under the ℓ_2 norm.

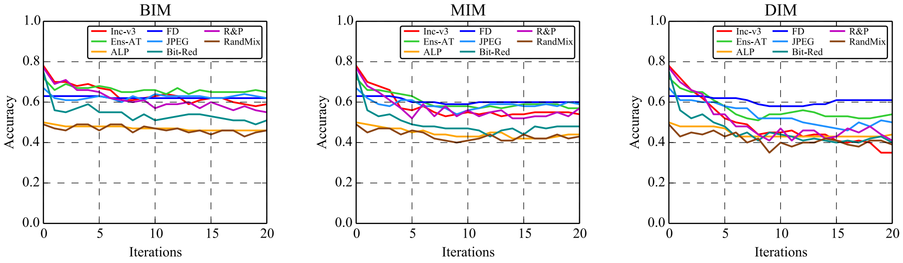

Figure 70: The *accuracy vs. attack strength* curves of the 8 models on ImageNet against targeted transfer-based attacks under the ℓ_2 norm.

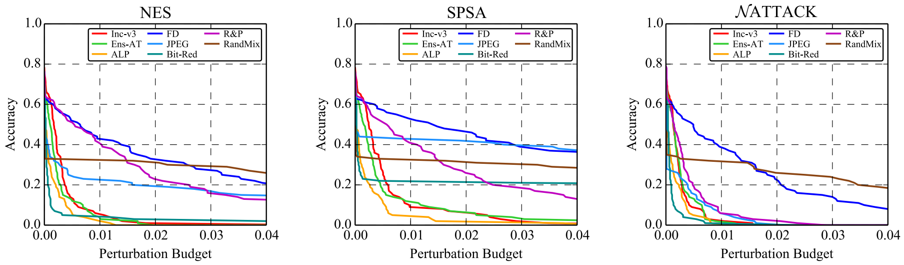

Figure 71: The *accuracy vs. perturbation budget* curves of the 8 models on ImageNet against untargeted score-based attacks under the ℓ_2 norm.

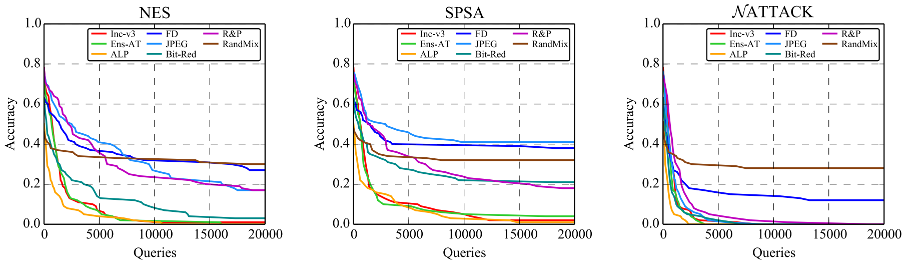

Figure 72: The *accuracy vs. attack strength* curves of the 8 models on ImageNet against untargeted score-based attacks under the ℓ_2 norm.

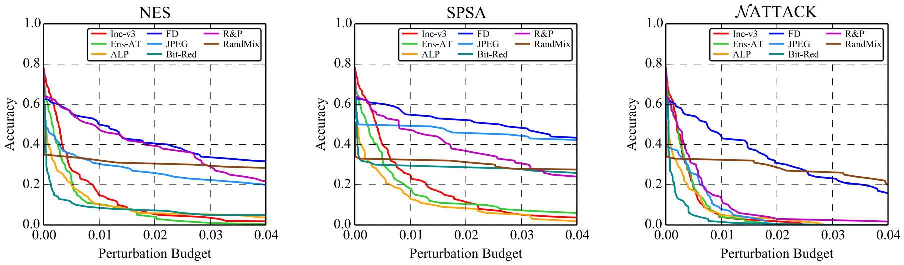

Figure 73: The *accuracy vs. perturbation budget* curves of the 8 models on ImageNet against targeted score-based attacks under the ℓ_2 norm.

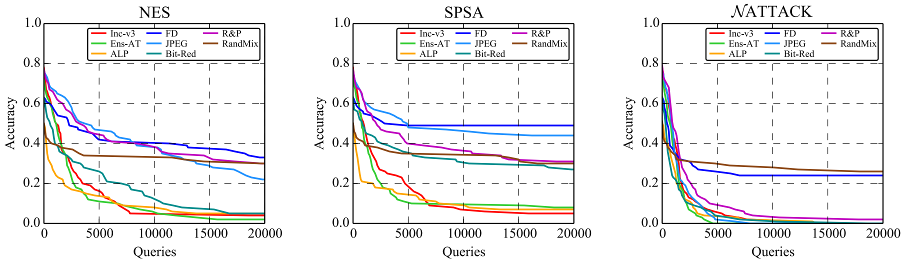

Figure 74: The *accuracy vs. attack strength* curves of the 8 models on ImageNet against targeted score-based attacks under the ℓ_2 norm.

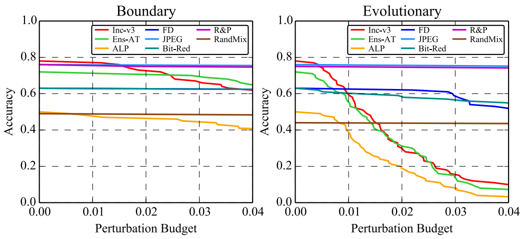

Figure 75: The *accuracy vs. perturbation budget* curves of the 8 models on ImageNet against targeted decision-based attacks under the ℓ_2 norm.

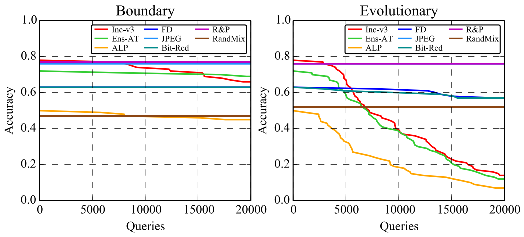

Figure 76: The *accuracy vs. attack strength* curves of the 8 models on ImageNet against targeted decision-based attacks under the ℓ_2 norm.

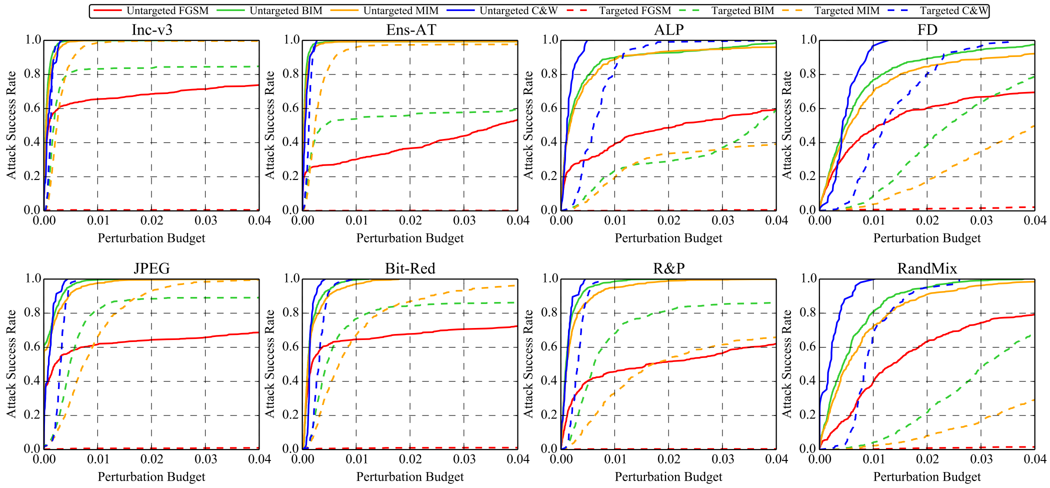

Figure 77: The *attack success rate vs. perturbation budget* curves of white-box attacks under the ℓ_2 norm on the 8 models on ImageNet.

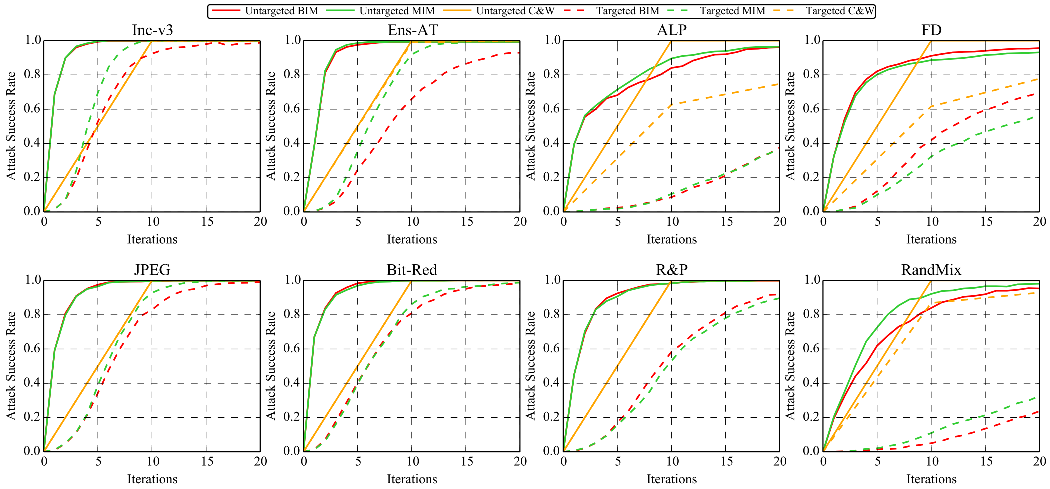

Figure 78: The *attack success rate vs. attack strength* curves of white-box attacks under the ℓ_2 norm on the 8 models on ImageNet.

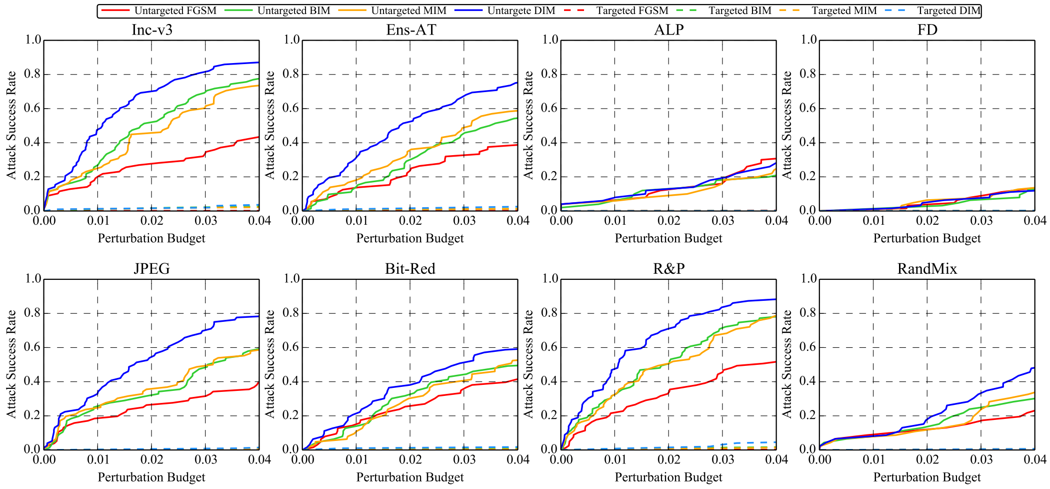

Figure 79: The *attack success rate vs. perturbation budget* curves of transfer-based attacks under the ℓ_2 norm on the 8 models on ImageNet.

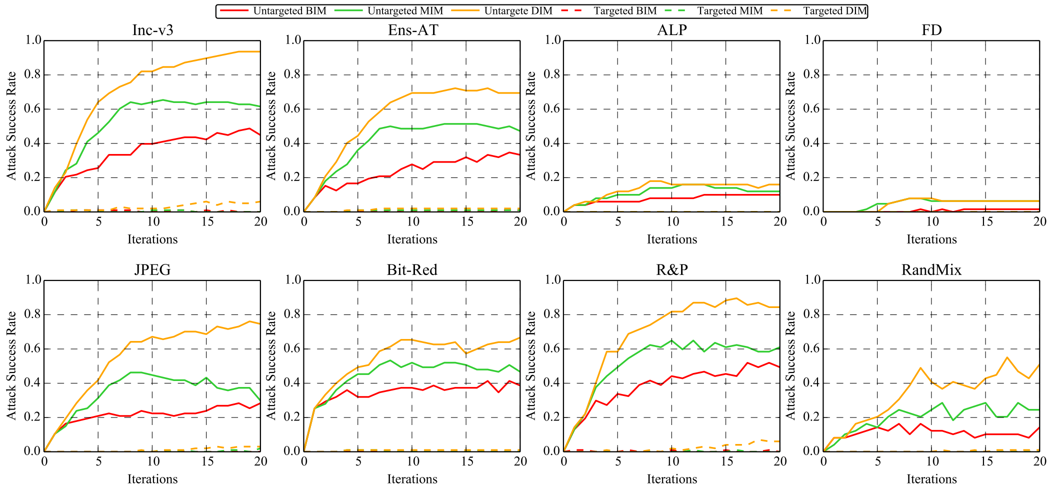

Figure 80: The *attack success rate vs. attack strength* curves of transfer-based attacks under the ℓ_2 norm on the 8 models on ImageNet.

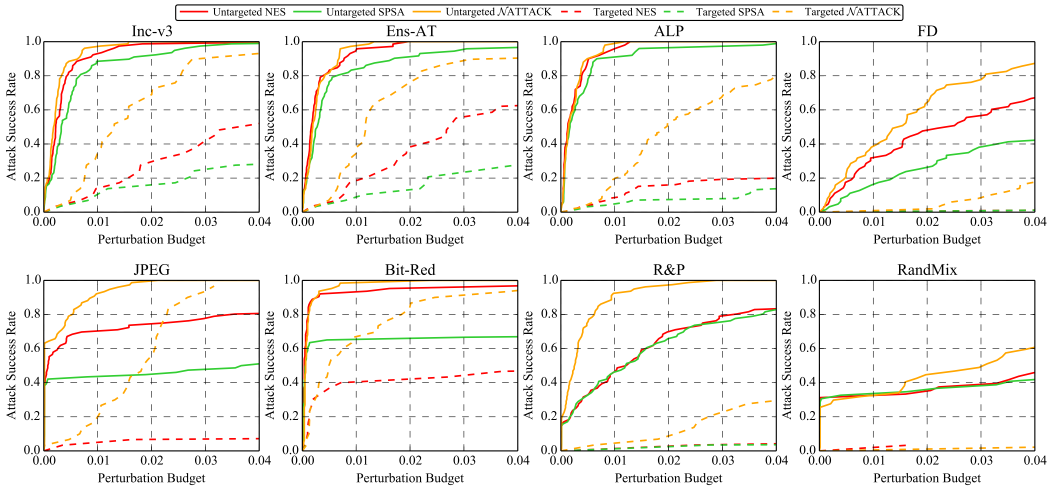

Figure 81: The *attack success rate vs. perturbation budget* curves of score-based attacks under the ℓ_2 norm on the 8 models on ImageNet.

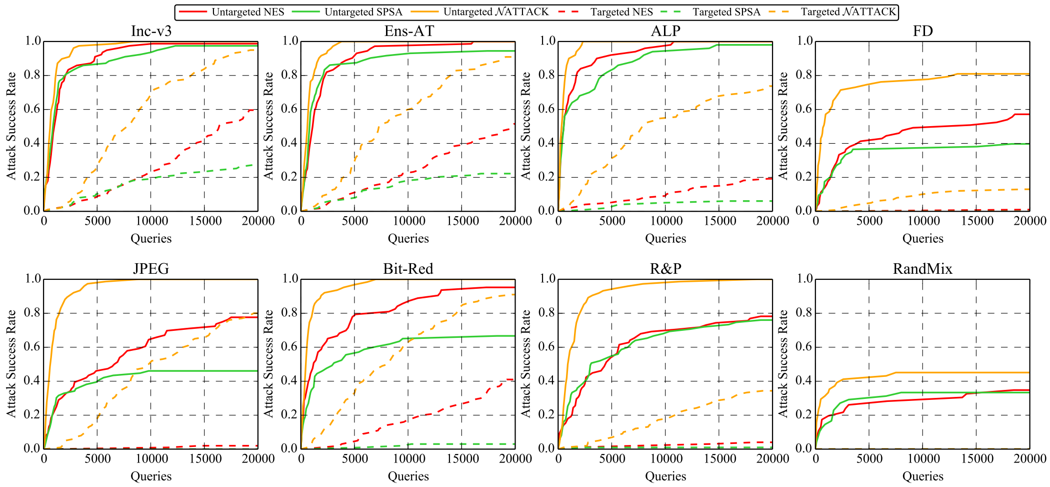

Figure 82: The *attack success rate vs. attack strength* curves of score-based attacks under the ℓ_2 norm on the 8 models on ImageNet.

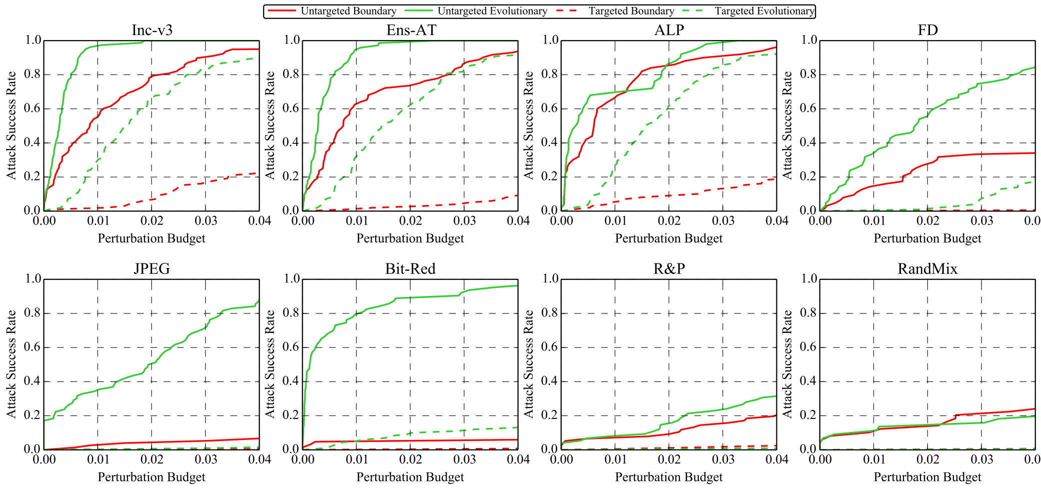

Figure 83: The *attack success rate vs. perturbation budget* curves of decision-based attacks under the ℓ_2 norm on the 8 models on ImageNet.

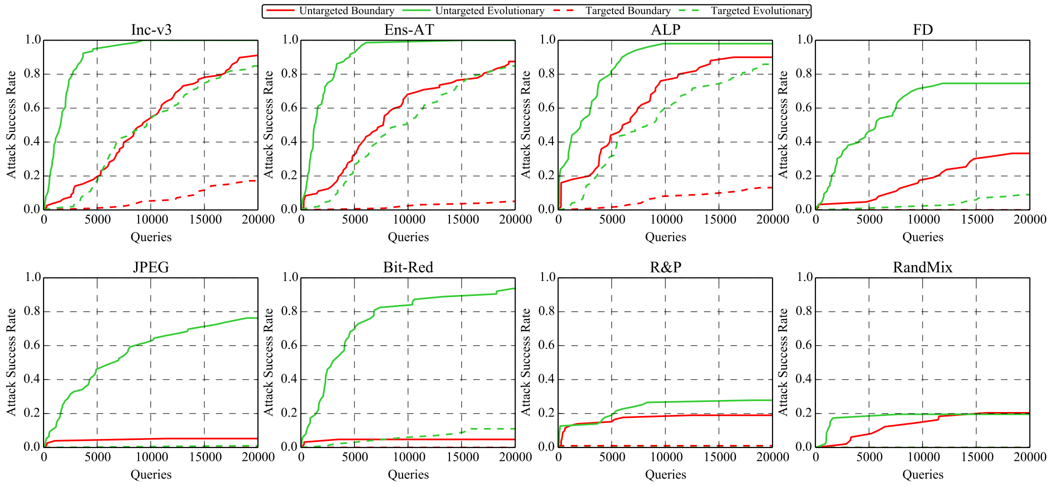

Figure 84: The *attack success rate vs. attack strength* curves of decision-based attacks under the ℓ_2 norm on the 8 models on ImageNet.