Appendix for: Benchmarking Adversarial Robustness on Image Classification

Yinpeng Dong[1], Qi-An Fu[1], Xiao Yang[1], Tianyu Pang[1], Hang Su[1*], Zihao Xiao[2], Jun Zhu[1*]
[1] Dept. of Comp. Sci. and Tech., BNRist Center, Institute for AI, THBI Lab
[1] Tsinghua University, Beijing, 100084, China [2] RealAI
{dyp17, fqa19, yangxiao19, pty17}@mails.tsinghua.edu.cn, {suhangss, dcszj}@tsinghua.edu.cn, zihao.xiao@realai.ai

A. Adversarial Robustness Platforms

There are several public platforms for adversarial machine learning, including CleverHans [23], Foolbox [24], ART [21], Advbox [9], *etc*. However, we observe that these platforms do not totally support our 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 [17], which also provides a uniform platform for adversarial robustness evaluation of DL models. However, as argued in [2], 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.

Our platform takes a modular implementation, which is easily extendable. It mainly consists of five parts, including datasets, attacks, backbone classifiers, defenses, and evaluations. Each part provides a uniform and orthogonal interface, which enables ourselves and other researchers to add new datasets, algorithms, and evaluations in a convenient way. We will maintain the platform and benchmark in the future.

B. Evaluation Details

In this section, we provide additional evaluation details. Table 4 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 [8] generates an untargeted adversarial example under the ℓ_∞ norm as

$$\boldsymbol{x}^{adv} = \boldsymbol{x} + \epsilon \cdot \mathrm{sign}(\nabla_{\boldsymbol{x}} \mathcal{J}(\boldsymbol{x}, y)), \tag{1}$$

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

$$\boldsymbol{x}^{adv} = \boldsymbol{x} + \epsilon \cdot \frac{\nabla_{\boldsymbol{x}} \mathcal{J}(\boldsymbol{x}, y)}{\|\nabla_{\boldsymbol{x}} \mathcal{J}(\boldsymbol{x}, y)\|_2}. \tag{2}$$

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 [15] extends FGSM by iteratively taking multiple small gradient updates as

$$\boldsymbol{x}_{t+1}^{adv} = \mathrm{clip}_{\boldsymbol{x},\epsilon}\big(\boldsymbol{x}_t^{adv} + \alpha \cdot \mathrm{sign}(\nabla_{\boldsymbol{x}} \mathcal{J}(\boldsymbol{x}_t^{adv}, y))\big), \tag{3}$$

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 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 [5] integrates a momentum term into BIM as

$$\boldsymbol{g}_{t+1} = \mu \cdot \boldsymbol{g}_t + \frac{\nabla_{\boldsymbol{x}} \mathcal{J}(\boldsymbol{x}_t^{adv}, y)}{\|\nabla_{\boldsymbol{x}} \mathcal{J}(\boldsymbol{x}_t^{adv}, y)\|_1}; \tag{4}$$

*Hang Su and Jun Zhu are corresponding authors.

CIFAR-10 [14]		ImageNet [25]	
Defense Model	Architecture	Defense Model	Architecture
Res-56 [10]	ResNet-56	Inc-v3 [26]	Inception v3
PGD-AT [19]	Wide ResNet-34-10	Ens-AT [27]	Inception v3
DeepDefense [34]	5-layer CNN	ALP [12]	ResNet-50
TRADES [35]	Wide ResNet-34-10	FD [31]	ResNet-152 with denoising layers
Convex [29]	ResNet	JPEG [7]	Inception v3
JPEG [7]	ResNet-56	Bit-Red [33]	Inception v3
RSE [18]	VGG	R&P [30]	Inception v3
ADP [22]	ResNet-110 $\times 3$	RandMix [36]	Inception v3

Table 4. 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 [29].

$$\boldsymbol{x}_{t+1}^{adv} = \text{clip}_{\boldsymbol{x},\epsilon}(\boldsymbol{x}_t^{adv} + \alpha \cdot \text{sign}(\boldsymbol{g}_{t+1})). \quad (5)$$

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 [20] 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 [3] is an optimization-based attack method, which generates an ℓ_2 adversarial example by solving

$$\begin{aligned} \boldsymbol{x}^{adv} = &\arg\min_{\boldsymbol{x}'} \{\|\boldsymbol{x}' - \boldsymbol{x}\|_2^2 \\ &+ c \cdot \max(Z(\boldsymbol{x}')_y - \max_{i \neq y} Z(\boldsymbol{x}')_i, 0)\}, \end{aligned} \quad (6)$$

where $Z(\boldsymbol{x}')$ is the logit output of the classifier and c is a constant. This optimization problem is solved by an Adam [13] optimizer. c is found by binary search. To get the accuracy (attack success rate) vs. perturbation budget curves, we optimize Eq. (6) for 100 iterations. To get the accuracy (attack success rate) vs. attack strength curves, we optimize Eq. (6) for 10, 20, 30, 40 iterations on CIFAR-10, and 10, 20 iterations on ImageNet to show the results.

DIM [32] 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 [4] has been proposed to optimize Eq. (6) in the black-box manner through queries. It estimates the gradient at each coordinate as

$$\hat{g}_i = \frac{\mathcal{L}(\boldsymbol{x} + \sigma \boldsymbol{e}_i, y) - \mathcal{L}(\boldsymbol{x} - \sigma \boldsymbol{e}_i, y)}{2\sigma} \approx \frac{\partial \mathcal{L}(\boldsymbol{x}, y)}{\partial x_i}, \quad (7)$$

where $\mathcal{L}$ is the objective in Eq. (6), 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 [11] and **SPSA** [28] adopt the update rule in Eq. (3) for adversarial example generation. Although the true gradient is unavailable, NES and SPSA give the full gradient estimation as

$$\hat{\boldsymbol{g}} = \frac{1}{q} \sum_{i=1}^{q} \frac{\mathcal{J}(\boldsymbol{x} + \sigma \boldsymbol{u}_i, y) - \mathcal{J}(\boldsymbol{x} - \sigma \boldsymbol{u}_i, y)}{2\sigma} \cdot \boldsymbol{u}_i, \quad (8)$$

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** [16] 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** [1] and **Evolutionary** [6] 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.

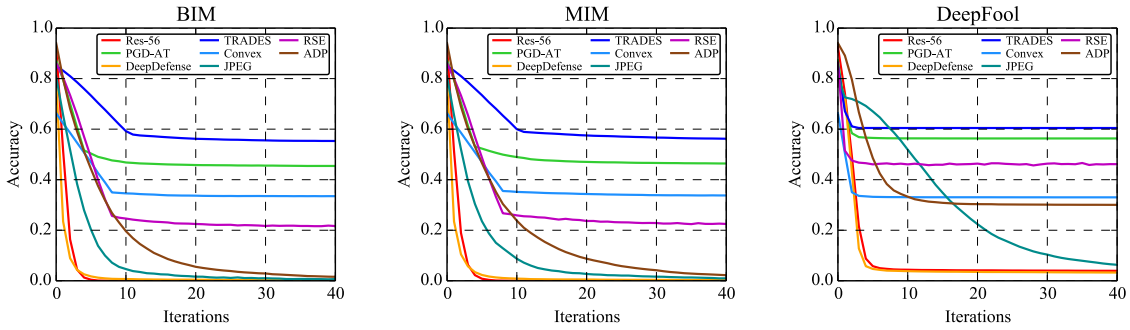

Figure 13. The *accuracy vs. attack strength* curves of the 8 models on CIFAR-10 against untargeted white-box attacks under the ℓ_∞ norm.

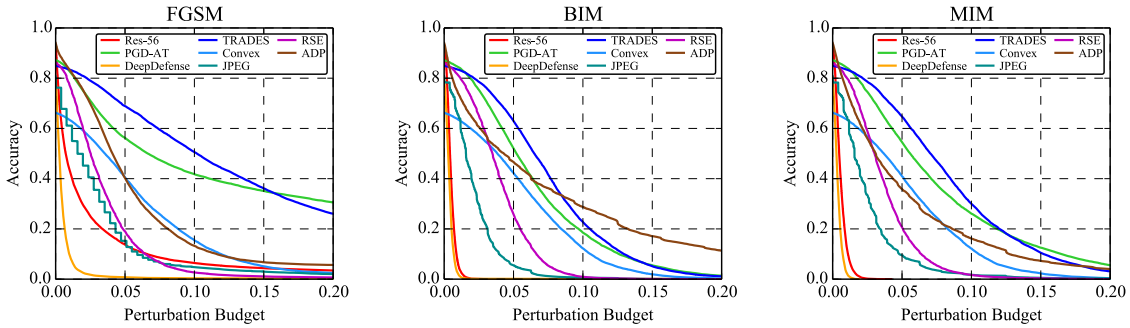

Figure 14. The *accuracy vs. perturbation budget* curves of the 8 models on CIFAR-10 against targeted white-box attacks under the ℓ_∞ norm.

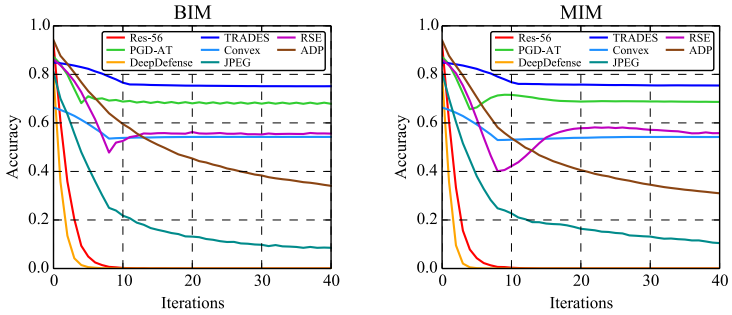

Figure 15. The *accuracy vs. attack strength* curves of the 8 models on CIFAR-10 against targeted white-box attacks under the ℓ_∞ norm.

C.1. Full Evaluation Results on CIFAR-10

Attacks under the ℓ_∞ norm: We have shown some of the accuracy curves of the defense models against untargeted attacks under the ℓ_∞ norm in Sec. 5.1. We next show the remaining curves of untargeted attacks under the ℓ_∞ norm, the curves of targeted attacks under the ℓ_∞ norm, and the attack success rate curves. Fig. 13 shows the *accuracy vs. attack strength* curves of the defenses on CIFAR-10 against untargeted white-box attacks under the ℓ_∞ norm. Fig. 14 and Fig. 15 show the accuracy curves of the defenses on CIFAR-10 against targeted white-box attacks under the ℓ_∞ norm. Fig. 16 shows the *accuracy vs. perturbation budget* curves of the defenses on CIFAR-10 against untargeted transfer-based attacks under the ℓ_∞ norm. Fig. 17 and Fig. 18 show the accuracy curves of the defenses on CIFAR-10 against targeted transfer-based attacks under the ℓ_∞ norm. Fig. 19 and Fig. 20 show the accuracy curves of the defenses on CIFAR-10 against targeted score-based attacks under the ℓ_∞ norm. Fig. 21 to Fig. 26 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. 27, Fig. 28, Fig. 29, and Fig. 30. 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. 31, Fig. 32, Fig. 33, and Fig. 34. We show the accuracy curves of the defenses on CIFAR-10 against untargeted and tar-

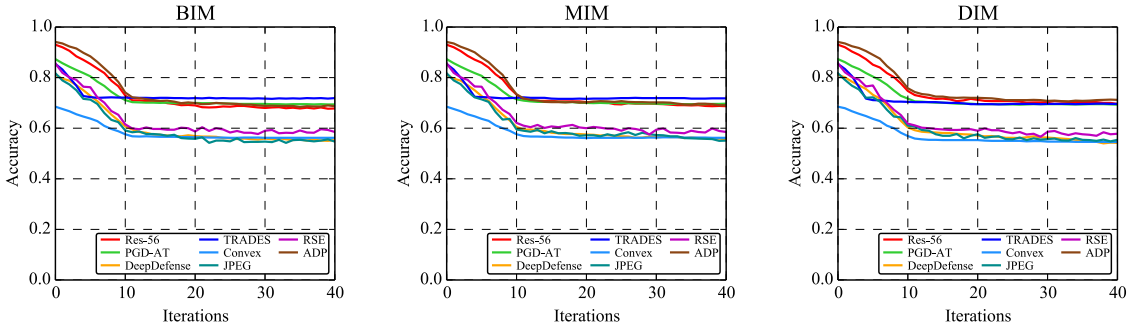

Figure 16. The *accuracy vs. attack strength* curves of the 8 models on CIFAR-10 against untargeted transfer-based attacks under the ℓ_∞ norm.

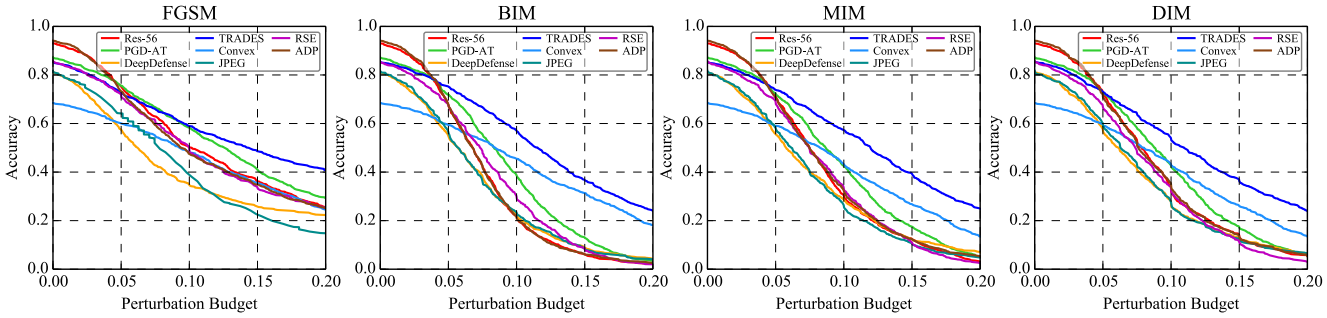

Figure 17. The *accuracy vs. perturbation budget* curves of the 8 models on CIFAR-10 against targeted transfer-based attacks under the ℓ_∞ norm.

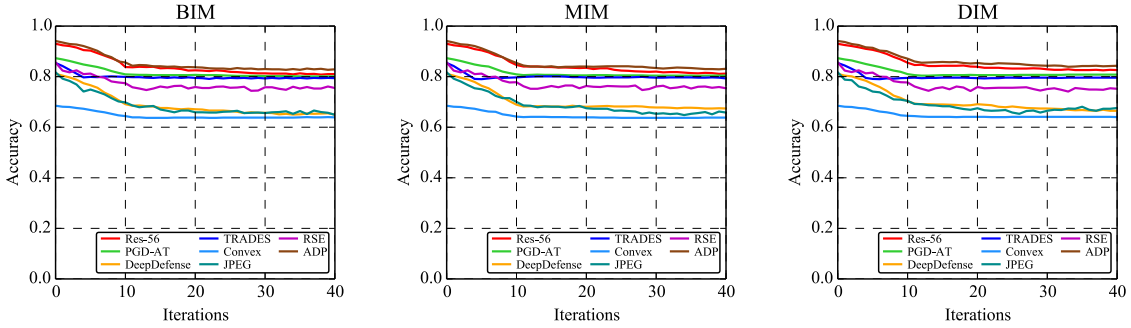

Figure 18. The *accuracy vs. attack strength* curves of the 8 models on CIFAR-10 against targeted transfer-based attacks under the ℓ_∞ norm.

geted score-based attacks under the ℓ_2 norm in Fig. 35, Fig. 36, Fig. 37, and Fig. 38. 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. 4, Fig. 6, Fig. 39, and Fig. 40. Fig. 41 to Fig. 48 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 the remaining untargeted attacks, targeted attacks under the ℓ_∞ norm, and the attacks success rate curves here. Fig. 49 shows the *accuracy vs. attack strength* curves of the defenses on ImageNet against untargeted white-box attacks under the ℓ_∞ norm. Fig. 50 and Fig. 51 show the accuracy curves of the defenses on ImageNet against targeted white-box attacks under the ℓ_∞ norm. Fig. 52 shows the *accuracy vs. attack strength* curves of the defenses on ImageNet against untargeted transfer-based 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*

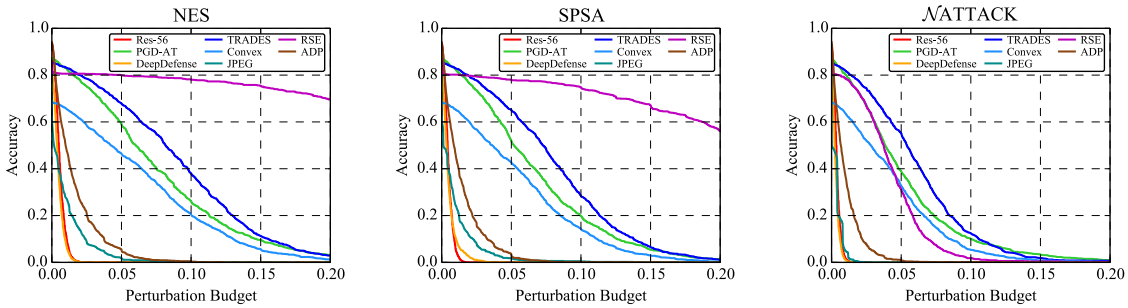

Figure 19. The *accuracy vs. perturbation budget* curves of the 8 models on CIFAR-10 against targeted score-based attacks under the ℓ_∞ norm.

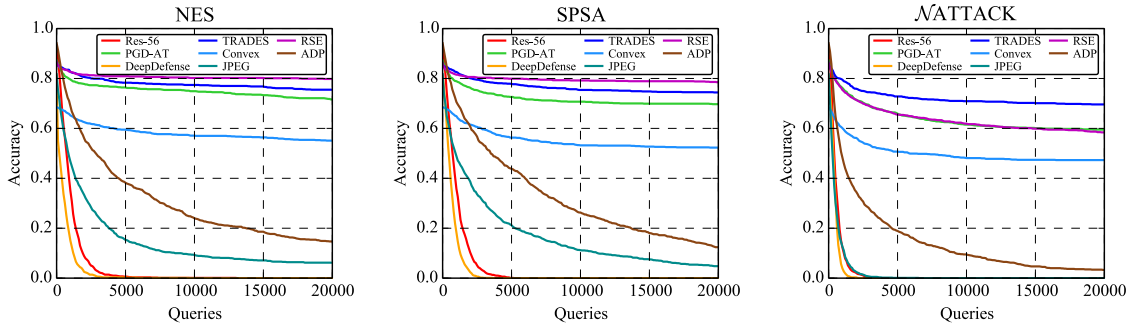

Figure 20. The *accuracy vs. attack strength* curves of the 8 models on CIFAR-10 against targeted score-based attacks under the ℓ_∞ norm.

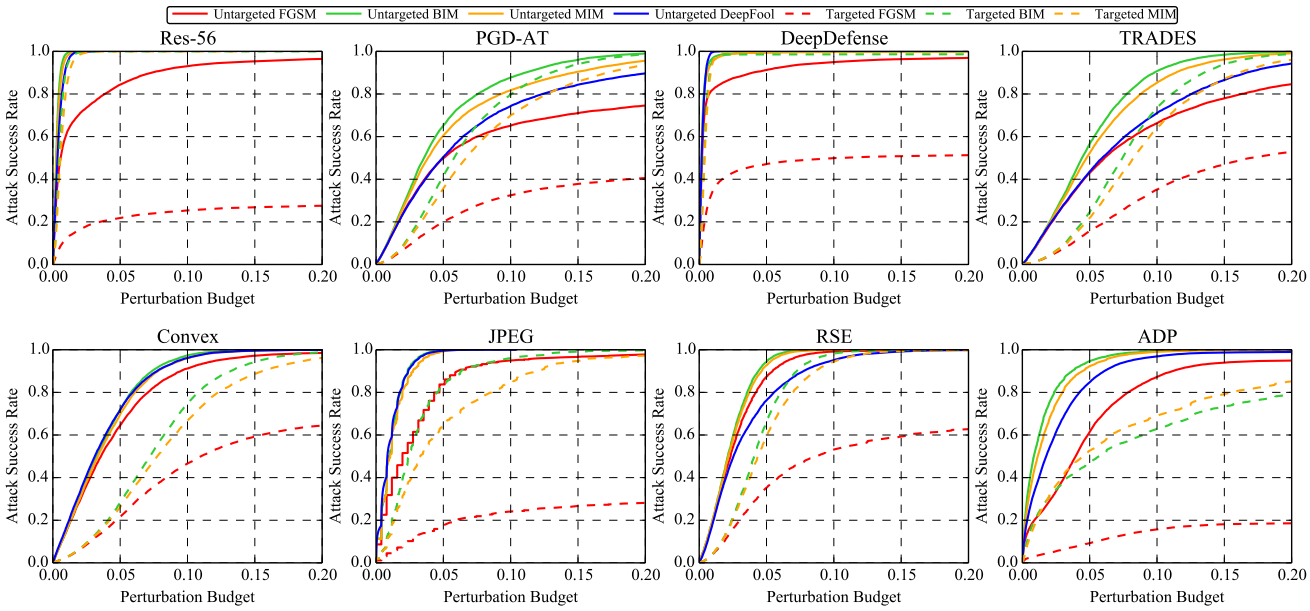

Figure 21. The *attack success rate vs. perturbation budget* curves of white-box attacks under the ℓ_∞ norm on the 8 models on CIFAR-10.

strength curves of white-box, transfer-based, and score-based attacks under the ℓ_∞ norm on the 8 models on ImageNet.

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,

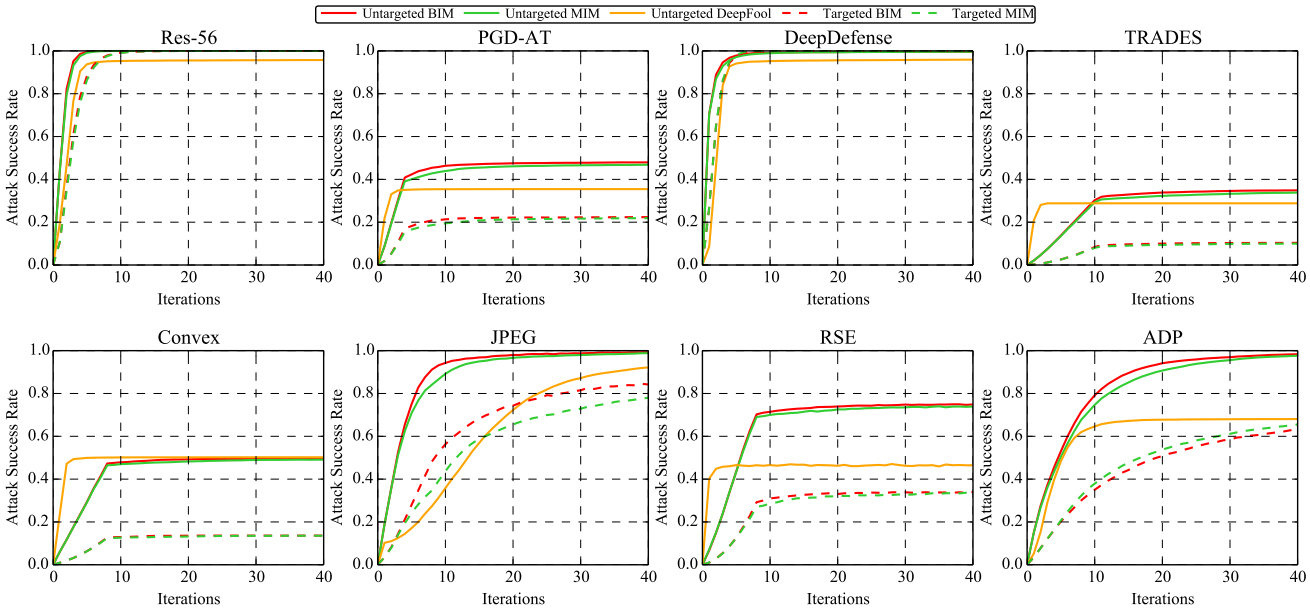

Figure 22. The *attack success rate vs. attack strength* curves of white-box attacks under the ℓ_∞ norm on the 8 models on CIFAR-10.

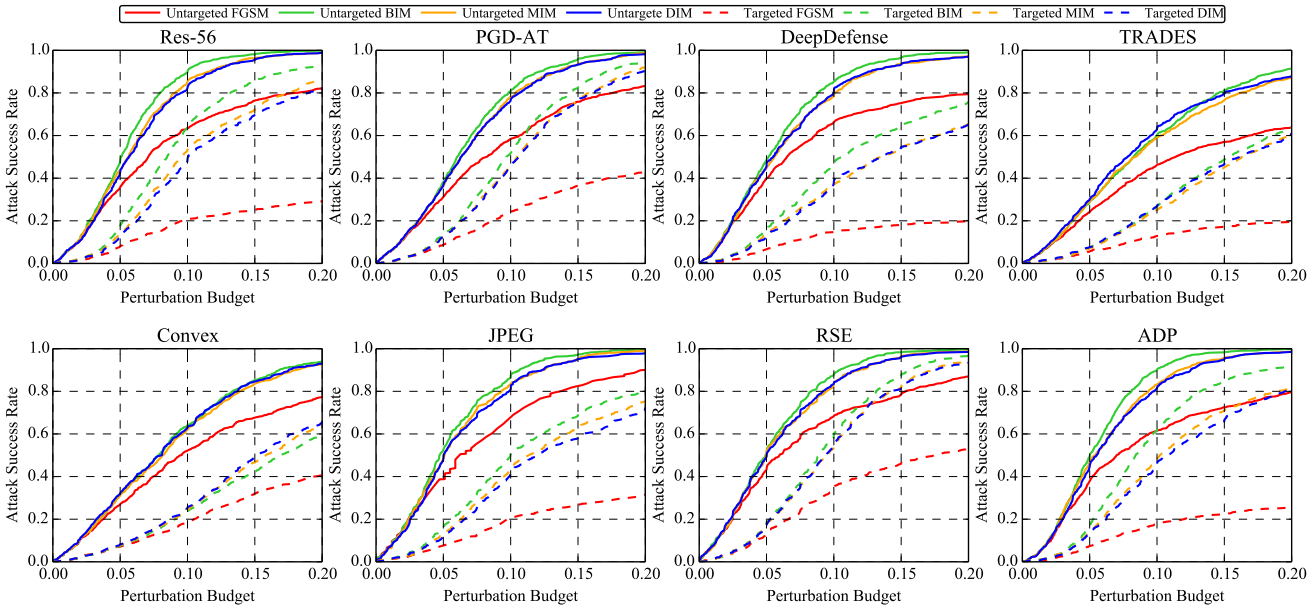

Figure 23. The *attack success rate vs. perturbation budget* curves of transfer-based attacks under the ℓ_∞ norm on the 8 models on CIFAR-10.

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. 10, Fig. 12, 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.

References

[1] 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)*, 2018. 2

[2] Nicholas Carlini. A critique of the deepsec platform for security analysis of deep learning models. *arXiv preprint arXiv:1905.07112*, 2019. 1

[3] Nicholas Carlini and David Wagner. Towards evaluating the

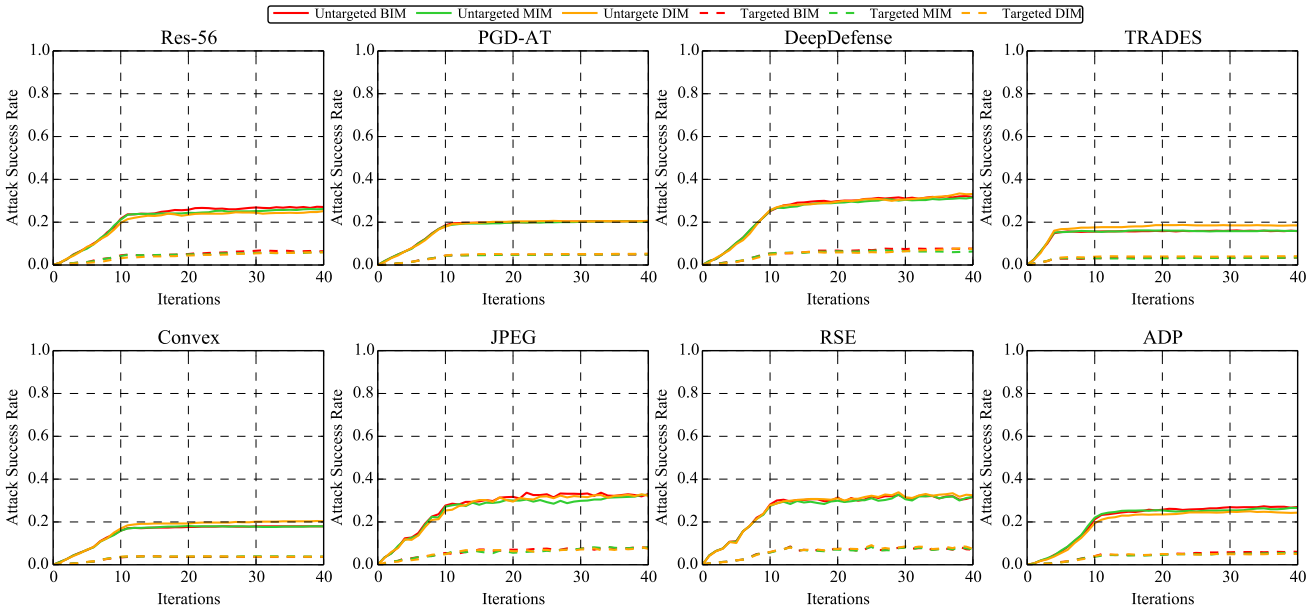

Figure 24. The *attack success rate vs. attack strength* curves of transfer-based attacks under the ℓ_∞ norm on the 8 models on CIFAR-10.

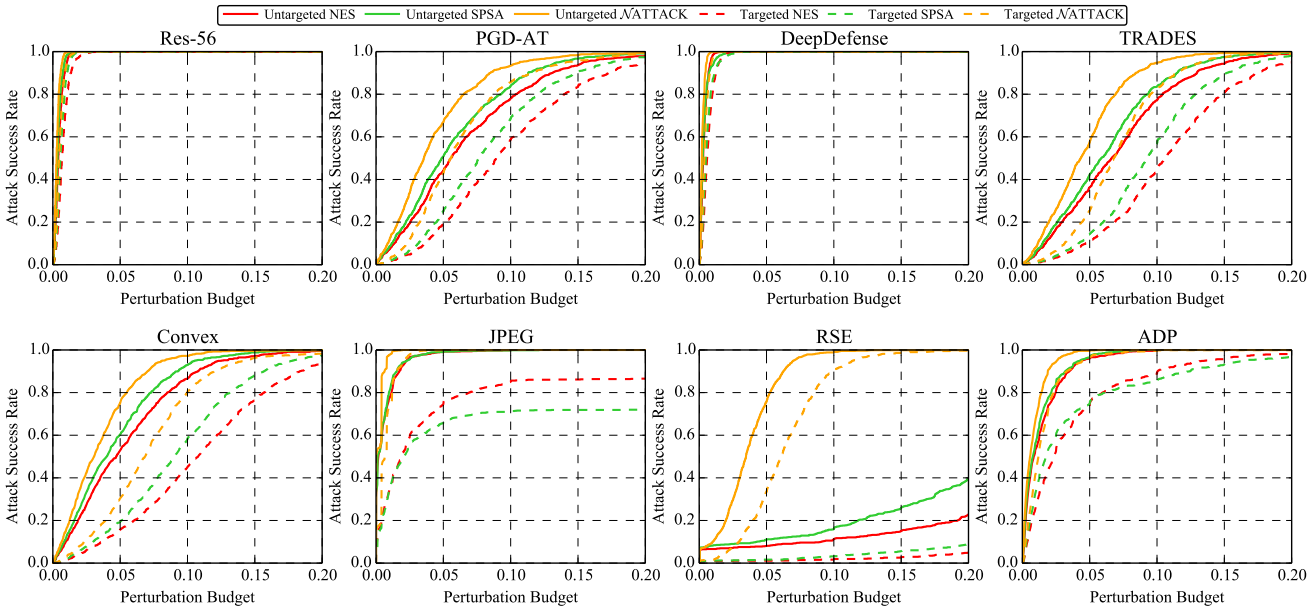

Figure 25. The *attack success rate vs. perturbation budget* curves of score-based attacks under the ℓ_∞ norm on the 8 models on CIFAR-10.

robustness of neural networks. In *IEEE Symposium on Security and Privacy*, 2017. 2

[4] 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*, pages 15–26. ACM, 2017. 2

[5] 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. 1

[6] 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. 2

[7] Gintare Karolina Dziugaite, Zoubin Ghahramani, and Daniel M Roy. A study of the effect of jpg compression on

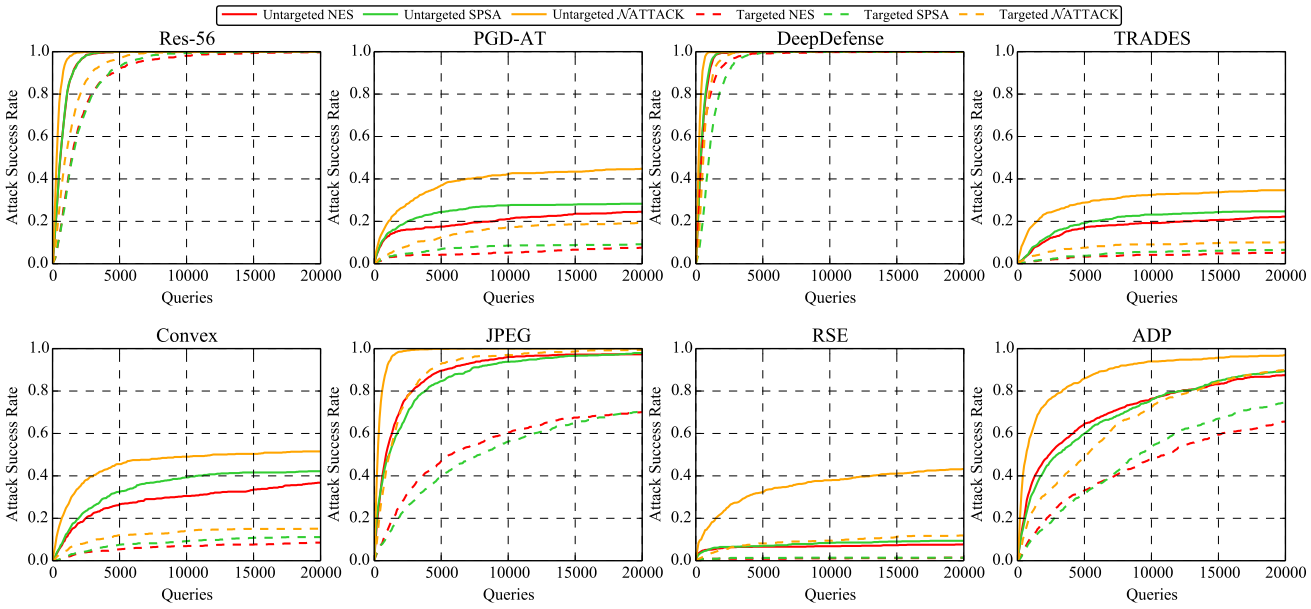

Figure 26. The *attack success rate vs. attack strength* curves of score-based attacks under the ℓ_∞ norm on the 8 models on CIFAR-10.

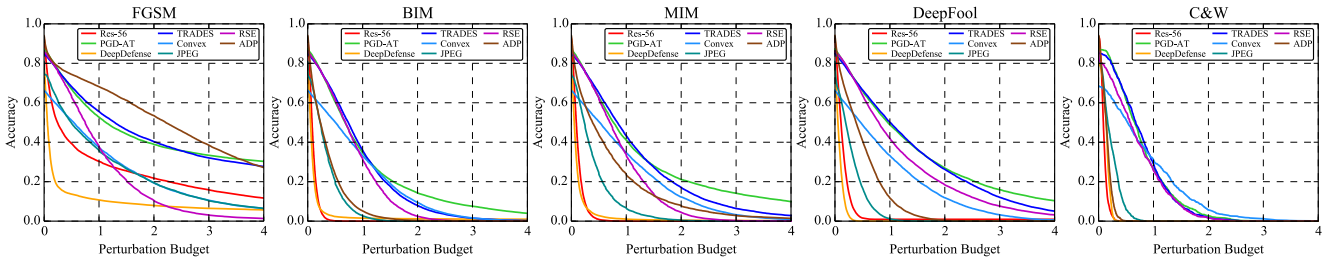

Figure 27. The *accuracy vs. perturbation budget* curves of the 8 models on CIFAR-10 against untargeted white-box attacks under the ℓ_2 norm.

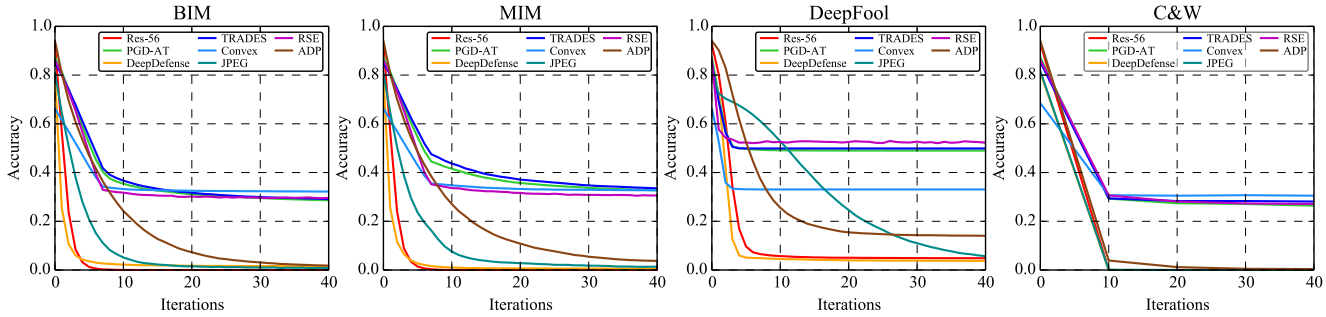

Figure 28. The *accuracy vs. attack strength* curves of the 8 models on CIFAR-10 against untargeted white-box attacks under the ℓ_2 norm.

adversarial images. *arXiv preprint arXiv:1608.00853*, 2016. 2

[8] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2015. 1

[9] Dou Goodman, Hao Xin, Wang Yang, Wu Yuesheng, Xiong Junfeng, and Zhang Huan. Advbox: a toolbox to generate adversarial examples that fool neural networks. *arXiv preprint arXiv:2001.05574*, 2020. 1

[10] 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. 2

[11] 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. 2

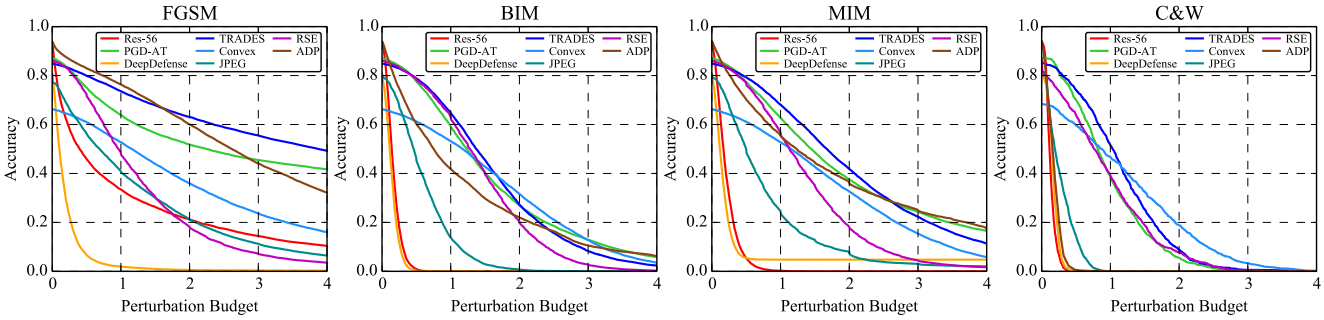

Figure 29. The *accuracy vs. perturbation budget* curves of the 8 models on CIFAR-10 against targeted white-box attacks under the ℓ_2 norm.

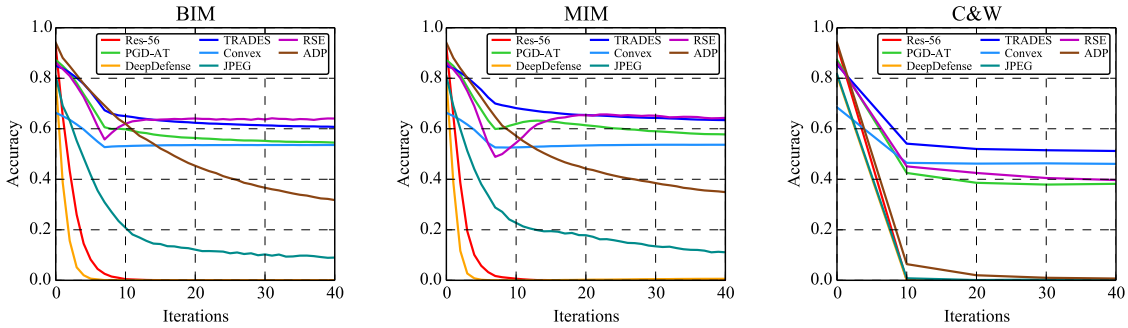

Figure 30. The *accuracy vs. attack strength* curves of the 8 models on CIFAR-10 against targeted white-box attacks under the ℓ_2 norm.

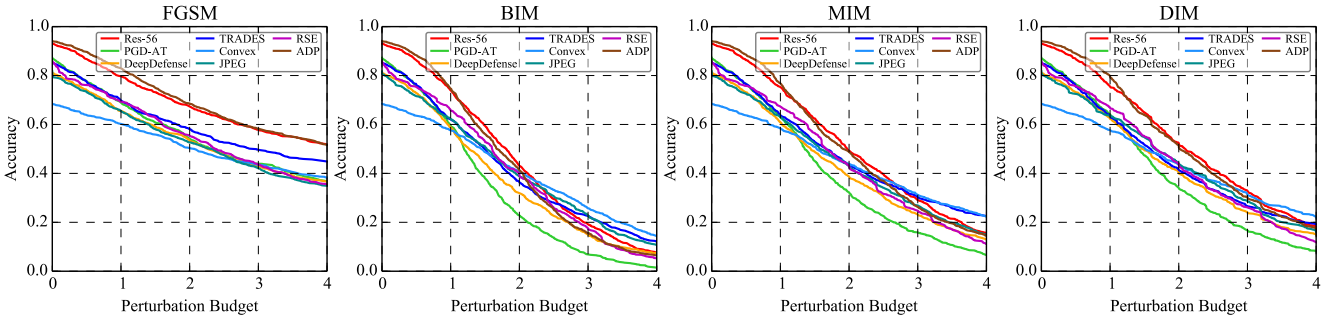

Figure 31. The *accuracy vs. perturbation budget* curves of the 8 models on CIFAR-10 against untargeted transfer-based attacks under the ℓ_2 norm.

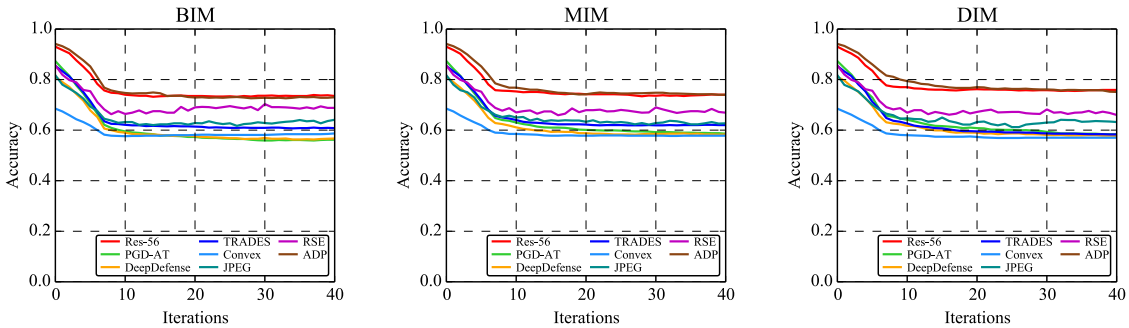

Figure 32. The *accuracy vs. attack strength* curves of the 8 models on CIFAR-10 against untargeted transfer-based attacks under the ℓ_2 norm.

[12] Harini Kannan, Alexey Kurakin, and Ian Goodfellow. Adversarial logit pairing. *arXiv preprint arXiv:1803.06373*, 2018. 2

[13] Diederik Kingma and Jimmy Ba. Adam: A method for

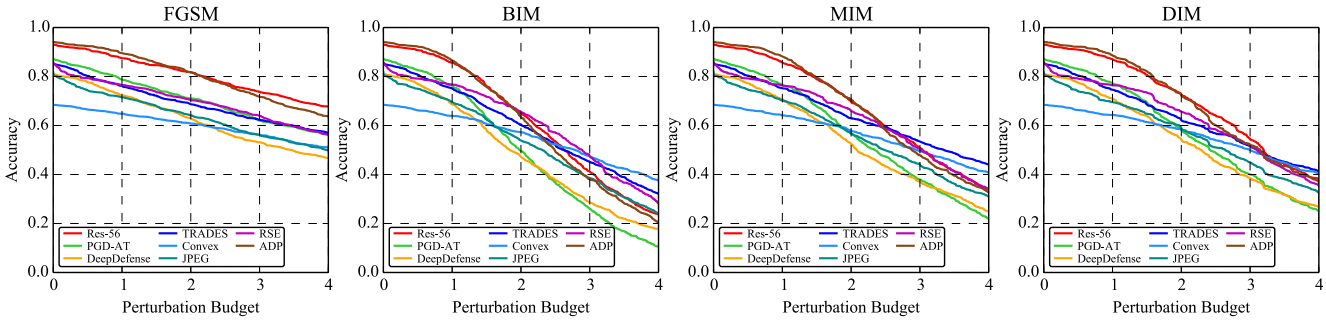

Figure 33. The *accuracy vs. perturbation budget* curves of the 8 models on CIFAR-10 against targeted transfer-based attacks under the ℓ_2 norm.

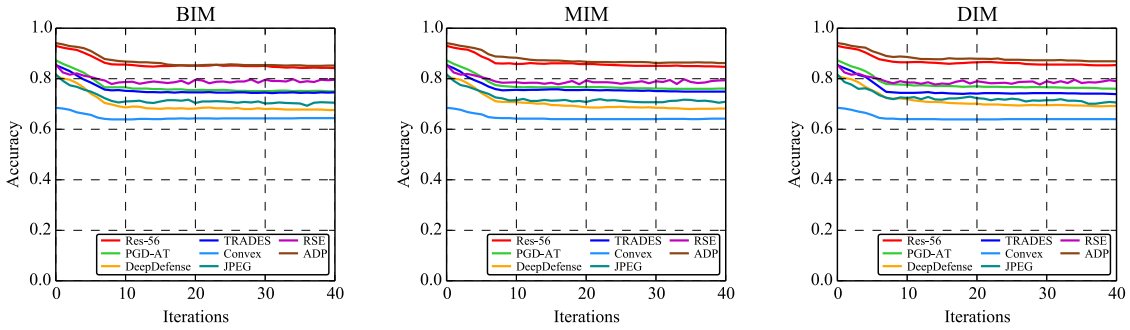

Figure 34. The *accuracy vs. attack strength* curves of the 8 models on CIFAR-10 against targeted transfer-based attacks under the ℓ_2 norm.

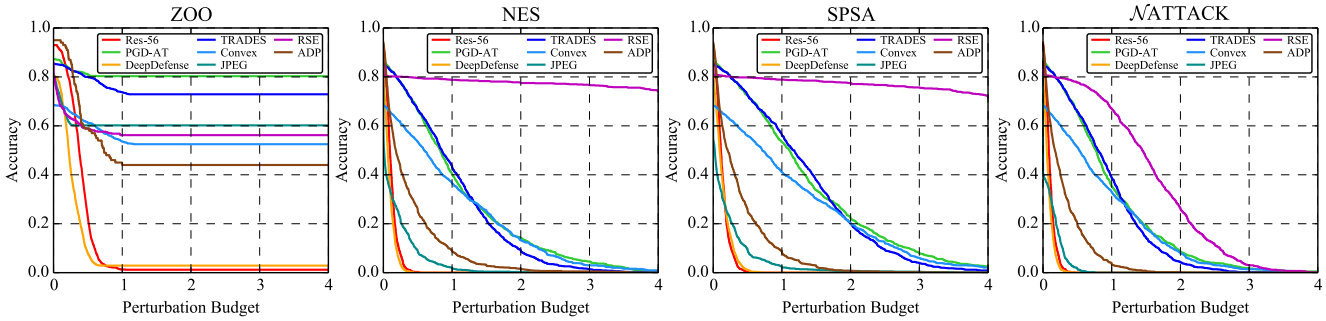

Figure 35. The *accuracy vs. perturbation budget* curves of the 8 models on CIFAR-10 against untargeted score-based attacks under the ℓ_2 norm.

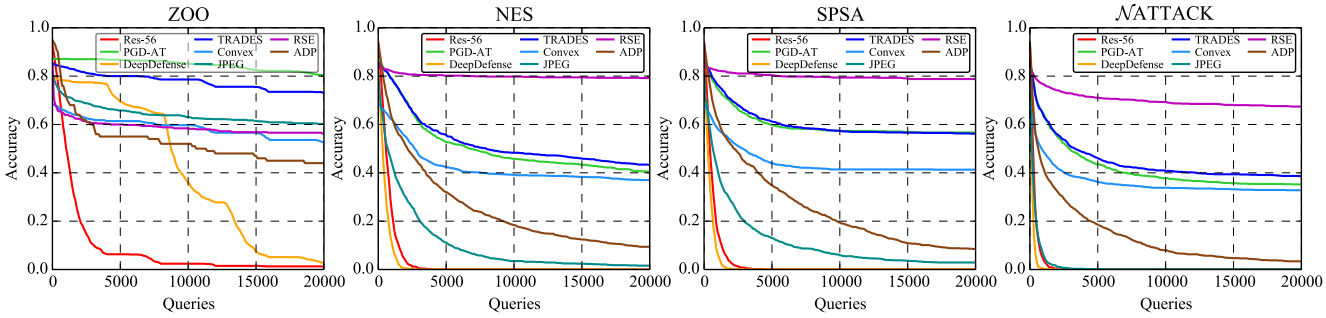

Figure 36. The *accuracy vs. attack strength* curves of the 8 models on CIFAR-10 against untargeted score-based attacks under the ℓ_2 norm.

stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. 2

[14] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. 2

[15] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Ad-

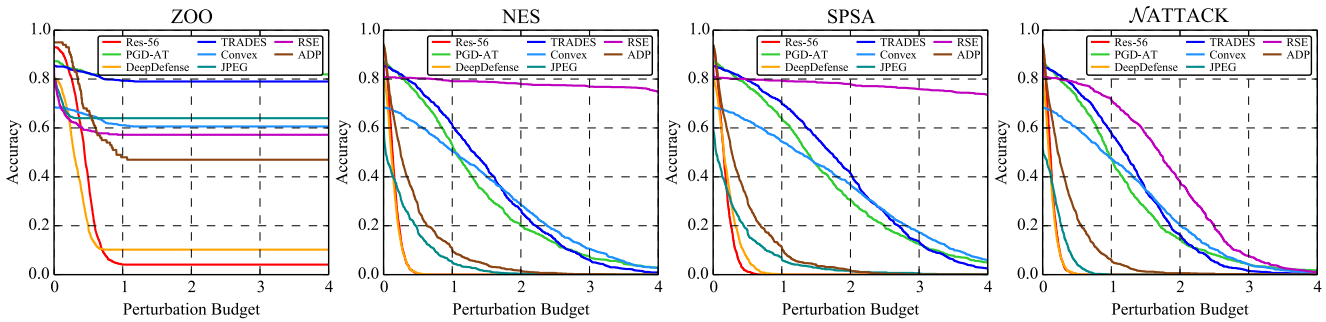

Figure 37. The *accuracy vs. perturbation budget* curves of the 8 models on CIFAR-10 against targeted score-based attacks under the ℓ_2 norm.

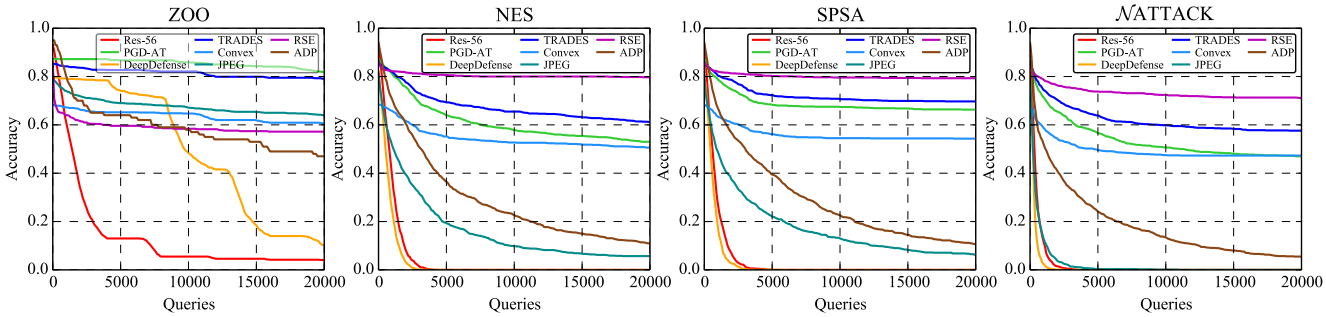

Figure 38. The *accuracy vs. attack strength* curves of the 8 models on CIFAR-10 against targeted score-based attacks under the ℓ_2 norm.

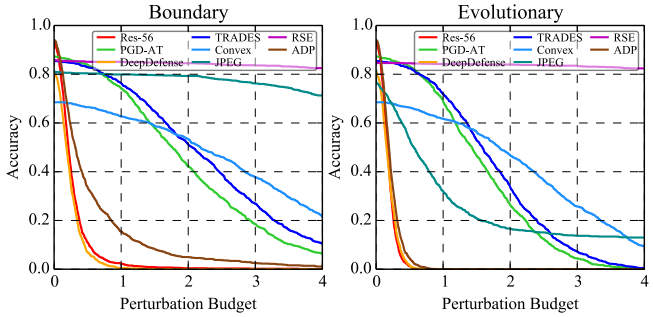

Figure 39. The *accuracy vs. perturbation budget* curves of the 8 models on CIFAR-10 against targeted decision-based attacks under the ℓ_2 norm.

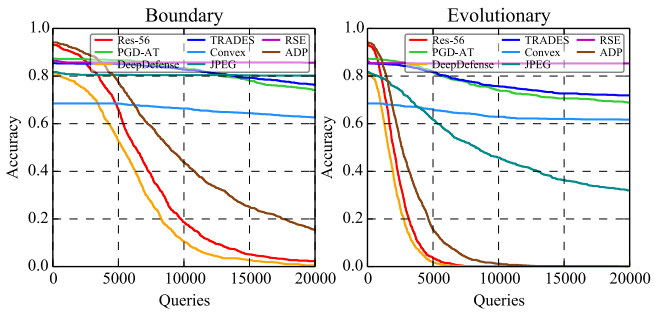

Figure 40. The *accuracy vs. attack strength* curves of the 8 models on CIFAR-10 against targeted decision-based attacks under the ℓ_2 norm.

versarial examples in the physical world. In *International Conference on Learning Representations (ICLR) Workshops*, 2017. 1

[16] 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. 2

[17] 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. 1

[18] 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. 2

[19] 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. 2

[20] 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. 2

[21] 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. 1

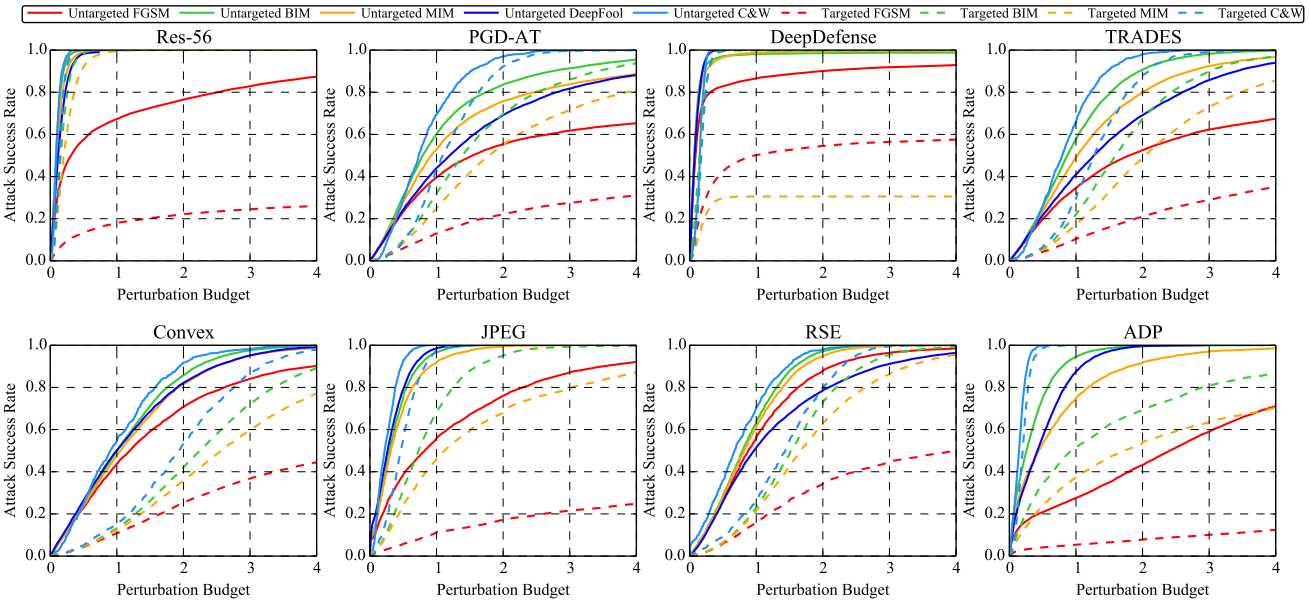

Figure 41. The *attack success rate vs. perturbation budget* curves of white-box attacks under the ℓ_2 norm on the 8 models on CIFAR-10.

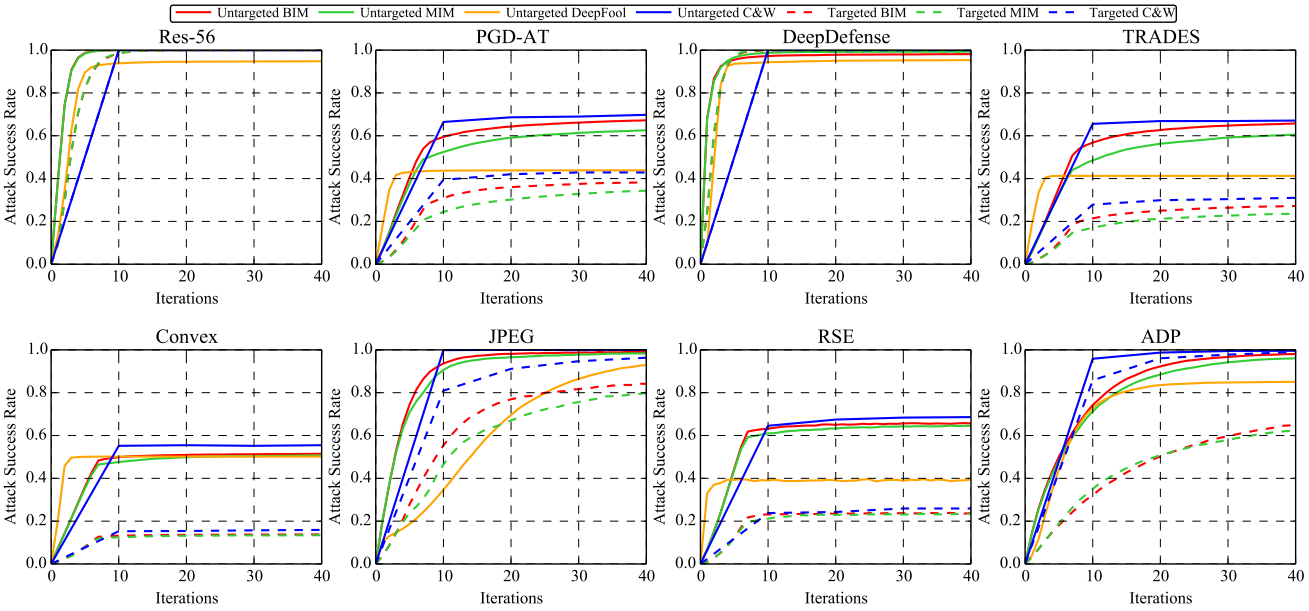

Figure 42. The *attack success rate vs. attack strength* curves of white-box attacks under the ℓ_2 norm on the 8 models on CIFAR-10.

[22] 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. 2

[23] 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*, 2016. 1

[24] 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. 1

[25] 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. 2

[26] 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*

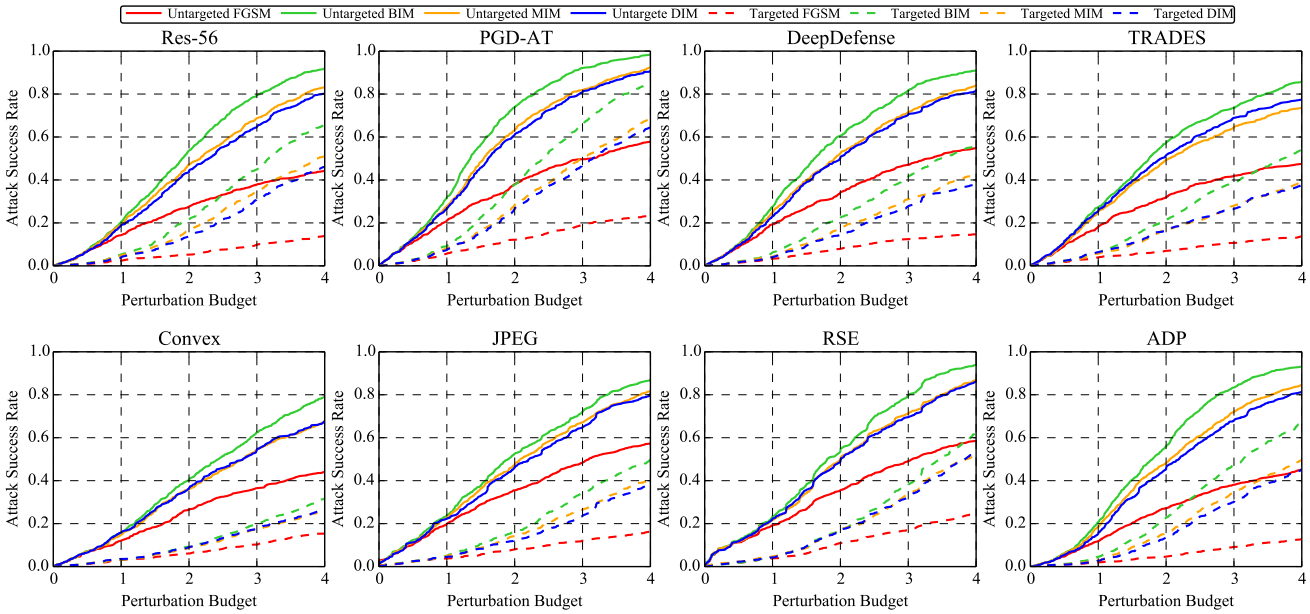

Figure 43. The *attack success rate vs. perturbation budget* curves of transfer-based attacks under the ℓ_2 norm on the 8 models on CIFAR-10.

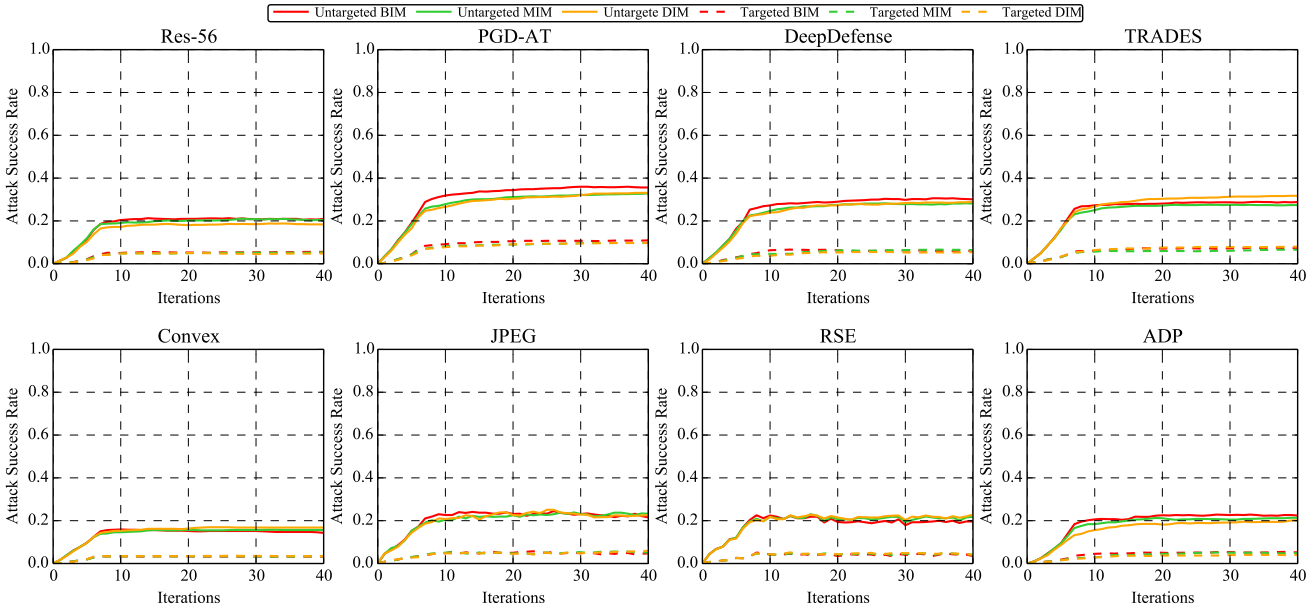

Figure 44. The *attack success rate vs. attack strength* curves of transfer-based attacks under the ℓ_2 norm on the 8 models on CIFAR-10.

(CVPR), 2016. 2

[27] 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. 2

[28] 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. 2

[29] Eric Wong, Frank Schmidt, Jan Hendrik Metzen, and J Zico Kolter. Scaling provable adversarial defenses. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 2

[30] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan Yuille. Mitigating adversarial effects through randomization. In *International Conference on Learning Representations (ICLR)*, 2018. 2

[31] 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)*, 2019. 2

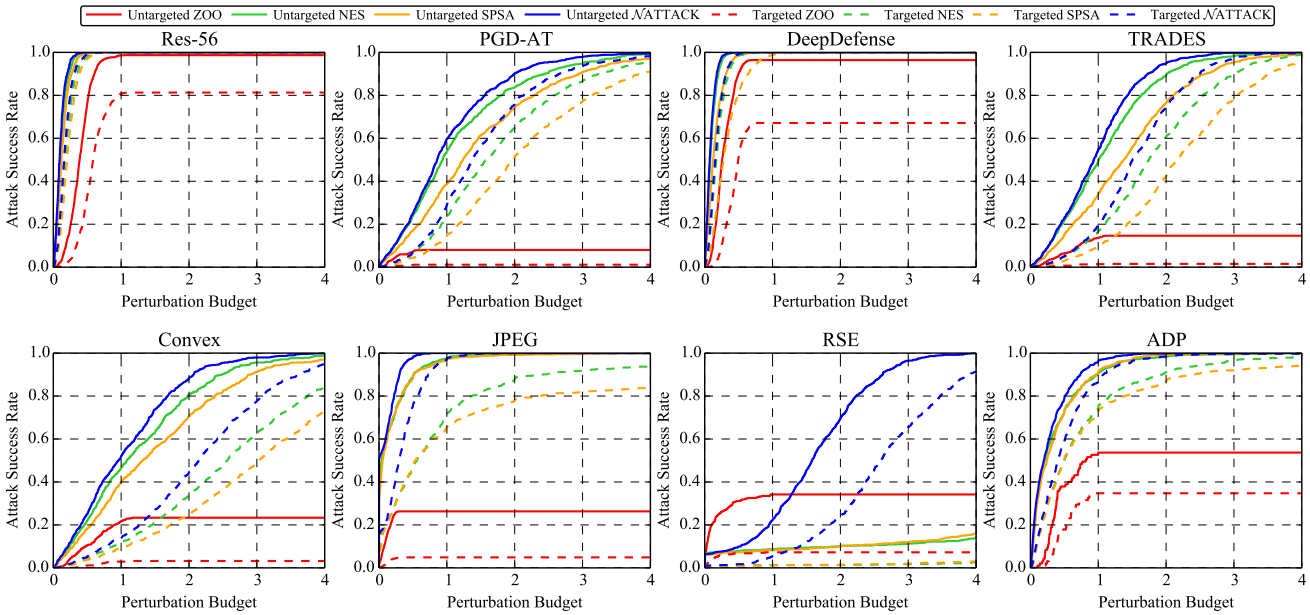

Figure 45. The *attack success rate vs. perturbation budget* curves of score-based attacks under the ℓ_2 norm on the 8 models on CIFAR-10.

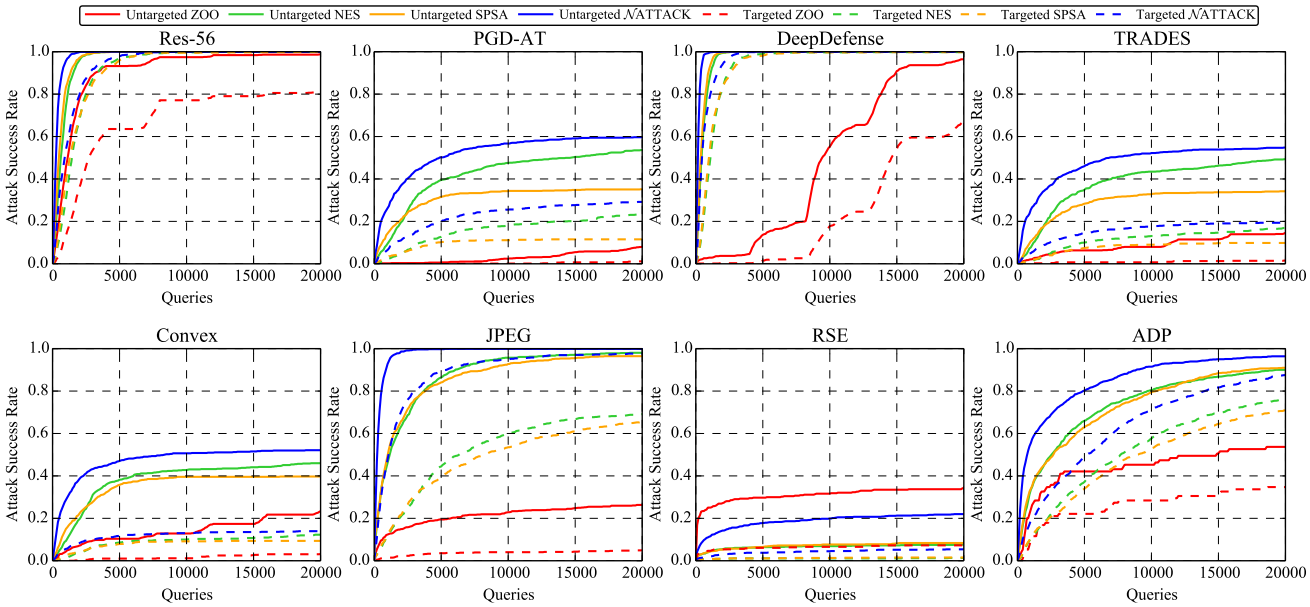

Figure 46. The *attack success rate vs. attack strength* curves of score-based attacks under the ℓ_2 norm on the 8 models on CIFAR-10.

[32] 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)*, 2019. 2

[33] 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. 2

[34] Ziang Yan, Yiwen Guo, and Changshui Zhang. Deep defense: Training dnns with improved adversarial robustness. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 2

[35] 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)*, 2019. 2

[36] Y. Zhang and P. Liang. Defending against whitebox adversarial attacks via randomized discretization. In *Artificial Intelligence and Statistics (AISTATS)*, 2019. 2

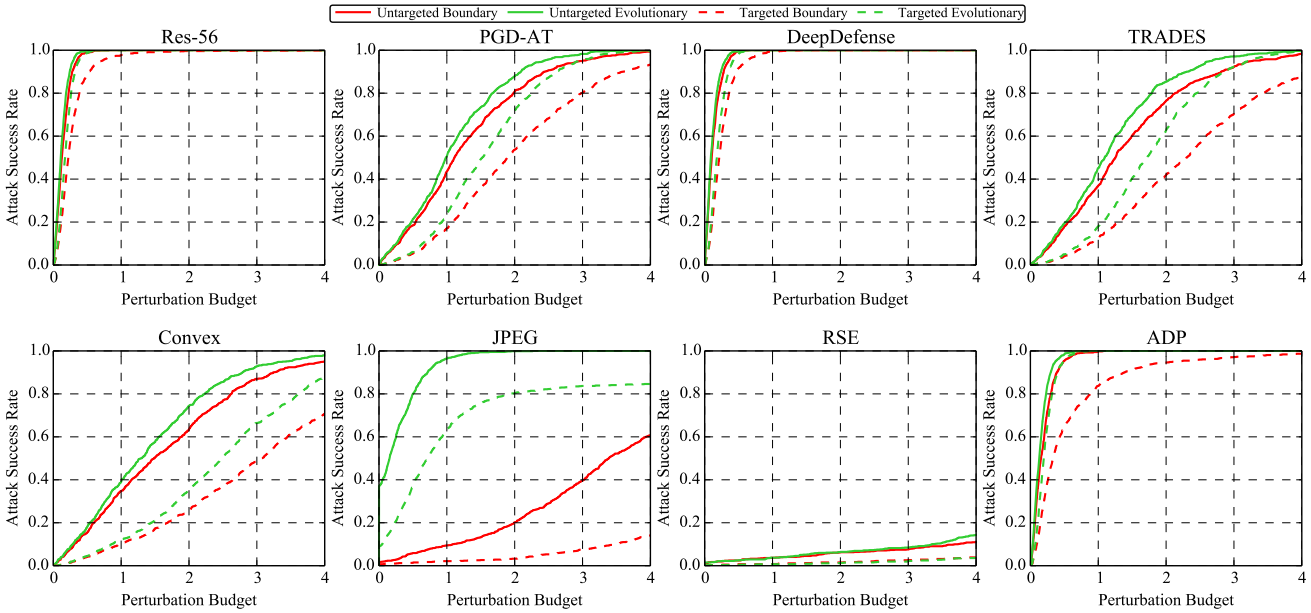

Figure 47. The *attack success rate vs. perturbation budget* curves of decision-based attacks under the ℓ_2 norm on the 8 models on CIFAR-10.

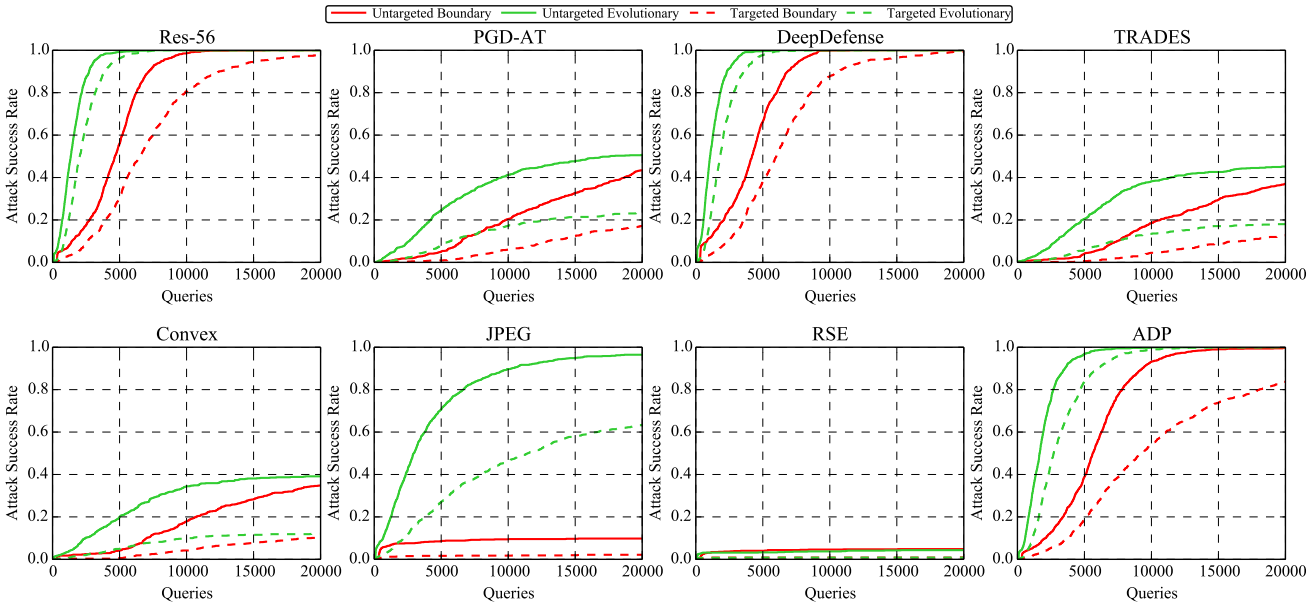

Figure 48. The *attack success rate vs. attack strength* curves of decision-based attacks under the ℓ_2 norm on the 8 models on CIFAR-10.

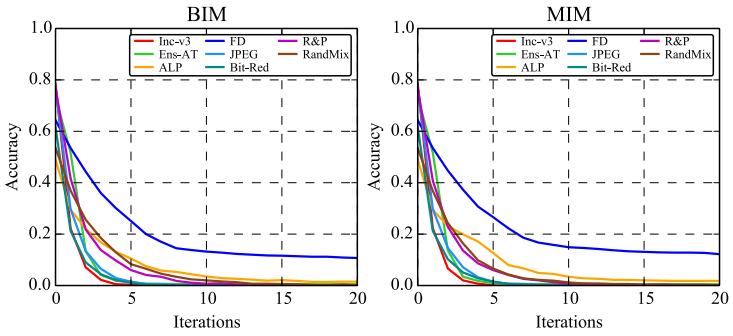

Figure 49. The *accuracy vs. attack strength* curves of the 8 models on ImageNet against untargeted white-box attacks under the ℓ_∞ norm.

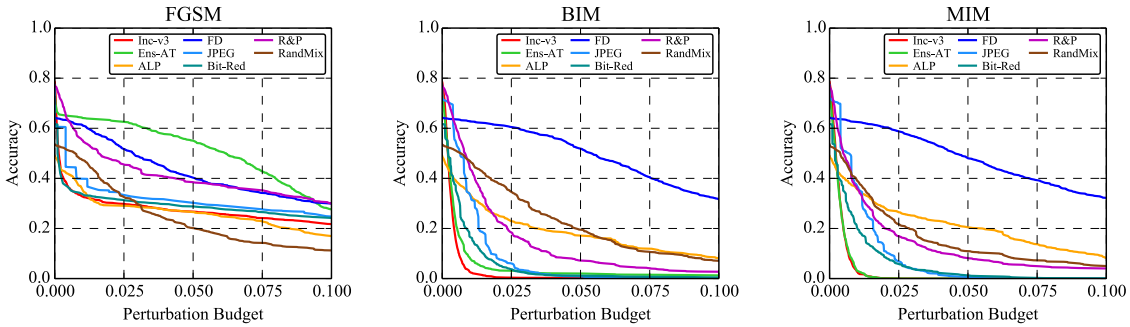

Figure 50. The *accuracy vs. perturbation budget* curves of the 8 models on ImageNet against targeted white-box attacks under the ℓ_∞ norm.

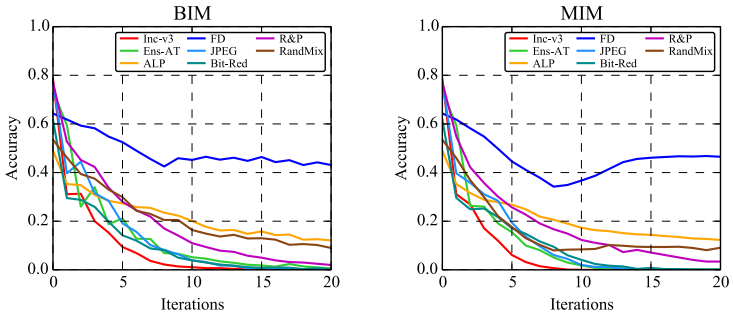

Figure 51. The *accuracy vs. attack strength* curves of the 8 models on ImageNet against targeted white-box attacks under the ℓ_∞ norm.

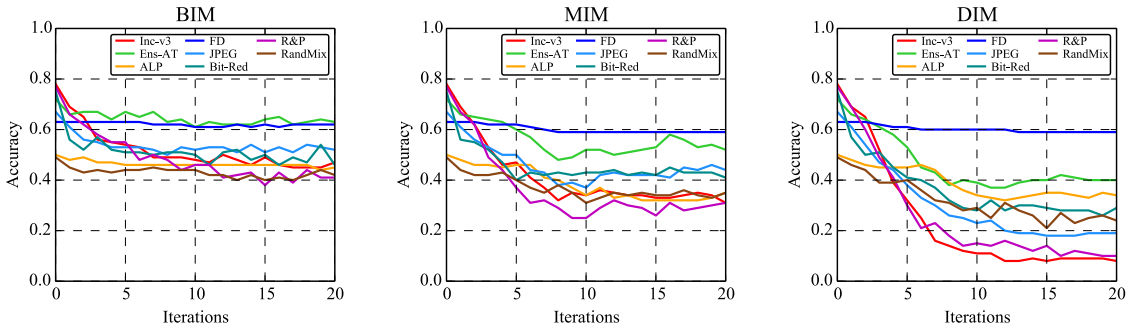

Figure 52. The *accuracy vs. attack strength* curves of the 8 models on ImageNet against untargeted transfer-based attacks under the ℓ_∞ norm.

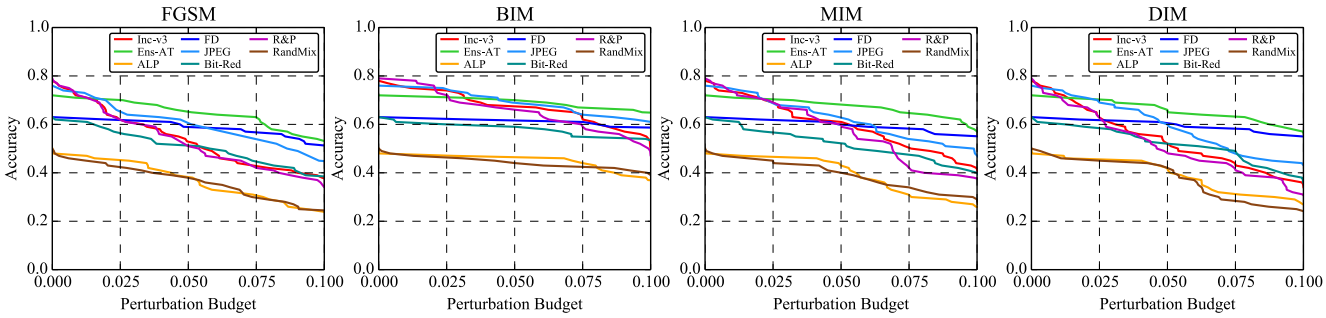

Figure 53. The *accuracy vs. perturbation budget* curves of the 8 models on ImageNet against targeted transfer-based attacks under the ℓ_∞ norm.

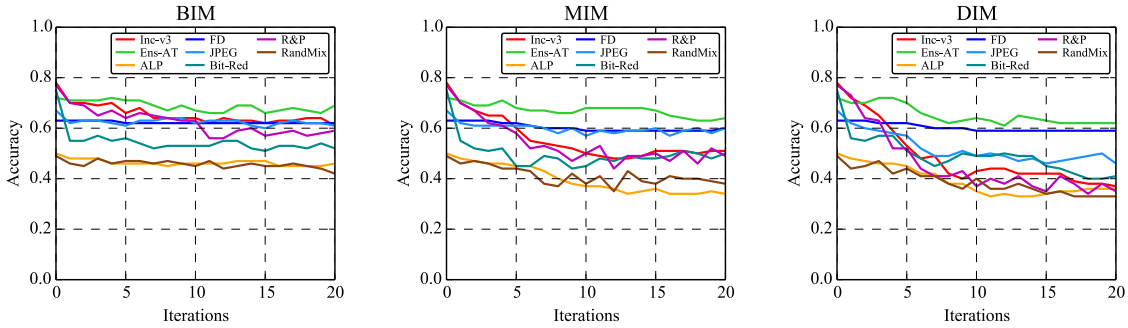

Figure 54. The *accuracy vs. attack strength* curves of the 8 models on ImageNet against targeted transfer-based attacks under the ℓ_∞ norm.

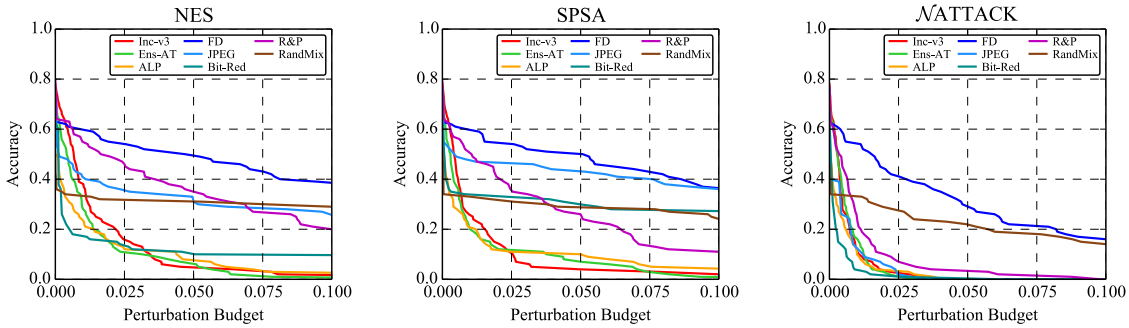

Figure 55. The *accuracy vs. perturbation budget* curves of the 8 models on ImageNet against targeted score-based attacks under the ℓ_∞ norm.

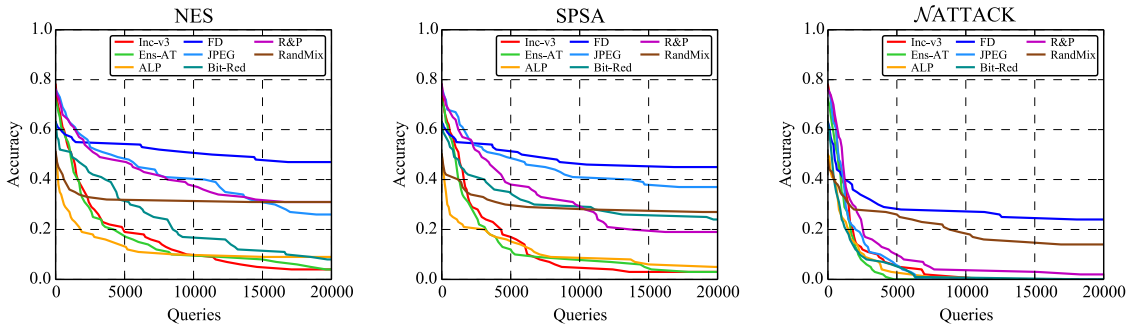

Figure 56. The *accuracy vs. attack strength* curves of the 8 models on ImageNet against targeted score-based attacks under the ℓ_∞ norm.

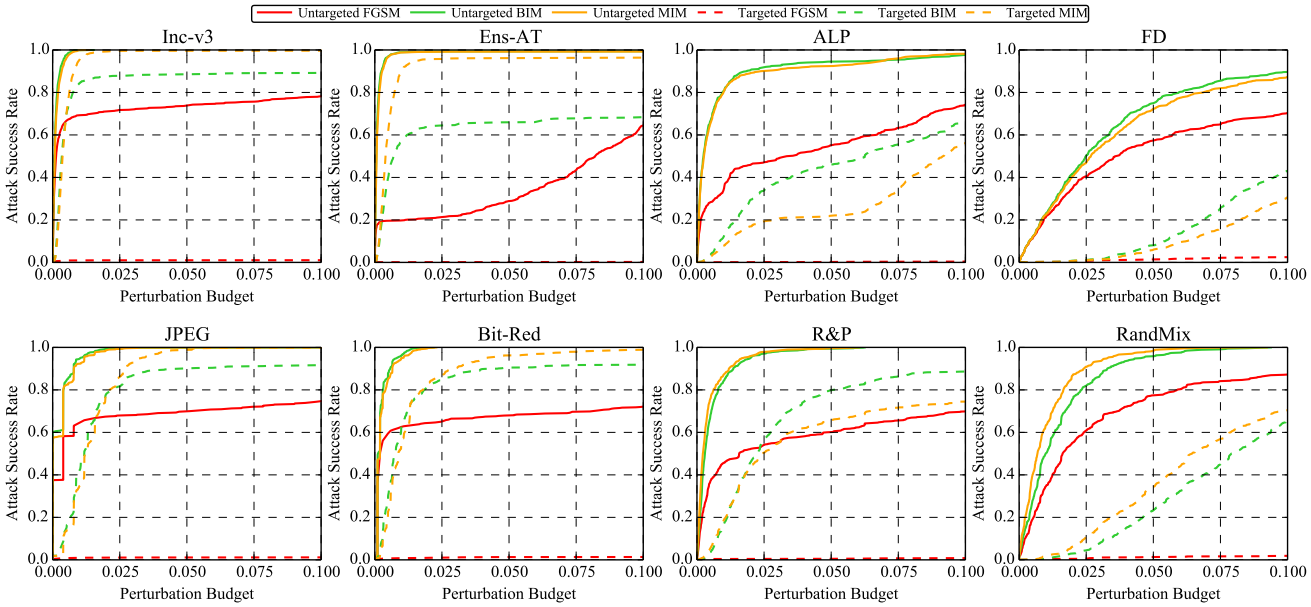

Figure 57. The *attack success rate vs. perturbation budget* curves of white-box attacks under the ℓ_∞ norm on the 8 models on ImageNet.

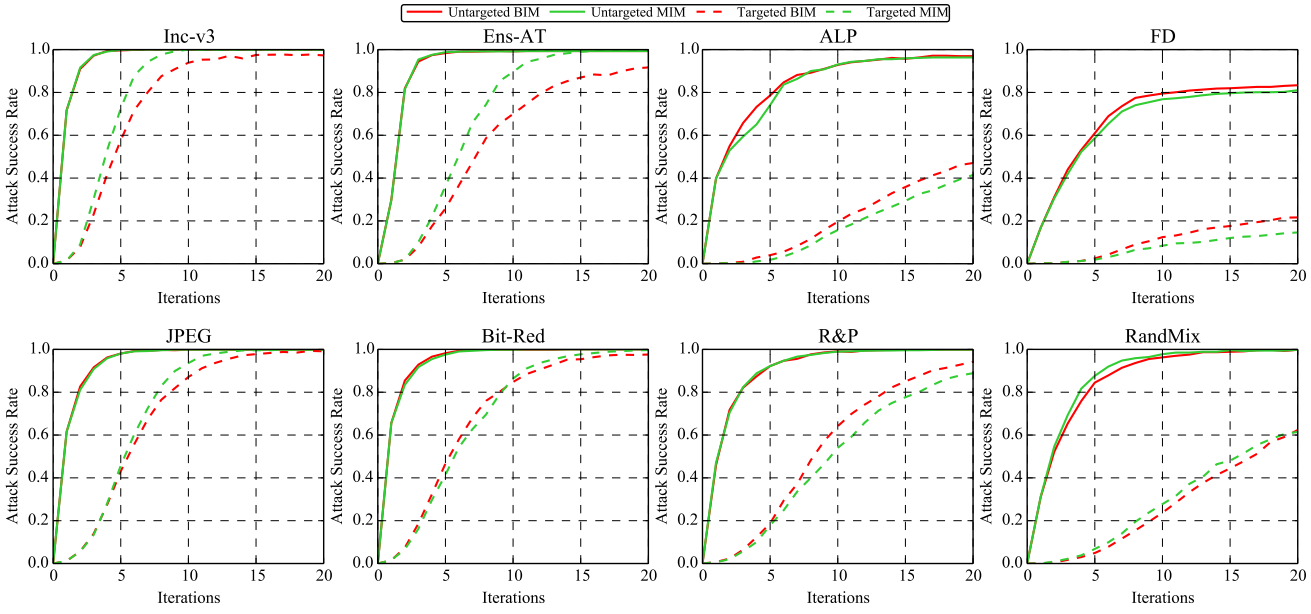

Figure 58. The *attack success rate vs. attack strength* curves of white-box attacks under the ℓ_∞ norm on the 8 models on ImageNet.

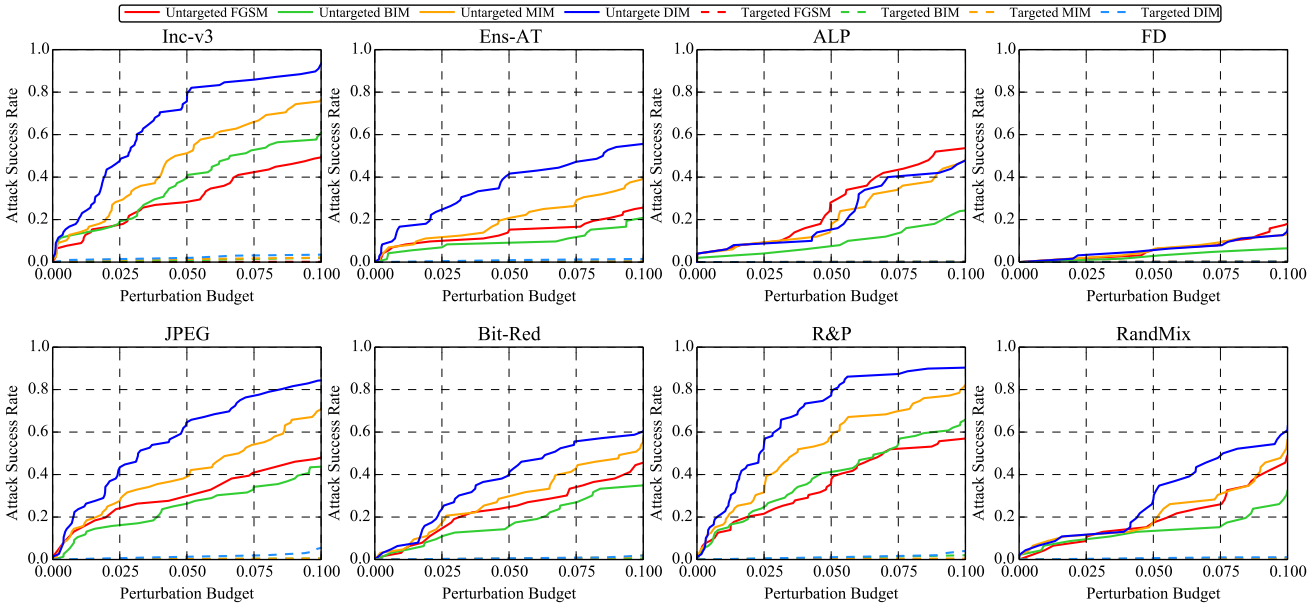

Figure 59. The *attack success rate vs. perturbation budget* curves of transfer-based attacks under the ℓ_∞ norm on the 8 models on ImageNet.

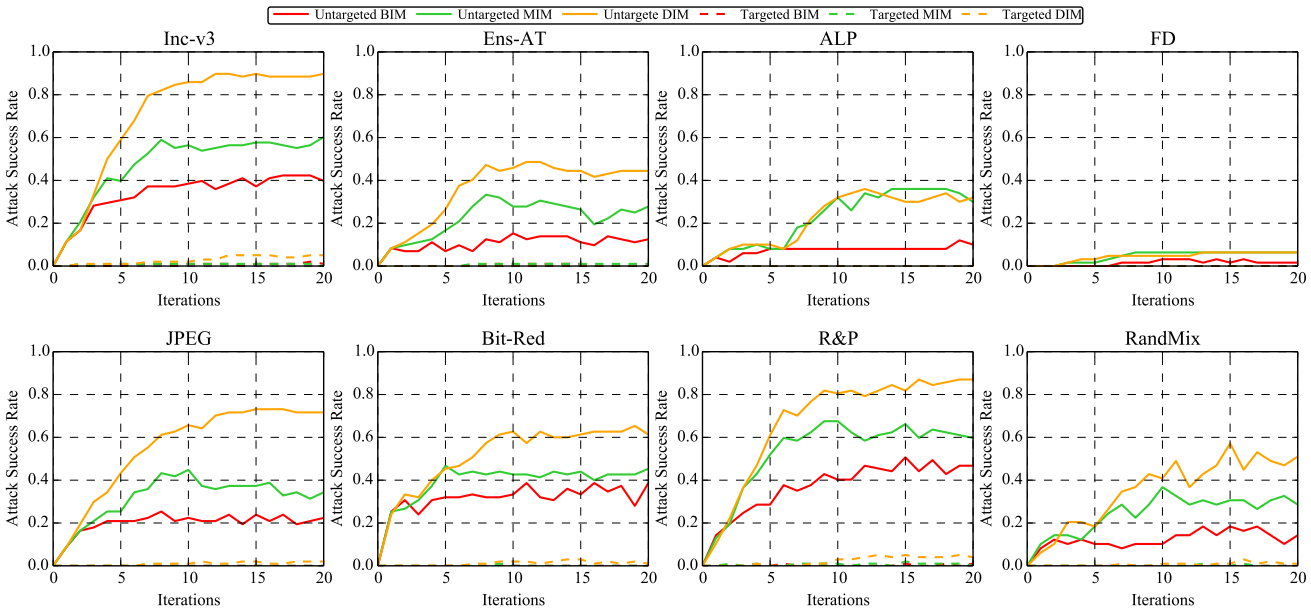

Figure 60. The *attack success rate vs. attack strength* curves of transfer-based attacks under the ℓ_∞ norm on the 8 models on ImageNet.

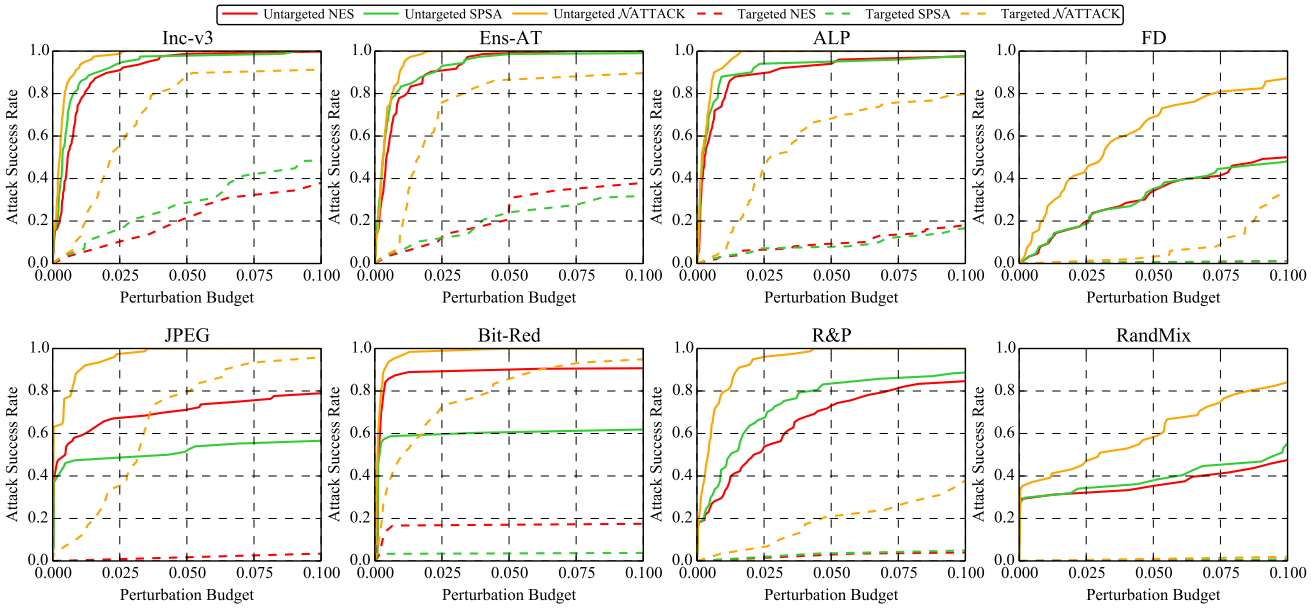

Figure 61. The *attack success rate vs. perturbation budget* curves of score-based attacks under the ℓ_∞ norm on the 8 models on ImageNet.

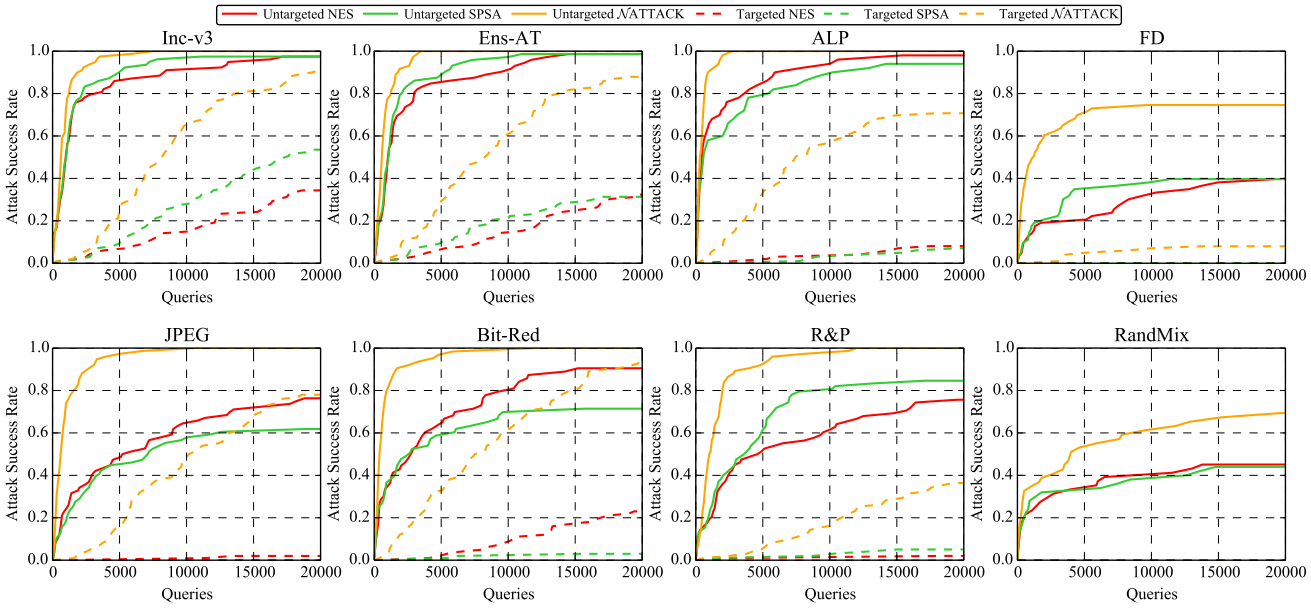

Figure 62. The *attack success rate vs. attack strength* curves of score-based attacks under the ℓ_∞ norm on the 8 models on ImageNet.

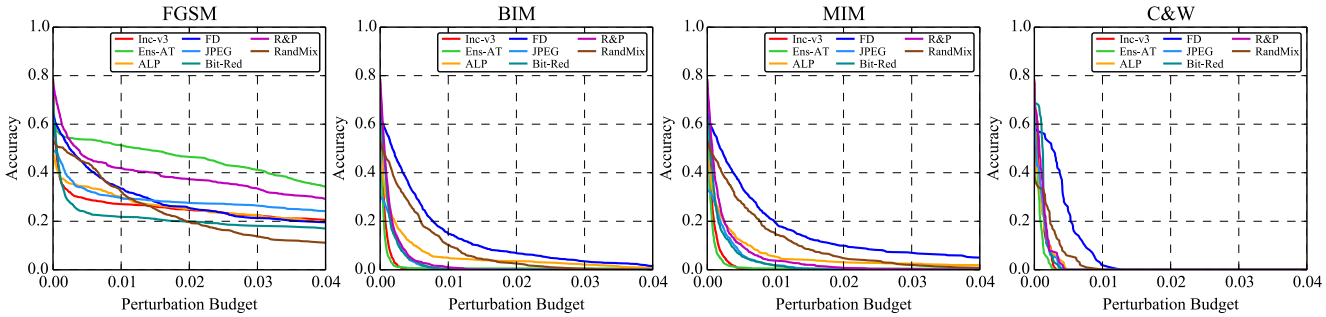

Figure 63. The *accuracy vs. perturbation budget* curves of the 8 models on ImageNet against untargeted white-box attacks under the ℓ_2 norm.

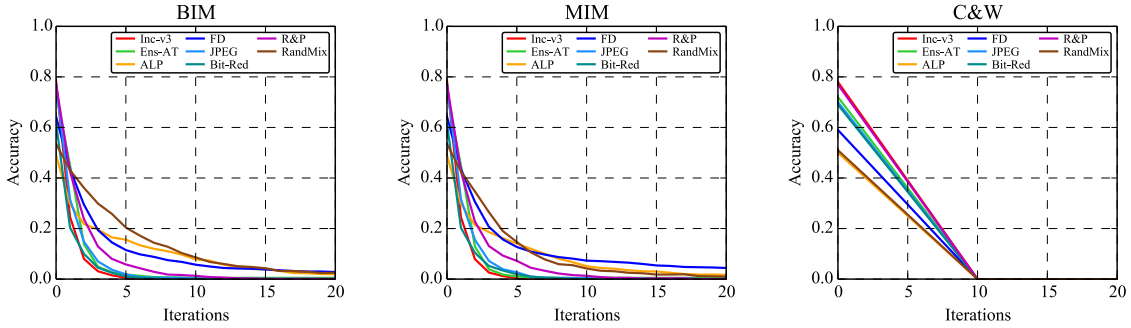

Figure 64. The *accuracy vs. attack strength* curves of the 8 models on ImageNet against untargeted white-box attacks under the ℓ_2 norm.

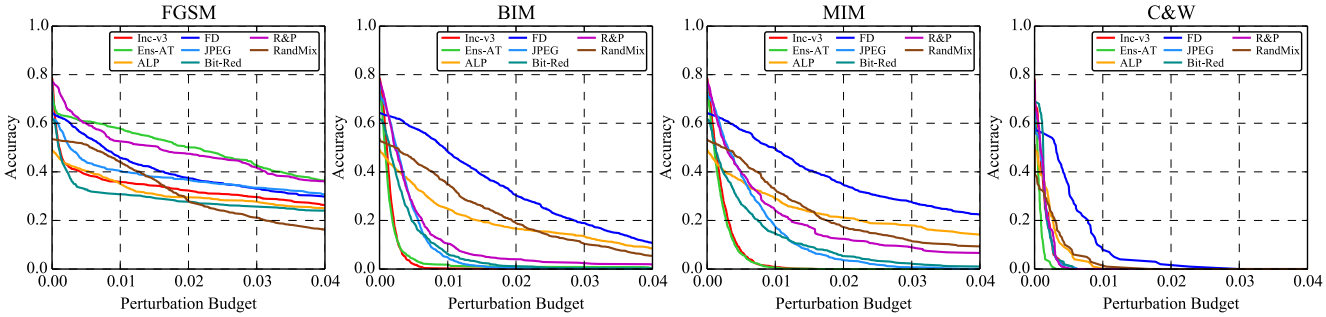

Figure 65. The *accuracy vs. perturbation budget* curves of the 8 models on ImageNet against targeted white-box attacks under the ℓ_2 norm.

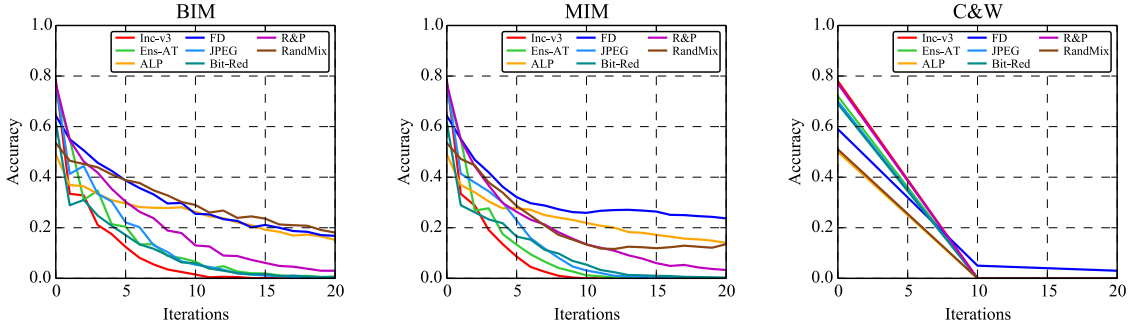

Figure 66. The *accuracy vs. attack strength* curves of the 8 models on ImageNet against targeted white-box attacks under the ℓ_2 norm.

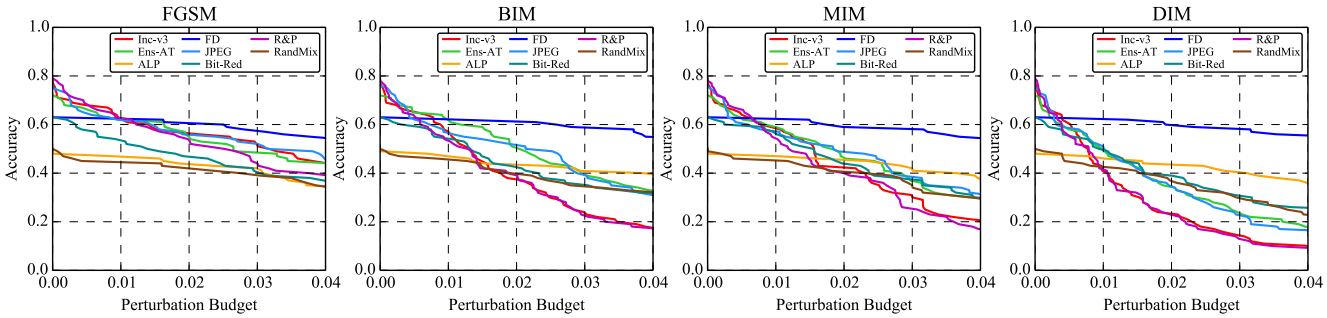

Figure 67. The *accuracy vs. perturbation budget* curves of the 8 models on ImageNet against untargeted transfer-based attacks under the ℓ_2 norm.

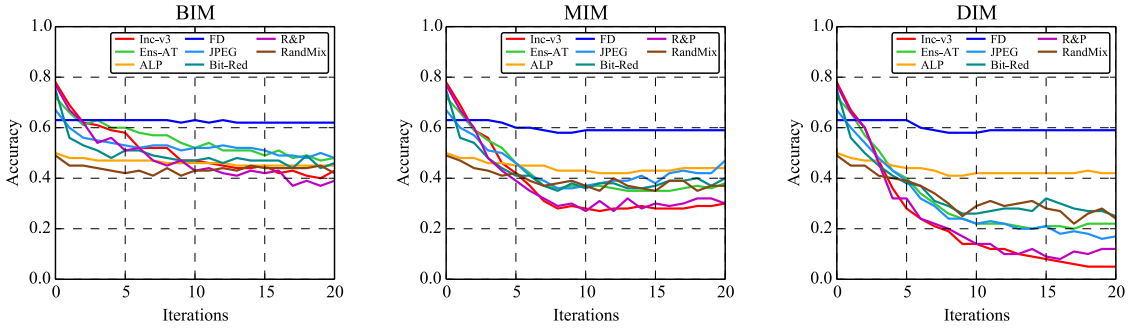

Figure 68. The *accuracy vs. attack strength* curves of the 8 models on ImageNet against untargeted transfer-based attacks under the ℓ_2 norm.

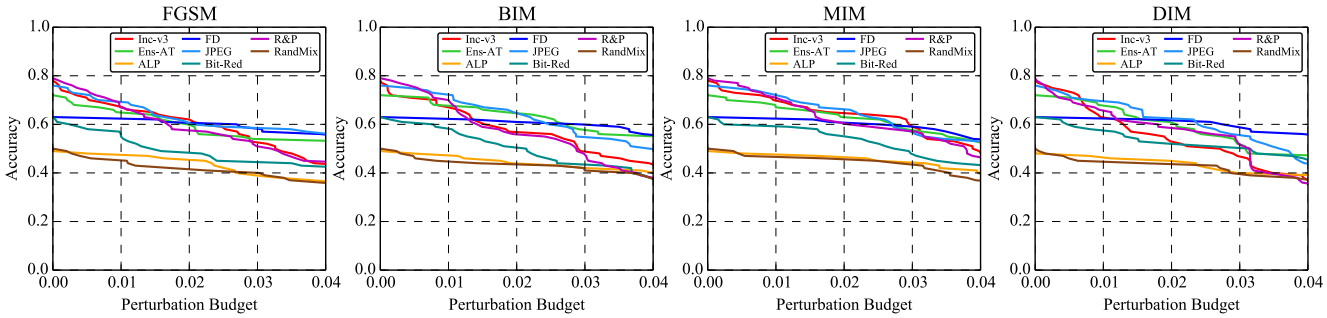

Figure 69. The *accuracy vs. perturbation budget* curves of the 8 models on ImageNet against targeted transfer-based attacks under the ℓ_2 norm.

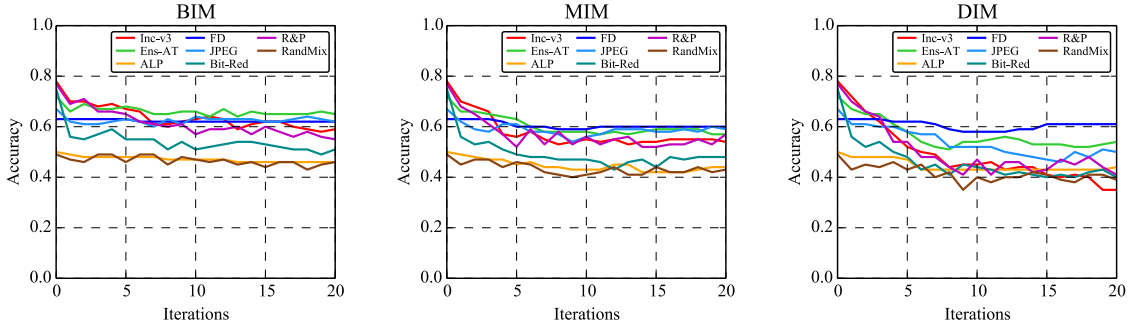

Figure 70. The *accuracy vs. attack strength* curves of the 8 models on ImageNet against targeted transfer-based attacks under the ℓ_2 norm.

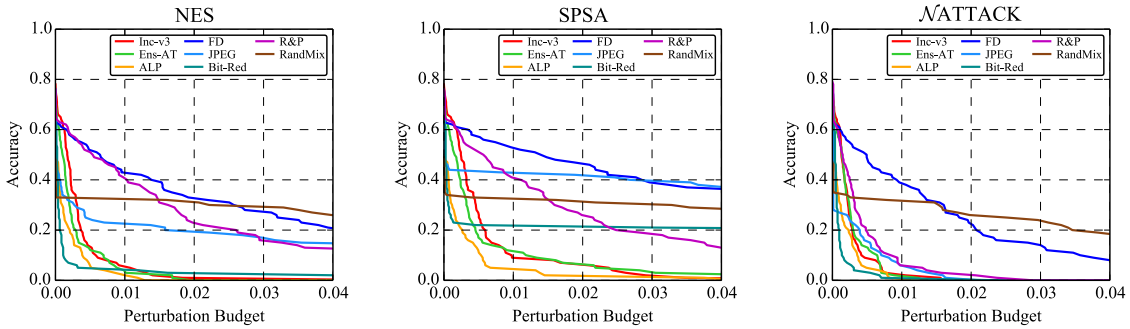

Figure 71. The *accuracy vs. perturbation budget* curves of the 8 models on ImageNet against untargeted score-based attacks under the ℓ_2 norm.

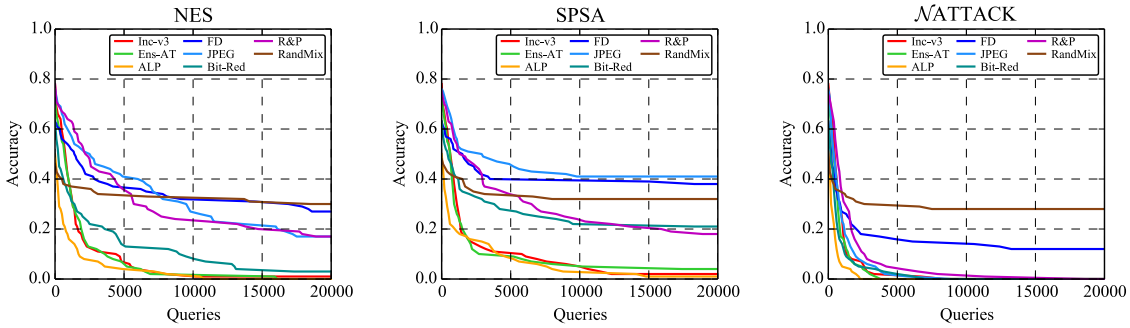

Figure 72. The *accuracy vs. attack strength* curves of the 8 models on ImageNet against untargeted score-based attacks under the ℓ_2 norm.

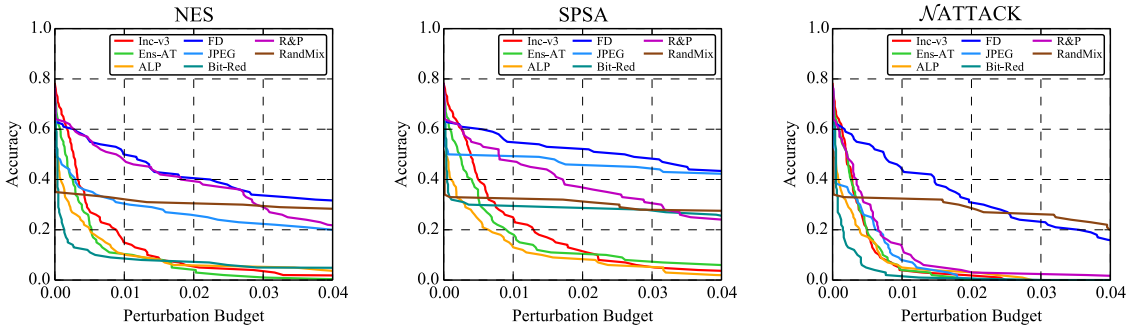

Figure 73. The *accuracy vs. perturbation budget* curves of the 8 models on ImageNet against targeted score-based attacks under the ℓ_2 norm.

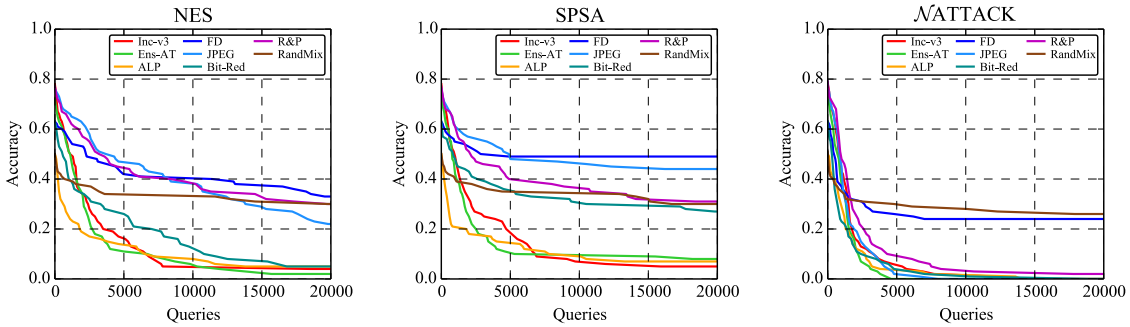

Figure 74. The *accuracy vs. attack strength* curves of the 8 models on ImageNet against targeted score-based attacks under the ℓ_2 norm.

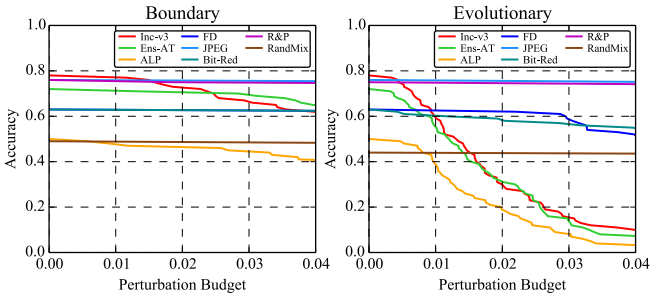

Figure 75. The *accuracy vs. perturbation budget* curves of the 8 models on ImageNet against targeted decision-based attacks under the ℓ_2 norm.

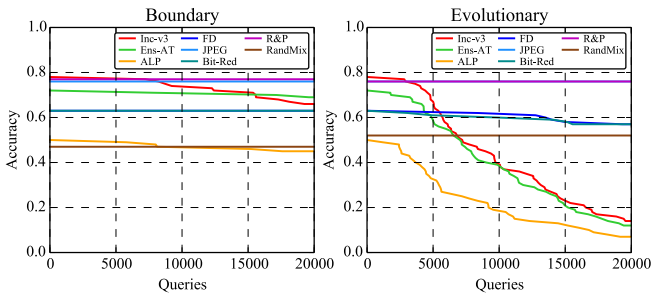

Figure 76. The *accuracy vs. attack strength* curves of the 8 models on ImageNet against targeted decision-based attacks under the ℓ_2 norm.

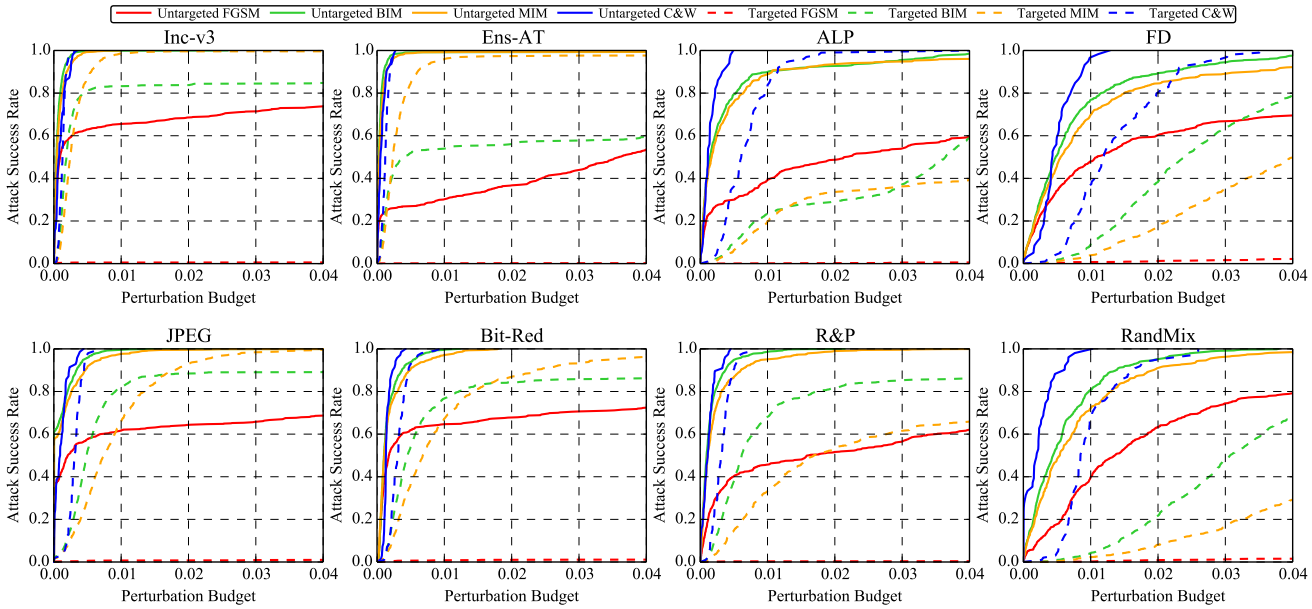

Figure 77. The *attack success rate vs. perturbation budget* curves of white-box attacks under the ℓ_2 norm on the 8 models on ImageNet.

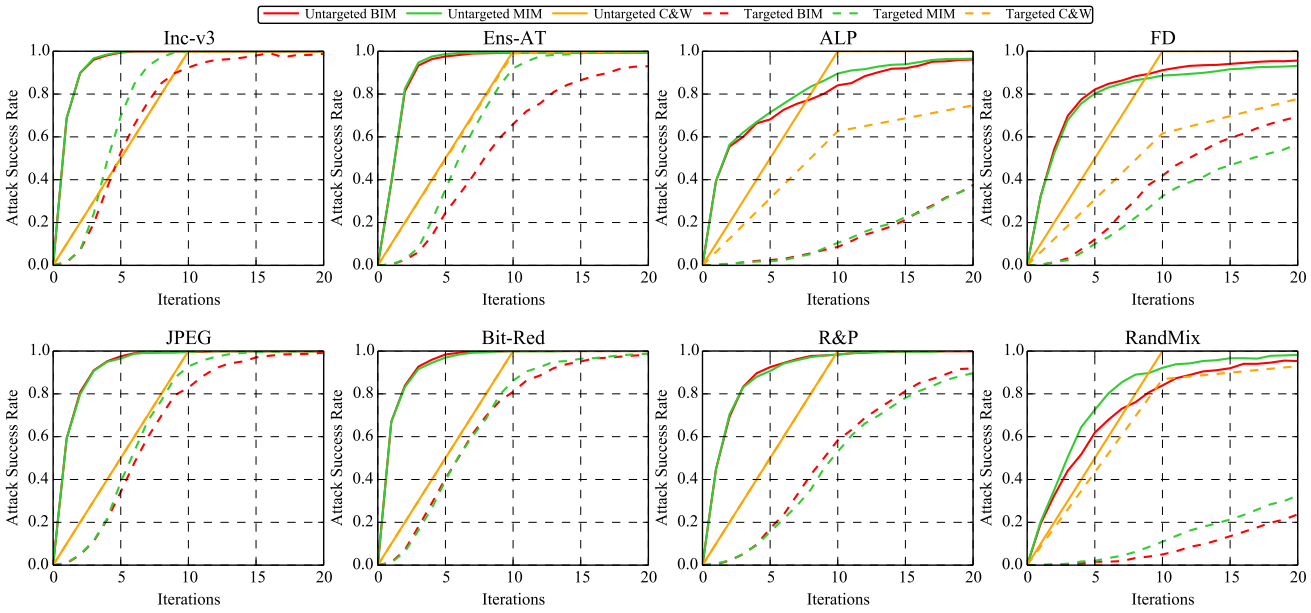

Figure 78. The *attack success rate vs. attack strength* curves of white-box attacks under the ℓ_2 norm on the 8 models on ImageNet.

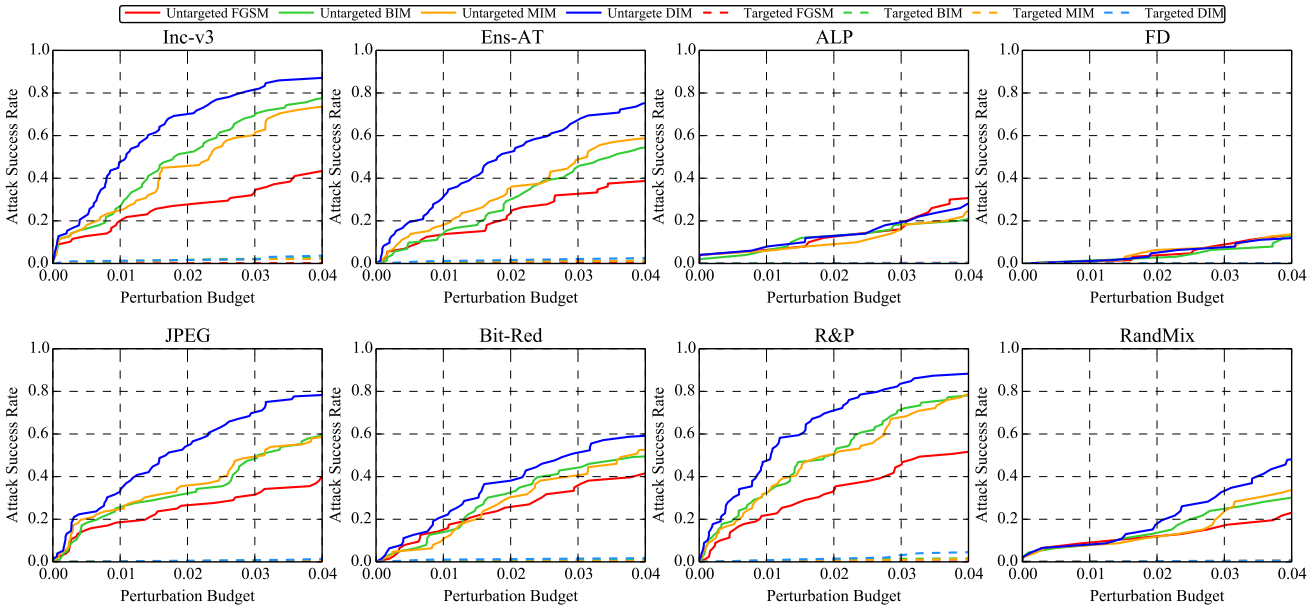

Figure 79. The *attack success rate vs. perturbation budget* curves of transfer-based attacks under the ℓ_2 norm on the 8 models on ImageNet.

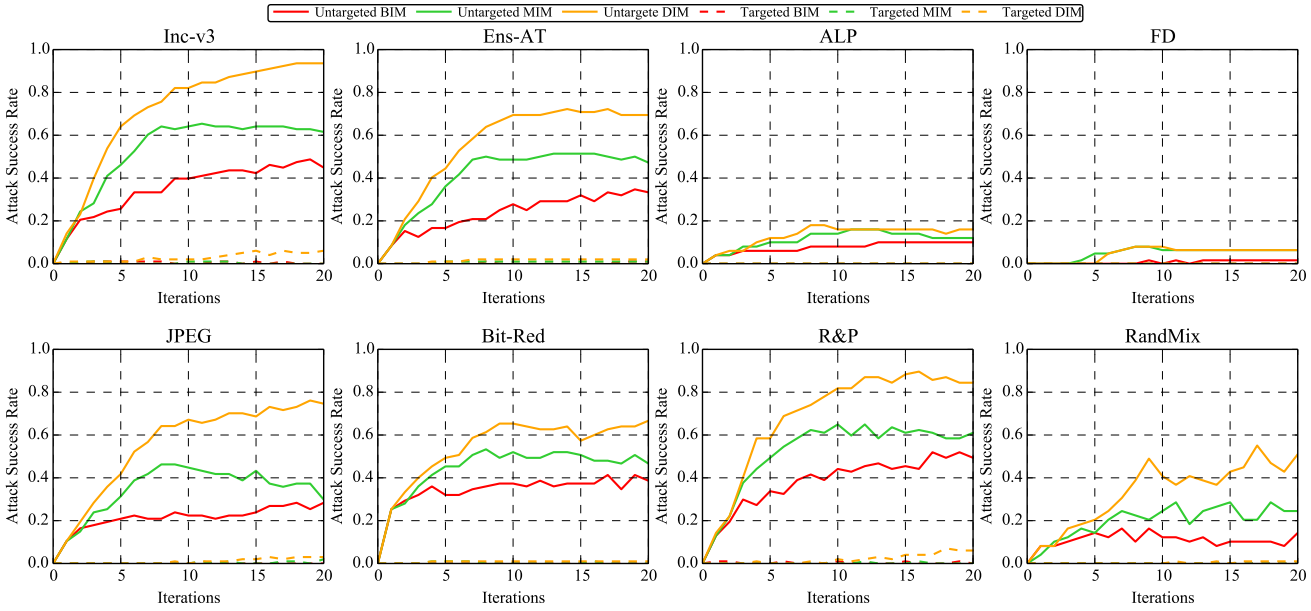

Figure 80. The *attack success rate vs. attack strength* curves of transfer-based attacks under the ℓ_2 norm on the 8 models on ImageNet.

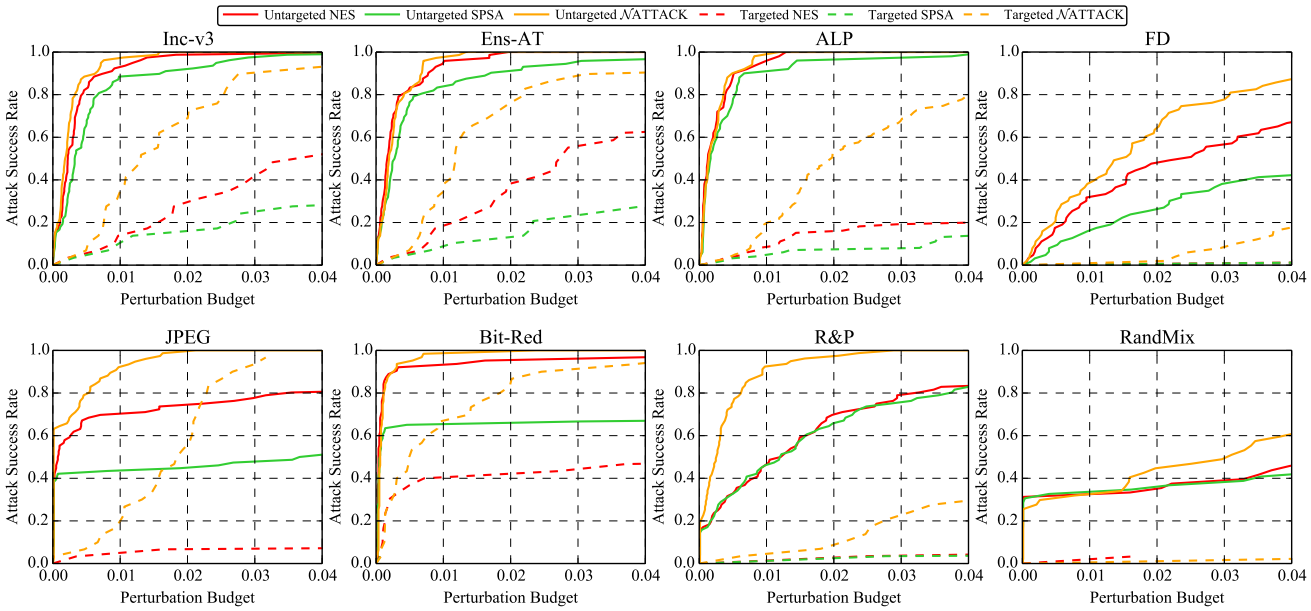

Figure 81. The *attack success rate vs. perturbation budget* curves of score-based attacks under the ℓ_2 norm on the 8 models on ImageNet.

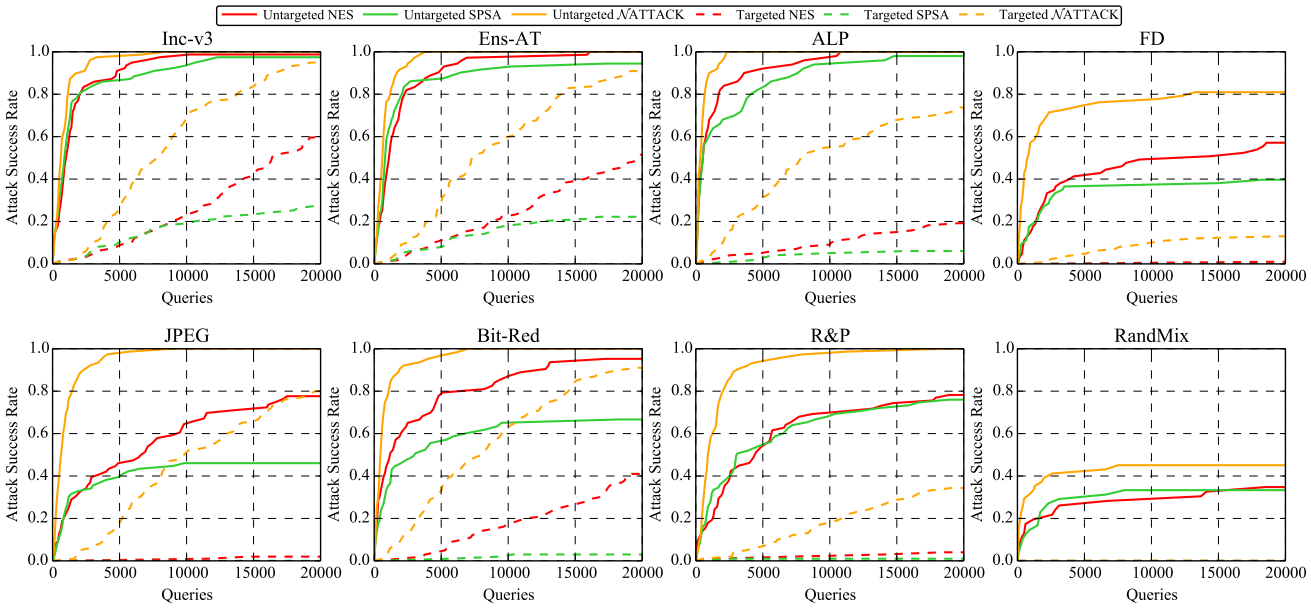

Figure 82. The *attack success rate vs. attack strength* curves of score-based attacks under the ℓ_2 norm on the 8 models on ImageNet.

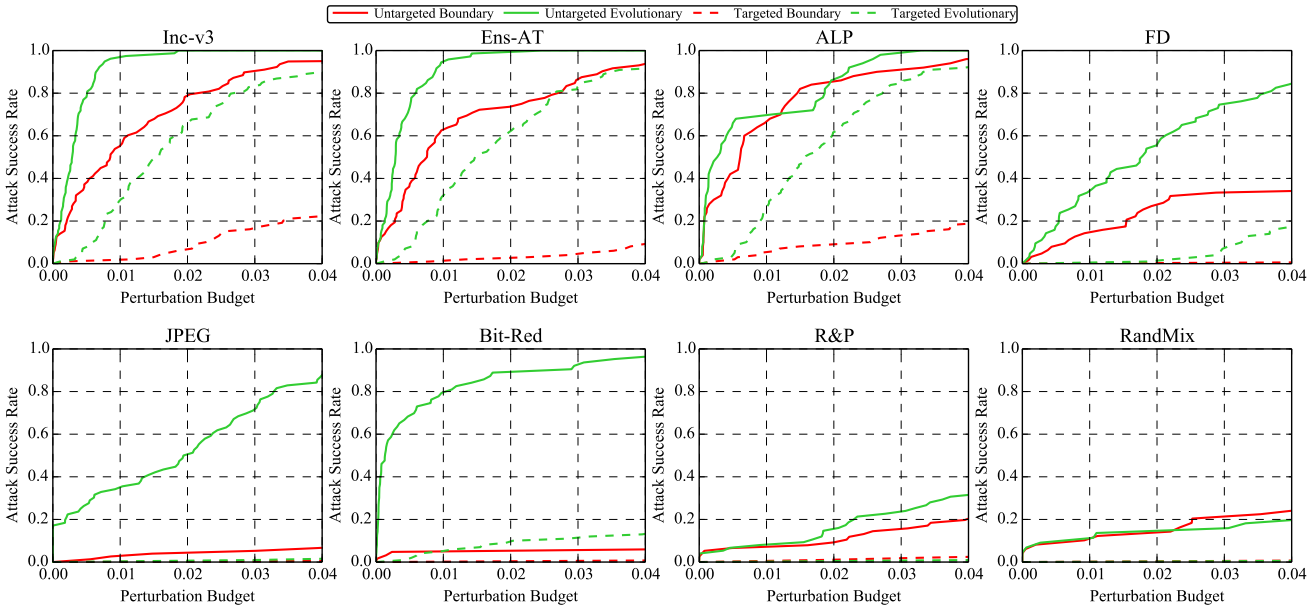

Figure 83. The *attack success rate vs. perturbation budget* curves of decision-based attacks under the ℓ_2 norm on the 8 models on ImageNet.

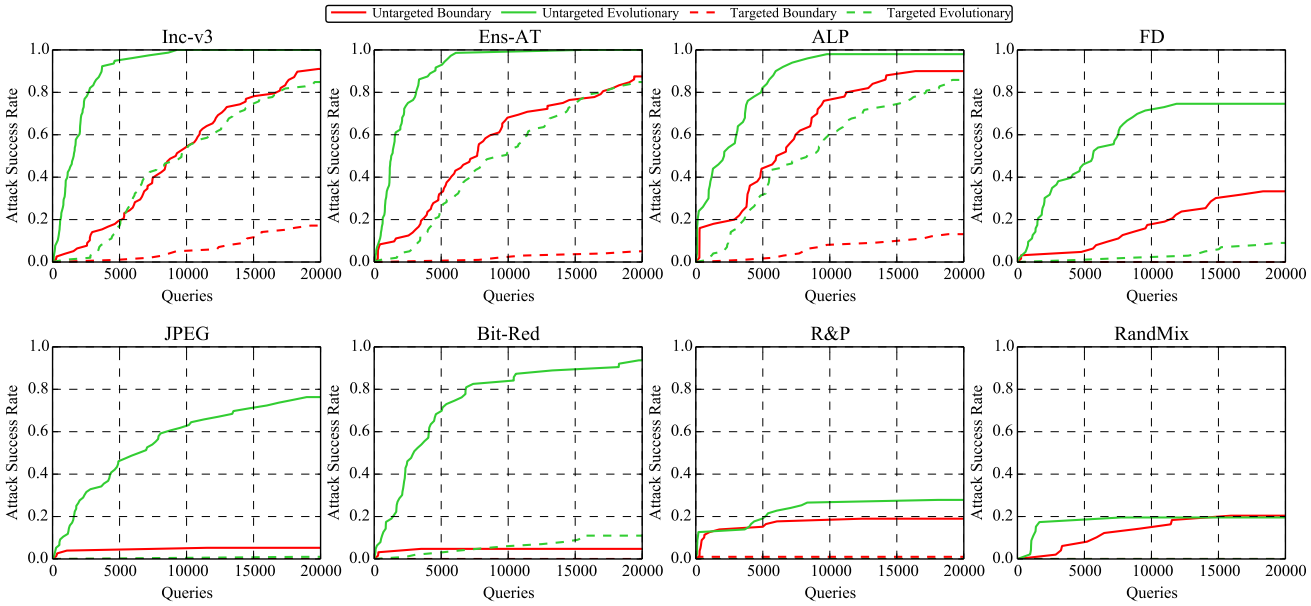

Figure 84. The *attack success rate vs. attack strength* curves of decision-based attacks under the ℓ_2 norm on the 8 models on ImageNet.