A. Comparison of Gradient-free Attacks

In our initial experiments, we evaluated several different gradient-free attacks, including SPSA, natural evolutionary strategies (NES) (Wierstra et al., 2008; Ilyas et al., 2017), and zero-order optimization (ZOO) (Chen et al., 2017). For ZOO, we used ZOO-ADAM, which uses coordinate descent with coordinate-wise Adam updates (Kingma and Ba, 2014), without the importance sampling or feature reduction tricks used on ImageNet.

All our initial experiments on ImageNet were run against a ResNet-50 model, which achieves 75% accuracy, with 100 randomly selected test images. All experiments on CIFAR-10 used a VGG-like fully convolutional model, which achieves 94.5% accuracy, with 1000 randomly selected test images.

Overall attack success rate: On CIFAR-10, all attacks drive accuracy to 0%, using less than 32768 model evaluations, and often much fewer, *e.g.* SPSA decreases accuracy to less than 5% after 2048 model evaluations. Attacks on ImageNet tended to require substantially more model evaluations to drive accuracy to 0%, likely due to the high dimensionality of the input images. Results on ImageNet are summarized below in Table 2.

Batch size	SPSA	NES	ZOO
256	29%	30%	72%
512	21%	22%	69%
2048	2%	2%	43%
8192	0%	0%	10%

Table 2: **Attack success rate with increasing computation** We show the accuracy of ResNet-50 against each adversary, with varying amounts of computation. Each attack is run for a maximum of 300 iterations, but with varying batch sizes. In other words, we vary the number of finite difference estimates used before applying each gradient estimate. With sufficient computation, both NES and SPSA decrease accuracy to 0% on ImageNet with $\epsilon = 2$.

Convergence speed: Figure 2 provides a different view on the same data and shows the amount of computation before the adversary finds a misclassified image.

For our main experiments, we use a fixed large batch size because our aim is to produce models robust to the strongest possible attacks. In particular, the hyperparameters are selected to reliably generate adversarial examples, rather than to be query-efficient, but in situations where query efficiency is a factor, the same algorithms can be tuned by, for example, reducing the batch size.

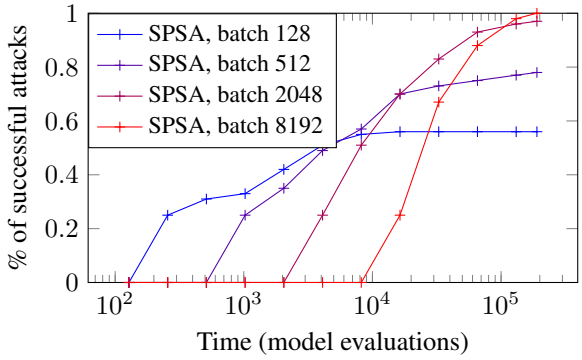


Figure 2: **Model evaluations before finding misclassified adversarial example**: Although attacks with larger batch sizes require longer before beginning to identify adversarial examples, over time, they more reliably identify adversarial examples. For fair comparison, attacks with smaller batch sizes uses more iterations so that the number of model evaluations is constant across attacks. We use a log scale axis since most attacks succeed significantly sooner than the maximum number of iterations.

B. Hyperparameters

We evaluated all attacks and defenses on 1000 images randomly sampled from either the CIFAR-10 or the ImageNet test set, with the exception of PixelDefend which we could only evaluate on 100 images due to a limited computational budget. We ran attacks for a maximum of 100 iterations, and stopped when the margin-based objective in Eq. (6) was less than -5.0. In many cases, the attack completes fairly quickly, in fewer than ten iterations. For all of our optimization-based attacks, we use random initializations by sampling a perturbation from the ℓ_∞ ball and projecting back onto the interval $[0, 255]$ (to ensure the resulting image is valid). Hyperparameters are provided below. We have additionally made our implementation avaliable through the open-source library *cleverhans* (Papernot et al., 2018).

Perturbation size δ	0.01
Adam learning rate	0.01
Maximum iterations	100
Batch size	8192

Table 3: **Hyperparameters for SPSA attack**

C. Discussion of Adversarial Training

As discussed in the paper, gradient-free adversaries allow us to investigate the degree of gradient masking in adversarially trained networks. Gradients estimated based on finite perturbations, rather than infinitesimal ones, may result in qualitatively different behavior (for example, in the case of high-frequency oscillations in the loss surface). Thus, SPSA may, in some cases, converge to better minima than techniques which purely use analytic gradients.

Section 6.4 verified that the expected surrogate adversarial risk (averaged over the test set) cannot be decreased with SPSA. We also investigated whether SPSA could identify better minima than projected gradient descent (PGD) for certain data points. We found that SPSA and PGD found similarly adversarial perturbations for almost all images, which provides further evidence that, in adversarially trained networks, perturbations found by PGD may be nearly worst-case possible perturbations, in which case the surrogate adversarial risk measured against a PGD adversary will be close to the true adversarial risk. These results are summarized in Figure 3.

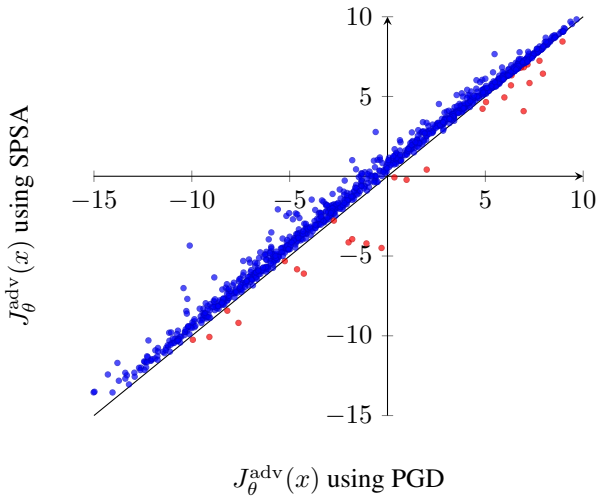
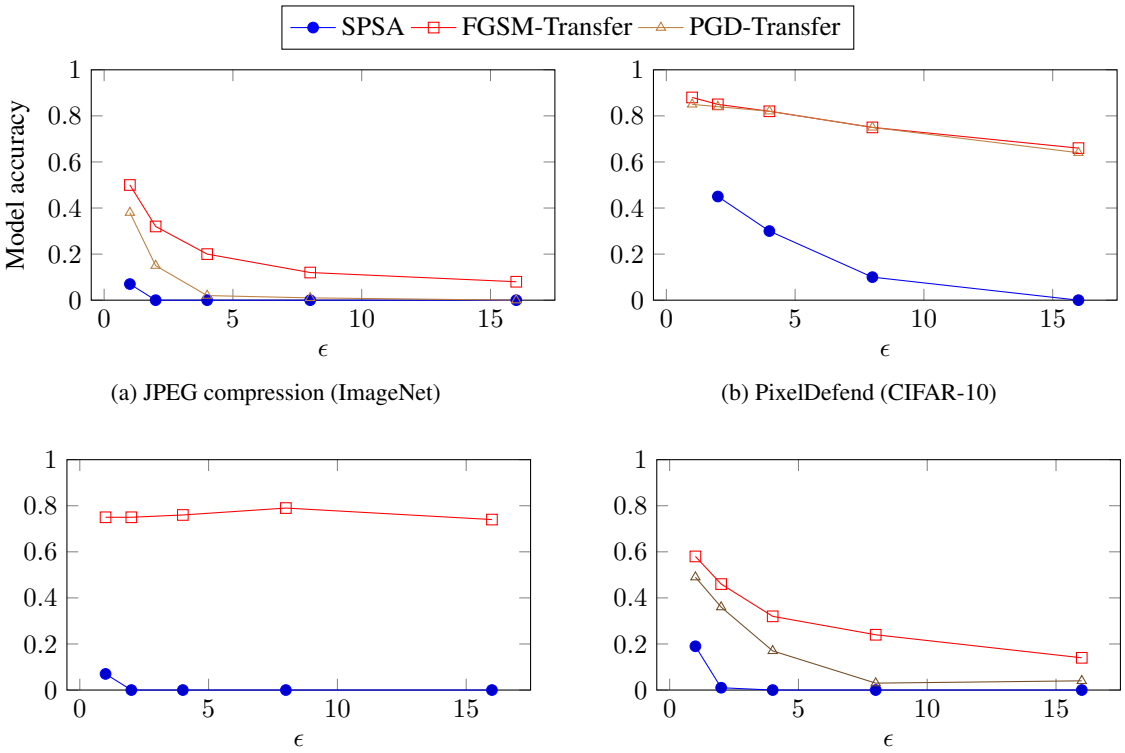


Figure 3: **Analysis of gradient-free masking in adversarially trained networks**: We compare the final values of the margin-based objective across different images, after using projected gradient descent or SPSA. Each point represents a single image, and is misclassified when $J_\theta^{\text{adv}}(x) < 0$. Points close to the line $y = x$ indicate that SPSA and PGD identified similarly adversarial perturbations. Points below the line, shown in red, indicate those for which SPSA identified stronger adversarial perturbation than PGD. Overall, SPSA and PGD identify comparably adversarial perturbations, and there are few points where SPSA identifies significantly stronger adversarial perturbations than PGD.

D. Additional Experiments

In Figure 4, we include complete experimental results evaluating each of the (non-adversarial training) defenses across a range of perturbation sizes for a number of different attack methods. Although all models shows significant robustness against the original evaluation adversaries, the accuracy of all models falls to near zero when evaluated against stronger attacks.



(a) JPEG compression (ImageNet)

(b) PixelDefend (CIFAR-10)

(c) High-level Guided Denoising (ImageNet)

(d) Stochasticity (ImageNet)

Figure 4: Models can be obscured to gradient-based and transfer-based attacks by adding effectively non-differentiable operations and using purification to change the model's decision boundaries compared to the original model. However, this does not remove all adversarial examples – using stronger attacks, we can reduce the accuracy of all defenses to near zero.