DISCO: Adversarial Defense with Local Implicit Functions

Anonymous Author(s)
Affiliation
Address
email

Abstract

The problem of adversarial defenses for image classification, where the goal is to robustify a classifier against adversarial examples, is considered. Inspired by the hypothesis that these examples lie beyond the natural image manifold, a novel *aDversarIal defenSe with local impliCit functiOns* (DISCO) is proposed to remove adversarial perturbations by localized manifold projections. DISCO consumes an adversarial image and a query pixel location and outputs a clean RGB value at the location. It is implemented with an encoder and a local implicit module, where the former produces per-pixel deep features and the latter uses the features in the neighborhood of query pixel for predicting the clean RGB value. Extensive experiments demonstrate that both DISCO and its cascade version outperform prior defenses, regardless of whether the defense is known to the attacker. DISCO is also shown to be data and parameter efficient and to mount defenses that transfers across datasets, classifiers and attacks.[1]

1 Introduction

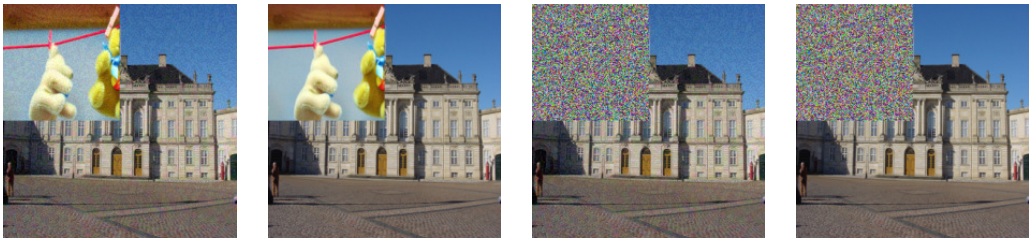
(a) Clean Image (b) Adversarial Image (c) DISCO Output

Figure 1: Qualitative performance of DISCO output of a randomly selected ImageNet [21] image.

It has long been hypothesized that vision is only possible because the natural world contains substantial statistical regularities, which are exploited by the vision system to overcome the difficulty of scene understanding [6, 34, 84, 27, 106, 28, 41, 92, 97, 72, 7, 102]. Under this hypothesis, natural images form a low-dimension manifold in image space, denoted as the *image manifold*, to which human vision is highly tuned. While deep neural networks (DNNs) [100, 37, 124, 101, 93] aim to classify natural images with human-like accuracy, they have been shown prone to adversarial attacks that, although imperceptible to humans, significantly decrease their performance. As shown

[1] All code will be made publicly available upon paper acceptance.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.



(a) (b) (c) (d)

Figure 2: DISCO is a conditional model of local image statistics. It performs a local manifold projection per pixel neighborhood, conditional on feature vectors of the adversarial image. This is critical to enable learning from limited data while achieving the model expressiveness needed to precisely control the manifold projection. (a) A mixed image of 2 adversarial images. (b) DISCO output for (a). (c) A mixed images of an adversarial image and noise. (d) DISCO output for (c).

in Fig. 1 (a) and (b), these attacks typically consist of adding an imperceptible perturbation, which can be generated in various manners [31, 64, 11, 57], to the image. Over the past few years, the introduction of more sophisticated attacks has exposed a significant vulnerability of DNNs to this problem [98, 23, 22, 17, 18, 89]. In fact, it has been shown that adversarial examples crafted with different classifiers and optimization procedures even transfer across networks [20, 83, 117, 43, 76].

A potential justification for the success of adversarial attacks and their transferability is that they create images that lie just barely outside the image manifold [95, 88, 58, 30, 51, 5]. We refer to these images as *barely outliers*. While humans have the ability to project these images into the manifold, probably due to an history of training under adverse conditions, such as environments with low-light or ridden with occlusions, this is not the case for current classifiers. A key factor to the success of the human projection is likely the accurate modeling of the image manifold. Hence, several defenses against adversarial attacks are based on models of natural image statistics. These are usually global image representations [66, 99, 88, 123] or conditional models of image pixel statistics [95, 5, 105, 86, 53]. For example, PixelDefend [95] and HPD [5] project malicious images into the natural image manifold using the PixelCNN model [53], which predicts a pixel value conditioned on the remainder of the image. However, these strategies [95, 5, 66, 99, 88] can be easily defeated by existing attacks. We hypothesize that this is due to the difficulty of learning generative image models, which require *global* image modelling, a highly complex task. It is well known that the synthesis of realistic natural images requires very large model sizes and training datasets [49, 8, 50]. Even with these, it is not clear that the manifold is modeled in enough detail to defend adversarial attacks.

In this work, we argue that, unlike image synthesis, the manifold projection required for adversarial defense is a *conditional* operation: the synthesis of a natural image *given* the perturbed one. Assuming that the attack does not alter the *global* structure of the image (which would likely not be *imperceptible* to humans) it should suffice for this function to be a conditional model of *local* image statistics. We argue that this conditional modeling can be implemented with an implicit function [13, 94, 68, 85, 71, 52, 46, 14], where the network learns a conditional representation of the image appearance in the neighborhood of each pixel, given a feature extracted at that pixel. This strategy is denoted *aDversarIal defenSe with local impliCit functiOns* (DISCO). Local implicit models have recently been shown to provide impressive results for 3D modeling [94, 68, 85, 71, 118, 52, 46, 14, 67, 77, 29, 115] and image interpolation [13]. We show that such models can be trained to project barely outliers into the image manifold, with much smaller parameter and dataset sizes than generative models, while enabling much more precise control of the manifold projection operation. This is illustrated in Fig. 1, which presents an image, its adversarial attack, and the output of the DISCO defense. To train DISCO, a dataset of adversarial-clean pairs is first curated. During training, DISCO inputs an adversarial image and a query pixel location, for which it predicts a new RGB value. This is implemented with a feature encoder and a local implicit function. The former is composed by a set of residual blocks with stacked convolution layers and produces a deep feature per pixel. The latter consumes the query location and the features in a small neighborhood of the query location. The implicit function is learned to minimize the L_1 loss between the predicted RGB value and that of the clean image.

The restriction of the manifold modeling to small image neighborhoods is a critical difference with respect to previous defenses based on the modeling of the natural image manifold. Note that, as shown in Fig. 2, DISCO does not project the entire image into the manifold, only each pixel

neighborhood. This considerably simplifies the modeling and allows a much more robust defense in a parameter and data efficient manner. This is demonstrated by evaluating the performance of DISCO under both the oblivious and adaptive settings [121, 82, 119]. Under the oblivious setting, the popular RobustBench [16] benchmark is considered, for both L_∞ and L_2 attacks with Autoattack [17]. DISCO achieves SOTA robust accuracy (RA) performance, e.g. outperforming the prior art on Cifar10, Cifar100 and ImageNet by 17%, 19% and 19% on L_∞ Autoattack. A study of the defense transferability across datasets, classifiers and attacks also shows that the DISCO defense maintains much of its robustness even when deployed in a setting that differs from that used for training by any combination of these three factors. Finally, the importance of the local manifold modeling is illustrated by experiments on ImageNet [21], where it is shown that even when trained with only 0.5% of the dataset DISCO outperforms all prior defenses. Under the adaptive setting, DISCO is evaluated using the BPDA [4] attack, known to circumvent most defenses based on image transformation [95, 88, 99, 119, 25]. While DISCO is more vulnerable under this setting, where the defense is known to the attacker, it still outperforms existing approaches by 46.76%. More importantly, we show that the defense can be substantially strengthened by cascading DISCO stages, which magnifies the gains of DISCO to 57.77%. This again leverages the parameter efficiency of the local modeling of image statistics, which allows the implementation of DISCO cascades with low complexity. The ability to cascade DISCO stages also allows a new type of defense, where the number of DISCO stages is randomized on a per image basis. This introduces some degree of uncertainty about the defense even under the adaptive setting and further improves robustness.

Overall, this work makes four contributions. First, it proposes the use of defenses based on the conditional modeling of local image statistics only. Second, it introduces a novel defense of this type, DISCO, based on local implicit functions. Third, it leverages the parameter efficiency of the local modeling to propose a cascaded version of DISCO that is shown robust even to adaptive attacks. Finally, DISCO is shown to outperform prior defenses on RobustBench [16] and other 11 attacks. Extensive ablations also demonstrate that DISCO has unmatched defense transferrability in the literature, across datasets, attacks and classifiers.

2 Related Work

Adversarial Attack and Defense. We briefly review adversarial attacks and defenses for classification and prior art related to our work. Please refer to [12, 74, 1] for more complete reviews.

Adversarial Attacks aim to fool the classifier by generating an imperceptible perturbation (under L_p norm constraint) that is applied to the clean image. Attack methods have evolved from simple addition of sign gradient, as in FGSM [31], to more sophisticated approaches [57, 11, 98, 104, 64, 17, 23, 22, 18]. While most white-box attacks assume access to the classifier gradient, BPDA [4] proposed a gradient approximation attack that can circumvent defenses built on obfuscated gradients. In general, these attacks fall into two settings, oblivious or adaptive, depending on whether the attacker is aware of the defense strategy [121, 82, 119]. DISCO is evaluated under both settings.

Adversarial Defenses can be categorized into adversarially trained and transformation based. The former are trained against adversarial examples generated on-the-fly during training [81, 33, 32, 80, 90, 75, 90], allowing the resulting robust classifier to defend against the adversarial examples. While adversarially trained defenses dominate the literature, they are bound together with the classifier. Hence, re-training is required if the classifier changes and the cost of adversarial training increases as for larger classifiers. Transformation based defenses [35, 119, 45, 95, 99] instead introduce an additional defense module, which can be applied to many classifiers. This module preprocesses the input image before passing it to the classifier. The proposed preprocessing steps include JPEG compression [19, 25, 63], bit reduction [35, 119, 45], pixel deflection [79] or applications of random transformations [116, 35] and median filters [73]. Beyond pixel space defenses, malicious images can also be reconstructed to better match natural image statistics using autoencoders [66, 99], GANs [88, 3, 123], or other generative models, such as the PixelCNN [86]. The latter is used to project the malicious image into the natural image manifold by methods like PixelDefend [95] or HPD [5]. These methods can only produce images of fixed size [66, 88, 99, 123] and model pixel likelihoods over the entire image [95, 5]. This is unlike DISCO, which models conditional local statistics and can produce outputs of various size.

Implicit Function refers to the use of a neural network to model a continuous function [94]. This has been widely used in applications involving audio [127, 36, 94], 2D images [13, 24] and 3D shapes [94, 68, 85, 71, 118, 52, 46, 14, 67, 77, 29, 115]. In the 3D literature, local implicit functions

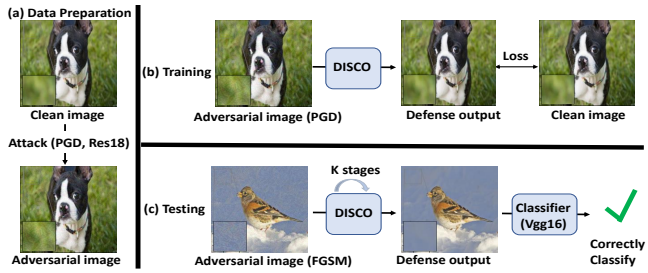


Figure 3: (a) Data preparation, (b) training and (c) testing phase of DISCO. DISCO supports different configurations of attack and classifier for training and testing. For cascade DISCO, $K > 1$.

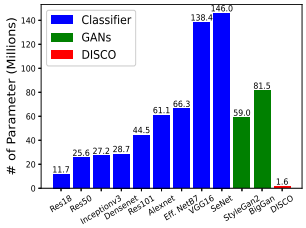


Figure 4: Number of parameters (Millions) of recent classifiers, GANs and DISCO.

have become popular models of object shape [71, 85, 77] or complex 3D scenes [68]. This also inspired 2D applications on super-resolution [13], image [48] and video generation [122]. In the adversarial attack literature, implicit functions have recently been proposed to restore adversarial point clouds of 3D shape, through the IF-Defense [115]. To the best of our knowledge, ours is the first paper to propose local implicit functions for 2D adversarial defense.

3 Method

In this section, we introduce DISCO.

3.1 Motivation

Under the hypothesis that natural images lie a low-dimension image manifold, classification networks can be robustified by learning to project barely outliers (i.e. adversarial images) into the manifold, a process that can be seen as *manifold thickening*. Images in a shell around the manifold are projected into it, leaving a larger margin to images that should be recognized as outliers. While this idea has been studied [30, 51, 61, 44], its success hinges on the ability of classification models to capture the complexities of the image manifold. This is a very hard problem, as evidenced by the difficulty of model inversion algorithms that aim to synthesize images with a classifier [107, 65, 70, 120]. These algorithms fail to produce images comparable to the state of the art in generative modeling, such as GANs [50, 49, 9]. Recently, however, it has been shown that it is possible to synthesize realistic images and 3D shapes with implicit functions, through the use of deep networks [29, 13, 94, 68, 85, 71, 46, 14], which basically memorizes images or objects as continuous functions. The assumption behind DISCO is that these implicit functions can capture the local statistics of images or 3D shapes, and can be trained projecting barely outliers into the image manifold.

3.2 Model Architecture and Training

Data Preparation To train DISCO, a dataset $D = \{(x^i_{cln}, x^i_{adv})\}_{i=1}^N$ containing a set of paired clean x^i_{cln} and adversarial x^i_{adv} images is curated. For this, a classifier P_{trn} and an attack procedure A_{trn} are first selected. For each image x^i_{cln}, the adversarial image x^i_{adv} is the generated by attacking the predictions $P_{trn}(x^i_{cln})$ using A_{trn}, as shown in Fig. 3(a).

Training As shown in Fig. 3(b), the DISCO defense is trained with pairs of random patches cropped at the same location of the images x_{cln} and x_{adv}. For example, random patches of size 48×48 are sampled from training pairs of the ImageNet dataset [21]. The network is then trained to minimize the L1 loss between the RGB values of the clean x_{cln} and defense output x_{def}.

Architecture To perform manifold thickening, DISCO takes an adversarial image $x_{adv} \in \mathbb{R}^{H \times W \times 3}$ and a query pixel location $p = [i, j] \in \mathbb{R}^2$ as input and predicts a clean RGB value $v \in \mathbb{R}^3$ at the query pixel location, ideally identical to the pixel value of the clean image $x_{cln} \in \mathbb{R}^{H \times W \times 3}$ at that location. The defense output $x_{def} \in \mathbb{R}^{H' \times W' \times 3}$ can then be synthesized by predicting a RGB value for each pixel locations in a grid of size $H' \times W' \times 3$. Note that it is not a requirement that the size of x_{def} be the same as that of x_{cln}. In fact, the size of x_{def} could be changed during inference.

To implement this, DISCO is composed of two modules, illustrated in Fig. 5. The first is an encoder E that extracts the per-pixel feature of an input image x. The encoder architecture resembles the design of EDSR [60], originally proposed for super-resolution. It contains a sequence of 15 residual blocks, each composed of a convolution layer, a ReLu layer and a second convolution layer. The

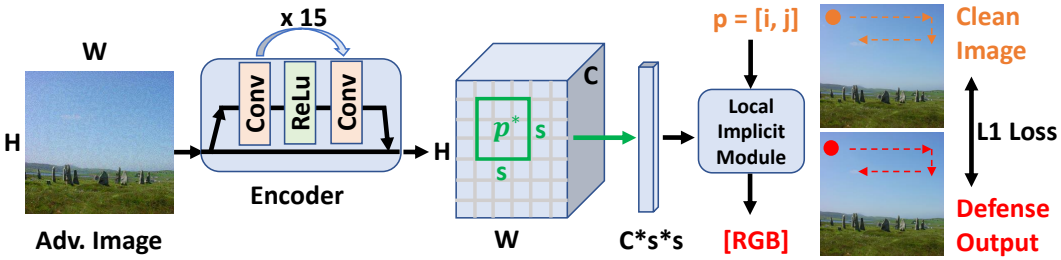


Figure 5: The DISCO architecture includes an encoder and a local implicit module. The network is trained to map Adversarial into Defense images, using an L_1 loss to Clean images.

encoder output is a feature map $f = E(x) \in \mathbb{R}^{H \times W \times C}$, with $C = 64$ channels. The feature at location $p = [i, j]$ is denoted as f_{ij}. The second module is the the local implicit module L, which is implemented by a MLP. Given query pixel location p, L first finds the nearest pixel value $p^* = [i^*, j^*]$ in the input image x and corresponding feature $f_{i^*j^*}$. L then takes the features in the neighborhood of p^* into consideration to predict the clean RGB value v. More specifically, let $\hat{f}_{i^*j^*}$ denote a concatenation of the features whose location is within the kernel size s of p^*. The local implicit module L takes the concatenated feature, the relative position $r = p - p^*$ between p and p^*, and the pixel shape as input, to predict a RGB value v. By default, the kernel size s is set to be 3. Since the network implements a continuous function, based only on neighboring pixels, the original grid size $H \times W$ is not important. The image coordinates can be normalized so that $(i, j) \in [-1, 1]^2$. This makes DISCO independent of the original image size or resolution.

3.3 Inference

For inference, DISCO takes either a clean or an adversarial image as input. Given a specified output size for x_{def}, DISCO loops over all the output pixel locations, predicting an RGB value per location. Note that this is not computationally intensive because the encoder feature map $f = E(x)$ is computed once and used to the predict the RGB values of all query pixel locations. Furthermore, while the training pairs are generated with classifier P_{trn} and attack A_{trn}, the inference time classifier P_{tst} and attack A_{tst} could be different. In the experimental section we show that DISCO is quite flexible, performing well when (1) P_{trn} and P_{tst} consume images of different input size and (2) the attack, classifier and dataset used for inference are different than those used for training. In fact, DISCO is shown to be more robust to these configuration changes than previous methods.

3.4 DISCO Cascades

DISCO is computationally very appealing because it disentangles the training of the defense from that of the classifier. This can be a big practical advantage, since classifier retraining is needed whenever training settings, such as architecture, hyper-parameters, or number of classes, change. Adversarial defenses require retraining on the entire dataset when this is the case, which is particularly expensive for large models (like SENet [38] or EfficientNet [101]) trained on large datasets (like ImageNet [21] and OpenImages [54]). Unsurprisingly, RobustBench [16], one of the largest adversarial learning benchmarks, reports more than 70 baselines for Cifar10, but less than 5 on ImageNet.

DISCO does not have this defense complexity, since DISCO is trained independently of the classifier. Furthermore, because DISCO is a model of local statistics, it is particularly parameter efficient. As shown in Fig. 4, DISCO has a lightweight design with only 1.6M parameters, which is significantly less than most recent classifier [101, 38, 93, 37] and GAN [50, 8] models with good performance for ImageNet-like images. This also leads to a computationally efficient defense. Our experiments show that DISCO can be trained with only 50,000 training pairs. In fact, we show that it can beat the prior SOTA using less than 0.5% of ImageNet as training data (Table 5). One major benefit of this efficiency is that it creates a *large unbalance between the costs of defense and attack.* Consider memory usage, which is dominated by the computation of gradients needed for either the attack or the backpropagation of training. Let N_d and N_c be the number of parameters of the DISCO network and classifier, respectively. The per image memory cost of training the DISCO defense is $O(N_d)$. On the other hand, the attack cost depends on the information available to the attacker. We consider two settings, commonly considered in the literature. In the *oblivious* setting, only the classifier is known to that attacker and the attack has cost $O(N_c)$. In the *adaptive* setting, both the classifier and

Table 1: Compare DISCO to the selected baselines on Cifar10 ($\epsilon_\infty = 8/255$).

Method	SA	RA	Avg.	Classifier
No Defense	**94.78**	0	47.39	WRN28-10
Rebuffi et al. [81]	92.23	66.58	79.41	WRN70-16
Gowal et al. [33]	88.74	66.11	77.43	WRN70-16
Gowal et al. [33]	87.5	63.44	75.47	WRN28-10
Bit Reduction [119]	92.66	1.04	46.85	WRN28-10
Jpeg [25]	83.9	50.73	67.32	WRN28-10
Input Rand. [116]	94.3	8.59	51.45	WRN28-10
AutoEncoder	76.54	67.41	71.98	WRN28-10
STL [99]	82.22	67.92	75.07	WRN28-10
DISCO	89.26	**85.56**	**87.41**	WRN28-10

Table 2: Compare DISCO to the selected baselines on Cifar10 ($\epsilon_2 = 0.5$).

Method	SA	RA	Avg.	Classifier
No Defense	**94.78**	0	47.39	WRN28-10
Rebuffi et al. [81]	95.74	82.32	**89.03**	WRN70-16
Gowal et al. [32]	94.74	80.53	87.64	WRN70-16
Rebuffi et al. [81]	91.79	78.8	85.30	WRN28-10
Bit Reduction [119]	92.66	3.8	48.23	WRN28-10
Jpeg [25]	83.9	69.85	76.88	WRN28-10
Input Rand. [116]	94.3	25.71	60.01	WRN28-10
AutoEncoder	76.54	71.71	74.13	WRN28-10
STL [99]	82.22	74.33	78.28	WRN28-10
DISCO	89.26	**88.47**	88.87	WRN28-10

the DISCO are exposed and backpropagation has memory cost $O(N_c + N_d)$. In experiments, we show that DISCO is quite effective against oblivious attacks. Adaptive attacks are more challenging. However, as shown in Fig 4, it is usually the case that $N_c > N_d$, making the complexity of the attack larger than that of the defense. This is unlike adversarial training, where attack and defense require backpropagation on the same model and thus have the same per-image cost.

This asymmetry between the memory cost of the attack and defense under DISCO can be *magnified* by cascading DISCO networks. If K identical stages of DISCO are cascaded, the defense complexity remains $O(N_d)$ but that of the attack raises to $O(N_c + KN_d)$. Hence the ratio of attack-to-defense cost raises to $O(K + N_c/N_d)$. Interestingly, our experiments (see Section 4.2) show that when K is increased the defense performance of the DISCO cascade increases as well. Hence, DISCO cascades combine high robust accuracy with a large ratio of attack-to-defense cost.

4 Experiments

In this section, we discuss experiments performed to evaluate the robustness of DISCO. Results are discussed for both the oblivious and adaptive settings [121, 82, 119] and each result is averaged over 3 trials. ϵ_p denotes the perturbation magnitude under the L_p norm. All experiments are conducted on a single Nvidia Titan Xp GPU with Intel Xeon CPU E5-2630 using Pytorch [78]. Please see appendix for more training details, quantitative results and visualizations.

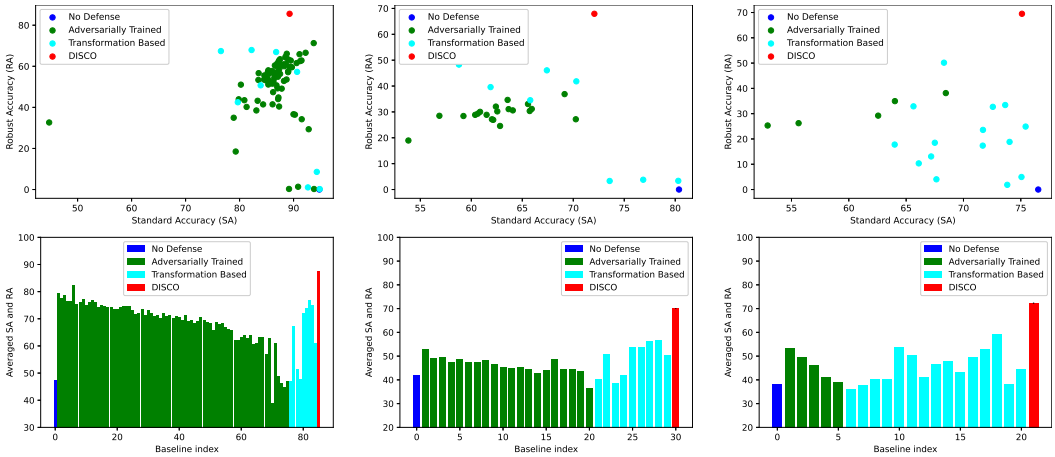
Training Dataset: The following training configuration is used unless noted. Three datasets are considered: Cifar10 [55], Cifar100 [56] and Imagenet [21]. For each, 50,000 adversarial-clean training pairs are curated. For Cifar10 and Cifar100, these are the images in the training set, while for ImageNet, 50 images are randomly sampled per class. Following RobustBench [16], the evaluation is then conducted on the test set of each dataset. To create training pairs, PGD [64] ($\epsilon_\infty = 8/255$) is used to attack a ResNet18 and a WideResNet28 on Cifar10/ImageNet and Cifar100, respectively.

Attack and Benchmark: DISCO is evaluated on RobustBench [16], which contains more than 100 baselines evaluated using Autoattack [17]. This is an ensemble of four sequential attacks, including the PGD [64] attack with two different optimization losses, the FAB attack [15] and the black-box Square Attack [2]. DISCO is compared to defense baselines under both L_∞ and L_2 norms. To study defense generalization, 11 additional attacks are considered, including FGSM [31], BIM [57], BPDA [4] and EotPgd [62]. Note that DISCO is not trained specifically for these attacks.

Metric: Standard (**SA**) and robust (**RA**) accuracy are considered. The former measures classification accuracy on clean examples, the latter on adversarial. The average of SA and RA is also used.

4.1 Oblivious Adversary

SOTA on RobustBench: DISCO achieves SOTA performance on RobustBench. Table 1 and 2 compare DISCO to the RobustBench baselines on Cifar10 under L_∞ and L_2 Autoattack, respectively. Baselines are categorized into (1) no defense (first block), (2) adversarially trained (second block) and (3) transformation based (third block). The methods presented in each table are those of highest RA performance in each category. The full table is given in the supplemental, together with those of Cifar100 and ImageNet. Note that the RobustBench comparison slightly favours the adversarially trained methods, which use a larger classifier. A detailed comparison to all RobustBench baselines is given in Fig. 6, for the three datasets. The upper row visualizes the trade-off between SA and RA. The bottom row plots the averaged SA/RA across baselines. Blue, green, cyan and red indicate no defense, adversarially trained baselines, transformation based baselines and DISCO, respectively.



(a) (b) (c)

Figure 6: Comparison of DISCO to No Defense, Adversarially Trained, and Transformation based baselines. (a) Cifar10, (b) Cifar100, and (c) ImageNet. Top-row: trade-off between SA and RA. Bottom row: average accuracy of each of the RobustBench baselines and DISCO.

These results show that, without a defense, the attack fools the classifier on nearly all examples. Adversarially trained baselines improve RA by training against the adversarial examples. Some of these [81, 33, 32, 47, 80, 40, 112, 96, 114] also leverage additional training data. Transformation based defenses require no modification of the pre-trained classifier and can generalize across attack strategies [88, 99]. While early methods (like Jpeg Compression [25] and Bit Reduction [119]) are not competitive, recent defenses [99] outperform adversarially trained baselines on Cifar100 and ImageNet. DISCO is an even more powerful transformation-based defense, which clearly outperforms the prior SOTA RA by a large margin (17 % on Cifar10, 19 % on Cifar100 and 19 % on ImageNet). In the upper row of Fig. 6, it clearly beats the prior Pareto front for the SA vs. RA trade-off. Table 1 and Table 2 also show that previous transformation based methods tend to perform better for L_2 than L_∞ Autoattack. DISCO is more robust, achieving similar RA for L_2 and L_∞ Autoattacks.

Improving SOTA Methods: While DISCO outperforms the SOTA methods on RobustBench, it can also be combined with the latter. Table 4 shows that adding DISCO improves the performance of top three ResNet50 robust baselines for ImageNet [87, 26, 110] by 16.77 (for RA) and 8.12 (for averaged SA/RA) on average. This demonstrates the plug-and-play ability of DISCO.

Dataset Size: Table 5 shows the SA and RA performance of DISCO when training pairs are sampled from a random subset of the ImageNet classes (100 and 500). Compared to the ImageNet SOTA [87] RA of 38.14% (See Appendix), DISCO outperforms the prior art by 21.7% (59.84 vs 38.14) when trained on about 0.5% of the ImageNet training data.

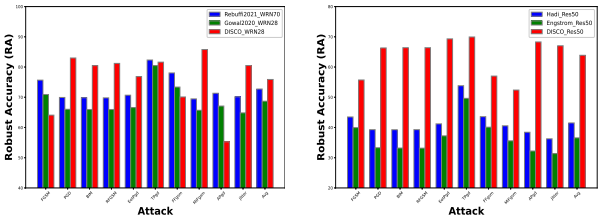
Defense Transferability of DISCO is investigated across attacks, classifiers and datasets.

Transfer across Attacks. RobustBench evaluates the model on Autoattack, which includes the PGD attack used to train DISCO. Table 3 summarizes the transfer performance of DISCO, trained with PGD attacks, when subject to ten different L_∞ attacks at inference. This is compared to the transfer peformance of the two top Cifar100 baselines on RobustBench. DISCO outperforms these baselines with an average RA gain greater than 24.49%. When compared to the baseline that uses the same classifier (WRN28-10), this gain increases to 29.5%. Fig. 7 visualizes the gains of DISCO (red bar) on Cifar10 and ImageNet. Among 3 datasets and 10 attacks, DISCO outperforms the baselines on 24 results. The average gains are largest on Cifar100 and ImageNet, where the RA of the prior approaches is lower. Note that the defense is more challenging on ImageNet, due to the higher dimensionality of its images [91]. The full table can be found in the supplemental.

We next evaluate the transfer ability of DISCO trained with the L_∞ PGD attacks to four L_2 norm inference attacks: FGSM [31], BIM [57], CW [11] and DeepFool [69]. Table 6 compares the defense transferability of DISCO to both adversarially trained (top block) and transformation baselines (lower block). DISCO generalizes well to L_2 attacks. It can defend more attacks than adversarially trained baselines (top block) and is more robust than the prior SOTA transformation based defenses [99, 123].

Method Classifier	Gowal [32] WRN70-16	Rebuffi [81] WRN28-10	DISCO WRN28-10
FGSM [31]	44.53	38.57	**50.4**
PGD [64]	40.46	36.09	**74.51**
BIM [57]	40.38	36.03	**72.25**
RFGSM [104]	40.42	35.99	**72.1**
EotPgd [62]	41.07	36.45	**74.8**
TPgd [125]	57.52	52.01	**74.06**
FFgsm [111]	47.61	41.47	**64.29**
MiFgsm [22]	42.37	37.31	**44.14**
APgd [18]	39.99	35.64	**77.33**
Jitter [89]	38.38	33.04	**73.75**
Avg.	43.27	38.26	**67.76**

Table 3: Defense transfer across ten L_∞ attacks ($\epsilon_\infty = 8/255$) on Cifar100.



(a) (b)

Figure 7: Defense transfer across L_∞ attacks on (a) Cifar10 and (b) ImageNet. (Blue, Green) Baselines, (Red) DISCO.

Table 4: Improving ResNet50 baselines on ImageNet.

Method	SA	RA	Avg.
Hadi et al. [87]	**64.02**	34.96	49.49
w/ DISCO	63.66	**50.6**	**57.13**
Engstrom et al. [26]	**62.56**	29.22	45.89
w/ DISCO	62.48	**49.44**	**55.96**
Wong et al. [110]	**55.62**	26.24	40.93
w/ DISCO	54.52	**40.68**	**47.6**

Table 5: Ablation of sampled classes.

Cls. #	Dataset size	SA	RA	Avg.
100	5000	**72.78**	59.84	66.31
500	25000	71.88	67.84	69.86
1000	50000	72.64	**68.2**	**70.42**

Table 6: Defense Transfer of L_∞ trained defenses to L_2 attacks on Cifar10. Top block: adversarially trained, Bottom block: transformation based.

Method	Classifier	Clean	FGSM	BIM	CW	DeepFool
Adv. FGSM	ResNet	91	**91**	**91**	7	0
Adv. BIM	ResNet	87	52	32	**42**	**48**
PixelDefend [95]	VGG	85	46	46	78	80
PixelDefend [95]	ResNet	82	62	61	79	76
PixelDefend (Adv.) [95]	ResNet	88	68	69	84	85
Feature Squeezing [119]	ResNet	84	20	0	78	N/A
EGC-FL [123]	ResNet	91.65	88.51	88.75	**90.03**	N/A
STL [99]	VGG16	90.11	87.15	88.03	89.04	88.9
DISCO	WRN28	89.26	**89.53**	**89.58**	89.3	**89.58**

Beyond different test attacks A_{tst} on a PGD-trained DISCO, we also evaluated the effect of changing the training attack A_{trn} used to generate the adversarial-clean pairs on which DISCO is trained. In rows 3-5 of Table 7, PGD [64], BIM [57] and FGSM [31] are used to generate training pairs, while Autoattack is used as testing attack. BIM and PGD have comparable results, which are stronger than those of FGSM. Nevertheless all methods outperform the SOTA RobustBench defense [81] for Autoattack on Cifar10, shown in the first row. These results suggests that DISCO is robust to many combinations of training and inference attacks.

Transfer across Classifiers. The first section of Table 7 shows the results when the testing classifier is different from the training classifier. While the ResNet18 is always used to curate the training pairs of DISCO, the testing classifier varies between ResNet18, WideResNet28 and VGG16. The small impact of the classifier used for inference on the overall RA shows that DISCO is classifier agnostic and can be applied to multiple testing classifiers once it is trained.

Transfer across Datasets. The evidence that adversarial attacks push images away from the natural image manifold [126, 59, 99, 25, 44, 61] and that attacks can be transferred across classifiers [20, 109, 113, 42, 108], suggest that it may be possible to transfer defenses across datasets. This, however, has not been studied in detail in the literature, partly because adversarially trained baselines entangle the defense and the classifier training. This is unlike DISCO and other transformation based baselines, which can be transferred across datasets. The bottom section of Table 7 shows the test performance on Cifar10 of DISCO trained on Cifar100 and ImageNet. Since Cifar100 images are more similar to those of Cifar10 than ImageNet ones, the Cifar100 trained DISCO transfers better than that trained on ImageNet. However, even the RA of the latter is 7.72% higher than the best RA reported on RobustBench [81]. Note that the DISCO trained on Cifar100 and ImageNet never see images from Cifar10 and the transfer is feasible because no limitation is imposed on the output size of DISCO.

4.2 Adaptive Adversary

The adaptive adversary assumes both the classifier and defense strategy are exposed to the attacker. As noted by [4, 99, 103], this setting is more challenging, especially for transformation based defenses. We adopt the BPDA [4] attack, which is known as an effective attack for transformation based defenses, such as DISCO. Fig. 8 compares the RA of DISCO trained with PGD attack to the results published for other methods in [99]. For fair comparison, DISCO is combined with a VGG16 classifier. The figure confirms that both prior transformation defenses and a single stage DISCO ($K = 1$) are vulnerable to an adaptive adversary. However, without training against BPDA, DISCO is 46.76% better than prior methods. More importantly, this gain increases substantially with K, saturating at RA of 57.77% for $K = 5$ stages, which significantly outperforms the SOTA by 57.35%.

Table 7: Defense transfer of DISCO across training attacks, classifiers, and datasets. In all cases the inference setting is: Cifar10 dataset with Autoattack. For comparison, the RobustBench SOTA [81] for no transfer is also shown.

	Transfer			Training			Testing			
	Classifier	Attack	Dataset	Attack	Classifier	Dataset	Classifier	SA	RA	Avg.
[81]				Autoattack	WRN70-16	Cifar10	WRN70-16	92.23	66.58	79.41
DISCO				PGD	Res18	Cifar10	Res18	89.57	76.03	82.8
	✓			PGD	Res18	Cifar10	VGG16	89.12	86.27	87.7
	✓			PGD	Res18	Cifar10	WRN28	89.26	85.56	87.41
	✓	✓		BIM	Res18	Cifar10	WRN28	91.96	84.92	88.44
	✓	✓		FGSM	Res18	Cifar10	WRN28	84.07	77.13	80.6
	✓	✓	✓	FGSM	Res18	Cifar100	WRN28	84.23	86.16	85.2
	✓	✓	✓	FGSM	Res18	ImageNet	WRN28	88.91	74.3	81.61

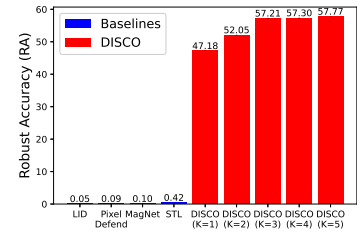


Figure 8: BPDA attack on Cifar10 under the adaptive setting.

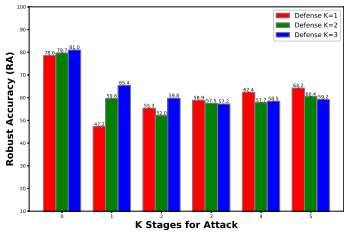


Figure 9: BPDA attack with cascade DISCO on Cifar10.

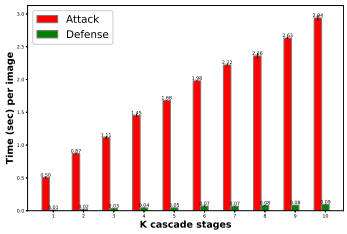


Figure 10: Attack and defense times vs DISCO cascade length.

4.3 Cascade DISCO

So far, we have considered the setting where the structure of the cascade DISCO is known to the attacker. DISCO supports a more sophisticated and practical setting, where the number K of DISCO stages used by the defense is randomized on a per-image basis. In this case, even if the use of DISCO is exposed to the attacker, there is still uncertainty about how many stages to use in the attack. We investigated the consequences of this uncertainty by measuring the defense performance when different values of K are use for attack and defense, denoted as K_{adv} and K_{def}, respectively. The oblivious setting has $K_{adv} = 0$ and $K_{def} \geq 1$, while $K_{adv} = K_{def}$ in the adaptive setting. We now consider the case where $K_{adv} \neq K_{def}$. Fig. 9 investigates the effectiveness of cascade DISCO trained with PGD attack when faced with the BPDA [4] attack in this setting, where RA(K_{adv},K_{def}) is the RA when K_{adv} and K_{def} are used, and $K_{adv} \in \{i\}_{i=0}^{5}$, $K_{def} \in \{i\}_{i=1}^{3}$. Under the setting of $K_{adv} \neq K_{def}$, the RA is higher than that of the adaptive setting. Take $K_{adv} = 2$ for example. Both RA(2,1)=55.3 and RA(2,3)=59.8 outperform RA(2,2)=52. In addition, Fig. 10 compares the time to generate a single adversarial example on Cifar10 and defend against it using DISCO. Clearly, the computational resources needed to generate an attack are significantly higher than those of the defense and the ratio of attack-to-defense cost raises with K. Both this and the good defense performance for mismatched Ks give the defender a strong advantage. It appears that the defense is more vulnerable when the attacker knows K (adaptive setting) and even there, as seen in the previous section, the defense can obtain the upper hand by casacading several DISCO stages.

5 Discussion, Societal Impact and Limitations

In this work, we have proposed the use of local implicit functions for adversarial defense. Given an input adversarial image and a query location, the DISCO model is proposed to project the RGB value of each image pixel into the image manifold, conditional on deep features centered in the pixel neighborhood. By training this projection with adversarial and clean images, DISCO learns to remove the adversarial perturbation. Experiments demonstrate DISCO's computational efficiency, its outstanding defense performance and transfer ability across attacks, datasets and classifiers. The cascaded version of DISCO further strengthens the defense with minor additional cost. **Limitations:** While DISCO shows superior performance on the attacks studied in this work, it remains to be tested whether it is robust to other type of attacks, such as one pixel attack [98] or patch attacks [10, 39]. In addition, since DISCO is a promising architecture for defense transferability, more evaluation configurations across attacks, datasets and classifiers will be investigated in the future. **Societal Impact:** We hope the idea of using local implicit functions can inspire better defenses and prevent the nefarious effects of deep learning attacks. Obviously, better defenses can also be leveraged by bad actors to improve resistance to the efforts of law enforcement, for example.

References

[1] Naveed Akhtar, Ajmal Mian, Navid Kardan, and Mubarak Shah. Advances in adversarial attacks and defenses in computer vision: A survey. *IEEE Access*, 9:155161–155196, 2021.

[2] Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square attack: a query-efficient black-box adversarial attack via random search. 2020.

[3] Ali ArjomandBigdeli, Maryam Amirmazlaghani, and Mohammad Khalooei. Defense against adversarial attacks using dragan. In *2020 6th Iranian Conference on Signal Processing and Intelligent Systems (ICSPIS)*, pages 1–5, 2020.

[4] Anish Athalye, Nicholas Carlini, and David A. Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *ICML*, pages 274–283, 2018.

[5] Yang Bai, Yan Feng, Yisen Wang, Tao Dai, Shutao Xia, and Yong Jiang. Hilbert-based generative defense for adversarial examples. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4783–4792, 2019.

[6] H. B. Barlow. The exploitation of regularities in the environment by the brain. *The Behavioral and brain sciences*, 24 4:602–7; discussion 652–71, 2001.

[7] Anthony J. Bell and Terrence J. Sejnowski. The "independent components" of natural scenes are edge filters. *Vision Research*, 37(23):3327–3338, 1997.

[8] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019.

[9] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *ArXiv*, abs/1809.11096, 2019.

[10] Tom B. Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. Adversarial patch. *ArXiv*, abs/1712.09665, 2017.

[11] Nicholas Carlini and David A. Wagner. Towards evaluating the robustness of neural networks. *2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57, 2017.

[12] Anirban Chakraborty, Manaar Alam, Vishal Dey, Anupam Chattopadhyay, and Debdeep Mukhopadhyay. Adversarial attacks and defences: A survey. *ArXiv*, abs/1810.00069, 2018.

[13] Yinbo Chen, Sifei Liu, and Xiaolong Wang. Learning continuous image representation with local implicit image function. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8628–8638, 2021.

[14] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5932–5941, 2019.

[15] F. Croce and M. Hein. Minimally distorted adversarial examples with a fast adaptive boundary attack. In *ICML*, 2020.

[16] Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. RobustBench: a standardized adversarial robustness benchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

[17] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *ICML*, 2020.

[18] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *ICML*, 2020.

[19] Nilaksh Das, Madhuri Shanbhogue, Shang-Tse Chen, Fred Hohman, Li Chen, Michael E. Kounavis, and Duen Horng Chau. Keeping the bad guys out: Protecting and vaccinating deep learning with jpeg compression. *ArXiv*, abs/1705.02900, 2017.

[20] Ambra Demontis, Marco Melis, Maura Pintor, Matthew Jagielski, Battista Biggio, Alina Oprea, Cristina Nita-Rotaru, and Fabio Roli. Why do adversarial attacks transfer? explaining transferability of evasion and poisoning attacks. In *USENIX Security Symposium*, 2019.

[21] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[22] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9185–9193, 2018.

[23] Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019.

[24] Emilien Dupont, Adam Golinski, Milad Alizadeh, Yee Whye Teh, and Arnaud Doucet. COIN: COmpression with implicit neural representations. In *Neural Compression: From Information Theory to Applications – Workshop @ ICLR 2021*, 2021.

[25] Gintare Karolina Dziugaite, Zoubin Ghahramani, and Daniel M. Roy. A study of the effect of jpg compression on adversarial images. *ArXiv*, abs/1608.00853, 2016.

[26] Logan Engstrom, Andrew Ilyas, Hadi Salman, Shibani Santurkar, and Dimitris Tsipras. Robustness (python library), 2019.

[27] David J. Field. Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America. A, Optics and image science*, 4 12:2379–94, 1987.

[28] David J. Field. What is the goal of sensory coding? *Neural Computation*, 6:559–601, 1994.

[29] Kyle Genova, Forrester Cole, Avneesh Sud, Aaron Sarna, and Thomas A. Funkhouser. Local deep implicit functions for 3d shape. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4856–4865, 2020.

[30] Justin Gilmer, Luke Metz, Fartash Faghri, Sam Schoenholz, Maithra Raghu, Martin Wattenberg, and Ian Goodfellow. Adversarial spheres, 2018.

[31] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

[32] Sven Gowal, Chongli Qin, Jonathan Uesato, Timothy Mann, and Pushmeet Kohli. Uncovering the limits of adversarial training against norm-bounded adversarial examples. *arXiv preprint arXiv:2010.03593*, 2020.

[33] Sven Gowal, Sylvestre-Alvise Rebuffi, Olivia Wiles, Florian Stimberg, Dan Andrei Calian, and Timothy A Mann. Improving robustness using generated data. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 4218–4233. Curran Associates, Inc., 2021.

[34] Daniel J. Graham and David J. Field. Statistical regularities of art images and natural scenes: spectra, sparseness and nonlinearities. *Spatial vision*, 21 1-2:149–64, 2007.

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

[36] Yudong Guo, Keyu Chen, Sen Liang, Yongjin Liu, Hujun Bao, and Juyong Zhang. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[37] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[38] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018.

[39] Yu-Chih-Tuan Hu, Bo-Han Kung, Daniel Stanley Tan, Jun-Cheng Chen, Kai-Lung Hua, and Wen-Huang Cheng. Naturalistic physical adversarial patch for object detectors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[40] Hanxun Huang, Yisen Wang, Sarah Monazam Erfani, Quanquan Gu, James Bailey, and Xingjun Ma. Exploring architectural ingredients of adversarially robust deep neural networks. In *NeurIPS*, 2021.

[41] Jinggang Huang and D. Mumford. Statistics of natural images and models. In *Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149)*, volume 1, pages 541–547 Vol. 1, 1999.

[42] Yi Huang and Adams Wai-Kin Kong. Transferable adversarial attack based on integrated gradients. In *International Conference on Learning Representations*, 2022.

[43] Zhichao Huang and Tong Zhang. Black-box adversarial attack with transferable model-based embedding. In *International Conference on Learning Representations*, 2020.

[44] Susmit Jha, Uyeong Jang, Somesh Jha, and Brian Jalaian. Detecting adversarial examples using data manifolds. In *MILCOM 2018 - 2018 IEEE Military Communications Conference (MILCOM)*, pages 547–552, 2018.

[45] Xiaojun Jia, Xingxing Wei, Xiaochun Cao, and Hassan Foroosh. Comdefend: An efficient image compression model to defend adversarial examples. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6077–6085, 2019.

[46] Chiyu Max Jiang, Avneesh Sud, Ameesh Makadia, Jingwei Huang, Matthias Nießner, and Thomas A. Funkhouser. Local implicit grid representations for 3d scenes. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6000–6009, 2020.

[47] QIYU KANG, Yang Song, Qinxu Ding, and Wee Peng Tay. Stable neural ODE with lyapunov-stable equilibrium points for defending against adversarial attacks. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

[48] Tero Karras, Miika Aittala, Samuli Laine, Erik Harkonen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. In *NeurIPS*, 2021.

[49] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4396–4405, 2019.

[50] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8107–8116, 2020.

[51] Marc Khoury and Dylan Hadfield-Menell. On the geometry of adversarial examples. *ArXiv*, abs/1811.00525, 2018.

[52] A. Kohli, V. Sitzmann, and G. Wetzstein. Semantic Implicit Neural Scene Representations with Semi-supervised Training. In *International Conference on 3D Vision (3DV)*, 2020.

[53] Alexander Kolesnikov and Christoph H. Lampert. Pixelcnn models with auxiliary variables for natural image modeling. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 1905–1914. JMLR.org, 2017.

[54] Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Andreas Veit, Serge Belongie, Victor Gomes, Abhinav Gupta, Chen Sun, Gal Chechik, David Cai, Zheyun Feng, Dhyanesh Narayanan, and Kevin Murphy. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://github.com/openimages*, 2017.

[55] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 (canadian institute for advanced research).

[56] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-100 (canadian institute for advanced research).

[57] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *ICLR Workshop*, 2017.

[58] Hyeungill Lee, Sungyeob Han, and Jungwoo Lee. Generative adversarial trainer: Defense to adversarial perturbations with gan. *ArXiv*, abs/1705.03387, 2017.

[59] Dongze Li, Wei Wang, Hongxing Fan, and Jing Dong. Exploring adversarial fake images on face manifold. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5785–5794, 2021.

[60] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.

[61] Wei-An Lin, Chun Pong Lau, Alexander Levine, Rama Chellappa, and Soheil Feizi. Dual manifold adversarial robustness: Defense against lp and non-lp adversarial attacks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 3487–3498. Curran Associates, Inc., 2020.

[62] 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*, 2019.

[63] Zihao Liu, Qi Liu, Tao Liu, Yanzhi Wang, and Wujie Wen. Feature distillation: Dnn-oriented jpeg compression against adversarial examples. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 860–868, 2019.

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

[65] Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5188–5196, 2015.

[66] Dongyu Meng and Hao Chen. Magnet: A two-pronged defense against adversarial examples. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, CCS '17, page 135–147, New York, NY, USA, 2017. Association for Computing Machinery.

[67] Lars M. Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4455–4465, 2019.

[68] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020.

[69] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: A simple and accurate method to fool deep neural networks. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2574–2582, 2016.

[70] Anh M Nguyen, Jeff Clune, Yoshua Bengio, Alexey Dosovitskiy, and Jason Yosinski. Plug & play generative networks: Conditional iterative generation of images in latent space. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3510–3520, 2017.

[71] Michael Niemeyer, Lars M. Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3501–3512, 2020.

[72] Bruno A. Olshausen and David J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609, 1996.

[73] Margarita Osadchy, Julio Hernandez-Castro, Stuart Gibson, Orr Dunkelman, and Daniel Pérez-Cabo. No bot expects the deepcaptcha! introducing immutable adversarial examples, with applications to captcha generation. *IEEE Transactions on Information Forensics and Security*, 12(11):2640–2653, 2017.

[74] Mesut Ozdag. Adversarial attacks and defenses against deep neural networks: A survey. *Procedia Computer Science*, 140:152–161, 2018. Cyber Physical Systems and Deep Learning Chicago, Illinois November 5-7, 2018.

[75] Tianyu Pang, Min Lin, Xiao Yang, Jun Zhu, and Shuicheng Yan. Robustness and accuracy could be reconcilable by (proper) definition, 2022.

[76] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z. Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, ASIA CCS '17, page 506–519, New York, NY, USA, 2017. Association for Computing Machinery.

[77] Jeong Joon Park, Peter R. Florence, Julian Straub, Richard A. Newcombe, and S. Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 165–174, 2019.

[78] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[79] Aaditya Prakash, Nick Moran, Solomon Garber, Antonella DiLillo, and James A. Storer. Deflecting adversarial attacks with pixel deflection. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8571–8580, 2018.

[80] Rahul Rade and Seyed-Mohsen Moosavi-Dezfooli. Helper-based adversarial training: Reducing excessive margin to achieve a better accuracy vs. robustness trade-off. In *ICML 2021 Workshop on Adversarial Machine Learning*, 2021.

[81] Sylvestre-Alvise Rebuffi, Sven Gowal, Dan Andrei Calian, Florian Stimberg, Olivia Wiles, and Timothy A. Mann. Fixing data augmentation to improve adversarial robustness. *ArXiv*, abs/2103.01946, 2021.

[82] Kui Ren, Tianhang Zheng, Zhan Qin, and Xue Liu. Adversarial attacks and defenses in deep learning. *Engineering*, 6(3):346–360, 2020.

[83] Luke E. Richards, André T. Nguyen, Ryan Capps, Steven Forsyth, Cynthia Matuszek, and Edward Raff. Adversarial transfer attacks with unknown data and class overlap. *CoRR*, abs/2109.11125, 2021.

[84] Daniel L. Ruderman. The statistics of natural images. *Network: Computation In Neural Systems*, 5:517–548, 1994.

[85] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2304–2314, 2019.

[86] Tim Salimans, Andrej Karpathy, Xi Chen, and Diederik P. Kingma. Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications. *ArXiv*, abs/1701.05517, 2017.

[87] Hadi Salman, Andrew Ilyas, Logan Engstrom, Ashish Kapoor, and Aleksander Madry. Do adversarially robust imagenet models transfer better? In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 3533–3545. Curran Associates, Inc., 2020.

[88] Pouya Samangouei, Maya Kabkab, and Rama Chellappa. Defense-GAN: Protecting classifiers against adversarial attacks using generative models. In *International Conference on Learning Representations*, 2018.

[89] Leo Schwinn, René Raab, An Nguyen, Dario Zanca, and Bjoern M. Eskofier. Exploring robust misclassifications of neural networks to enhance adversarial attacks. 2021.

[90] Vikash Sehwag, Saeed Mahloujifar, Tinashe Handina, Sihui Dai, Chong Xiang, Mung Chiang, and Prateek Mittal. Robust learning meets generative models: Can proxy distributions improve adversarial robustness? In *International Conference on Learning Representations*, 2022.

[91] Ali Shafahi, W. Ronny Huang, Christoph Studer, Soheil Feizi, and Tom Goldstein. Are adversarial examples inevitable? In *International Conference on Learning Representations*, 2019.

[92] Eero P. Simoncelli and Bruno A. Olshausen. Natural image statistics and neural representation. *Annual review of neuroscience*, 24:1193–216, 2001.

[93] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2015.

[94] Vincent Sitzmann, Julien N. P. Martel, Alexander W. Bergman, David B. Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. *ArXiv*, abs/2006.09661, 2020.

[95] 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*, 2018.

[96] Kaustubh Sridhar, Oleg Sokolsky, Insup Lee, and James Weimer. Robust learning via persistency of excitation. *ArXiv*, abs/2106.02078, 2021.

[97] A. Srivastava, A. B. Lee, Eero P. Simoncelli, and S.-C. Zhu. On advances in statistical modeling of natural images. *Journal of Mathematical Imaging and Vision*, 18:17–33, 2004.

[98] Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 23:828–841, 2019.

[99] Bo Sun, Nian hsuan Tsai, Fangchen Liu, Ronald Yu, and Hao Su. Adversarial defense by stratified convolutional sparse coding. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11439–11448, 2019.

[100] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, 2016.

[101] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *ICML*, 2019.

[102] Antonio Torralba and Aude Oliva. Statistics of natural image categories. *Network: Computation in Neural Systems*, 14:391 – 412, 2003.

[103] Florian Tramèr, Nicholas Carlini, Wieland Brendel, and Aleksander Mądry. On adaptive attacks to adversarial example defenses. NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc.

[104] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. In *International Conference on Learning Representations*, 2018.

[105] Aäron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. *ArXiv*, abs/1601.06759, 2016.

[106] Manuela Vasconcelos and Nuno Vasconcelos. Natural image statistics and low-complexity feature selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2):228–244, 2009.

[107] Pei Wang, Yijun Li, Krishna Kumar Singh, Jingwan Lu, and Nuno Vasconcelos. Imagine: Image synthesis by image-guided model inversion. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3680–3689, 2021.

[108] Xiaosen Wang and Kun He. Enhancing the transferability of adversarial attacks through variance tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1924–1933, June 2021.

[109] Xiaosen Wang, Xu He, Jingdong Wang, and Kun He. Admix: Enhancing the transferability of adversarial attacks. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16138–16147, 2021.

[110] Eric Wong, Leslie Rice, and J. Zico Kolter. Fast is better than free: Revisiting adversarial training. In *International Conference on Learning Representations*, 2020.

[111] Eric Wong, Leslie Rice, and J. Zico Kolter. Fast is better than free: Revisiting adversarial training. In *International Conference on Learning Representations*, 2020.

[112] Boxi Wu, Jinghui Chen, Deng Cai, Xiaofei He, and Quanquan Gu. Do wider neural networks really help adversarial robustness? In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

[113] Dongxian Wu, Yisen Wang, Shu-Tao Xia, James Bailey, and Xingjun Ma. Skip connections matter: On the transferability of adversarial examples generated with resnets. In *International Conference on Learning Representations*, 2020.

[114] Dongxian Wu, Shu-Tao Xia, and Yisen Wang. Adversarial weight perturbation helps robust generalization. In *NeurIPS*, 2020.

[115] Ziyi Wu, Yueqi Duan, He Wang, Qingnan Fan, and Leonidas J. Guibas. If-defense: 3d adversarial point cloud defense via implicit function based restoration. *ArXiv*, abs/2010.05272, 2020.

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

[117] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan Yuille. Improving transferability of adversarial examples with input diversity. In *Computer Vision and Pattern Recognition*. IEEE, 2019.

[118] Qiangeng Xu, Weiyue Wang, Duygu Ceylan, Radomir Mech, and Ulrich Neumann. Disn: Deep implicit surface network for high-quality single-view 3d reconstruction. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 492–502. Curran Associates, Inc., 2019.

[119] Weilin Xu, David Evans, and Yanjun Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. *ArXiv*, abs/1704.01155, 2018.

[120] Hongxu Yin, Pavlo Molchanov, Zhizhong Li, José Manuel Álvarez, Arun Mallya, Derek Hoiem, Niraj Kumar Jha, and Jan Kautz. Dreaming to distill: Data-free knowledge transfer via deepinversion. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8712–8721, 2020.

[121] Xuwang Yin, Soheil Kolouri, and Gustavo K Rohde. Gat: Generative adversarial training for adversarial example detection and robust classification. In *International Conference on Learning Representations*, 2020.

[122] Sihyun Yu, Jihoon Tack, Sangwoo Mo, Hyunsu Kim, Junho Kim, Jung-Woo Ha, and Jinwoo Shin. Generating videos with dynamics-aware implicit generative adversarial networks. In *International Conference on Learning Representations*, 2022.

[123] Jianhe Yuan and Zhihai He. Ensemble generative cleaning with feedback loops for defending adversarial attacks. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 578–587, 2020.

[124] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In Edwin R. Hancock Richard C. Wilson and William A. P. Smith, editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 87.1–87.12. BMVA Press, September 2016.

[125] 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*, 2019.

[126] Jianli Zhou, Chao Liang, and Jun Chen. Manifold projection for adversarial defense on face recognition. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX*, page 288–305, Berlin, Heidelberg, 2020. Springer-Verlag.

[127] Jan Zuiderveld, Marco Federici, and Erik J Bekkers. Towards lightweight controllable audio synthesis with conditional implicit neural representations. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.