LEVERAGING IMAGE REPRESENTATIONS FOR BOUNDED ADVERSARIAL ATTACKS AND ROBUSTNESS

Anonymous authors
Paper under double-blind review

ABSTRACT

Both classical and learned image transformations such as the discrete wavelet transforms (DWTs) and flow-based generative models provide semantically meaningful representations of images. In this paper, we propose a general method for robustness exploiting the expressiveness of image representations by targeting substantially low-dimensional subspaces inside the L^∞ box. Experiments with DCT, DWTs and Glow produce adversarial examples that are significantly more similar to the original than those found considering the full L^∞ box. Further, through adversarial training we show that robustness under the introduced constraints transfers better to robustness against a broad class of common image perturbations compared to the standard L^∞ box, without a major sacrifice of natural accuracy.

1 INTRODUCTION

The deployment of deep neural networks for image classification in critical decision-making processes has raised concerns about their robustness. Despite often stellar test set accuracies, these models have also shown to be brittle in various ways in which the human vision is not. For example, a network can be fooled by suitably designed malicious perturbations that look non-suspicious or even undetectable by a human. Further, networks are also not robust when faced by real-world image corruptions such as images taken under different weather conditions.

Adversarial robustness. Given a neural network that makes accurate predictions on clean data, adversarial attacks (Biggio et al., 2013; Szegedy et al., 2014; Papernot et al., 2016a) compute a suitable choice of additive noise to produce erroneous predictions. For images, the noise is typically measured and bounded by an L^p norm. The seminal method of projected gradient descent (PGD) (Madry et al., 2018) is a prominent example. Learning-based methods also have been used to build adversarial attacks either by leaning an embedding space using neural networks (Huang & Zhang, 2020; Baluja & Fischer, 2018), using latent space of generative adversarial networks (GAN) (Xiao et al., 2018; Wang & Yu, 2019) or using flow-based models to attack in black-box settings (Dolatabadi et al., 2020). Many approaches have been proposed to detect if an input is adversarial (Xu et al., 2018; Ma et al., 2018; Feinman et al., 2017; Metzen et al., 2017) and defend against it (Gu & Rigazio, 2014; Papernot et al., 2016b; Liao et al., 2018; Xie et al., 2019; Zhou et al., 2021). However, most of these defenses can again be broken by suitable adaptive attacks (Tramèr et al., 2020; Carlini & Wagner, 2017).

More importantly, adversarial attacks (assuming they are fast enough) can be used for adversarial training to increase robustness by first generating adversarial examples from clean training data, and then either performing standard training on these (Madry et al., 2018) or combining them with clean data to define a loss that, when minimized, better preserves the natural accuracy (Kannan et al., 2018; Zhang et al., 2019) or other variants such as (Chen et al., 2021; Rebuffi et al., 2021; Jiang et al., 2023). Further, adversarial training can be used to obtain provably robust models (Salman et al., 2019; Müller et al., 2022). However, typically the price of adversarial training is a significant drop of the classification accuracy on the unperturbed, clean data. On the other hand, provable adversarial robustness can be provided through randomized smoothing (Salman et al., 2019; Carlini et al., 2022), a sampling-based approach that scales to large models regardless of their complexity.

Corruption robustness. Arguably more important for practical applications, and a longstanding goal in neural network design, is robustness against distribution shifts between training data and

application data (Mintun et al., 2021; Pan & Yang, 2010; Farahani et al., 2020). One class of such shifts, and the one considered in this paper, are image corruptions. Examples include digital effects such as compression or weather conditions such as fog. Other forms of distribution shift are studied by applying abstract changes in structure and style (Hendrycks et al., 2021) to images or by sampling new versions of datasets (Recht et al., 2019).

Training networks to be robust against common image corruptions has become an active research topic, especially after the introduction of dedicated benchmarks such as ImageNet-C (Hendrycks & Dietterich, 2019). Approaches include again suitable data augmentation (Geirhos et al., 2018; Erichson et al., 2022; Zhang et al., 2017; Hendrycks et al., 2019; Park et al., 2022; Yin et al., 2022; Liang et al., 2023) in training or the use of transformed image representations. For example, training techniques relying on the discrete cosine transform (DCT) are found effective to generalize to unseen image distortions, for example, by extending the dropout technique to DCT coefficients as a form of regularization (Hossain et al., 2019). Similarly, (Duan et al., 2021) defined constraints in the DCT domain to generate adversarial examples that are less affected by JPEG compression than those obtained by pixel-based attacks.

Interestingly, (Hendrycks & Dietterich, 2019; Ford et al., 2019; Xie et al., 2020; Kang et al., 2019; Kireev et al., 2022) found that adversarial training using L^p norms (without any further constraints) also yields good accuracy to several common perturbations such as blur and weather.

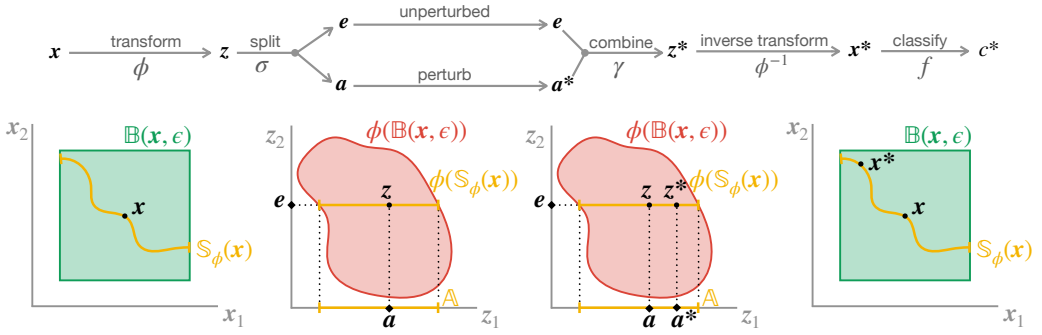
Our contribution. In this paper we offer progress in the quest for corruption robustness by presenting a powerful novel adversarial attack and associated adversarial training. We will demonstrate that the latter can yield networks with both only a small drop in accuracy on unperturbed test data and better robustness across common categories of corruptions. The key idea is to perform an adversarial attack in a meaningful subspace of a transformed image representation (e.g., the details in a wavelet-transformed image) while, at the same time, obeying the L^∞ box in the image domain, i.e., staying close in pixels. In other words, our attacks operate exclusively in a meaningful subspace (as defined by the chosen transform) of the L^∞ box.

Doing so is not possible with prior attacks such as PGD since the needed projections are not available in closed form for such a complex perturbation space. Instead, we use the *barrier method* from non-linear programming (Chachuat, 2007; Nocedal & Wright, 2006) to compute perturbations while satisfying the constraints without the need for projections. Doing so makes our approach efficient enough for attacks and to be integrated in adversarial training and for a wide class of image transformations including classical linear ones such as the discrete wavelet transform (DWT) or non-linear learned transforms such as flow-based models. Specifically, we contribute:

- A novel white-box attack that efficiently computes adversarial perturbations in a predefined transformed representation subspace while obeying the L^∞ pixel constraint at the same time.
- Instantiations of our approach for the two classical linear transforms DCT and DWT and the nonlinear learned flow-based model Glow (Kingma & Dhariwal, 2018).
- An evaluation of our attacks against prior work on ImageNet and CIFAR-10. In particular, given the same L^∞ box, we show that adversarial images found by our approach present significantly higher similarity to the originals, as verified by the learned perceptual image patch similarity metric (LPIPS) (Zhang et al., 2018).
- We show that using our attacks for adversarial training can yield excellent robustness on the image corruption benchmark CIFAR-10-C, up to on average 12.17% more accurate, with only a small drop in natural accuracy, than those trained under the full L^∞ box.

2 Adversarial Attack

In this section we explain our adversarial attack based on perturbing an image in a transformed representation, while, at the same time, obeying the classical box constraint in the pixel domain. The approach is depicted as cartoon for two-dimensional images in Fig. 1 and leverages the barrier method (Chachuat, 2007; Nocedal & Wright, 2006) from non-linear programming for the occurring optimization problem.



(a) Considered perturbations. (b) (a) in the ϕ-domain. (c) Possible perturbation $\boldsymbol{z}^*$. (d) (c) in the image domain.

Figure 1: High-level depiction (in two dimensions) of our approach for finding an adversarial example of an image $\boldsymbol{x}$ using a chosen transform ϕ and split operator σ. (a) shows the considered perturbations $\mathbb{S}_\phi(\boldsymbol{x})$, (b) the same in the ϕ-domain, (c) a possible perturbation $\boldsymbol{z}^*$ in the ϕ-domain by perturbing $\boldsymbol{a}$ but maintaining $\boldsymbol{e}$, and (d) the result in the image domain.

2.1 Problem Statement

Let ϕ be an image transformation that maps a pixel image $\boldsymbol{x} \in \mathbb{R}^n$ to a meaningful representation of the same dimension $\boldsymbol{z} = \phi(\boldsymbol{x})$. We assume that ϕ is bijective and (almost everywhere) differentiable. We aim to perturb some coordinates of $\boldsymbol{z}$ while leaving others unperturbed. To do so we define the split operator σ that divides $\boldsymbol{z}$ into two vectors: $\boldsymbol{e} \in \mathbb{R}^p$ (called essential) collects the coordinates to be maintained, and $\boldsymbol{a} \in \mathbb{R}^q$ (called auxiliary) those to be perturbed. Thus, $p + q = n$. Formally,

$$\sigma(\boldsymbol{z}) = (\boldsymbol{e}, \boldsymbol{a}) \quad \text{(split)}, \quad \boldsymbol{z} = \sigma^{-1}(\boldsymbol{e}, \boldsymbol{a}) = \gamma(\boldsymbol{e}, \boldsymbol{a}) \quad \text{(combine)}. \tag{1}$$

For example, if ϕ is the DCT at the heart of JPEG compression, $\boldsymbol{e}$ could collect the lowest frequencies that are most important for image recovery and $\boldsymbol{a}$ the remaining higher ones.

Let f be a classification model (e.g., a neural net) that correctly predicts the label c of the image $\boldsymbol{x}$. After transforming $\boldsymbol{x}$ to $\phi(\boldsymbol{x}) = \boldsymbol{z}$ and applying a chosen σ to obtain $\boldsymbol{e}$ and $\boldsymbol{a}$, we aim to perturb $\boldsymbol{a}$ to $\boldsymbol{a}^*$ such that $\boldsymbol{x}^* = \phi^{-1}\big(\gamma(\boldsymbol{e}, \boldsymbol{a}^*)\big)$ gets misclassified: $f(\boldsymbol{x}^*) = c^* \neq c$ (top row in Fig. 1). The set of these perturbations yields a (not necessarily linear) subspace of dimension q in the image (pixel) domain. In addition, we impose an L^∞ constraint on these perturbations in the image domain.

In summary, the perturbation space we consider is given by the intersection

$$\mathbb{S}_\phi(\boldsymbol{x}) = \phi^{-1}\big(\gamma(\boldsymbol{e}, \mathbb{R}^n)\big) \cap \mathbb{B}(\boldsymbol{x}, \epsilon) \tag{2}$$

and shown in yellow in Fig. 1a with the box depicted in green. A possible perturbation $\boldsymbol{x}^*$ is shown Fig. 1d. In the transformed ϕ-domain, $\phi(\mathbb{B}(\boldsymbol{x}, \epsilon))$ has some irregular shape (Fig. 1b), whereas the perturbations of $\boldsymbol{a}$ constitute a linear subspace.

2.2 Attack Description

The only free parameter in our perturbation space is $\boldsymbol{a}^*$. Thus, finding the corresponding adversarial example $\boldsymbol{x}^* = \phi^{-1}\big(\gamma(\boldsymbol{e}, \boldsymbol{a}^*)\big)$ amounts to solving a constrained optimization problem of the form

$$\min_{\boldsymbol{a}^* \in \mathbb{A}} \mathcal{L}(\boldsymbol{a}^*), \tag{3}$$

where $\mathcal{L}$ is a function that promotes misclassification when minimized. Several examples have been used in the literature (Carlini & Wagner, 2017). We use the negative cross entropy $-H$:

$$\mathcal{L}(\boldsymbol{a}^*) = -H(f(\phi^{-1}(\gamma(\boldsymbol{e}, \boldsymbol{a}^*))), c). \tag{4}$$

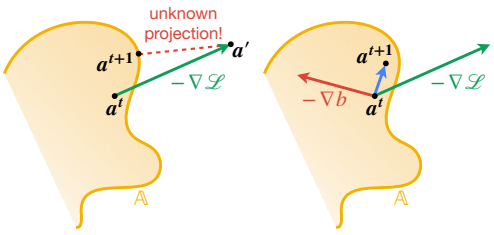
The set $\mathbb{A} \subset \mathbb{R}^q$ in (3) represents the allowed perturbation of $\boldsymbol{a}^*$ depicted in Fig. 1b. Formally,

$$\mathbb{A} = \{\boldsymbol{a}^* \in \mathbb{R}^q : \gamma(\boldsymbol{e}, \boldsymbol{a}^*) \in \phi(\mathbb{B}(\boldsymbol{x}, \epsilon))\} = \{\boldsymbol{a}^* \in \mathbb{R}^q : \big\|\phi^{-1}\big(\gamma(\boldsymbol{e}, \boldsymbol{a}^*)\big) - \boldsymbol{x}\big\|_\infty \leq \epsilon\}. \tag{5}$$

Solving this problem by a projected gradient descent (PGD) scheme, analogous to (Madry et al., 2018), would amount to iterating over two phases: updating $\boldsymbol{a}^*$ in the direction that minimizes $\mathcal{L}$ to promote misclassification and then projecting the updated $\boldsymbol{a}^*$ back into $\mathbb{A}$ as illustrated in Fig. 2a. Unfortunately, deriving this needed projection is practically unfeasible due to irregular shape of the perturbation shape for $q \geq 2$.[1] Thus, we need a fundamentally different approach to solve (3).

2.3 THE BARRIER METHOD

To remove the need for projection, we propose a method entirely different from the PGD approach from (Madry et al., 2018). It is based on the so-called barrier method from nonlinear programming (Chachuat, 2007; Nocedal & Wright, 2006). In the context of adversarial attacks, the barrier method was used before by (Finlay et al., 2019) to enforce a decision boundary constraint, which is fundamentally different from the subspace constraint that we are targeting.



(a) PGD (b) Barrier

Figure 2: Comparison between one update step of PGD vs. the barrier method.

To apply the barrier method, we first rewrite (3) into an inequality to obtain the standard form of nonlinear programming problems. This is straightforward using the definition of $\mathbb{A}$ in (5):

$$\min_{\boldsymbol{a}^* \in \mathbb{R}^q} \mathcal{L}(\boldsymbol{a}^*) \text{ subject to } g(\boldsymbol{a}^*) \leq 0, \text{ where } g(\boldsymbol{a}^*) = \left\| \phi^{-1}\big(\gamma(\boldsymbol{e}, \boldsymbol{a}^*)\big) - \boldsymbol{x} \right\|_\infty - \epsilon. \tag{6}$$

Problem translation. The barrier method translates problem (6) into the form

$$\min_{\boldsymbol{a}^*} \theta(\mu) \quad \text{s.t. } \mu \geq 0, \tag{7}$$

where $\theta(\mu) = \inf\{\mathcal{L}(\boldsymbol{a}^*) + \mu b(\boldsymbol{a}^*)) : g(\boldsymbol{a}^*) < \boldsymbol{0}\}$. The barrier function b is intended to take the value zero on $\mathbb{A}$, and the value ∞ on its boundary. This guarantees that $\boldsymbol{a}^*$ does not leave $\mathbb{A}$, and consequently the solution $\boldsymbol{x}^*$ does not leave $\mathbb{S}_\phi(\boldsymbol{x})$ provided that the minimization problem starts at an interior point. However, this discontinuity poses difficulties for gradient-based solvers. Therefore, a more realistic construction of b would be non-negative and continuous inside $\mathbb{A}$ and approach infinity as the boundary of $\mathbb{A}$ is approached. We adopt this choice: $b(\boldsymbol{a}^*) = -\frac{1}{g(\boldsymbol{a}^*)}$.

As a result, if we minimize the function $\mathcal{L}(\boldsymbol{a}^*) + \mu b(\boldsymbol{a}^*)$ starting from a point in the interior of $\mathbb{A}$, the term $b(\boldsymbol{a}^*)$ approaches infinity as $\boldsymbol{a}^*$ moves near the boundary preventing the violation of the constraint $g(\boldsymbol{a}^*) \leq \boldsymbol{0}$.

The concrete algorithm. Usually, the minimization in (7) is performed by a second-order Newton or quasi-Newton solver (Chachuat, 2007). However, we opt for a fast first-order update rule, which we found more practical in our setting (after setting $\boldsymbol{a}^0$ to $\boldsymbol{a}$):

$$\boldsymbol{a}^{t+1} = \boldsymbol{a}^t - \eta \cdot \operatorname{sign}\big(\nabla_{\boldsymbol{a}} \mathcal{L}(\boldsymbol{a}^t) + \mu \nabla_{\boldsymbol{a}} b(\boldsymbol{a}^t)\big). \tag{8}$$

The idea of this update is that the gradient of the barrier function $\nabla_{\boldsymbol{a}} b(\boldsymbol{a}^t)$ pushes back when $\boldsymbol{a}^t$ approaches the boundary of $\mathbb{A}$ from the interior (see Fig. 2b for an illustration). Since this gradient has very small values on points that are far from the boundary (as b is flat around the center of $\mathbb{A}$), the step size η should therefore be small enough to allow $\boldsymbol{a}^{t+1}$ to progress slowly toward the boundary where $\nabla_{\boldsymbol{a}} b(\boldsymbol{a}^t)$ shows its effect, instead of causing a large leap that might drive $\boldsymbol{a}^{t+1}$ out of $\mathbb{A}$ as in PGD.

After T iterations, we report the modified image found in this subspace $\boldsymbol{x}^* = \phi^{-1}(\gamma(\boldsymbol{e}, \boldsymbol{a}^T))$. Just as in the original PGD attack, there is no optimality guarantee of this solution for an arbitrary classifier f.

Dealing with the discontinuity of the L^∞ norm. Computing the gradient of the barrier function b in the iterative update (8) using the chain rule involves computing the gradient of g, i.e., the gradient

[1] Fig. 1b is misleading here since $\mathbb{A}$ has only one dimension: $q = 1$.

of the L^∞ norm $\nabla||.||_\infty$. The latter is highly sparse as only one dimension is +1 or -1 (the one with the maximum absolute value) and all the other dimensions are 0. As a result, using it during optimization causes oscillation issues leaning to a poor convergence (see Sec.VI.C of (Carlini & Wagner, 2017) for a numerical example).

In our work, we eliminate this issue by replacing g with another function $\tilde{g} : \mathbb{R}^q \to \mathbb{R}^{2n}$ that equivalently characterizes the set $\mathbb{A} = \{\boldsymbol{a}^* \in \mathbb{R}^q : \tilde{g}(\boldsymbol{a}^*) \leq \mathbf{0}\}$ and is defined as follows:

$$\tilde{g}(\boldsymbol{a}^*)_k = \begin{cases} \phi^{-1}(\gamma(\boldsymbol{e}, \boldsymbol{a}^*))_k - \boldsymbol{x}_k - \epsilon, & \text{for } k \leq n, \\ -\phi^{-1}(\gamma(\boldsymbol{e}, \boldsymbol{a}^*))_{k-n} + \boldsymbol{x}_{k-n} - \epsilon, & \text{otherwise.} \end{cases}$$

Hence, the barrier function b is replaced by $\tilde{b}(\boldsymbol{a}^*) = -\sum_{k=1}^{2n} \frac{1}{\tilde{g}(\boldsymbol{a}^*)_k}$ and then used in (8).

In the implementation, we also consider the natural range of pixels $[0, 1]^n$. That is by enforcing the inequalities $\boldsymbol{x}_k \geq 0$ and $\boldsymbol{x}_k \leq 1$ for all $k = 1, .., n$ through the same procedure detailed above.

3 Instantiation for Different Image Representations

Our attack can be used with any image transformation ϕ that satisfies the conditions stated in Section 2.1 and any choice of split operator σ. In this paper we consider three instantiations of ϕ: the two classical linear DCT and DWT from the JPEG and JPEG2000 standards (Wallace, 1992; Adams, 2001), and a learned transform based on the flow-based model Glow.

DCT. As in JPEG, we apply the DCT on 8×8 blocks and first convert from RGB to the YCbCr color space[2]. We determine the auxiliary $\boldsymbol{a}$ in the DCT domain by inspecting the JPEG quantization tables; namely $\boldsymbol{a}$ collects the frequencies that are most severally reduced in the JPEG compression step (entries with large values in the quantization table, refer to Appendix A). Those are 12 out of 64 luminance frequencies and 51 out of 64 chrominance frequencies for each block. The others are assigned to the essential $\boldsymbol{e}$. As a result, $p \approx 0.4n$.

DWT. As for the DCT, and in JPEG2000, Similarly, we first perform a color conversion before applying the two-dimensional wavelet to the entire image. As DWT we use the Cohen-Daubechies-Feauveau (CDF) 9/7 lowpass and highpass filters [3]. The result is a downscaled version of the image that we consider the essential $\boldsymbol{e}$ of dimension $p = n/4$, plus horizontal, vertical, and diagonal details that we assign to the auxiliary $\boldsymbol{a}$.

Glow. Many deep learning techniques provide meaningful representation of images such as variational auto-encoders (VAEs) (Kingma & Welling, 2013). As a third instantiation for our attack, we chose the flow-based model Glow (Kingma & Dhariwal, 2018) because it is bijective with an exact formula for the inverse, unlike other flow-based models for which one can compute the inverse only iteratively such as iResNets (Behrmann et al., 2018).

Glow is a normalizing flow (Papamakarios et al., 2021), a sequence of invertible mappings that transform images $\boldsymbol{x} \in \mathbb{R}^n$ drawn from a complex intractable probability distribution, that is accessed through sampling, to latent vectors with the same dimension belonging to a Gaussian distribution $\boldsymbol{z} \in \mathbb{R}^n$.

Even when Glow is trained only on images without labels, the latent space has been shown to be useful for down-stream tasks (Kingma & Dhariwal, 2018; Peychev et al., 2022). Further, we are particularly interested in the class-conditional variant of Glow (Kingma & Dhariwal, 2018), where a classification loss is introduced to effectively predict the label of the input image using only one quarter of components of the latent vector $\boldsymbol{z}$. This can be viewed as a way to force these component to contain the most essential features needed to identify objects within images. More details about how we trained this model are provided in Section 4. The essential $\boldsymbol{e}$ collects the aforementioned quarter of $\boldsymbol{z}$ with $p = n/4$, while the rest is the auxiliary $\boldsymbol{a}$.

4 Experimental evaluation

[2] Y is the luminance component and Cb and Cr are the chrominance components of the blue and red difference

[3] as defined in `https://ch.mathworks.com/help/wavelet/ref/dwtfilterbank.html`

	Proposed attacks on the subspaces $\mathbb{S}_\phi$			Baseline attacks on the full box $\mathbb{B}$				
	barrier-glow	*barrier-dwt*	*barrier-dct*	*barrier*	*pgd*	*apgd-ce*	*apgd-dlr*	*square*
$\epsilon = 0.025$								
avg. L^∞	0.003089	0.0137	0.01138	0.02282	0.025	0.025	0.025	0.02447
avg. L^2	0.02929	0.1752	0.117	0.7427	0.9015	1.134	1.012	1.346
avg. LPIPS	7.371e-06	1.62e-05	2.903e-06	0.0007075	0.0008824	0.001483	0.0009306	0.00866
success rate (%)	1.82	44.64	26.38	89.06	100	100	100	97.88
$\epsilon = 0.05$								
avg. L^∞	0.01043	0.03459	0.03118	0.04995	0.05	0.05	0.05	0.04999
avg. L^2	0.09889	0.4258	0.327	1.465	1.504	2.07	1.994	2.738
avg. LPIPS	5.066e-05	7.237e-05	1.546e-05	0.002828	0.002967	0.006253	0.004671	0.03012
success rate (%)	4.09	77.35	50.55	99.86	100	100	100	99.98
$\epsilon = 0.1$								
avg. L^∞	0.0345	0.06623	0.06067	0.1	0.1	0.1	0.1	0.1
avg. L^2	0.3329	0.759	0.6359	2.554	2.579	3.838	3.898	5.425
avg. LPIPS	0.0003574	0.000223	5.293e-05	0.01036	0.01055	0.02591	0.02164	0.07852
success rate (%)	9.41	97.5	81.54	100	100	100	100	100
$\epsilon = 0.15$								
avg. L^∞	0.06477	0.09185	0.08248	0.15	0.15	0.15	0.15	0.15
avg. L^2	0.6381	1.01	0.8672	3.558	3.581	5.547	5.726	8.045
avg. LPIPS	0.001051	0.0004038	9.732e-05	0.02137	0.02159	0.05413	0.04825	0.1303
success rate (%)	15.53	99.82	93.95	100	100	100	100	100
$\epsilon = 0.2$								
avg. L^∞	0.09581	0.1141	0.1008	0.2	0.2	0.2	0.2	0.2
avg. L^2	0.9616	1.224	1.066	4.53	4.552	7.171	7.616	10.58
avg. LPIPS	0.002097	0.0006111	0.0001472	0.03471	0.03506	0.0849	0.08265	0.1774
success rate (%)	21.95	99.98	97.94	100	100	100	100	100

Table 1: Evaluation of our three proposed attacks against five baseline attacks for various L^∞ box radii ϵ on the correctly classified images of CIFAR10 testset (9546 images) using $T = 30$ iterations. We show the average L^∞, L^2, and LIPS distance of the obtained adversarial examples compared to the original (lower is better). Further, we show the success rate of the attacks. An analysis of the similarity-success rate trade-off is provided in Appendix. B

In this section, we first examine our adversarial attack on naturally trained classification models, namely a DenseNet121 (Huang et al., 2016) for CIFAR-10 and a vision transformer (ViT-B-16) (Dosovitskiy et al., 2020) for ImageNet. Since our attacks operate within predefined L^∞ boxes, we compare them against state-of-the-art L^∞-based attacks. Then we leverage our attacks for adversarial training and evaluate the robustness of the obtained networks against common image corruptions. All of our code and scripts to reproduce the experiments will be made available under a GPLv2 license.

The three instantiations of our method from Section 3 are called *barrier-dct*, *barrier-dwt*, and *barrier-glow* with associated perturbation spaces (see (2)) $\mathbb{S}_{dct}$, $\mathbb{S}_{dwt}$, and $\mathbb{S}_{glow}$, respectively. We also implemented an attack *barrier* in the pixel domain with standard box constraint based on the barrier method as a sanity check for comparison to PGD. Following (Kingma & Dhariwal, 2018), the class-conditional Glow architecture is composed of 3 flow levels; the depth of each level is 32, trained on CIFAR-10 for 1600 epochs with a batch size of 512.

Figure 3: A sample from ImageNet under the same settings as Fig. 4 (the Glow instantiation is omitted due to the high training cost on 256x256 images).

4.1 Comparison of attacks

We compare our attacks against four baselines: standard PGD, the two variants of the automatic projected gradient descent attack (APGD) (Croce & Hein, 2020), and the square attack (Andriushchenko et al., 2020). We ran all the mentioned attacks on the correctly classified images of CIFAR-10 testset (9546 images) and report averages in Table 1 for various choice of box bounds ϵ, showing L^∞ distance, L^2 distance, attack success rates and distance using the LPIPS similarity metric that relies on deep features learned in supervised/self-supervised/unsupervised regimes proven effective in capturing similarity between images (Zhang et al., 2018).

Further, we randomly selected images to show the found adversarial examples and the associated difference to the original in Figs. 4 (CIFAR-10) and 3 (ImageNet). Finally, for $\epsilon = 0.1$, Fig. 5 shows the interplay between LPIPS distance, L^∞ distance, and attack success rate.

First, Fig. 4 visually shows that the adversarial images found by our attacks are significantly less visually impacted compared to those reported by the pixel-based attacks, even when targeting relatively large L^∞ box radii. Table 1 confirms this higher visual similarity by a consistently lower values in both LPIPS similarity distance and L^2 distance across all experiments. We also observe that attacks based on Glow produce results that are adapted to the depicted scenery. This is not the case for DCT and DWT which modulate details. Fig. 3 shows the same behavior on image net where we consider in the instantiation of our method for classical transforms *barrier-dct* and *barrier-dwt* without the learning-based instantiation *barrier-glow* due the high cost of training this generative model on 256 by 256 images (even experiments of the paper proposing Glow (Kingma & Dhariwal, 2018) down-scaled ImageNet images to 32x32 or 64x64). We notice that *barrier-dct* introduces block boundary artifacts. In the L^∞ metric we observe that our proposed attacks use the freedom provide by ϵ but, unlike all benchmarks, typically do not find adversarial examples at the boundary of the box. This is also explained by the lower dimension of the perturbation subspace that we consider, and is not an intrinsic consequence of using the barrier method, since, when applied in the image domain on the entire box (our sanity check, first column of baseline attacks in Table 1), it operates similar to the PGD attack.

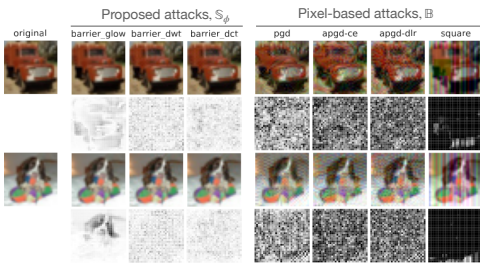


Figure 4: Two randomly selected images from CIFAR-10 and adversarial examples found by our proposed attacks compared to prior attacks, all run with L^∞ box radius value of $\epsilon = 0.1$. Each row of adversarial images is followed by a row of heatmaps representing the pixelwise difference w.r.t the corresponding clean image in the first column: white $= 0$ and black $= \epsilon$.

4.2 Adversarial training for practical robustness

The high similarity and semantic nature (in the sense of the transform being used) of the adversarial examples produced by our attacks motivates their use as proxies to achieve robustness against another class of image perturbations that preserves visual similarity: common image corruptions. To do so, we use our attacks for adversarial training (AT), a technique in which the neural network is trained on adversarial examples aiming to increase robustness against this adversary.

Specifically, we adopt the AT technique TRADES (Zhang et al., 2019), which is a heuristic algorithm based on multi-class calibrated loss theory that balances the trade-off between robustness and accuracy. We train for robustness under four types of constraints: L^∞ box $\mathbb{B}$, and DCT subspace $\mathbb{S}_{\texttt{dct}}$, DWT subspace $\mathbb{S}_{\texttt{dwt}}$ and Glow subspace $\mathbb{S}_{\texttt{glow}}$ against their corresponding adversaries: the standard *pgd*, *barrier-dct*, *barrier-dwt*, and *barrier-glow*, respectively. All ATs are granted the same number of iterations $T = 10$ to fetch adversarial examples for each training epoch using a common choice of $\epsilon = 0.05$. The naturally trained vanilla model is the same as Sec. 4.1. All models are trained without any additional data. All models are DensNet121 (Huang et al., 2016) which is a multigrade architecture found to resist noise corruptions more effectively than ResNets (Hendrycks & Dietterich, 2019).

We consider the CIFAR-10-C dataset (Hendrycks & Dietterich, 2019), a benchmark constructed by applying common image corruptions to the CIFAR-10 test set. These corruptions are only used for evaluation and not to augment the data during training. The results for all categories in CIFAR-10-C are reported in Table 2. The first data column show the accuracy on the clean, unperturbed images. The latter columns show the accuracy on corrupted images, considering 18 different types of corruptions: different forms of blurring, digital, noise corruptions, and weather related ones.

First, networks trained under our subspace constraints achieve higher accuracy on corrupted images compared to the model trained under the box constraint across all types of corruptions. They do so while suffering only a minor reduction in natural accuracy compared to the vanilla network. The performance of AT with our DWT subspace constraint $\mathbb{S}_{dwt}$ stand out in particular. On average, it is

Table 2: Accuracies on different image corruption categories of networks adversarially trained under different constraints.

	Natural	Blur				Digital				
		Gauss	Motion	Defocus	Zoom	Contrast	Elastic	JPEG	Pixel	Saturate
Vanilla (no AT)	**95.46**	72.36	81.22	83.81	77.74	**83.01**	85.04	81.07	74.83	**92.46**
Full L^∞ box $\mathbb{B}$	81.23	75.36	73.77	77.12	76.30	45.47	75.54	79.30	79.23	77.75
Proposed subspace $\mathbb{S}_{\texttt{dct}}$	94.84	76.36	79.06	85.01	80.81	78.45	86.06	88.02	77.13	92.23
Proposed subspace $\mathbb{S}_{\texttt{dwt}}$	93.83	**87.95**	**85.55**	**89.91**	**88.79**	67.28	**88.42**	**89.44**	**92.09**	88.93
Proposed subspace $\mathbb{S}_{\texttt{glow}}$	95.41	77.56	84.27	85.93	81.75	82.81	86.73	82.36	75.31	92.23
	Natural	Noise				Weather				
		Gauss	Impulse	Speckle	Shot	Snow	Fog	Frost	Bright	Spatter
Vanilla (no AT)	**95.46**	48.50	60.21	65.09	61.09	84.33	89.15	81.55	**94.00**	87.99
Full L^∞ box $\mathbb{B}$	81.23	75.61	73.78	76.67	76.68	75.15	59.69	71.02	77.92	77.24
Proposed subspace $\mathbb{S}_{\texttt{dct}}$	94.84	68.83	68.71	76.56	75.83	85.37	86.84	83.60	93.62	87.86
Proposed subspace $\mathbb{S}_{\texttt{dwt}}$	93.83	**81.83**	**73.98**	**84.93**	**85.24**	**88.89**	80.34	**88.79**	92.05	**89.78**
Proposed subspace $\mathbb{S}_{\texttt{glow}}$	95.41	43.77	57.32	60.33	56.15	85.50	**90.81**	82.44	93.99	87.28

7.31% more accurate than the vanilla model and 12.27% more accurate than the model trained under the standard box constraint $\mathbb{B}$. The sacrifice in natural accuracy is only about 1.5%. We note that these results are in line with recent advances in machine learning interpretability where the wavelet domain also provides better performance than the pixel-based methods (Kolek et al., 2022).

Finally, we note that AT with $\mathbb{S}_{glow}$ does not perform well on corruption, possibly since the features learned by the flow-based model are semantically at a higher level, and thus not compatible with the considered corruptions that are closer related to the DCT and DWT frequency representations. However, AT with $\mathbb{S}_{glow}$ practically maintains the natural accuracy.

Limitations and discussion. For the experiment in Table 2 there are techniques that achieve better corruption robustness. These are not based on adversarial attacks but on other data augmentation techniques (Geirhos et al., 2018; Erichson et al., 2022; Zhang et al., 2017; Hendrycks et al., 2019; Park et al., 2022; Yin et al., 2022; Liang et al., 2023). They are specifically targeting this benchmark (whereas our approach is oblivious to it) and usually train substantially larger networks (e.g., WideResNet-28-4 used by NoisyMix (Erichson et al., 2022)) and require pre-training on larger datasets.

Our goal was to expand the tool set of adversarial attacks and to also make progress on the link between adversarial robustness and corruption robustness as a followup to the findings of (Hendrycks & Dietterich, 2019; Ford et al., 2019; Xie et al., 2020; Kang et al., 2019; Kireev et al., 2022). The precise specification of our perturbation space makes porting of state-of-the-art certification techniques (either approximation-based (Singh et al., 2019; Müller et al., 2022) or probabilistic (Cohen et al., 2019; Carlini et al., 2022)) to operate under the proposed constraints possible. The generality of our approach in the choice of transform ϕ and associated subspace to be perturbed invites further exploration.

5 Related work

We cited a number of related work in the introduction and throughout the paper. Here we focus on prior uses of transformed image representations. In particular, discrete linear transforms have been used in machine learning for different purposes. For example, (Gueguen et al., 2018; dos Santos & Almeida, 2021) proposed DCT-based architectures operating directly on the JPEG format to avoid decompression before inference. Furthermore, (Kolek et al., 2022) have extended the rate-distortion framework (MacDonald et al., 2019) to the wavelet domain to build a state-of-the-art explanation method for DNN. The remainder of this section is focused on previous works related to robustness.

Discrete transforms for robustness. Most prior work using discrete transforms aimed at defending against pixel-based adversarial attacks or improving the generalization of neural networks towards common image corruptions. The work of (Dziugaite et al., 2016; Das et al., 2017; Guo et al., 2018) aims to filter out noise from the adversarial examples by adjusting various quality factor values during JPEG compression/decompression, which amounts to reducing the magnitude of the DCT coefficients. Closely related, (Bafna et al., 2018) sought L^0 robustness through projecting the

largest DCT coefficients. These defenses have been shown to be breakable through adaptive attacks (Shin & Song, 2017; Tramèr et al., 2020), specifically, by approximating the non-differentiable rounding operator of the JPEG compression and running a gradient-based attack. Other fast and iterative rounding schemes have been proposed in (Shi et al., 2021b). (Yin et al., 2019; Guo et al., 2019) considers L^2 perturbations that preserve norms due to orthogonality of the used transforms, discrete Fourier transform (DFT) and DCT respectively. The work in (Duan et al., 2021) generates adversarial examples by removing information in the DCT domain. The L^∞ box used is on the JPEG quantization matrix instead of the input image. Since the DCT coefficients of the clean image are element-wise divided by this matrix before rounding, larger box radii allow their technique to eliminate more frequencies from the image. In the same direction, (Hossain et al., 2019) preceded the neural network by a DCT based layer that randomly crops some DCT coefficients during training. This can be interpreted as an extension of the dropout technique aiming at its regularization effects. (Yahya et al., 2020) propose a gradient-free method that obtains adversarial examples by mixing the frequencies of a clean image with the frequencies of another auxiliary image that they call watermark. In addition to DFT and DCT, they make use of two wavelets: Haar and Daubechies 3. (Sharma et al., 2019) applies masks to selectively perturb low and high frequencies. Much like (Deng & Karam, 2020; Shi et al., 2021a), all these works do not provide any guarantee on the L^∞ bounds in the pixel space, which is the primary contribution in our work. We can also target low dimensional spaces which (Long et al., 2022) cannot. Yuan et al. (2022) proposes a DCT-based attack and uses the fact that DCT is linear and orthogonal, where our method only needs invertible and differentiable (e.g. Glow) since we do not need projections due to the barrier method. (Luo et al., 2022; Wang et al., 2021; Laidlaw et al., 2021; Kireev et al., 2022) explicitly uses a similarity distance in the optimization problem formulation in the pursuit of semantically similar adversarial examples.

In contrast to all prior work we show how to perform attacks on a subspace of an image representation, linear or not, that also enforces the L^∞ box in the image domain.

Learning-based methods for robustness. Flow-based generative models themselves are prone to adversarial attacks that manipulate their likelihood scores (Pope et al., 2020). That is a different focus from our work as we are used a flow-based generative model (in one of our 3 instances) to define a meaningful subspace rather than a likelihood estimation. (Huang & Zhang, 2020; Baluja & Fischer, 2018) trained NNs to produced perturbations under the L^∞ and L^2 constraints that can operate in the black-box settings. Adversarial generative models (GANs) also have been trained to generate adversarial examples in semi-whitebox and black-box settings. (Dolatabadi et al., 2020) used a pre-trained flow-based model (RealNVP (Dinh et al., 2017)) to craft adversarial examples in black-box settings under the L^∞ constraint. They used an additive noise in the latent space where they faced a similar projection problem as this work. They solve it by going back and forth to the pixel space to project using the PGD formula. In contrast, we removed the projection by using the barrier method in a way that is fast enough to be incorporated in adversarial training. More importantly, we operate in a subspace within the box instead of the full box. We show that doing so is effective in AT to produce networks robust to common image corruptions. Some works altered the semantic features of images through conditional generative models (Joshi et al., 2019) or conditional image editing(Qiu et al., 2020), but those are not bounded to a norm.

6 Conclusion

We have expanded the toolbox of adversarial attacks, and associated adversarial training, with a general, and powerful novel method. The novelty is twofold. First, in the ability to attack semantically (in a sense associated with the chosen transform) in a suitable transformed image representation space while preserving proximity in the pixel space. Second, in using the barrier method needed to enable such an attack when projections are not available. Thus our approach fuses two prior lines of research that attack in either of these spaces. We emphasize that the transform used does not need to be linear, only invertible, as we show by also considering a learned transform. The benefit of our approach is best visible when used for adversarial training: in particular with the DWT a major improvement in accuracy on a broad range of corruptions with only a small drop in natural accuracy.

The generality of our work in both chosen transform and chosen subspace should invite further exploration. In particular, by leveraging decades of research on image representations for better defining constraints under which adversarial robustness is studied.

REFERENCES

Michael D. Adams. The jpeg-2000 still image compression standard, 2001.

Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square attack: a query-efficient black-box adversarial attack via random search. In *European Conference on Computer Vision*, pp. 484–501. Springer, 2020.

Mitali Bafna, Jack Murtagh, and Nikhil Vyas. Thwarting adversarial examples: An l_0-robust sparse fourier transform. *Advances in Neural Information Processing Systems*, 31, 2018.

Shumeet Baluja and Ian Fischer. Learning to attack: Adversarial transformation networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Jens Behrmann, Will Grathwohl, Ricky T. Q. Chen, David Duvenaud, and Jörn-Henrik Jacobsen. Invertible residual networks. 2018. doi: 10.48550/ARXIV.1811.00995. URL https://arxiv.org/abs/1811.00995.

Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In *Joint European conference on machine learning and knowledge discovery in databases*, pp. 387–402. Springer, 2013.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 39–57, 2017. doi: 10.1109/SP.2017.49.

Nicholas Carlini, Florian Tramer, Krishnamurthy Dj Dvijotham, Leslie Rice, Mingjie Sun, and J Zico Kolter. (certified!!) adversarial robustness for free! *arXiv preprint arXiv:2206.10550*, 2022.

Benoit Chachuat. *Nonlinear and Dynamic Optimization: From Theory to Practice*. 2007. URL https://infoscience.epfl.ch/record/111939?ln=fr.

Jinghui Chen, Yu Cheng, Zhe Gan, Quanquan Gu, and Jingjing Liu. Efficient robust training via backward smoothing, 2021. URL https://openreview.net/forum?id=49V11oUejQ.

Jeremy M Cohen, Elan Rosenfeld, and J. Zico Kolter. Certified adversarial robustness via randomized smoothing, 2019. URL https://arxiv.org/abs/1902.02918.

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

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 preprint arXiv:1705.02900*, 2017.

Yingpeng Deng and Lina J. Karam. Frequency-tuned universal adversarial perturbations. In Adrien Bartoli and Andrea Fusiello (eds.), *Computer Vision – ECCV 2020 Workshops*, pp. 494–510, Cham, 2020. Springer International Publishing. ISBN 978-3-030-68238-5.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real NVP. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=HkpbnH9lx.

Hadi Mohaghegh Dolatabadi, Sarah Erfani, and Christopher Leckie. Advflow: Inconspicuous black-box adversarial attacks using normalizing flows. In *Proceedings of the Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

Samuel Felipe dos Santos and Jurandy Almeida. Less is more: Accelerating faster neural networks straight from jpeg. In *Iberoamerican Congress on Pattern Recognition*, pp. 237–247. Springer, 2021.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Ranjie Duan, Yuefeng Chen, Dantong Niu, Yun Yang, A Kai Qin, and Yuan He. Advdrop: Adversarial attack to dnns by dropping information. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7506–7515, 2021.

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

N. Benjamin Erichson, Soon Hoe Lim, Winnie Xu, Francisco Utrera, Ziang Cao, and Michael W. Mahoney. Noisymix: Boosting model robustness to common corruptions, 2022. URL https://arxiv.org/abs/2202.01263.

Abolfazl Farahani, Sahar Voghoei, Khaled Rasheed, and Hamid R. Arabnia. A brief review of domain adaptation, 2020. URL https://arxiv.org/abs/2010.03978.

Reuben Feinman, Ryan R. Curtin, Saurabh Shintre, and Andrew B. Gardner. Detecting adversarial samples from artifacts. In *International Conference on Machine Learning*, 2017.

Chris Finlay, Aram-Alexandre Pooladian, and Adam Oberman. The logbarrier adversarial attack: Making effective use of decision boundary information. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4861–4869, 2019. doi: 10.1109/ICCV.2019.00496.

Nicolas Ford, Justin Gilmer, and Ekin D. Cubuk. Adversarial examples are a natural consequence of test error in noise, 2019. URL https://openreview.net/forum?id=S1xoy3CcYX.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness, 2018. URL https://arxiv.org/abs/1811.12231.

Shixiang Gu and Luca Rigazio. Towards deep neural network architectures robust to adversarial examples. *arXiv preprint arXiv:1412.5068*, 2014.

Lionel Gueguen, Alex Sergeev, Ben Kadlec, Rosanne Liu, and Jason Yosinski. Faster neural networks straight from jpeg. *Advances in Neural Information Processing Systems*, 31, 2018.

Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens van der Maaten. Countering adversarial images using input transformations. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=SyJ7ClWCb.

Chuan Guo, Jared S. Frank, and Kilian Q. Weinberger. Low frequency adversarial perturbation. In Amir Globerson and Ricardo Silva (eds.), *Proceedings of the Thirty-Fifth Conference on Uncertainty in Artificial Intelligence, UAI 2019, Tel Aviv, Israel, July 22-25, 2019*, volume 115 of *Proceedings of Machine Learning Research*, pp. 1127–1137. AUAI Press, 2019. URL http://proceedings.mlr.press/v115/guo20a.html.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HJz6tiCqYm.

Dan Hendrycks, Norman Mu, Ekin D. Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty, 2019. URL https://arxiv.org/abs/1912.02781.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. *ICCV*, 2021.

Md Tahmid Hossain, Shyh Wei Teng, Dengsheng Zhang, Suryani Lim, and Guojun Lu. Distortion robust image classification using deep convolutional neural network with discrete cosine transform. In *2019 IEEE International Conference on Image Processing, ICIP 2019, Taipei, Taiwan, September 22-25, 2019*, pp. 659–663. IEEE, 2019. doi: 10.1109/ICIP.2019.8803787. URL https://doi.org/10.1109/ICIP.2019.8803787.

Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks, 2016. URL https://arxiv.org/abs/1608.06993.

Zhichao Huang and Tong Zhang. Black-box adversarial attack with transferable model-based embedding. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SJxhNTNYwB.

Yulun Jiang, Chen Liu, Zhichao Huang, Mathieu Salzmann, and Sabine Susstrunk. Towards stable and efficient adversarial training against l_1 bounded adversarial attacks. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 15089–15104. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/jiang23f.html.

Ameya Joshi, Amitangshu Mukherjee, Soumik Sarkar, and Chinmay Hegde. Semantic adversarial attacks: Parametric transformations that fool deep classifiers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4773–4783, 2019.

Daniel Kang, Yi Sun, Tom Brown, Dan Hendrycks, and Jacob Steinhardt. Transfer of adversarial robustness between perturbation types. *arXiv preprint arXiv:1905.01034*, 2019.

Harini Kannan, Alexey Kurakin, and Ian Goodfellow. Adversarial logit pairing, 2018. URL https://arxiv.org/abs/1803.06373.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2013. URL https://arxiv.org/abs/1312.6114.

Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/d139db6a236200b21cc7f752979132d0-Paper.pdf.

Klim Kireev, Maksym Andriushchenko, and Nicolas Flammarion. On the effectiveness of adversarial training against common corruptions. In *The 38th Conference on Uncertainty in Artificial Intelligence*, 2022. URL https://openreview.net/forum?id=BcU_UIIjqg9.

Stefan Kolek, Duc Anh Nguyen, Ron Levie, Joan Bruna, and Gitta Kutyniok. Cartoon explanations of image classifiers, 2022. URL https://openreview.net/forum?id=RYTBAtyXqJ.

Cassidy Laidlaw, Sahil Singla, and Soheil Feizi. Perceptual adversarial robustness: Defense against unseen threat models. In *ICLR*, 2021.

Wen Liang, Youzhi Liang, and Jianguo Jia. Miamix: Enhancing image classification through a multi-stage augmented mixed sample data augmentation method, 2023.

Fangzhou Liao, Ming Liang, Yinpeng Dong, Tianyu Pang, Xiaolin Hu, and Jun Zhu. Defense against adversarial attacks using high-level representation guided denoiser. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1778–1787, 2018.

Yuyang Long, Qilong Zhang, Boheng Zeng, Lianli Gao, Xianglong Liu, Jian Zhang, and Jingkuan Song. Frequency domain model augmentation for adversarial attack. In *Computer Vision – ECCV 2022*, pp. 549–566, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-19772-7.

Cheng Luo, Qinliang Lin, Weicheng Xie, Bizhu Wu, Jinheng Xie, and Linlin Shen. Frequency-driven imperceptible adversarial attack on semantic similarity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Xingjun Ma, Bo Li, Yisen Wang, Sarah M. Erfani, Sudanthi N. R. Wijewickrema, Michael E. Houle, Grant Schoenebeck, Dawn Song, and James Bailey. Characterizing adversarial subspaces using local intrinsic dimensionality. *CoRR*, abs/1801.02613, 2018. URL http://arxiv.org/abs/1801.02613.

Jan MacDonald, Stephan Wäldchen, Sascha Hauch, and Gitta Kutyniok. A rate-distortion framework for explaining neural network decisions. *CoRR*, abs/1905.11092, 2019. URL http://arxiv.org/abs/1905.11092.

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. URL https://openreview.net/forum?id=rJzIBfZAb.

Jan Hendrik Metzen, Tim Genewein, Volker Fischer, and Bastian Bischoff. On detecting adversarial perturbations. *ICLR*, 2017.

Eric Mintun, Alexander Kirillov, and Saining Xie. On interaction between augmentations and corruptions in natural corruption robustness, 2021. URL https://arxiv.org/abs/2102.11273.

Mark Niklas Müller, Franziska Eckert, Marc Fischer, and Martin Vechev. Certified training: Small boxes are all you need, 2022. URL https://arxiv.org/abs/2210.04871.

Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer, New York, NY, USA, 2e edition, 2006.

S.J. Pan and Q. Yang. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.

George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *J. Mach. Learn. Res.*, 22(1), jan 2021. ISSN 1532-4435.

Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *2016 IEEE European symposium on security and privacy (EuroS&P)*, pp. 372–387. IEEE, 2016a.

Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *2016 IEEE symposium on security and privacy (SP)*, pp. 582–597. IEEE, 2016b.

Chanwoo Park, Sangdoo Yun, and Sanghyuk Chun. A unified analysis of mixed sample data augmentation: A loss function perspective. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=SLdfxFdIFeN.

Momchil Peychev, Anian Ruoss, Mislav Balunovic, Maximilian Baader, and Martin Vechev. Latent space smoothing for individually fair representations, 2022. URL https://openreview.net/forum?id=DqJgzrcA8lH.

Phillip Pope, Yogesh Balaji, and Soheil Feizi. Adversarial robustness of flow-based generative models. In *International Conference on Artificial Intelligence and Statistics*, pp. 3795–3805. PMLR, 2020.

Haonan Qiu, Chaowei Xiao, Lei Yang, Xinchen Yan, Honglak Lee, and Bo Li. Semanticadv: Generating adversarial examples via attribute-conditioned image editing. In *ECCV*, 2020.

Sylvestre-Alvise Rebuffi, Sven Gowal, Dan Andrei Calian, Florian Stimberg, Olivia Wiles, and Timothy Mann. Data augmentation can improve robustness. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet?, 2019. URL https://arxiv.org/abs/1902.10811.

Hadi Salman, Jerry Li, Ilya Razenshteyn, Pengchuan Zhang, Huan Zhang, Sebastien Bubeck, and Greg Yang. Provably robust deep learning via adversarially trained smoothed classifiers. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/3a24b25a7b092a252166a1641ae953e7-Paper.pdf.

Yash Sharma, Weiguang Ding, and Marcus Brubaker. On the effectiveness of low frequency perturbations. pp. 3389–3396, 08 2019. doi: 10.24963/ijcai.2019/470.

Mengte Shi, Sheng Li, Zhaoxia Yin, Xinpeng Zhang, and Zhenxing Qian. On generating jpeg adversarial images. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, 2021a. doi: 10.1109/ICME51207.2021.9428243.

Mengte Shi, Sheng Li, Zhaoxia Yin, Xinpeng Zhang, and Zhenxing Qian. On generating jpeg adversarial images. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6. IEEE, 2021b.

Richard Shin and Dawn Song. Jpeg-resistant adversarial images. In *NIPS 2017 Workshop on Machine Learning and Computer Security*, volume 1, 2017.

Gagandeep Singh, Timon Gehr, Markus Püschel, and Martin Vechev. An abstract domain for certifying neural networks. *Proc. ACM Program. Lang.*, 3(POPL), jan 2019. doi: 10.1145/3290354. URL https://doi.org/10.1145/3290354.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. January 2014. 2nd International Conference on Learning Representations, ICLR 2014 ; Conference date: 14-04-2014 Through 16-04-2014.

Florian Tramèr, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. On adaptive attacks to adversarial example defenses. *CoRR*, abs/2002.08347, 2020. URL https://arxiv.org/abs/2002.08347.

G.K. Wallace. The jpeg still picture compression standard. *IEEE Transactions on Consumer Electronics*, 38(1):xviii–xxxiv, 1992. doi: 10.1109/30.125072.

Huaxia Wang and Chun-Nam Yu. A direct approach to robust deep learning using adversarial networks. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=S1lIMn05F7.

Yajie Wang, Shangbo Wu, Wenyi Jiang, Shengang Hao, Yu-an Tan, and Quanxin Zhang. Demiguise attack: Crafting invisible semantic adversarial perturbations with perceptual similarity. *CoRR*, abs/2107.01396, 2021. URL https://arxiv.org/abs/2107.01396.

Chaowei Xiao, Bo Li, Jun-Yan Zhu, Warren He, Mingyan Liu, and Dawn Song. Generating adversarial examples with adversarial networks. *arXiv preprint arXiv:1801.02610*, 2018.

Cihang Xie, Yuxin Wu, Laurens van der Maaten, Alan L Yuille, and Kaiming He. Feature denoising for improving adversarial robustness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 501–509, 2019.

Cihang Xie, Mingxing Tan, Boqing Gong, Jiang Wang, Alan L. Yuille, and Quoc V. Le. Adversarial examples improve image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Weilin Xu, David Evans, and Yanjun Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. In *25th Annual Network and Distributed System Security Symposium, NDSS 2018, San Diego, California, USA, February 18-21, 2018*. The Internet Society, 2018. URL http://wp.internetsociety.org/ndss/wp-content/uploads/sites/25/2018/02/ndss2018_03A-4_Xu_paper.pdf.

Zakia Yahya, Muhammad Hassan, Shahzad Younis, and Muhammad Shafique. Probabilistic analysis of targeted attacks using transform-domain adversarial examples. *IEEE Access*, 8:33855–33869, 2020. doi: 10.1109/ACCESS.2020.2974525.

Dong Yin, Raphael Gontijo Lopes, Jon Shlens, Ekin Dogus Cubuk, and Justin Gilmer. A fourier perspective on model robustness in computer vision. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/b05b57f6add810d3b7490866d74c0053-Paper.pdf.

Hao Yin, Dongyu Cao, and Ying Zhou. Randommix: An effective framework to protect user privacy information on ethereum. In *2022 IEEE 22nd International Conference on Software Quality, Reliability, and Security Companion (QRS-C)*, pp. 764–765, 2022. doi: 10.1109/QRS-C57518.2022.00124.

Zheng Yuan, Jie Zhang, and Shiguang Shan. Adaptive image transformations for transfer-based adversarial attack. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part V*, pp. 1–17, Berlin, Heidelberg, 2022. Springer-Verlag. ISBN 978-3-031-20064-9. doi: 10.1007/978-3-031-20065-6_1. URL https://doi.org/10.1007/978-3-031-20065-6_1.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pp. 7472–7482. PMLR, 2019.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization, 2017. URL https://arxiv.org/abs/1710.09412.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

Dawei Zhou, Tongliang Liu, Bo Han, Nannan Wang, Chunlei Peng, and Xinbo Gao. Towards defending against adversarial examples via attack-invariant features. In *International Conference on Machine Learning*, pp. 12835–12845. PMLR, 2021.

A SELECTING DCT COEFFICIENTS FOR THE SPLIT OPERATOR σ

These are the quantization tables used in JPEG for the luminance $\boldsymbol{Q}_Y$ and the two chrominance channels $\boldsymbol{Q}_C$. We select the DCT coefficients corresponding to entries below 99 (in bold) to the essential vector $\boldsymbol{e}$ while the rest is the auxiliary vector $\boldsymbol{a}$.

$$\boldsymbol{Q}_Y = \begin{bmatrix} \mathbf{16} & \mathbf{11} & \mathbf{10} & \mathbf{16} & \mathbf{24} & \mathbf{40} & \mathbf{51} & \mathbf{61} \\ \mathbf{12} & \mathbf{12} & \mathbf{14} & \mathbf{19} & \mathbf{26} & \mathbf{58} & \mathbf{60} & \mathbf{55} \\ \mathbf{14} & \mathbf{13} & \mathbf{16} & \mathbf{24} & \mathbf{40} & \mathbf{57} & \mathbf{69} & \mathbf{56} \\ \mathbf{14} & \mathbf{17} & \mathbf{22} & \mathbf{29} & \mathbf{51} & \mathbf{87} & \mathbf{80} & \mathbf{62} \\ \mathbf{18} & \mathbf{22} & \mathbf{37} & \mathbf{56} & \mathbf{68} & 109 & 103 & \mathbf{77} \\ \mathbf{24} & \mathbf{35} & \mathbf{55} & \mathbf{64} & \mathbf{81} & 104 & 113 & \mathbf{92} \\ \mathbf{49} & \mathbf{64} & \mathbf{78} & \mathbf{87} & 103 & 121 & 120 & 101 \\ \mathbf{72} & \mathbf{92} & \mathbf{95} & \mathbf{98} & 112 & 100 & 103 & 99 \end{bmatrix} \tag{9}$$

$$\boldsymbol{Q}_C = \begin{bmatrix} \mathbf{17} & \mathbf{18} & \mathbf{24} & \mathbf{47} & 99 & 99 & 99 & 99 \\ \mathbf{18} & \mathbf{21} & \mathbf{26} & \mathbf{66} & 99 & 99 & 99 & 99 \\ \mathbf{24} & \mathbf{26} & \mathbf{56} & 99 & 99 & 99 & 99 & 99 \\ \mathbf{47} & \mathbf{66} & 99 & 99 & 99 & 99 & 99 & 99 \\ 99 & 99 & 99 & 99 & 99 & 99 & 99 & 99 \\ 99 & 99 & 99 & 99 & 99 & 99 & 99 & 99 \\ 99 & 99 & 99 & 99 & 99 & 99 & 99 & 99 \\ 99 & 99 & 99 & 99 & 99 & 99 & 99 & 99 \end{bmatrix} \tag{10}$$

B THE SIMILARITY-SUCCESS RATE TRADE-OFF

Adversarial examples exhibit high similarity. The trade-off for this higher similarity is a lower success rate for very small ϵ, whereas the prior benchmarks almost always succeed as visualized in Fig. 5. This is due to the fact that, despite being bounded by the same L^∞, our methods operate only on a q-dimensional subspace of the n-dimensional box, where $q \approx 0.6n$ for *barrier-dct*, and $q = 0.75n$ for *barrier-dwt* and *barrier-glow*.

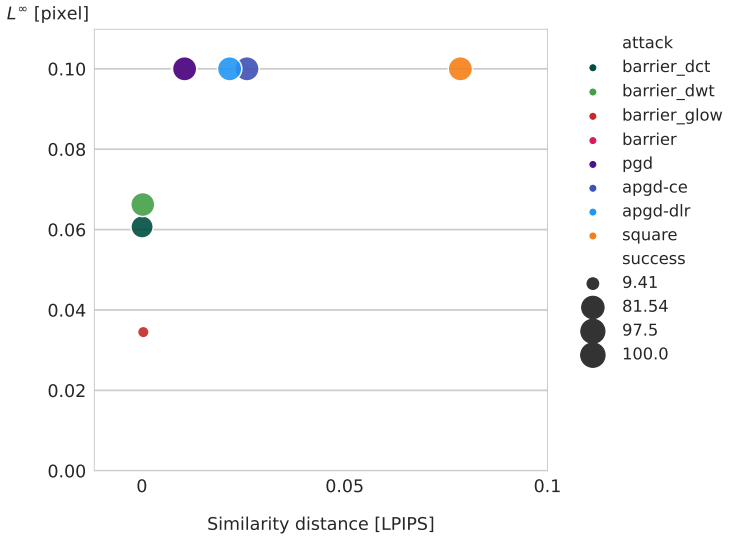


Figure 5: Interplay of LPIPS distance, L^∞ distance, and success rate (encoded by marker size) for $\epsilon = 0.1$ based on the numbers in Table 1.