Wavelet-based Low Frequency Adversarial Attacks

Shpresim Sadiku sadiku@zib.de
AI in Society, Science, and Technology & Institute of Mathematics
Zuse Institute Berlin & Technische Universität Berlin
Berlin, Germany

Moritz Wagner wagner@zib.de
AI in Society, Science, and Technology & Institute of Mathematics
Zuse Institute Berlin & Technische Universität Berlin
Berlin, Germany

Sebastian Pokutta pokutta@zib.de
AI in Society, Science, and Technology & Institute of Mathematics
Zuse Institute Berlin & Technische Universität Berlin
Berlin, Germany

Abstract

Despite their impressive success in various machine learning tasks, deep neural networks are vulnerable to adversarial attacks. Through the addition of imperceptible levels of distortion to a given image, such attacks can cause a learned network to quite spectacularly misclassify the perturbed input. Several defense approaches including adversarial training and methods manipulating basis function representations of images such as JPEG compression, PCA, wavelet denoising, and soft-thresholding have shown success. The former defense works well in defending against small ℓ_p norm attacks in the pixel representation, whereas the latter methods rely on removing high frequency signal. We show that both training-based and basis-manipulation defense methods are significantly less effective if we restrict the generation of adversarial attacks to the low frequency discrete wavelet transform (DWT) domain, thus providing new insights into vulnerabilities of deep learning models.

1 Introduction

As machine learning models become more widespread in real-life applications, their security becomes a more relevant aspect to consider. Despite the ability of deep neural networks to outperform humans in many tasks (Silver et al., 2016), several instabilities of such architectures to *adversarial attacks* have been discovered recently. Much research has since gone into the topic of developing adversarial examples and several defense methods have been proposed (Akhtar and Mian, 2018; Tramer et al., 2020).

Defense methods against adversarial attacks can be categorized into two main types. Approaches of the first type modify the training procedure or architecture of the model, usually with the aim of making the function learned by the neural network smoother, for instance by augmenting the training data with adversarial examples (Goodfellow et al., 2015; Shaham et al., 2018b; Cohen et al., 2019). However, such defenses are only effective against first-order attacks that are of the same type as the ones used during the training, for instance, small ℓ_p norm attacks in the pixel representation. A-priori, there is no restriction on the attacker operating only in the pixel domain. Given access to the input image, the attacker has the ability to operate in other natural representations of images, such as the one given by the *discrete cosine transform (DCT)* basis. The perturbations generated in the new basis are still *imperceptible* but do circumvent adversarially trained networks due to the large ℓ_p norm in the pixel basis (Awasthi et al., 2021).

Defenses of the second type do not modify the training procedure or the architecture, but rather modify the data, aiming to detect or remove adversarial attacks often by smoothing the input data. Shaham et al.

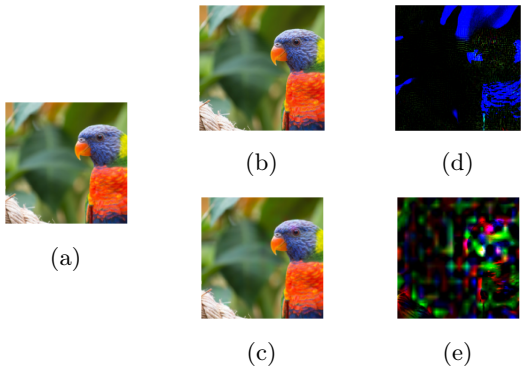

Figure 1: Example of successful pixel domain and low frequency C&W ℓ_2 (DWT scale 4) white-box attacks. Original image (a), pixel domain C&W ℓ_2 adversarial image (b), pixel domain perturbation (c), low frequency C&W ℓ_2 adversarial image (d), and low frequency perturbation (e). MobileNet V3 Small model, image from NIPS2017[1].

(2018a) investigate several such denoising techniques using PCA, JPEG compression, wavelet approximation, and soft-thresholding of wavelet coefficients. Many of these methods represent the data using a subset of its basis functions corresponding to the first principal components, in the case of PCA, or low frequency terms, in the case of JPEG and wavelet-based methods. The defenses are applied as a pre-processing step only at test time, and they are evaluated by their successful classification of benign and adversarial data. Such defenses usually rely on removing high frequency signal (Dziugaite et al., 2016; Guo et al., 2018a; Xu et al., 2017) and have shown ineffective against attacks whose search space is constrained to certain Fourier frequencies or when smoothing steps are taken in the attack algorithm (Guo et al., 2018b; Sharma et al., 2019; Zhang et al., 2020; Dabouei et al., 2020).

On the attack side, Guo et al. (2018b) construct a *black-box* attack by restricting the search space to a low frequency DCT subspace. They claim that the density of adversarial examples in this subspace is higher than in the whole image space, thus their attack needs fewer model queries than previous black-box attacks. Santamaria-Pang et al. (2021) restrict their search space to the high frequency wavelet subbands obtained by the *discrete wavelet transform (DWT)* basis to alter only local horizontal, diagonal, and vertical edge content. They argue that this restriction helps disrupt early convolutional layers in CNNs.

In this work, we address the limitations of existing deep learning defense methods to adversarial attacks generated in a low frequency domain, given by the DWT basis. The inability of DCT to localize an attack, thus changing unnecessary information in an image, becomes a key factor to departure from attacks generated in the *time-domain* (low and high frequencies) (Li and Li, 2017; Borji, 2019). In contrary to the Fourier transformation, *time-scale* representations given by the DWT do not lose the spatial content coherence. Moreover, considering that low frequency patterns are crucial for the state-of-the-art models to extract class-specific information from images, we exploit a new pitfall of deep learning models by generating perturbations in the low frequency wavelet domain (Wang et al., 2020). Our framework consists of three parts. First, we use *multiresolution analysis* to decompose an image into low and high frequencies. Next, we adjust popular adversarial attack techniques to design attacks that perturb only the low frequency coefficients. Finally, we reconstruct the images from the perturbed low frequency coefficients and the original high frequency part and investigate the susceptibility of adversarial training and image denoising techniques against our *white-box* low frequency attacks. Figure 1 shows sample adversarial images and their perturbations in the pixel domain and the low frequency wavelet domain produced by the C&W ℓ_2 attack.

Contributions. To the best of our knowledge, our work is the first to design adversarial attacks in the low frequency subspace given by the DWT basis. Our main contributions are the following.

1. We establish the vulnerability of neural networks to adversarial attacks generated in the low frequency representation given by the DWT basis.

[1]https://www.kaggle.com/c/nips-2017-defense-against-adversarial-attack/data

2. We design almost imperceptible low frequency adversarial attacks in the wavelet domain from three popular white-box attacks and show that adversarial training[2] and image processing methods, such as JPEG compression, PCA denoising, soft-thresholding, and wavelet denoising, are significantly less effective against such attacks.

2 Background

Among many possibilities to represent real-world data, a popular representation in the context of images is the two-dimensional Discrete Wavelet Transform (DWT) basis (Daubechies, 1988). The advantage of DWT is that it captures both frequency and location information, unlike, for example, the Fourier Transform. Moreover, it is well known that signals when represented in the DWT basis have approximately sparse representations (Kutyniok and Lim, 2011). This representation will allow us to separate the low frequency part of an image and restrict adversarial attacks to operate only in the low frequency coefficients.

Multiresolution Analysis (MRA): We give a brief overview of multiresolution analysis (Mallat, 1999).

Definition 2.1. *An orthonormal multiresolution analysis (MRA) of $L^2(\mathbb{R})$ (Hilbert space of square-integrable complex valued functions on $\mathbb{R}$) is an ordered chain of closed subspaces $\cdots \subseteq V_{-1} \subseteq V_0 \subseteq V_1 \subseteq \cdots$, where V_j contains features of scale j, satisfying the following three conditions*

1. *Completeness:*
   $\overline{\lim_{j\to\infty} V_j} = L^2(\mathbb{R})$ *and* $\lim_{j\to-\infty} V_j = \{0\}$.
2. *Dyadic Similarity:*
   $u(x) \in V_j$ *if and only if* $u(2x) \in V_{j+1}$.
3. *Translation Seed:*
   There exists $\varphi \in V_0$ *such that* $(\varphi(x-k))_{k\in\mathbb{Z}}$ *is an orthonormal basis (ONB) of* V_0.

Then φ is defined as a *father wavelet* or scaling function if φ generates an MRA.

Lemma 2.2. *Let $\{V_i\}_{i\in\mathbb{Z}}$ denote an MRA of $L^2(\mathbb{R})$. Then for $\varphi_{j,k}(x) := 2^{\frac{j}{2}}\varphi(2^j x - k), j, k \in \mathbb{Z}$, the $\{\varphi_{j,k}\}_{k\in\mathbb{Z}}$ form an ONB of V_j.*

Hence, scaled translations of φ are sufficient to represent all of L^2. Indeed, by the completeness of an MRA, any signal $u \in L^2(\mathbb{R})$ can be approximated to any desired precision by its projection $u_j = P_j u = \sum_k \langle u, \varphi_{j,k}\rangle \varphi_{j,k}$ onto V_j. Next, let us describe the fine detail that we obtain when moving from a coarser space V_j to a finer scale space V_{j+1}. An orthogonal projection $P_j : V_{j+1} \to V_j$ wipes out fine details, thus the *space of details* is given by $W_j := \{(I - P_j)u_{j+1} | u_{j+1} \in V_{j+1}\}$, where I is the identity. Hence, $P_j W_j = \{0\}$ and $V_{j+1} = V_j \oplus W_j$. which means that an image $u \in L^2(\mathbb{R})$ is the accumulated effect of its details.

The dyadic similarity is faithfully inherited, i.e., $\eta(x) \in W_j$ if and only if $\eta(2x) \in W_{j+1}$. From the completeness condition we have

$$L^2(\mathbb{R}) = \overline{V_0 \oplus \left(\bigoplus_{j=0}^{\infty} W_j\right)},$$

which means that an element $u \in L^2(\mathbb{R})$ is the accumulated effect of its details.

A *mother wavelet* is a function $\psi \in W_0$ orthogonal to the father wavelet such that $\{\psi(x-k)\}_{k\in\mathbb{Z}}$ form an ONB of W_0. By the second condition of an MRA, $\{\psi_{j,k} = 2^{j/2}\psi(2^j x - k) | k \in \mathbb{Z}\}$ is an ONB of W_j and $\{\psi_{j,k} = 2^{j/2}\psi(2^j x - k) | j, k \in \mathbb{Z}\}$ is an ONB of $L^2(\mathbb{R})$, the so-called *wavelet basis* (Mallat, 1999).

[2]Throughout the text, adversarial training will denote adversarial training based on small ℓ_p norm perturbations.

Then any $u \in L^2(\mathbb{R})$ can be represented in terms of the father and mother wavelet

$$u(x) = \sum_k \langle u, \varphi_{0,k}\rangle \varphi_{0,k}(x) + \sum_{j=0}^{\infty}\sum_k \langle u, \psi_{j,k}\rangle \psi_{j,k}(x).$$

The coefficients of the first term, so-called *approximation coefficients*, capture the main signal content while the coefficients of the second term, the *detail coefficients*, capture the local details. Level j wavelet approximation results in an approximation image of resolution which is coarser as j grows, containing 2^{-2j} of the pixels of the original image.

2D Discrete Wavelet Transform (DWT): A direct generalization of the 1D MRA into $L^2(\mathbb{Z}^2)$ gives the 2D Discrete Wavelet Transform (DWT). Let φ denote a scaling function whose corresponding wavelet is given by ψ. By defining three wavelets $\psi^{(1)} = \varphi\psi, \psi^{(2)} = \psi\varphi$, $\psi^{(3)} = \psi\psi$, and for $k \in \{1, 2, 3\}$

$$\psi^{(k)}_{j,(n_1,n_2)}(t_1, t_2) := 2^{\frac{j}{2}}\psi^{(k)}\left(\frac{2^j n_1 - t_1}{2^j}, \frac{2^j n_2 - t_2}{2^j}\right),$$

then the family $\{\psi^{(1)}_{j,n}, \psi^{(2)}_{j,n}, \psi^{(3)}_{j,n}\}_{n\in\mathbb{Z}^2}$ with $n = (n_1, n_2)$ is an ONB of W_j^2 and $\{\psi^{(1)}_{j,n}, \psi^{(2)}_{j,n}, \psi^{(3)}_{j,n}\}_{j\in\mathbb{Z},n\in\mathbb{Z}^2}$ is an ONB of $L^2(\mathbb{Z}^2)$ (Santamaria-Pang et al., 2021). Low frequencies are represented by the approximation coefficients $\langle u, \hat{\varphi}_{j,n}\rangle$, for $\hat{\varphi} = \varphi\varphi$, whereas the detail coefficients $\langle u, \psi^{(k)}_{j,n}\rangle, \forall k \in \{1, 2, 3\}$ are associated with high frequencies at horizontal, vertical, and diagonal orientations.

The scaling function φ and the mother wavelet ψ can further be represented as quadrature mirror filter banks G_0 and G_1 (Burros et al., 1998). Then, the 1D DWT of a signal x with N samples $x[n], n \in 0, ..., N-1$ is computed by passing it through two filters. By denoting $H_0[n] = G_0[N - n]$, then the low frequency part of the signal is given by a convolution of x and H_0

$$x_l[n] = \sum_{k=0}^{N-1} H_0[k]x[n - k].$$

However, since half of the frequencies have been removed, half of the samples can be discarded. We denote removing every other sample by the down-sampling operator $\downarrow$ followed by a 2, i.e., $x[n] \downarrow 2 = x[2n]$. The down-sampled low frequency part of x is then given by

$$x_L[n] = x_l[n] \downarrow 2.$$

In order to derive the high frequency part of the signal, let $H_1[n] = G_1[N - n]$. Then we can obtain the high frequency parts of the signal via

$$x_H[n] = x_h[n] \downarrow 2 = \left(\sum_{k=0}^{N-1} H_1[k]x[n - k]\right) \downarrow 2.$$

For a 2D signal $\mathbf{x} \in \mathbb{Z}^{n_i\times n_j}$, we initially filter each row to obtain $\mathbf{x}_L = [\mathbf{x}_{1L}, ..., \mathbf{x}_{n_iL}]$ and $\mathbf{x}_H = [\mathbf{x}_{1H}, ..., \mathbf{x}_{n_iH}]$ and then we pass again the columns of $\mathbf{x}_L$ and $\mathbf{x}_H$ through the filters H_0 and H_1. The four down-sampled resulting signals $\mathbf{x}_{HH}, \mathbf{x}_{HL}, \mathbf{x}_{LH}, \mathbf{x}_{LL}$ are the DWT coefficients of $\mathbf{x}$ for a decomposition of scale 1. Here $\mathbf{x}_{LL}$ gives an approximation of the signal and $\mathbf{x}_{HH}, \mathbf{x}_{HL}$ and $\mathbf{x}_{LH}$ contain diagonal, vertical, and horizontal details, respectively. Applying again the same DWT decomposition to the signal $\mathbf{x}_{LL}$, which was passed through the low-pass filter twice, gives the DWT coefficients for decomposition of scale 2. This decomposition is illustrated in Figure 2.

Next, let $\mathcal{R}$ denote a 2D DWT map that applies appropriately chosen filters as described above. Given an image $\mathbf{x} \in [0, 1]^{n\times c}$, its 2D DWT coefficients are written as

$$\mathcal{R}(\mathbf{x}) = \left[\begin{array}{c|c} \mathbf{x}_{LL} & \mathbf{x}_{LH} \\ \hline \mathbf{x}_{HL} & \mathbf{x}_{HH} \end{array}\right] \in \mathbb{R}^{n\times c}.$$

We can invert the 2D DWT coefficients by up-sampling and convolving with the filters G_0 and G_1 in the inverse order in which we applied H_0 and H_1. Since we only apply down-sampling, up-sampling, and linear convolution with fixed filter coefficients, 2D DWT is a linear transformation.

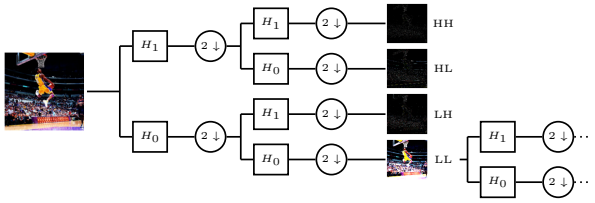

Figure 2: The DWT decomposition tree for a basketball image from ImageNet dataset (Russakovsky et al., 2015).

3 Wavelet-based Low Frequency Adversarial Attacks Modeling

Image compression techniques have long utilized the fact that low frequency signal consists of the most crucial content-defining information in natural images, whereas high frequency spectrum often represents the noise (Wallace, 1992b). Moreover, accuracy tests on deep learning models (see Figure 6 in Appendix A.2) show that low frequency patterns are critical for the model to extract class-specific information from images. Thus, we directly target the class-defining information by designing white-box attacks that alter the approximation coefficients while preserving the detail coefficients of a given image $\mathbf{x}$. We explicitly model the adversarial perturbation from three popular attacks in the low frequency representation given by the DWT basis.

3.1 Problem Formulation

Assume we are given an input image $\mathbf{x} \in \mathcal{X} := [0,1]^{n \times c}$ of $n := n_i \times n_j$ pixels and c colour channels, flattened in a given order of the spatial components. A neural network classifier $f_\theta : [0,1]^{n \times c} \to \mathbb{R}^k$ maps the input $\mathbf{x}$ to $\mathbf{y}$ containing the *logits* of k classes. The network is traditionally followed by softmax and cross-entropy loss at supervised training or by arg max at test time. An input $\mathbf{x}$ with logits $\mathbf{y} = f_\theta(\mathbf{x})$ is correctly classified if the *prediction* $p(\mathbf{x}) = \arg\max_i y_i$ equals the true label of $\mathbf{x}$.

Let $L(\theta, \mathbf{x}, t)$ denote a *classification loss* function (for instance cross-entropy loss), defined on an output logit vector $\mathbf{y} = f_\theta(\mathbf{x})$ with respect to the original true label t. We consider a white-box attack setting, where f_θ is known. Adversaries have the ability to modify a given image $\mathbf{x} \in \mathcal{X}$ of correct label $t \in \{1, ..., k\}$ into an adversarial instance $\hat{\mathbf{x}} \in \mathcal{X}$, which may be wrongly classified by the network with label $l \in \{1, ..., k\}, l \neq t$, although the two images look visually indistinguishable. The latter is often measured by a small ℓ_p *distortion* $\|\hat{\mathbf{x}} - \mathbf{x}\|_p$.

Hence, the ultimate goal of an adversary is to succeed under minimal distortion

$$\max_{\hat{\mathbf{x}} \in \mathcal{X} : \|\hat{\mathbf{x}} - \mathbf{x}\|_p \leq \varepsilon} L(\theta, \hat{\mathbf{x}}, t). \tag{1}$$

Following the original formulation of Szegedy et al. (2014), we can explicitly express the objective as a function of variable $\mathbf{r} := \hat{\mathbf{x}} - \mathbf{x}$

$$\max_{\|\mathbf{r}\|_p \leq \varepsilon} L(\theta, \mathbf{x} + \mathbf{r}, t), \tag{2}$$

where the box constraint $\mathbf{x} + \mathbf{r} \in \mathcal{X}$ is satisfied by clipping the adversarial example to be in $\mathcal{X}$.

In this manuscript, we experiment with three popular adversarial attacks.

Fast Gradient Sign Method (FGSM) Goodfellow et al. (2015) aim to solve (2) for $\mathbf{r}$ such that its ℓ_∞-norm is smaller than some $\varepsilon > 0$. For small enough ε, the first-order approximation to this problem is given by

$$\delta \approx \arg\max_{\|\mathbf{r}\|_\infty \leq \varepsilon} L(\theta, \mathbf{x}, t) + \mathbf{r}^\top \nabla_{\mathbf{x}} L(\theta, \mathbf{x}, t).$$

The resulting approximate maximal perturbation is

$$\delta = \varepsilon \ \text{sign}(\nabla_{\mathbf{x}} L(\theta, \mathbf{x}, t)),$$

and the adversarial example is given by

$$\hat{\mathbf{x}} = \text{clip}_{\mathcal{X}}(\mathbf{x} + \delta).$$

Iterative Fast Gradient Sign Method (I-FGSM) Kurakin et al. (2021) repeatedly apply the FGSM perturbation for J iterations

$$\begin{aligned}\hat{\mathbf{x}}^{(0)} &= \mathbf{x},\\ \hat{\mathbf{x}}^{(j)} &= \text{clip}_{\mathcal{X},\varepsilon}\Big(\hat{\mathbf{x}}^{(j-1)} + \alpha\ \text{sign}(\nabla_{\mathbf{x}} L(\theta, \hat{\mathbf{x}}^{(j-1)}, t))\Big),\end{aligned}$$

and set the adversarial image to be $\hat{\mathbf{x}} = \mathbf{x}^{(J)}$, the output of the last iteration. The step size is typically set to $\alpha = \frac{\varepsilon}{J}$, where ε is the maximal total perturbation.

Auto-PGD Croce and Hein (2020) compute an intermediate iterate

$$\mathbf{z}^{(j)} = \text{clip}_{\mathcal{X},\varepsilon}\Big(\hat{\mathbf{x}}^{(j-1)} + \eta\ \text{sign}(\nabla_{\mathbf{x}} L(\theta, \hat{\mathbf{x}}^{(j-1)}, t))\Big)$$

as in I-FGSM. The perturbed image at iteration j is then obtained by

$$\mathbf{x}^{(j)} = \text{clip}_{\mathcal{X},\varepsilon}\left(\mathbf{x}^{(j-1)} + \alpha\left(\mathbf{z}^{(j)} - \mathbf{x}^{(j-1)}\right) + (1-\alpha)\left(\mathbf{x}^{(j-1)} - \mathbf{x}^{(j-2)}\right)\right),$$

where α is typically set to 0.75. At each iteration in a set of checkpoints $\{w_0, ..., w_n\}$, the step size η is decreased by half if one of the conditions

$$\begin{gathered}\sum_{i=w_{j-1}}^{w_j - 1} 1_{L(\theta,\mathbf{x}^{(i+1)},t) > L(\theta,\mathbf{x}^{(i)},t)} < \rho(w_j - w_{j-1}),\\ \eta^{(w_{j-1})} \equiv \eta^{(w_j)} \quad \text{and} \quad L_{\max}^{(w_{j-1})} \equiv L_{\max}^{(w_j)}\end{gathered}$$

is met. $L_{\max}^{(w_j)}$ is the maximum loss up to iteration w_j.

Carlini-Wagner (C&W) Carlini and Wagner (2017) approximate the constrained problem (1) by its Lagrangian formulation and define a new loss function

$$L_{cw}(\theta, \mathbf{x}, t) = \max(\max_{i \neq t}(f_\theta(\hat{\mathbf{x}})_i) - f_\theta(\hat{\mathbf{x}})_t, -\kappa)$$

whose value is low when $p(\hat{\mathbf{x}}) \neq t$. To ensure the box constraint $\mathbf{x} + \delta \in \mathcal{X}$ is satisfied, they apply a change of variables and optimize

$$\min_{\mathbf{w}} \{\|\frac{1}{2}(\tanh(\mathbf{w}) + 1) - \mathbf{x}\|_2^2 + c \cdot L_{cw}(\theta, \frac{1}{2}(\tanh(\mathbf{w}) + 1), t)\}, \tag{3}$$

where κ controls the desired confidence of the model, and c is a trade-off parameter. The second part of (3) is minimized when the logit of at least one class exceeds that of the correct class, by κ or more. We set $\kappa = 0$ in our experiments. Given w, the adversarial example is obtained via

$$\hat{\mathbf{x}} = \frac{1}{2}(\tanh(\mathbf{w}) + 1).$$

3.2 Wavelet-based Low Frequency Adversarial Attacks

Instead of the pixel domain, let us now consider a representation space with a corresponding map given by $\mathcal{R}$, the 2D DWT basis of Daubechies mother wavelet.

The FGSM problem in the DWT space aims to solve

$$\delta' = \underset{\|\mathbf{r}\|_\infty \leq \varepsilon}{\arg\max}\, L(\theta, \mathcal{R}^{-1}(\mathcal{R}(\mathbf{x}) + \mathbf{r}), t),$$

whose first-order approximation is given by

$$\delta' \approx \underset{\|\mathbf{r}\|_\infty \leq \varepsilon}{\arg\max} \left\{ L(\theta, \mathcal{R}^{-1}(\mathcal{R}(\mathbf{x})), t) + \mathbf{r} \nabla_{\mathcal{R}(\mathbf{x})} L(\theta, \mathcal{R}^{-1}(\mathcal{R}(\mathbf{x})), t) \right\}.$$

Thus, we can derive the maximal perturbation

$$\delta' = \varepsilon \ \text{sign}(\nabla_{\mathcal{R}(\mathbf{x})} L(\theta, \mathcal{R}^{-1}(\mathcal{R}(\mathbf{x})), t)),$$

which in the case of a linear $\mathcal{R}$ is simply given by

$$\delta' = \varepsilon \ \text{sign}\left(\mathcal{R}\left(\frac{\partial L(\theta, \mathbf{x}, t)}{\partial \mathbf{x}} \right) \right). \tag{4}$$

Low frequency FGSM To get a low frequency attack from (4), the perturbation is obtained by simply applying $\mathcal{R}$ to the gradient and dropping the high frequency coefficients

$$\delta' = \varepsilon \ \text{sign}\left(\left[\begin{array}{c|c} \left[\mathcal{R}\left(\frac{\partial L(\theta, \mathbf{x}, t)}{\partial \mathbf{x}} \right) \right]_{LL} & 0 \\ \hline 0 & 0 \end{array} \right] \right).$$

Finally, the adversarial example is given by

$$\hat{\mathbf{x}} = \text{clip}_{[0,1]}(\mathbf{x} + \mathcal{R}^{-1}(\delta')).$$

The dimensionality of the space of the low frequency coefficients decreases with the increase of the DWT scale. Thus, for higher scale DWT, low frequency attacks are weaker on average than normal FGSM attacks. Figure 3 illustrates low frequency FGSM attack for the first decomposition level.

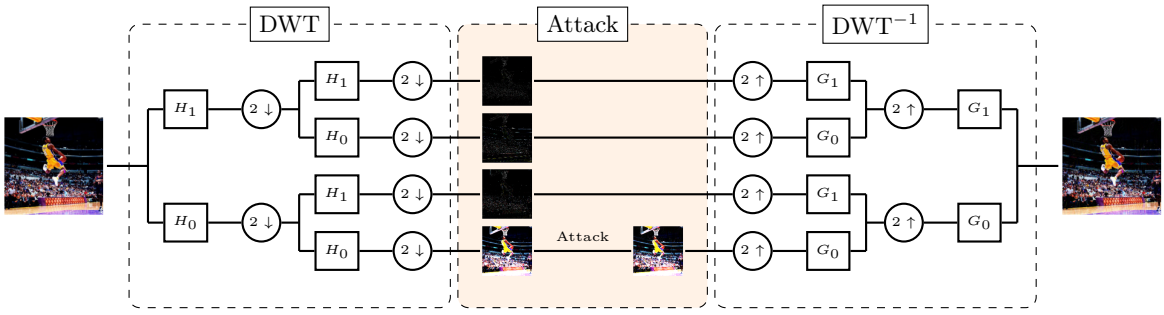

Figure 3: The low frequency FGSM attack with DWT of scale 1 for a basketball image from ImageNet.

Low frequency I-FGSM From the low frequency FGSM, the low frequency I-FGSM is derived as

$$\hat{\mathbf{x}}^{(0)} = \mathbf{x},$$
$$\hat{\mathbf{x}}^{(n)} = \text{clip}_{\mathbf{x},\varepsilon}\left(\text{clip}_{[0,1]}\left(\hat{\mathbf{x}}^{(n-1)} - \mathcal{R}^{-1}\left(\delta^{(n)} \right) \right) \right),$$

with

$$\delta^{(n)} = \varepsilon \ \text{sign}\left(\left[\begin{array}{c|c} \left[\mathcal{R}\left(\frac{\partial L(\theta, \hat{\mathbf{x}}^{(n-1)}, t)}{\partial \hat{\mathbf{x}}^{(n-1)}} \right) \right]_{LL} & 0 \\ \hline 0 & 0 \end{array} \right] \right),$$

where we again use clipping to keep our images in the feasible set of inputs for the attacked model.

Low frequency Auto-PGD We compute the intermediate iterate as

$$\mathbf{z}^{(j)} = \text{clip}_{\mathcal{X},\varepsilon}\left(\mathcal{R}(\mathbf{x}^{(j-1)}) + \eta \left[\begin{array}{c|c} \left[\mathcal{R}\left(\frac{\partial L(\theta,\mathbf{x}^{(j-1)},t)}{\partial \mathbf{x}^{(j-1)}}\right)\right]_{LL} & 0 \\ \hline 0 & 0 \end{array}\right]\right),$$

and the perturbed image at iteration j is given by

$$\mathbf{x}^{(j)} = \mathcal{R}^{-1}\left(\text{clip}_{\mathcal{X},\varepsilon}\left(\mathcal{R}(\mathbf{x}^{(j-1)}) + \alpha\left(\mathbf{z}^{(j)} - \mathcal{R}(\mathbf{x}^{(j-1)})\right) + (1-\alpha)\left(\mathcal{R}(\mathbf{x}^{(j-1)}) - \mathcal{R}(\mathbf{x}^{(j-2)})\right)\right)\right).$$

The step size decrease is identical to the pixel domain attack.

Low frequency C&W ℓ_2 To convert C&W ℓ_2 into a low frequency attack, we define $\tilde{\mathbf{x}} = \mathcal{R}(\tanh^{-1}(2\mathbf{x} - 1))$ and

$$\hat{\mathbf{w}} = \left[\begin{array}{c|c} \mathbf{w} & \tilde{\mathbf{x}}_{LH} \\ \hline \tilde{\mathbf{x}}_{HL} & \tilde{\mathbf{x}}_{HH} \end{array}\right].$$

For the perturbation δ, we choose

$$\delta = \mathcal{R}\left(\frac{1}{2}\left(\tanh\left(\mathcal{R}^{-1}\left(\hat{\mathbf{w}}\right)\right) + 1\right)\right) - \mathcal{R}(\mathbf{x}).$$

Note that $\mathcal{R}^{-1}(\mathcal{R}(\mathbf{x}) + \delta) \in [0,1]^{n\times m}$. The new objective function is given by

$$\min \quad \left\{\|\mathcal{R}(\frac{1}{2}(\tanh(\mathcal{R}^{-1}(\hat{\mathbf{w}})) + 1)) - \mathcal{R}(\mathbf{x})\|_2^2 + c \cdot f(\frac{1}{2}(\tanh(R^{-1}(\hat{\mathbf{w}})) + 1))\right\},$$

which we optimize over $\mathbf{w}$. The final adversarial example $\hat{\mathbf{x}} = \frac{1}{2}(\tanh(\mathcal{R}^{-1}(\hat{\mathbf{w}})) + 1)$ has the same high frequency coefficients in the DWT domain as the original image.

3.3 Defenses against Adversarial Attacks

When adversarial examples were first pointed out, they were only generated in the pixel domain. With the development of adversarial training techniques (Goodfellow et al., 2015; Madry et al., 2018), adversarial attacks in different representation spaces came into play. Such attacks, even though imperceptible for the human eye, are able to circumvent adversarial training (Dabouei et al., 2020), as they result in perturbations of large ℓ_p norm in the pixel space while maintaining a small distance to the original image in a different representation space. Yet however, the structure of such attacks is often made up of high frequency noise. In such cases, traditional image processing techniques, in particular the celebrated JPEG compression, have shown to be effective defenses (Guo et al., 2018a).

Recent works have further exploited vulnerabilities of deep learning models by constraining the search space of black-box attacks to certain Fourier frequencies or generating smooth attacks through smoothing steps in the attack algorithm (Guo et al., 2018b; Zhang et al., 2020; Dabouei et al., 2020). In the realm of *robust machine learning*, we examine the effectiveness of five popular defense techniques; namely adversarial training, JPEG compression, PCA denoising, soft-thresholding, and wavelet denoising against our wavelet-based low frequency adversarial attacks. Let us first give a brief overview of such defense methods.

Adversarial Training (Madry et al., 2018) aims to solve the following robust objective function

$$\min_{\theta} \mathbb{E}_{(\mathbf{x},t)\sim D}\left[\max_{\hat{\mathbf{x}}\in\mathcal{X}:\|\hat{\mathbf{x}}-\mathbf{x}\|_p\leq\varepsilon} L(\theta,\hat{\mathbf{x}},t)\right].$$

To solve the above problem, the projected gradient descent (PGD) method proposes an alternated scheme, by performing gradient ascent to maximize the inner objective and gradient descent to minimize the outer objective.

Table 1: Accuracy of the model that has been adversarially trained with PGD for ℓ_∞ robustness, attacked by FGSM in pixel domain, DWT domain, and low frequency DWT domain, tested on 10,000 images from the CIFAR10 test set. J denotes the DWT scale for the low frequency attack.

pixel ℓ_∞	**l.f. $J = 1$ ℓ_∞**	**DWT ℓ_∞**	**Nat. Acc.**
31.62%	27.80%	36.34%	72.88%

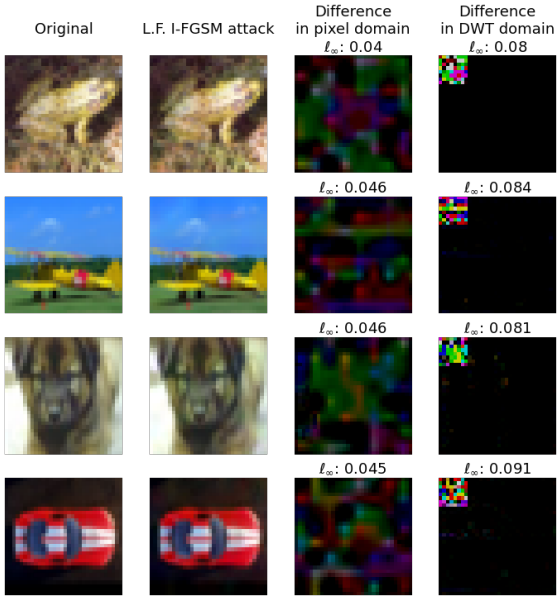

Figure 4: The figure shows examples of images from the CIFAR-10 dataset with their corresponding adversarial examples generated by I-FGSM in the low frequency DWT domain (with a scale of 2), as well as their differences in the pixel and DWT domain. The perturbation values are scaled by 10 for visibility.

However, this method trains classifiers to only defend against small ℓ_p norm attacks in the pixel domain. One could envision generating imperceptible examples by perturbing the image in the low frequency DWT basis that are far away from the original image in the original pixel basis, thus fooling an adversarially trained classifier in the pixel domain. For instance, in Table 1 we show the performance of an adversarially trained neural network in the pixel domain. Note that the trained network has poor robustness against attacks which were not considered during training, such as scale 1 low frequency attacks in the DWT domain.

Figure 4 shows examples of images from the CIFAR-10 dataset (Krizhevsky et al., 2009) and the corresponding adversarial examples generated by I-FGSM in the low frequency DWT domain of scale 2, as well as their differences in the pixel and the DWT domain. Most of the work on the construction of imperceptible adversarial examples for CIFAR-10 has been conducted with ℓ_∞ perturbations up to a magnitude of $\varepsilon = 0.03$ in the pixel basis. However, the low frequency I-FGSM adversarial images exhibited in Figure 4, while being imperceptible, are far from the original images in terms of ℓ_∞ norm in the pixel representation.

JPEG Compression The traditional JPEG compression technique (Wallace, 1992a; Shaham et al., 2018a) consists of five processing steps from which in theory only two are lossy.

1. Conversion from RGB to luminance, blue, and red chrominance (YCbCr) colour space.
2. 2×2 lowpass filtering and subsampling of the Cb and Cr channels.
3. Splitting each channel into 8×8 blocks and applying 2D discrete cosine transform to each block (Wallace, 1992a, Section 4.1).

4. Quantization of the frequencies, i.e., dividing each frequency coefficient by a constant (determined by the quality setting) and rounding to the nearest integer.

5. Lossless compression entropy coding was omitted in our implementation.

PCA denoising This pre-processing method is performed for each of the c colour channels of an image $\mathbf{x}$ separately. Concretely, considering an image as a matrix X of size $n_{\text{i}} \times n_{\text{j}}$, PCA denoising represents it by its low-rank approximation, while keeping as much variance as possible in the original matrix. This procedure can be summarized as $X_{\text{pca}} = XUU^T$, where U is a $n_{\text{j}} \times k$ matrix containing the eigenvectors corresponding to the k most dominant eigenvalues of the $n_{\text{j}} \times n_{\text{j}}$ covariance matrix $\frac{1}{n_{\text{i}}}(X - \bar{X})^T(X - \bar{X})$.

Soft-thresholding (ST) and Wavelet Denoising (WD) Mallat (1999) applies soft-thresholding in the wavelet domain. Concretely, if we let $\mathcal{R}$ denote a DWT and $\mathcal{S}_\lambda$ the soft-thresholding operator using λ as the threshold, then the processed image is given by $\hat{\mathbf{x}} = \mathcal{R}^{-1}(\mathcal{S}_\lambda(\mathcal{R}(\mathbf{x})))$.[3] Similarly, wavelet denoising is thresholding in the wavelet domain. We use the VisuShrink wavelet denoising procedure which uses the universal threshold $\lambda = \sigma\sqrt{2 \log I}$, where σ is the estimated noise variance and I is the number of pixels (Donoho and Johnstone, 1994).

4 Experiments

We next demonstrate the effectiveness of the above-mentioned defense techniques against our wavelet-based low frequency adversarial attacks.

4.1 Setup

We experiment with the CIFAR-10 dataset, which consists of images of dimensionality $32 \times 32 \times 3$ and is split into a training set of 50,000 images, on which we train a CNN classifier as in (Carlini and Wagner, 2017), and a test set of 10,000 images, each with a mini-batch size of 100.

For adversarial training, we use 40 iterations of projected gradient descent with a maximal ℓ_∞ distortion of 0.03. The quality of JPEG compression is set to 25%, PCA is performed by retaining the largest 10 principal components using `PCA` from the sklearn decomposition package, wavelet denoising is performed using `denoise_wavelet` from the skimage restoration package, soft-thresholding in the wavelet domain is performed using the `pytorch_wavelets` package (Cotter, 2019).

For C&W we do 1000 iterations per search step. To compute the data for Table 2 we use 10 search steps. To generate Figure 9 (c) and (f) we set $c = 10^{-4}$ and to control the magnitude we multiply the perturbations with some $\varepsilon \geq 1$ (Guo et al., 2018a). For (low frequency) FGSM and I-FGSM in Figure 9 (a), (b), (d), and (e) we use $\varepsilon \in [0, 0.1]$ in increments of 0.005. For Table 2 we perform binary search on ε for FGSM and I-FGSM, as well as on c for C&W, to find the smallest perturbation for which an adversarial attack can be found. The step size for I-FGSM is set to $\alpha = \varepsilon/s$, where s denotes the number of steps.

4.2 Evaluation

The average normalized ℓ_2 similarity (Guo et al., 2018a) between benign images $\mathbf{x}_1, ..., \mathbf{x}_m$ and their corresponding adversarial examples $\hat{\mathbf{x}}_1, ..., \hat{\mathbf{x}}_m$ is defined as

$$\frac{1}{m} \sum_{i=1}^{m} \frac{\|\mathbf{x}_i - \hat{\mathbf{x}}_i\|_2}{\|\mathbf{x}_i\|_2}.$$

Note that this evaluation criterion will be identical on the pixel basis and the wavelet basis. The accuracy of a model is measured by only considering the label with the highest confidence for each image, i.e., via *top-1 accuracy*.

[3] Note that DWT is surjective, i.e., $\mathcal{R}^{-1}$ exists.

Table 2: Fooling ratio performance comparison for attacks in normal and defense models. ND - no defense.

Method	**ND**	**JPEG**	**WD**	**ST**	**PCA**
C&W	1.00	0.37	0.35	0.43	0.40
l.f. C&W	0.98	0.46	0.46	0.70	0.58
FGSM	0.90	0.37	0.35	0.46	0.38
l.f. FGSM	0.62	0.43	0.40	0.54	0.45
I-FGSM	0.99	0.35	0.32	0.44	0.35
l.f. I-FGSM	0.82	0.40	0.42	0.52	0.44

4.3 Results

We examine the accuracy of the attacked model which has undergone a defense method and use this as a measure to compare the performance of the adversarial attacks in the pixel domain and the ones in the low frequency wavelet domain. Concretely, we generate perturbations using FGSM, I-FGSM, and C&W ℓ_2 in the pixel basis and in the low frequency DWT basis. We apply each of the defenses to the adversarial examples, feed them back to the model, and measure the top 1 accuracy against the normalized ℓ_2 similarity of the adversarial examples and the original images.

Figure 9 (a), (b), and (c) show that soft-thresholding is the worst-performing method when clean data is used. As we increase the distortion of perturbations attacking the model in the pixel domain, soft-thresholding and JPEG compression quickly increase their performance, but are still surpassed by adversarial training. However, when the FGSM and I-FGSM attacks are performed in the low frequency DWT domain, soft-thresholding is close to the worst-performing defense method across all tested maximal distortion values (see Figure 9 (d)-(f)).

Most importantly, in Figure 9 (d), (e), and (f) we observe that all defense methods are significantly less effective at defending against the low frequency attacks compared to the original attacks. Note that the low frequency attacks are weaker than the original attacks since we significantly decrease the dimensionality of the search space for perturbations when attacking the low frequency DWT domain. In the cases of (low frequency) FGSM and I-FGSM (Figure 9 (a), (b), (d), (e)), most defenses do not increase the model accuracy at all when the attacks are performed in the low frequency DWT domain. In the case of (low frequency) C&W ℓ_2 (Figure 9 (c) and (f)), the effectiveness of attacks in the pixel domain and the low frequency DWT domain is the same for an undefended model as well as for three of the five defenses. However, JPEG Compression and soft-thresholding perform much worse when defending against C&W ℓ_2 attack in the low frequency DWT domain.

From Table 2 we conclude that, if we allow binary search, low frequency C&W ℓ_2 attack outperforms the original C&W ℓ_2 attack in circumventing all tested defense methods. Moreover, using binary search over ε, FGSM and I-FGSM in the low frequency DWT domain, while being weaker than their pixel domain counterparts when attacking an undefended model, do also outperform original attacks at circumventing all tested defenses.

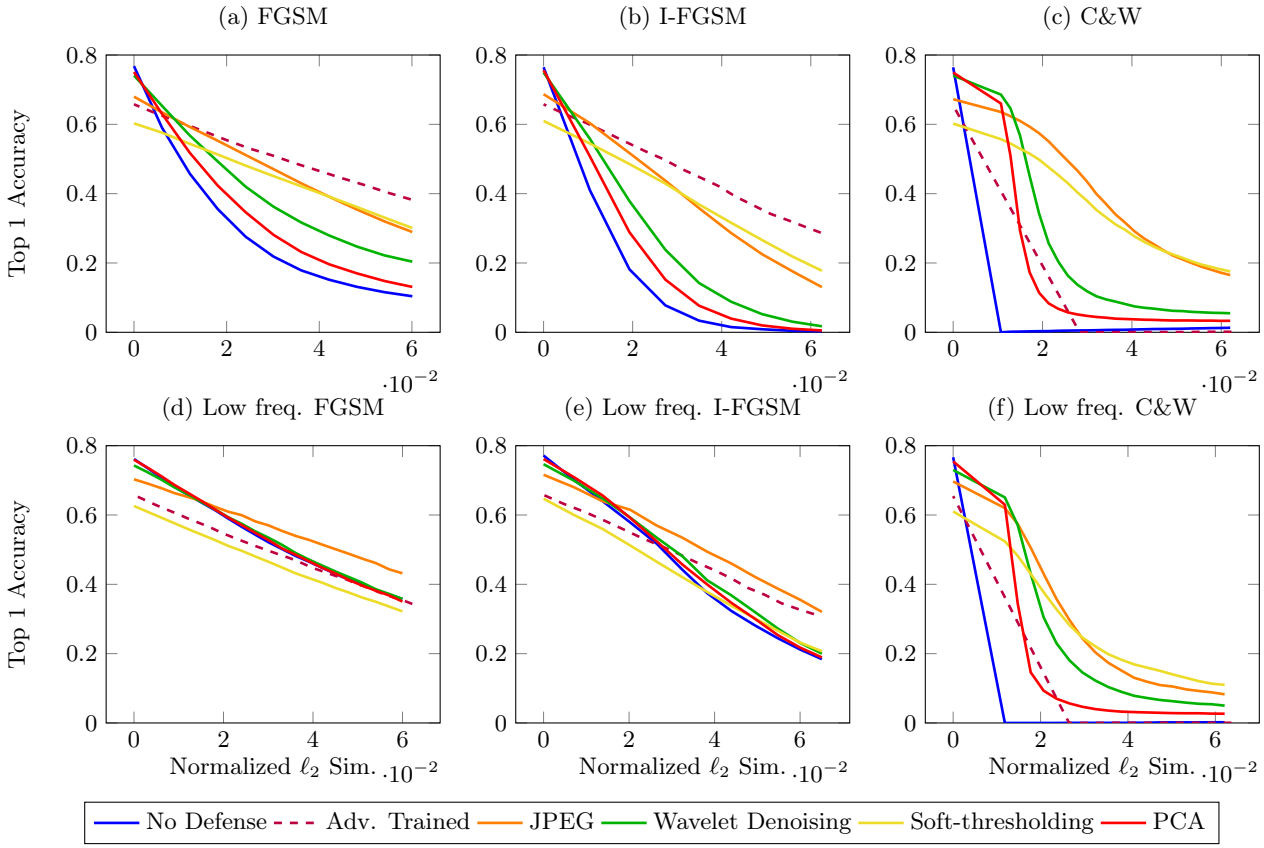

Figure 5: Model accuracy with pre-processing defenses attacked by FGSM, I-FGSM and C&W ℓ_2 in pixel domain (a), (b), (c), and low frequency DWT domain (d), (e), (f). Tested on 10,000 images from the CIFAR-10 dataset. The DWT scale was set to 1 for attacking the adversarially trained model, and 2 for all other plots.

5 Conclusions and Future Work

We examined limitations of existing deep learning defense methods, which either guarantee robustness to small ℓ_p norm attacks in the pixel domain or rely on removing high frequency signal. We demonstrated vulnerabilities of such defense techniques from the perspective of almost imperceptible attacks generated in the low frequency representation given by the DWT basis. We designed practical low frequency adversarial attacks in the wavelet domain from three popular white-box attacks against whom we employed traditional defense methods such as adversarial training and image processing methods, such as JPEG compression, PCA denoising, soft-thresholding, and wavelet denoising. We showed that, while being weaker at a given maximal distortion, our attacks were able to outperform the original attacks at circumventing defense methods. Given this vulnerability of neural networks, we wish to study in future work how low frequency attacks can help in designing state-of-the-art defense strategies.

References

Naveed Akhtar and Ajmal Mian. Threat of adversarial attacks on deep learning in computer vision: A survey. *IEEE Access*, 2018.

Pranjal Awasthi, George Yu, Chun-Sung Ferng, Andrew Tomkins, and Da-Cheng Juan. Adversarial robustness across representation spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

Ali Borji. Pros and cons of gan evaluation measures. *Computer Vision and Image Understanding*, 179, 2019.

C Burros, R Goliath, and Haitao Guo. Introduction to wavelets and wavelet transforms. *La Recherche*, 1998.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE symposium on security and privacy (sp)*. IEEE, 2017.

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning*. PMLR, 2019.

F Cotter. Uses of complex wavelets in deep convolutional neural networks. *Doctoral Thesis*, 2019.

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

Ali Dabouei, Sobhan Soleymani, Fariborz Taherkhani, Jeremy Dawson, and Nasser Nasrabadi. Smoothfool: An efficient framework for computing smooth adversarial perturbations. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020.

Ingrid Daubechies. Orthonormal bases of compactly supported wavelets. *Communications on Pure and Applied Mathematics*, 41, 1988.

David L Donoho and Iain M Johnstone. Ideal Spatial Adaptation by Wavelet Shrinkage. *Biometrika*, 81, 1994.

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.

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

Chuan Guo, Moustapha Cisse, and Laurens Van Der Maaten. Countering adversarial images using input transformations. *International Conference on Learning Representations*, 2018a.

Chuan Guo, Jared S Frank, and Kilian Q Weinberger. Low frequency adversarial perturbation. *arXiv preprint arXiv:1809.08758*, 2018b.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. *University of Toronto*, 2009.

Alexey Kurakin, Ian Goodfellow, Samy Bengio, et al. Adversarial examples in the physical world. *International Joint Conferences on Artificial Intelligence*, 2021.

Gitta Kutyniok and Wang-Q Lim. Compactly supported shearlets are optimally sparse. *Journal of Approximation Theory*, 163, 2011.

Xin Li and Fuxin Li. Adversarial examples detection in deep networks with convolutional filter statistics. In *Proceedings of the IEEE international conference on computer vision*, 2017.

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

Stéphane Mallat. *A wavelet tour of signal processing*. Elsevier, 1999.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115, 2015.

Alberto Santamaria-Pang, Jianwei Qiu, Aritra Chowdhury, James Kubricht, Peter Tu, Iyer Naresh, and Nurali Virani. Adversarial attacks with time-scale representations. *arXiv preprint arXiv:2107.12473*, 2021.

Uri Shaham, James Garritano, Yutaro Yamada, Ethan Weinberger, Alex Cloninger, Xiuyuan Cheng, Kelly Stanton, and Yuval Kluger. Defending against adversarial images using basis functions transformations. *arXiv preprint arXiv:1803.10840*, 2018a.

Uri Shaham, Yutaro Yamada, and Sahand Negahban. Understanding adversarial training: Increasing local stability of supervised models through robust optimization. *Neurocomputing*, 307, 2018b.

Yash Sharma, Gavin Weiguang Ding, and Marcus Brubaker. On the effectiveness of low frequency perturbations. *arXiv preprint arXiv:1903.00073*, 2019.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529, 2016.

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

Florian Tramer, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. On adaptive attacks to adversarial example defenses. *Advances in Neural Information Processing Systems*, 2020.

G.K. Wallace. The jpeg still picture compression standard. *IEEE Transactions on Consumer Electronics*, 38, 1992a.

Gregory K Wallace. The jpeg still picture compression standard. *IEEE transactions on consumer electronics*, 38, 1992b.

Haohan Wang, Xindi Wu, Zeyi Huang, and Eric P Xing. High-frequency component helps explain the generalization of convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8684–8694, 2020.

Weilin Xu, David Evans, and Yanjun Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. *arXiv preprint arXiv:1704.01155*, 2017.

Hanwei Zhang, Yannis Avrithis, Teddy Furon, and Laurent Amsaleg. Smooth adversarial examples. *EURASIP Journal on Information Security*, 2020.

A Appendix

A.1 Missing Proofs

In this section, we present the detailed proof of Lemma 2.2.

Proof. Since $\varphi(x-k) \in V_0$, applying dyadic similarity j times gives $2^{j/2}\varphi(2^j x - k) \in V_j$. Orthonormality

$$\langle \varphi_{j,k}, \varphi_{j,l} \rangle = \int 2^j \overline{\varphi(2^j x - k)} \varphi(2^j x - l) dx = \delta_{k,l},$$

follows by substituting $y = 2^j x$ and using translation seed of $(\varphi(x-k))_{k\in\mathbb{Z}}$. Finally, given $\psi \in V_j$, dyadic similarity induces $\psi(2^{-j}x) = \sum_{k\in\mathbb{Z}} \alpha_k \varphi(x-k)$, hence $\psi(y) = \sum_{k\in\mathbb{Z}} \tilde{\alpha_k} \varphi_{j,k}(x)$ for $y = 2^{-j}x$ and $\tilde{\alpha_k} = 2^{-j/2}\alpha_k$. □

A.2 Additional Experiments

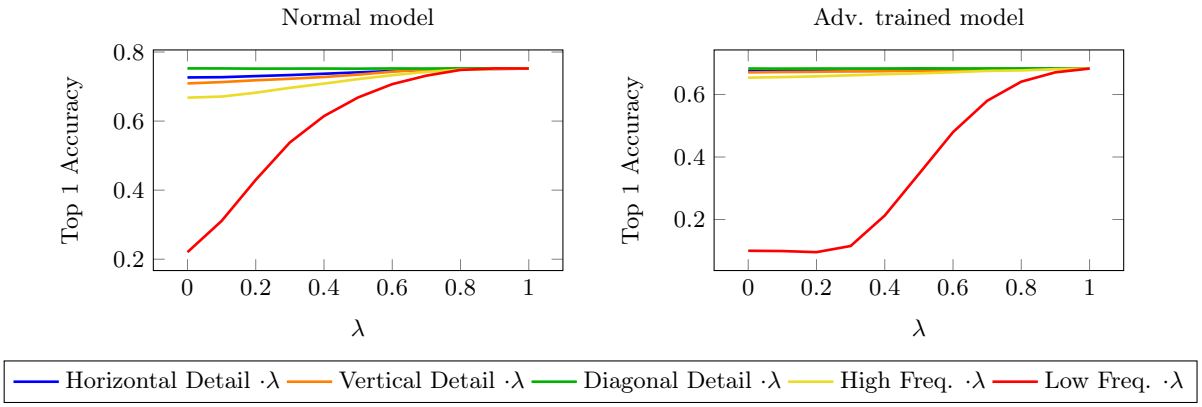

Figure 6: Accuracy of model trained on clean data and adversarially trained model. Some wavelet coefficients of the test images are multiplied by $0 \leq \lambda \leq 1$. Either the low frequency, HL, LH, HH, or all high frequency coefficients are multiplied by λ.

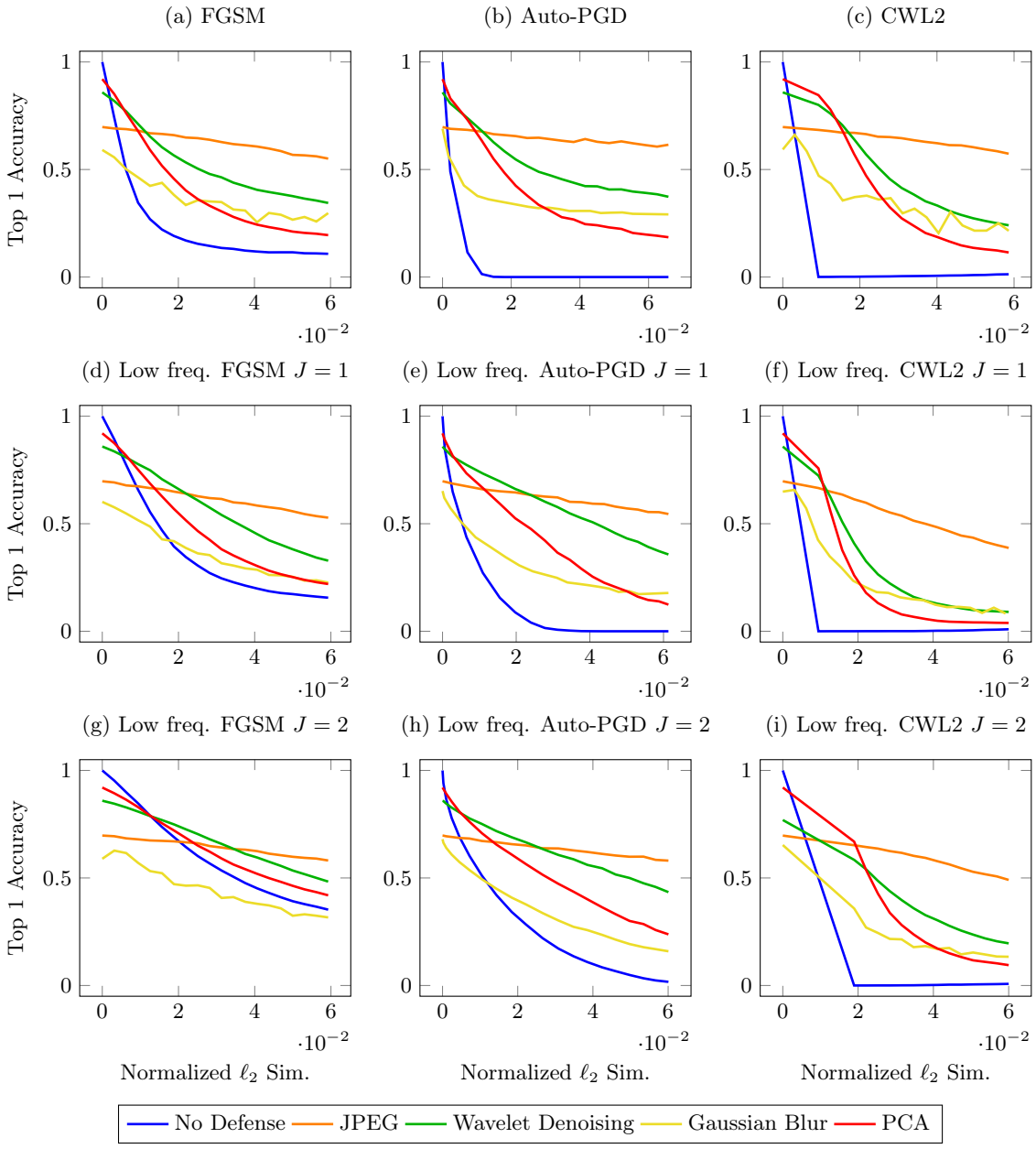

Figure 7: ResNet20 model accuracy with pre-processing defenses attacked by FGSM, I-FGSM, and Auto-PGD in pixel domain (a), (b), (c), low-frequency DWT domain with decomposition scale $J = 1$ (d), (e), (f), and $J = 2$ (g), (h), (i). Tested on 10,000 images from the CIFAR-10 dataset.

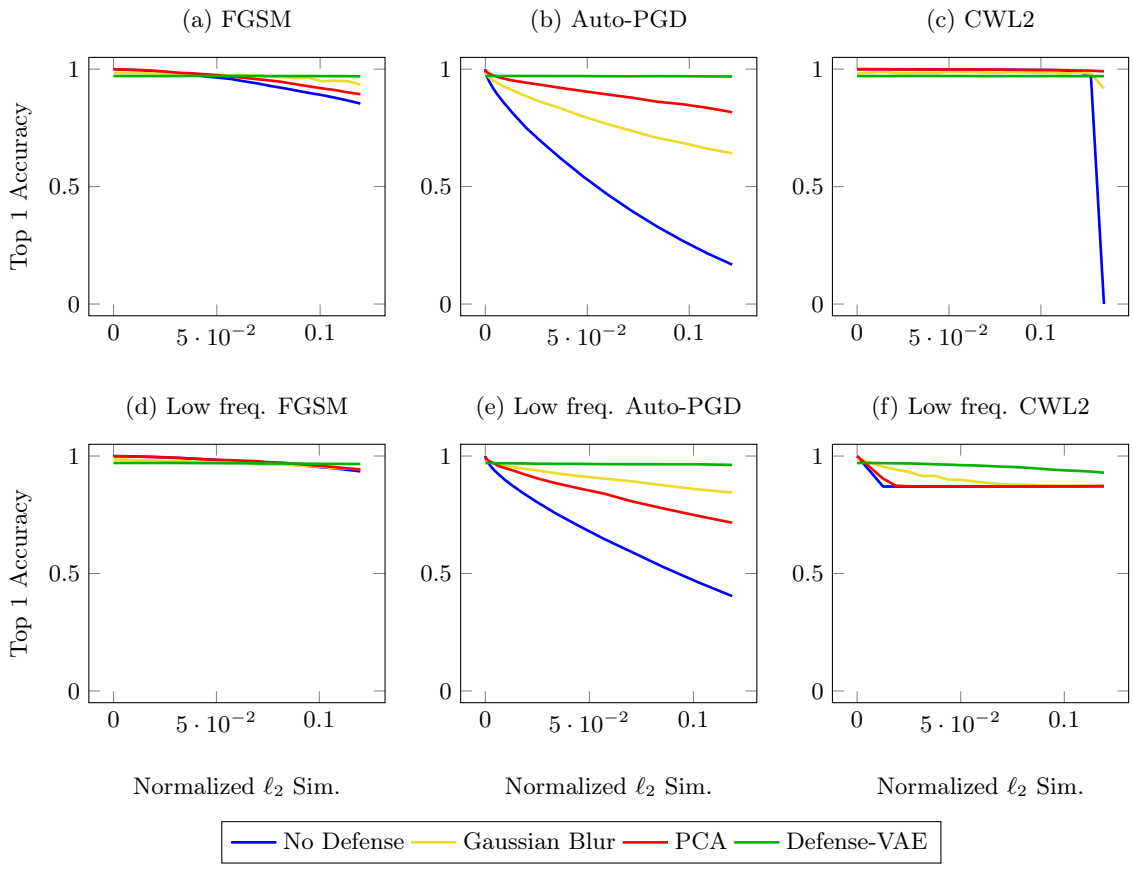

Figure 8: Model accuracy with pre-processing defenses attacked by FGSM, I-FGSM, and Auto-PGD in pixel domain (a), (b), (c), and low-frequency DWT domain (d), (e), (f). Tested on 10,000 images from the MNIST dataset.

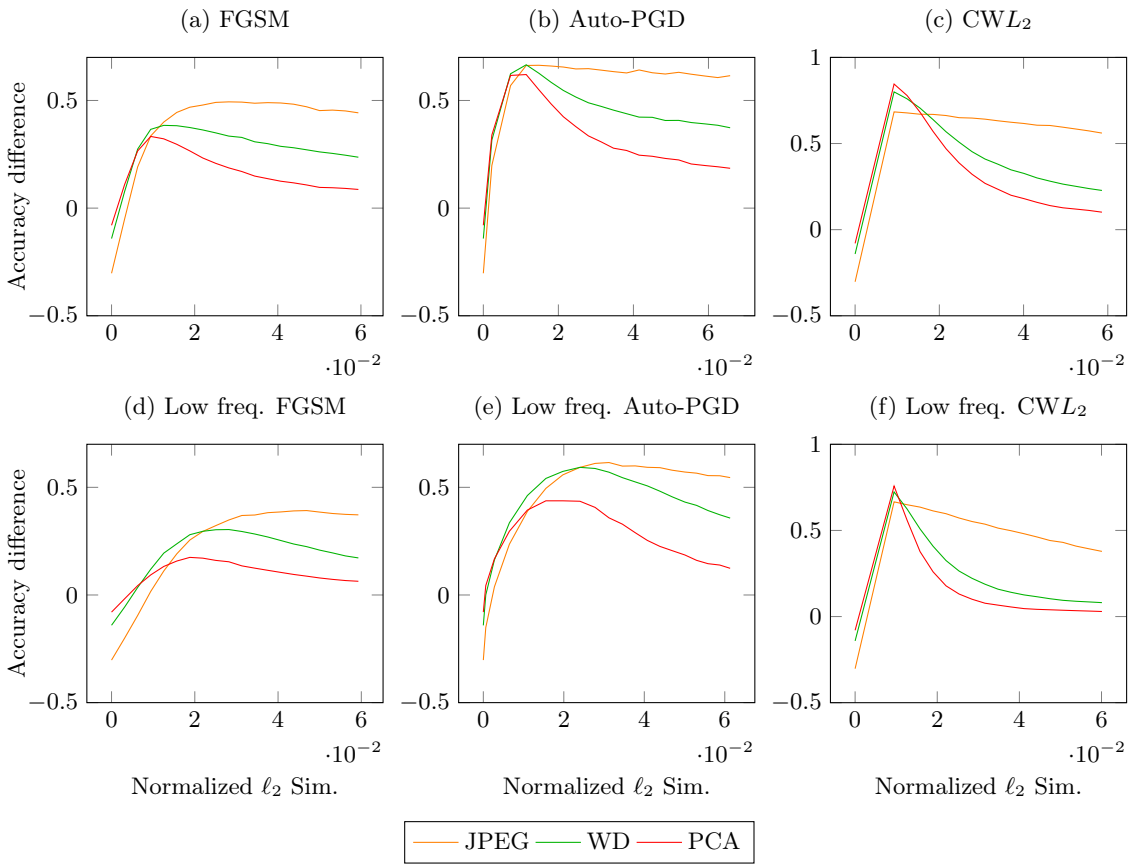

Figure 9: Difference of ResNet20 model accuracy with pre-processing defenses and without defense. The model is attacked by FGSM, I-FGSM, and Auto-PGD in pixel domain (a), (b), (c), and low-frequency DWT domain (d), (e), (f). Tested on 10,000 images from the CIFAR-10 dataset.