A Fourier Perspective of Feature Extraction and Adversarial Robustness

Liangqi Zhang[1], **Yihao Luo**[2], **Haibo Shen**[1] and **Tianjiang Wang**[1]*
[1]Huazhong University of Science and Technology
[2]Yichang Testing Technique R&D Institute
{zhangliangqi, luoyihao, shenhaibo, tjwang}@hust.edu.cn

Abstract

Adversarial robustness and interpretability are longstanding challenges of computer vision. Deep neural networks are vulnerable to adversarial perturbations that are incomprehensible and imperceptible to humans. However, the opaqueness of networks prevents one from theoretically addressing adversarial robustness. As a human-comprehensible approach, the frequency perspective has been adopted in recent works to investigate the properties of neural networks and adversarial examples. In this paper, we investigate the frequency properties of feature extraction and analyze the stability of different frequency features when attacking different frequencies. Therefore, we propose an attack method, $\mathcal{F}$**-PGD**, based on the projected gradient descent to attack the specified frequency bands. Utilizing this method, we find many intriguing properties of neural networks and adversarial perturbations. We experimentally show that contrary to the low-frequency bias of neural networks, the effective features of the same class are distributed across all frequency bands. Meanwhile, the high-frequency features often dominate when the neural networks make conflicting decisions on different frequency features. Furthermore, the attack experiments show that the low-frequency features are more robust to the attacks on different frequencies, but the interference to the high frequencies makes the network unable to make the right decision. These properties indicate that the decision-making process of neural networks tends to use as few low-frequency features as possible and cannot integrate features of different frequencies.

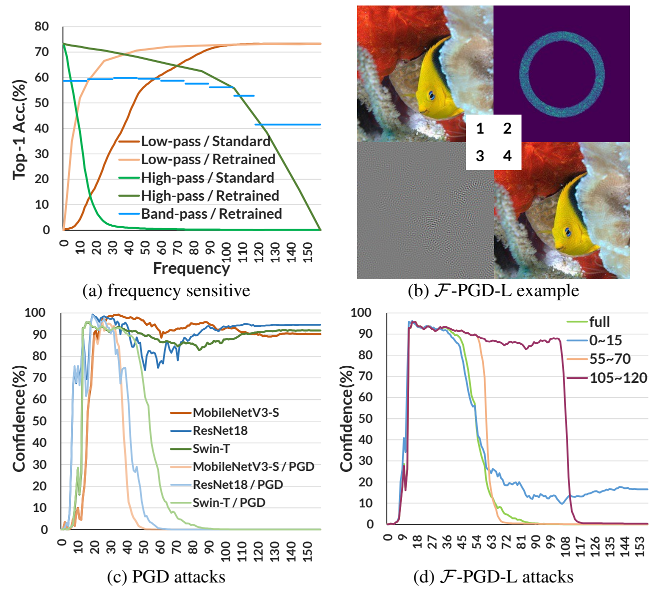


Figure 1: (a) The standard trained neural networks have a low-frequency bias, but the retrained networks indicate that all frequency bands have rich effective features. (b) 1: original sample; 2: perturbations on frequency domain; 3: adversarial perturbations; 4: adversarial example. (c) PGD attacks on sample-(b). (d) Attacks on different frequency bands. The low-frequency features are more robust to attacks on different frequencies, but attacks at all frequencies can seriously affect high-frequency features.

1 Introduction

Although deep neural networks (DNNs) have shown considerable promises in various visual applications, these neural networks are actually quite brittle. In particular, the models trained using standard methods are vulnerable to *adversarial perturbations* [Szegedy *et al.*, 2014; Goodfellow *et al.*, 2015]. However, the lack of interpretability of neural networks leads to the inability to address these drawbacks theoretically, which makes us eager to understand their feature extraction and decision-making processes. Although many specialized methods have been proposed to find the decision basis of neural networks in recent years, such as *feature visualization* [Erhan *et al.*, 2009; Yosinski *et al.*, 2015] and *attribution methods* [Zeiler and Fergus, 2014; Simonyan *et al.*, 2014; Springenberg *et al.*, 2015; Smilkov *et al.*, 2017], there is still a huge gap in understanding the feature extraction and decision process.

Besides, *data augmentation* [Cubuk *et al.*, 2019; Zhong *et al.*, 2020] can improve the robustness of neural networks in many aspects, but these methods are not defensive against ad-

*Corresponding Author

versarial attack methods. In practice, one can often generate imperceptible perturbations of the input images and cause the model to make highly-confident but erroneous predictions. The vulnerability of models trained using standard methods to adversarial perturbations makes it clear that the paradigm of adversarially robust learning differs from the classic learning setting. Consequently, various adversarial defense methods have been proposed, among which adversarial training has proven to be the most effective means of adversarial defense. However, adversarial training consumes considerable computational resources and can not fully address adversarial robustness.

Many works have interpreted the adversarial examples from aspects such as networks themselves or datasets to understand them deeply. Goodfellow *et al.* [2015] argue that adversarial examples are due to the high-dimensional linear nature of neural networks. Ilyas *et al.* [2019] thought that adversarial examples are non-robust features, the models can still have good performance and generalizations by only learning the non-robust features but will become vulnerable to various attacks. Furthermore, the frequency perspective has been adopted in recent works to investigate their properties. Some works [Rahaman *et al.*, 2019; Xu *et al.*, 2019; Yin *et al.*, 2019; Xu and Zhou, 2021] found that DNNs favor low frequencies and learn them first. Moreover, Wang *et al.* [2020] thought that DNNs also could exploit the high-frequency components that are not perceivable to humans, and these high-frequency features would improve the generalization of the network. While it is a consensus that adversarial perturbations are high-frequency noise, the recent work [Maiya *et al.*, 2021] shows that adversarial examples are neither in high-frequency nor in low-frequency components, but are simply dataset dependent. These beg a natural question:

How do neural networks make classification decisions using features of different frequencies, and will robust low-frequency features improve the adversarial robustness?

Consequently, we investigate the frequency properties of feature extraction and analyze the stability of different frequency features when attacking different frequencies. Therefore, we propose an attack method based on the projected gradient descent to attack the specified frequency bands. Our experiments are mainly conducted on the ImageNet dataset [Deng *et al.*, 2009], which has a wider frequency range and is closer to the real world observed by human eyes than MNIST and CIFAR10 datasets. Our major contributions are:

- The information contained in different frequency bands of images has no essential difference, and the accuracies of most classes in different frequency bands are almost similar to that of the full frequency band;
- For the standard training network, low-frequency features are the basis for classification decisions. When the networks can not make a decision using low-frequency features, it will progressively use higher-frequency features;
- When the low-frequency feature and high-frequency feature conflict, the high-frequency feature will play a dominant role;
- The low-frequency features have high adversarial robustness, but the attacks of all frequencies will cause great interference to the high-frequency features, so the network cannot show adversarial robustness.

2 Related Work

2.1 Robustness

Robustness is a long-standing and challenging goal of computer vision. Although data augmentation can improve the robustness of neural networks in many aspects, it cannot defend against adversarial examples. Recent works have proven that adversarial training is an effective method of defending against adversarial attacks.

Adversarial Robustness. Szegedy *et al.* [2014] discover neural networks are vulnerable to adversarial examples, various *adversarial attack algorithms* [Goodfellow *et al.*, 2015; Carlini and Wagner, 2017; Madry *et al.*, 2018; Guo *et al.*, 2019b; Su *et al.*, 2019; Croce and Hein, 2020] have been proposed to investigate the vulnerability of machine learning models. Adversarial perturbations are almost imperceptible changes to the input that cause neural networks to make erroneous predictions. Goodfellow *et al.* [2015] proposed the Fast Gradient Sign Method (FGSM) to generate perturbations with a single gradient step. It is an attack for an l_∞-bounded adversary and computes an adversarial example as

$$\boldsymbol{x} + \epsilon \cdot \operatorname{sign}\left(\nabla_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}, y; \boldsymbol{\theta})\right), \tag{1}$$

where sign operation makes perturbations meet the l_∞-norm bound as soon as possible.

The more powerful adversary is the multi-step variant, which is essentially Projected Gradient Descent (PGD) [Carlini and Wagner, 2017] on the negative loss function

$$\boldsymbol{x}^{t+1} = \operatorname{clip}_{\boldsymbol{x}^t, \epsilon}\left(\boldsymbol{x}^t + \alpha \cdot \operatorname{sign}\left(\nabla_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}, y; \boldsymbol{\theta})\right)\right), \tag{2}$$

where

$$\operatorname{clip}_{\boldsymbol{x}, \epsilon}(\cdot) = \min\left(\max\left(\cdot, \boldsymbol{x} - \epsilon\right), \boldsymbol{x} + \epsilon\right). \tag{3}$$

PGD is a very stable and widely used adversarial attack method, and our subsequent analyses are based on PGD-style attacks.

Adversarial Training. The adversarial training method has proven to be an effective method of defending against adversarial attacks. It was first proposed by [Goodfellow *et al.*, 2015], which is the most successful approach for building robust models so far for defending adversarial examples. Adversarial training can be formulated as solving a robust optimization problem [Shaham *et al.*, 2015]

$$\min_{\boldsymbol{\theta}} \underset{(\boldsymbol{x}, y) \sim \mathcal{D}}{\mathbb{E}} \left[\max_{\delta} \mathcal{L}(\boldsymbol{x} + \boldsymbol{\delta}, y; \boldsymbol{\theta})\right], \tag{4}$$

where $\mathcal{L}(\cdot, \cdot; \cdot)$ is the chosen loss function and $\boldsymbol{\theta}$ denotes the parameters of the neural network; the data pair $(\boldsymbol{x}, y)$ is sample from the data distribution $\mathcal{D}$ and δ denotes the corresponding adversarial perturbation. The inner maximization is approximated by adversarial examples generated by various adversarial attack methods.

2.2 Frequency Perspective

According to the convolution theorem, convolutional neural networks (CNNs) have the natural ability to separate different frequency information. And various experiments [Geirhos *et al.*, 2019; Brendel and Bethge, 2019] have demonstrated that neural networks have different sensitivities to various frequency components of the input image, standard CNNs make their predictions rely on the local textures rather than long-range dependencies encoded in the shape of objects. Some works [Rahaman *et al.*, 2019; Xu *et al.*, 2019] find empirical evidence of a spectral bias that lower frequencies are learned first and then the higher frequencies are captured slowly. Zhang *et al.* [2023] further argue that frequency bias is also data-dependent. Different classes or samples have different frequency biases, and the bias is related to the scale of the image classification target.

The frequency perspective has also been employed in the analysis of adversarial examples. For a long time, it was thought that adversarial perturbations were mostly high-frequency perturbations, but recent work [Maiya *et al.*, 2021] has demonstrated that adversarial perturbations are neither in high-frequency nor in low-frequency components but are dataset dependent. Other experiments [Guo *et al.*, 2019a; Sharma *et al.*, 2019; Tancik *et al.*, 2020; Long *et al.*, 2022] have found that training methods that improve the robustness of the network, such as data augmentation or adversarial training, make the neural networks prefer lower frequencies. And this also leads to the opinion that the robustness of low-frequency features is higher than that of high-frequency features.

3 Preliminaries

Let us consider a standard classification task with an underlying data distribution $\mathcal{D}$ over pairs of examples $\boldsymbol{x} \in X$ and corresponding labels $y \in Y$. We also assume that we are given a suitable loss function $\mathcal{L}(\boldsymbol{x}, y; \boldsymbol{\theta})$, for instance the cross-entropy loss for a neural network. As usual, $\boldsymbol{\theta}$ is the set of model parameters. The goal is to find model parameters $\boldsymbol{\theta}$ that minimize the risk

$$\min_{\boldsymbol{\theta}} \underset{(\boldsymbol{x},y)\sim\mathcal{D}}{\mathbb{E}} \left[\mathcal{L}(\boldsymbol{x}, y; \boldsymbol{\theta})\right]. \tag{5}$$

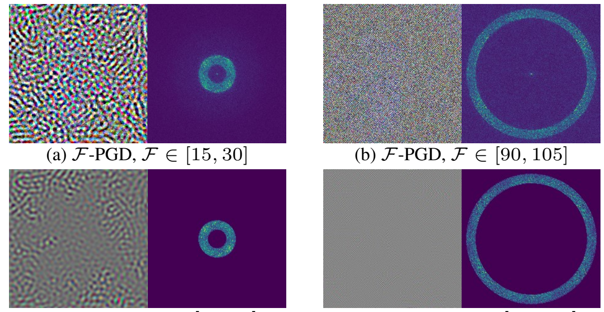
3.1 Statistics and Features

Before moving on to the Fourier perspective, let us consider what information a neural network needs to extract. According to the Information Bottleneck (IB) [Tishby *et al.*, 2000] neural networks extract relevant information that an input random variable X contains about an output random variable Y, the relevant information is defined as the *mutual information* $I(X;Y)$. In statistical terms, the relevant information of X with respect to Y, denoted by T, is a *minimal sufficient statistics* of X with respect Y. In this case, Y implicitly determines the relevant and irrelevant features in Y. Since exact minimal sufficient statistics only exist for special distributions, the information bottleneck method [Tishby *et al.*, 2000] relaxed this optimization problem by allowing the map to be stochastic, defined as an encoder $P(T|X)$, and capture as much as possible of $I(X,Y)$, not necessarily all of it. Due to the limited capacity of neural networks, $P(T|X)$ will always compress the information, the Equation 5 will be implicitly equivalent to the information bottleneck problem. And it is formulated by the following optimization problem with the Markov chain: $Y \to X \to T$:

$$\min\{I(X;T) - \beta I(T;Y)\} \tag{6}$$

where the hyperparameter β controls the information loss ratio.

In practice, due to the limited model capability and dataset, it can only guarantee to extract the statistic T related to the output Y from the input data, but not to the classified objects O^Y in the images. In other words, there are no constraints to make the networks learn to distinguish between various statistics and human-comprehensible robust features (or concepts) and tends to learn to use less information for classification purposes. This is consistent with subsequent experiments showing that the networks tend to use a small amount of low-frequency information to make decisions.



(a) $\mathcal{F}$-PGD, $\mathcal{F} \in [15, 30]$ (b) $\mathcal{F}$-PGD, $\mathcal{F} \in [90, 105]$
(c) $\mathcal{F}$-PGD-L, $\mathcal{F} \in [15, 30]$ (d) $\mathcal{F}$-PGD-L, $\mathcal{F} \in [90, 105]$

Figure 2: Perturbations and their spectrums of Fourier-based PGD attack. $\mathcal{F}$-PGD always passes information into other frequencies, but $\mathcal{F}$-PGD-L will not.

3.2 Fourier-Based PGD Attack

To evaluate the effect of the adversarial attack on different frequencies, we use *Fast Fourier Transform* (FFT) to constrain the frequencies of adversarial perturbations. Let F and F^{-1} represent the forward Fast Fourier Transform and its corresponding inverse.

Our algorithm is based on the PGD and removes specific frequency components of the on perturbation $\boldsymbol{\delta}^t$ by applying a mask to its frequency spectrum $\mathrm{FFT}(\boldsymbol{\delta}^t)$, and reconstruct the gradient by applying the IFFT on the masked spectrum. Specifically, the mask, $\mathbf{M} \in \{0,1\}^{d\times d}$, is a two-dimensional matrix, and the mask operation is done by element-wise product $\odot$. In our work, we consider the l_∞-norm and the algorithm that attack $f \in \mathcal{F}$ frequencies performs T-step attack with a small step size $\alpha = \epsilon/T$:

$$\boldsymbol{\delta}^t = \nabla_{\boldsymbol{x}^t}\mathcal{L}(\boldsymbol{x}^t, y; \boldsymbol{\theta}) \tag{7}$$

$$\boldsymbol{\delta}^t_f = \mathrm{F}^{-1}(\mathrm{F}(\boldsymbol{\delta}^t) \odot \mathbf{M}) \tag{8}$$

$$\boldsymbol{x}^{t+1} = \mathrm{clip}_{\boldsymbol{x},\epsilon}\left(\boldsymbol{x}^t + \alpha\cdot \mathrm{sign}(\boldsymbol{\delta}^t_f)\right). \tag{9}$$

Note that the non-linear sign and clip operators alias some passed information into other frequencies, and so the perturbations are not strictly contained in the frequency band, as

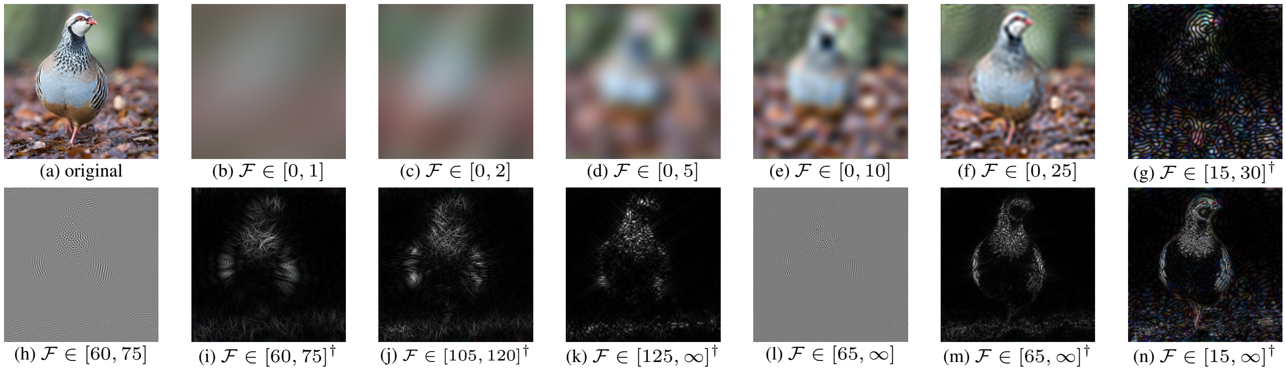


Figure 3: Information of the image in different frequency bands. It is difficult to recognize images lacking low frequencies or in narrower frequency bands for humans, but neural networks can easily extract information for classification tasks. †: normalized.

Low-pass						High-pass						Band-pass					
$\mathcal{F}$	Pre-trained		Retrained			$\mathcal{F}$	Pre-trained		Retrained			$\mathcal{F}$	Pre-trained		Retrained		
	Clean	$\mathcal{F}^{\in}$-PGD	Clean	$\mathcal{F}^{\in}$-PGD	$\mathcal{F}^{\notin}$-PGD		Clean	$\mathcal{F}^{\in}$-PGD	Clean	$\mathcal{F}^{\in}$-PGD	$\mathcal{F}^{\notin}$-PGD		Clean	$\mathcal{F}^{\in}$-PGD	Clean	$\mathcal{F}^{\in}$-PGD	$\mathcal{F}^{\notin}$-PGD
0, 145	73.2	0.1	73.2	0.2	52.5	145, 159	0.1	58.3	15.9	0.0	0.0	120, 159	0.1	16.6	41.5	0.0	0.0
0, 125	73.2	0.1	73.2	0.2	22.4	125, 159	0.1	23.3	38.5	0.0	0.0	105, 159	0.1	4.9	55.8	0.0	0.0
0, 105	72.8	0.2	72.9	0.2	9.6	105, 159	0.1	4.5	55.8	0.0	0.0	105, 120	0.1	6.8	52.8	0.0	0.0
0, 85	69.4	0.3	72.5	0.2	2.4	85, 159	0.2	1.2	62.5	0.0	0.0	90, 105	0.2	3.9	56.2	0.0	0.0
0, 65	63.1	0.5	72.1	0.3	1.7	65, 159	0.4	0.6	65.2	0.0	0.0	75, 90	0.2	3.0	57.6	0.0	0.0
0, 45	53.1	1.3	70.6	0.2	0.8	45, 159	0.8	0.4	68.1	0.0	0.1	60, 75	0.2	3.6	58.7	0.0	0.0
0, 25	29.3	5.0	66.5	0.2	0.3	25, 159	4.3	0.2	70.6	0.0	6.8	45, 60	0.2	5.1	59.5	0.0	0.0
0, 15	11.4	17.5	58.6	0.3	0.0	15, 159	19.9	0.2	71.5	0.0	25.3	30, 45	0.3	5.2	59.8	0.0	0.0
0, 10	4.6	31.1	52.0	0.5	0.0	10, 159	39.4	0.2	72.1	0.0	42.9	15, 30	0.4	4.8	59.3	0.0	0.0
0, 5	1.2	50.6	34.9	0.7	0.0	5, 159	57.1	0.1	72.5	0.0	56.5	0, 15	11.4	17.5	58.6	0.3	0.0
0, 2	0.5	65.5	13.0	0.5	0.0	2, 159	67.7	0.1	72.9	0.1	66.7	-	-	-	-	-	-
0, 1	0.4	68.8	6.8	0.4	0.1	1, 159	70.1	0.1	73.1	0.1	68.2	-	-	-	-	-	-

Table 1: Comparison of full-frequency pre-trained and retrained models at different r_L. These experiments show that different frequency bands of image data have a large number of effective features. And training a network using only the low-frequency bands does not improve robustness but decreases it. Clean: The pre-trained models are tested on the full band and the retrained models are tested on the passed bands.

shown in Fig 2. These non-linear operations lead to large effects in many cases, so we relax the l_∞-norm by restricting the mean of perturbation to be equal to α at each update:

$$\boldsymbol{x}^{t+1} = \boldsymbol{x}^t + \alpha \cdot \frac{\mathrm{card}(\boldsymbol{g}_f^t)}{\|\boldsymbol{g}_f^t\|_1} \cdot \boldsymbol{g}_f^t \tag{10}$$

where $\mathrm{card}(\cdot)$ count the number of matrix elements.

We refer to the standard method as $\mathcal{F}$-PGD and the sign- and clip-free method as $\mathcal{F}$-PGD-L in the rest of the paper. Besides, the $\mathcal{F}^{\in}$-PGD denotes an attack on frequencies belonging to $\mathcal{F}$; the $\mathcal{F}^{\notin}$-PGD denotes an attack on frequencies that do not belong to $\mathcal{F}$. $\mathcal{F}$-PGD-L also can generate imperceptible adversarial examples, but it is more pronounced than $\mathcal{F}$-PGD at lower frequencies.

4 Frequency Properties of Feature Extraction

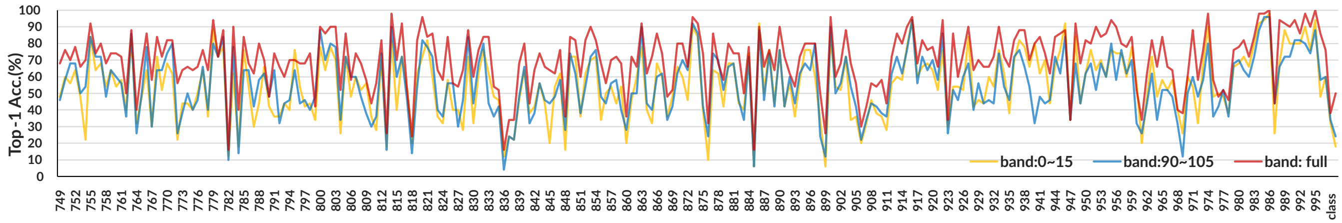
To analyze the frequency properties of the feature extraction, we need to evaluate the performance of the networks in different frequency bands. Furthermore, we analyze the frequency properties of the networks over the entire dataset as well as for each class. Therefore, we transform the images to the frequency domain and then use ideal low-pass filters (LPF), high-pass filters (HPF), and band-pass filters (BPF) to remove specified frequency information, respectively. Then all the experiments are conducted on ImageNet [Deng *et al.*, 2009]. The tested models include both CNNs and transformer-based networks, such as ResNets [He *et al.*, 2016], MobileNetV1 [Howard *et al.*, 2017], MobileNetV3 [Howard *et al.*, 2019], EfficientNets [Tan and Le, 2019], ViT [Dosovitskiy *et al.*, 2021] and Swin [Liu *et al.*, 2021]. All the tested images are at the resolution of 224×224.

Since directly removing frequencies from the images would lead to inconsistent distribution between test and training data, we preprocess the data with ideal low-pass, high-pass, and band-pass filters and retrain the MobileNetV1 $\times 1.0$ model on ImageNet from scratch, respectively.

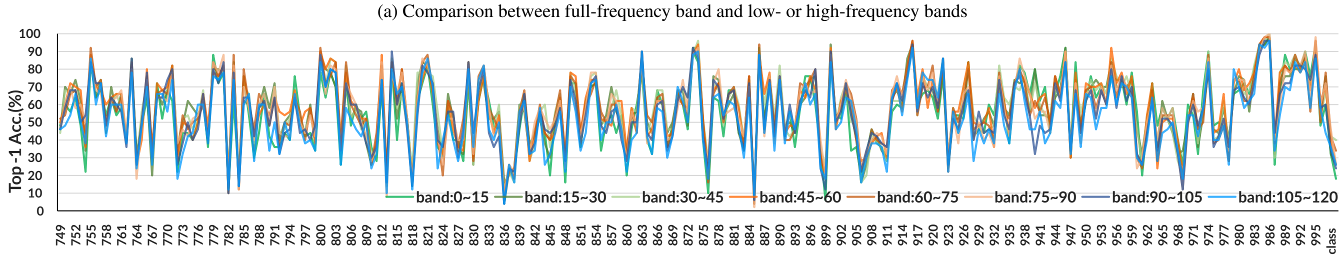
4.1 Distribution of Features

We first test the distribution of features in different frequency bands (bandwidth of 15) using band-pass filters. As shown in Table 1 and Figure 1a, all models retrained with band-pass filters achieve more than 50% accuracy in the frequency range [0,120] (note that the information decreases rapidly when the filter radius exceeds 112). There is no essential difference in each band for the classification models, which indicates that each band is rich in effective features.

Then, we test the feature distribution of each class. As shown in Figure 4, most of the classes have similar accuracy in different frequency bands, rather than different classes achieving different accuracy in different frequency bands. It is only significantly different from the accuracy achieved in



(a) Comparison between full-frequency band and low- or high-frequency bands



(b) Comparison between different frequency bands

Figure 4: Accuracy of the classes in different bands. The accuracy of the classes is similar at high- or low-frequency bands, most classes either achieve high accuracy or low accuracy in all bands.

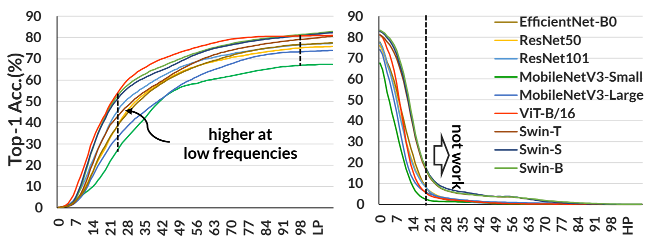


Figure 5: Frequency sensitivity analysis. It seems that low-frequency components are more important than high-frequency components for standard-trained models.

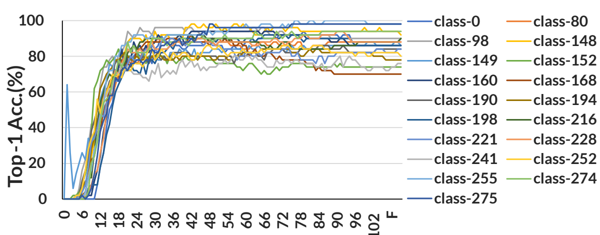


Figure 6: Low-frequency biased classes in MobileNetV3-Large. There are 115 classes that achieve more than 70% accuracy in the frequency band of $[0, 30]$.

the full frequency band. We hold the opinion that these components of the same feature at different frequencies are redundant, and the neural networks do not need to extract the information from all frequency bands.

As shown in Figure 3, images with a frequency bandwidth of 15 are also indistinguishable to humans. This further demonstrates that the classification task does not require the models to extract robust features and can rely only on the i.i.d statistics to achieve high accuracy.

4.2 Importance of Low-Frequency Features

As shown in Figure 5, with the increase of high-frequency information, the accuracy of the models increases gradually; With the absence of low-frequency information, the accuracy of the models decreases sharply, and the models cannot work at all when some low-frequency information is missing. In addition, the higher-accuracy network performs better in the low-frequency band than other networks. This indicates that the classification models rely more on low-frequency components and are vulnerable to differences in data distribution due to low-frequency loss. And the models cannot rely on high-frequency information alone to make correct classifications.

On the one hand, the retrained models are significantly higher than the standard trained models at most frequencies, as shown in Table 1. In particular, the retrained model achieve 52% top-1 accuracy at $r_{\mathcal{LP}} = 10$, which is much higher than the 4.6% top-1 accuracy of the standard training model. And it achieve 72.1% with only 1% accuracy lower than the full-frequency trained model at $r_{\mathcal{LP}} = 65$, which also indicates that for the standard training models, high frequencies are less important for classification prediction.

On the other hand, the gap between the accuracy of the network retrained with high-frequency information and the standard trained network is even more prominent. These results indicate that the importance of low-frequency information does not show up in the features it contains but in the low-frequency bias. The lack of low frequencies will have a huge impact on the image structure and the distribution of the data set. For the standard trained networks, the low-frequency components are the basis for further utilization of the high-frequency components.

4.3 Importance of High-Frequency Features

As shown in Table 1 and Figure 6, high-frequency features seem complementary to low-frequency features to achieve better generalization. However, the adversarial attack results

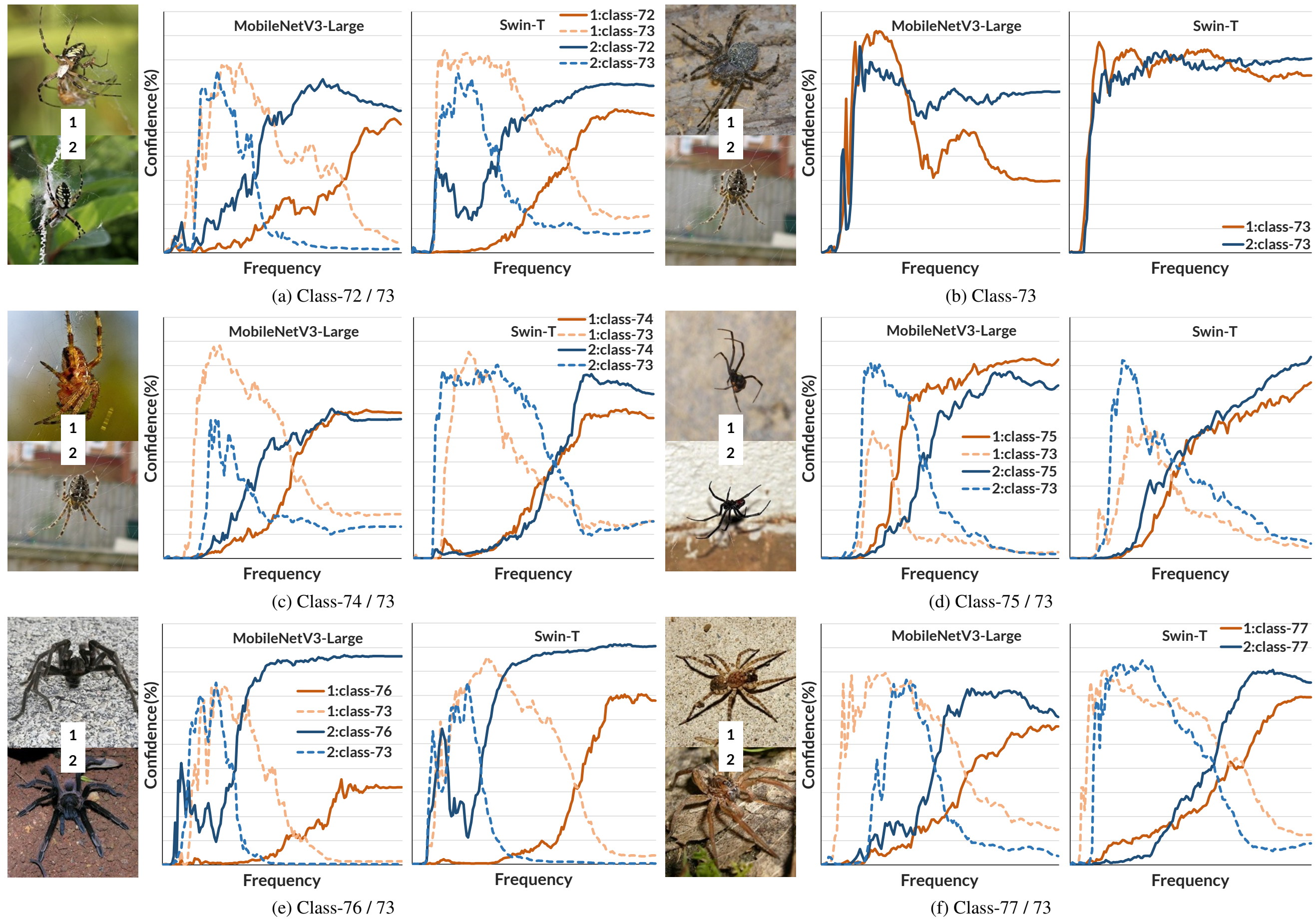


Figure 7: Low-frequency vs. High-frequency Features. When the low-frequency features and high-frequency features conflict, the higher-frequency feature will play the dominant role.

show that high-frequency features are more vulnerable to adversarial perturbations, causing networks to fail to make correct decisions, as shown in Figure 1c and 1d.

The significant impact of perturbed high-frequency features on decision-making indicates that high-frequency features are no less important than low-frequency features. We hypothesize that the neural network will preferentially select low-frequency features as the basis for its decision, and when it cannot decide, it will gradually use higher-frequency features. When the network needs to use both low-frequency and high-frequency features, high-frequency features are more important to the decision result than low-frequency features. Furthermore, our experimental results support this hypothesis.

4.4 Low-Frequency vs. High-Frequency Features

As shown in Figure 6, in addition to some classes that can be used as decisions only depending on low-frequency features, other classes require the participation of low- and high-frequency features. In the experiment, we found that the accuracy of some classes will gradually decrease with the addition of high-frequency features, especially among some classes with the same low-frequency features. We believe that the conflict between low-frequency and high-frequency features causes this, and it is the high-frequency features that ultimately play a decisive role.

Conflicts between high-frequency and low-frequency features will appear in similar classes on the ImageNet, such as between different species of spiders. As shown in Figure 7, networks using high- and low-frequency features always produce different predictions in classes between 72 and 77. These samples are always predicted to be class 73 at low frequencies and gradually classified into the correct class with the increase of high-frequency information. This indicates that these classes have the same low-frequency features and that these features belong to class 73. However, the high-frequency features will dominate the final classification results. The high-frequency feature-dominated classification results extracted by the standard trained models are also shown in the adversarial attack, which we analyze in the next section.

5 Frequency-Based Attack

To verify the effectiveness of adversarial attacks on different frequencies, we performed frequency attacks on both the standard-trained and the retrained models on different fre-

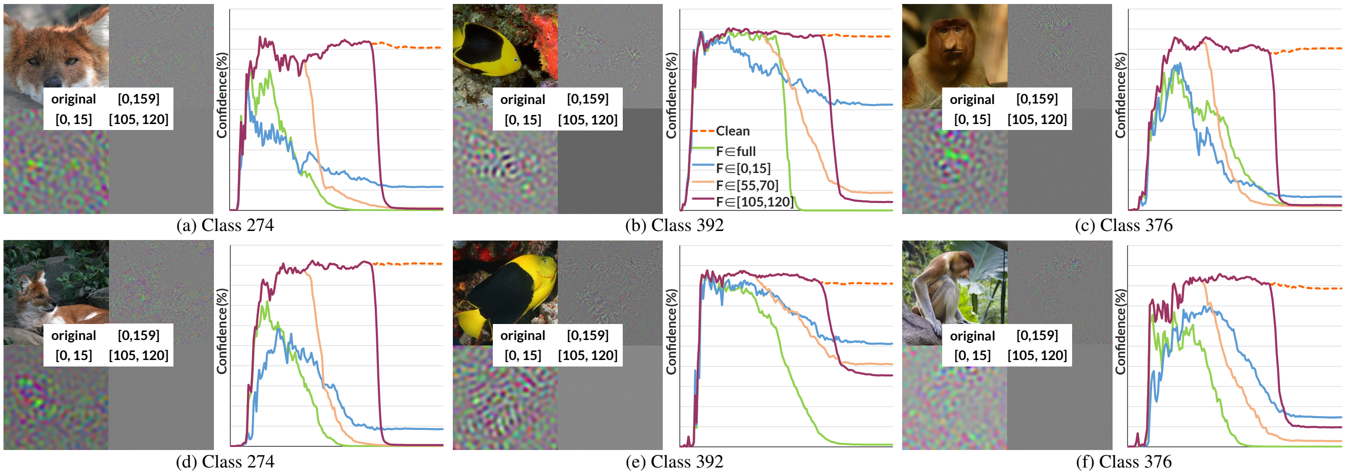


Figure 8: $\mathcal{F}$-PGD-L attack on samples which have stable low-frequency features, $\epsilon = 3/255$. Such low-frequency features are concentrated in a very narrow low-frequency band and are inherently more stable against adversarial attacks at various frequencies. However, the final classification results will still be dominated by high-frequency features, which are easily attacked by all frequencies.

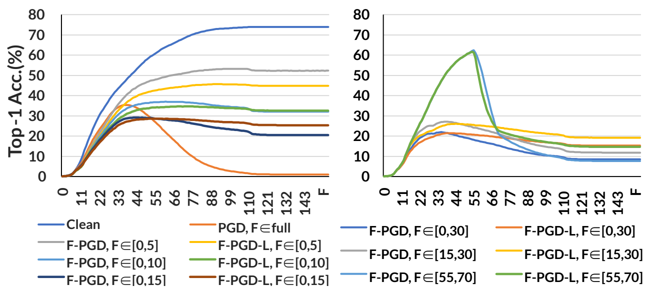


Figure 9: $\mathcal{F}$-PGD attack on MobileNetV3-Large. The adversarial attacks on low-frequency bands also affect high-frequency features, and the low-frequency features show higher robustness than the high-frequency features.

quency bands. The results of the attacks are shown in Table 1, the models retrained in the low-frequency band are less stable than the standard trained models when attacked by $\mathcal{F}^{\in}$-PGD. This indicates that models forced to be trained with low frequencies do not become robust. Similarly, models trained using only high frequencies or using a certain frequency band are also less robust than the standard trained models. Notably, the attack using $\mathcal{F}^{\notin}$-PGD shows that the models are not robust to the features in frequency bands that do not appear during training.

Further, we examined the stability of different frequency features under adversarial attacks, as shown in Figure 9, low-frequency features showed some stability under attacks at different frequencies. However, high-frequency features were more affected under both low-frequency and high-frequency attacks, and the influence of high-frequency features dominated the final classification results.

5.1 Attack on Low-Frequency Features

As shown in Figure 9, low-frequency features have better stability than high-frequency features, which explains why some blurring operations can be effective against attacks. As shown in Figure 6, many classes in the networks that can be correctly classified by ultra-low-frequency information alone. However, such stable low-frequency features could not improve the adversarial robustness of networks.

We perform $\mathcal{F}$-PGD-L attacks on low-frequency dependent samples at different frequency bands. As shown in Figure 8, low-frequency features have some robustness against attacks of various frequencies, and the networks could make correct decisions when using only low-frequency features. Moreover, the attacks on the low-frequency bands will affect the features of all frequency bands, while attacks on the high-frequency band will hardly interfere with the low-frequency features. The effect of adversarial attacks on low-frequency bands is higher than on high-frequency bands. The disturbed high-frequency features dominate the networks to make the final wrong decision. This indicates that stable low-frequency features cannot improve the robustness, and the model cannot unify low-frequency and high-frequency features.

6 Conclusion

In this paper, we analyzed the feature extraction and adversarial robustness from the Fourier perspective. We experimentally show that low- and high-frequency features are both important for decision-making. Low-frequency features are the basis for classification decisions, but high-frequency features often dominate when the neural networks make conflicting decisions on different frequency features. The attack experiments show that the low-frequency features are more robust to the attacks on different frequencies. However, since the higher frequency features dominate the decision-making and are very vulnerable to interference from various frequency attacks, the robust low-frequency features will not improve the adversarial robustness of CNNs. Next, we will explore whether reducing the sensitivity of networks to high-frequency features and improving the use of broader frequency features will improve the robustness of the network in future research.

References

[Brendel and Bethge, 2019] Wieland Brendel and Matthias Bethge. Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

[Carlini and Wagner, 2017] Nicholas Carlini and David A. Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017*, pages 39–57. IEEE Computer Society, 2017.

[Croce and Hein, 2020] 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, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 2206–2216. PMLR, 2020.

[Cubuk *et al.*, 2019] Ekin D. Cubuk, Barret Zoph, Dandelion Mané, Vijay Vasudevan, and Quoc V. Le. Autoaugment: Learning augmentation strategies from data. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 113–123. Computer Vision Foundation / IEEE, 2019.

[Deng *et al.*, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 248–255. IEEE Computer Society, 2009.

[Dosovitskiy *et al.*, 2021] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[Erhan *et al.*, 2009] Dumitru Erhan, Yoshua Bengio, Aaron Courville, and Pascal Vincent. Visualizing higher-layer features of a deep network. *Bernoulli*, pages 1–13, 2009.

[Geirhos *et al.*, 2019] 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. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

[Goodfellow *et al.*, 2015] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[Guo *et al.*, 2019a] Chuan Guo, Jared S. Frank, and Kilian Q. Weinberger. Low frequency adversarial perturbation. In Amir Globerson and Ricardo Silva, editors, *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*, pages 1127–1137. AUAI Press, 2019.

[Guo *et al.*, 2019b] Chuan Guo, Jacob R. Gardner, Yurong You, Andrew Gordon Wilson, and Kilian Q. Weinberger. Simple black-box adversarial attacks. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 2484–2493. PMLR, 2019.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016.

[Howard *et al.*, 2017] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *CoRR*, abs/1704.04861, 2017.

[Howard *et al.*, 2019] Andrew Howard, Ruoming Pang, Hartwig Adam, Quoc V. Le, Mark Sandler, Bo Chen, Weijun Wang, Liang-Chieh Chen, Mingxing Tan, Grace Chu, Vijay Vasudevan, and Yukun Zhu. Searching for mobilenetv3. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 1314–1324. IEEE, 2019.

[Ilyas *et al.*, 2019] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 125–136, 2019.

[Liu *et al.*, 2021] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 9992–10002. IEEE, 2021.

[Long *et al.*, 2022] 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 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part IV*, volume 13664 of *Lecture Notes in Computer Science*, pages 549–566. Springer, 2022.

[Madry *et al.*, 2018] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

[Maiya *et al.*, 2021] Shishira R. Maiya, Max Ehrlich, Vatsal Agarwal, Ser-Nam Lim, Tom Goldstein, and Abhinav Shrivastava. A frequency perspective of adversarial robustness. *CoRR*, abs/2111.00861, 2021.

[Rahaman *et al.*, 2019] Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred A. Hamprecht, Yoshua Bengio, and Aaron C. Courville. On the spectral bias of neural networks. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 5301–5310. PMLR, 2019.

[Shaham *et al.*, 2015] Uri Shaham, Yutaro Yamada, and Sahand Negahban. Understanding adversarial training: Increasing local stability of neural nets through robust optimization. *arXiv*, 2015.

[Sharma *et al.*, 2019] Yash Sharma, Gavin Weiguang Ding, and Marcus A. Brubaker. On the effectiveness of low frequency perturbations. In Sarit Kraus, editor, *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 3389–3396. ijcai.org, 2019.

[Simonyan *et al.*, 2014] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *2nd International Conference on Learning Representations, ICLR 2014 - Workshop Track Proceedings*, pages 1–8, 2014.

[Smilkov *et al.*, 2017] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda B. Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *CoRR*, abs/1706.03825, 2017.

[Springenberg *et al.*, 2015] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *3rd International Conference on Learning Representations, ICLR 2015 - Workshop Track Proceedings*, pages 1–14, 2015.

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

[Szegedy *et al.*, 2014] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In Yoshua Bengio and Yann LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

[Tan and Le, 2019] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *36th International Conference on Machine Learning, ICML 2019*, 2019-June:10691–10700, 2019.

[Tancik *et al.*, 2020] Matthew Tancik, Pratul P. Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T. Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. volume 2020-December, 2020.

[Tishby *et al.*, 2000] Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. 4 2000.

[Wang *et al.*, 2020] Haohan Wang, Xindi Wu, Zeyi Huang, and Eric P. Xing. High-frequency component helps explain the generalization of convolutional neural networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 8681–8691. Computer Vision Foundation / IEEE, 2020.

[Xu and Zhou, 2021] Zhiqin John Xu and Hanxu Zhou. Deep frequency principle towards understanding why deeper learning is faster. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35:10541–10550, 5 2021.

[Xu *et al.*, 2019] Zhi-Qin John Xu, Yaoyu Zhang, and Yanyang Xiao. Training behavior of deep neural network in frequency domain. In Tom Gedeon, Kok Wai Wong, and Minho Lee, editors, *Neural Information Processing - 26th International Conference, ICONIP 2019, Sydney, NSW, Australia, December 12-15, 2019, Proceedings, Part I*, volume 11953 of *Lecture Notes in Computer Science*, pages 264–274. Springer, 2019.

[Yin *et al.*, 2019] Dong Yin, Raphael Gontijo Lopes, Jonathon Shlens, Ekin Dogus Cubuk, and Justin Gilmer. A fourier perspective on model robustness in computer vision. In *NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13255–13265, 2019.

[Yosinski *et al.*, 2015] Jason Yosinski, Jeff Clune, Anh Mai Nguyen, Thomas J. Fuchs, and Hod Lipson. Understanding neural networks through deep visualization. *CoRR*, abs/1506.06579, 2015.

[Zeiler and Fergus, 2014] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. *Lecture Notes in Computer Science*, 8689 LNCS:818–833, 2014.

[Zhang *et al.*, 2023] Liangqi Zhang, Yihao Luo, Xiang Cao, Haibo Shen, and Tianjiang Wang. Frequency and scale perspectives of feature extraction. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023.

[Zhong *et al.*, 2020] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 13001–13008. AAAI Press, 2020.