How Does Frequency Bias Affect the Robustness of Neural Image Classifiers against Common Corruption and Adversarial Perturbations?

Alvin Chan[1,2*] , **Yew-Soon Ong**[2,1] , **Clement Tan**[1,2]
[1]Nanyang Technological University, Singapore
[2]Agency for Science, Technology and Research, Singapore

Abstract

Model robustness is vital for the reliable deployment of machine learning models in real-world applications. Recent studies have shown that data augmentation can result in model over-relying on features in the low-frequency domain, sacrificing performance against low-frequency corruptions, highlighting a connection between frequency and robustness. Here, we take one step further to more directly study the frequency bias of a model through the lens of its Jacobians and its implication to model robustness. To achieve this, we propose Jacobian frequency regularization for models' Jacobians to have a larger ratio of low-frequency components. Through experiments on four image datasets, we show that biasing classifiers towards low (high)-frequency components can bring performance gain against high (low)-frequency corruption and adversarial perturbation, albeit with a tradeoff in performance for low (high)-frequency corruption. Our approach elucidates a more direct connection between the frequency bias and robustness of deep learning models.

1 Introduction

Recent research has shown that model performances can drop drastically when natural test images are altered [Szegedy *et al.*, 2013; Hendrycks *et al.*, 2021]. One of these scenarios is when common image corruptions are added to the images. These corruptions include noise attributed to weather conditions, noisy environments, blurring effects and digital artifacts [Hendrycks and Dietterich, 2019]. Another case where models can fail is adversarial examples where a malicious party can craft imperceptible perturbations to images to influence the model's prediction [Szegedy *et al.*, 2013; Croce and Hein, 2019]. While research to build models more robust to these two situations started independently, recent studies have started to draw a connection between them. Interestingly, models trained to be robust against adversarial examples have shown mixed results in common corruptions: improving accuracy for some corruptions types while doing poorly for others.

Though studies have shown a link between data augmentation strategies and robustness against corruptions with different frequency components [Yin *et al.*, 2019], there is no study on how *direct* changes to a model's Fourier profile would affect its robustness. Here, we aim to directly alter the Fourier profile of a model to study its direct effect on both the model's adversarial and corruption robustness. To achieve this, we investigate the model's Jacobian which represents a visual map of pixel importance in a particular input image. Intuitively, a model would be relying on low (high)-frequency features when its Jacobians have a large ratio of low (high)-frequency components. While observing the Fourier spectra of natural images such as SVHN, CIFAR-10 and CIFAR-100, and the Jacobians of standard-trained models (Figure 1), these images have a much larger component of low-frequency features than their standard-trained models. This mismatch of frequency profiles motivates us to train models to bias towards low-frequency features through its Jacobians and study its effect on robustness.

To quantify the frequency profile of a model, we propose a frequency bias term that computes a scalar value from a Fourier spectrum of 2-D inputs such as an image or Jacobian to improve frequency evaluation beyond the visual inspection of Fourier spectra. Through this differentiable frequency bias term, we can use Jacobian frequency regularization (JaFR) to explicitly train a model to bias more heavily on low- or high-frequency features. Through our experiments, we find that a more direct change in a model's frequency profile towards low-frequency regions to match the frequency profile of the training data can boost clean accuracy, adversarial robustness and common corruptions in certain settings while trading off performance against low-frequency noise. Conversely, regularizing the model towards high-frequency regions can boost performance against low-frequency noise while sacrificing accuracy under high-frequency noise and adversarial examples. All in all, the core contributions of this paper are[1]:

- We propose a frequency bias term to measure the frequency bias of a Fourier spectrum.
- We show that biasing the Jacobians of models towards

[*] Corresponding author: `guoweialvin.chan@ntu.edu.sg`

[1] Full version with the Appendix in arXiv. Code available at: `https://github.com/alvinchangw/JaFR_IJCAI2022`

low or high frequency have implications on model robustness against adversarial robustness and an array of corruptions.

- To achieve this, we propose Jacobian frequency regularization (JaFR) to train model's Jacobians to have a larger or smaller weightage of low-frequency components.
- We conduct experiments on SVHN, CIFAR-10, CIFAR-100 and TinyImageNet to show how low-frequency bias in Jacobians can improve robustness against adversarial and high-frequency corruptions, albeit with tradeoffs in performance for low-frequency corruptions.

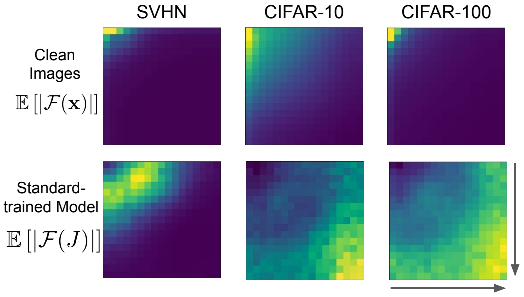


Figure 1: Fourier spectra of image datasets and the Jacobians of models trained on them, showing a mismatch between the frequency profiles of the images and models that are trained on them. Arrows show the direction of increasing frequency. Clean images contain large ratios of low-frequency components, shown by the intensity on the top left of the spectra (top row) while their corresponding standard-trained models rely heavily on high-frequency features (bottom row).

2 Background and Related Work

Adversarial Robustness. Adversarial robustness measures how well a model is resistant to attacks by malicious actors. In such attacks, imperceptible perturbations could be crafted to form adversarial examples with the aim to control the prediction of neural networks [Szegedy *et al.*, 2013]. This threat could undermine the deployment of deep learning models in mission-critical applications. In a classification task, a model (f) parameterized by θ takes an input $\mathbf{x}$ to predict the probabilities for k classes, i.e., $f(\mathbf{x};\theta):\mathbf{x}\mapsto\mathbb{R}^k$. In supervised setting of empirical risk minimization (ERM), given training samples $(\mathbf{x},\mathbf{y})\sim D$, the model's parameters are trained to minmize the standard cross-entropy loss:

$$\mathcal{L}(\mathbf{x},\mathbf{y})=\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim D}\left[-\mathbf{y}^\top\log f(\mathbf{x})\right] \tag{1}$$

where $\mathbf{y}\in\mathbb{R}^k$ is the one-hot label for the input $\mathbf{x}$. Though ERM can train models that display high accuracy on a holdout test set, their performance degrades under adversarial examples. Given an adversarial perturbation of magnitude ε, we say a model is robust against this attack if

$$\arg\max_{i\in C} f_i(\mathbf{x};\theta)=\arg\max_{i\in C} f_i(\mathbf{x}+\delta;\theta)\ , \quad \forall\delta\in B_p(\varepsilon)=\delta:\|\delta\|_p\leq\varepsilon \tag{2}$$

where $p=\infty$ in this paper and is the most widely studied scenario.

One of the most effective approaches to train models robust against adversarial examples is adversarial training (AT) [Goodfellow *et al.*, 2014; Andriushchenko and Flammarion, 2020; Wu *et al.*, 2020]. More details of related work in adversarial training are deferred to the Appendix.

Corruption Robustness. In contrast to adversarial robustness, corruption robustness [Hendrycks and Dietterich, 2019] entails studying the performance of models when input data are corrupted with common-occurring noise, not necessarily those created by a malicious actor to control models' prediction. CIFAR-10-C and CIFAR-100-C are two benchmark datasets that study 19 corruption types, each with 5 levels of severity. These 19 corruption types can be grouped under the 'Noise', 'Blur', 'Weather', 'Digital' categories. There is a line of work that seek to improve performances under these common corruptions by mostly improving the training data augmentation [Geirhos *et al.*, 2018; Vasconcelos *et al.*, 2020]. Some other works assemble expert models whose performance are finetuned for subsets of the corruptions [Saikia *et al.*, 2021] to boost overall performance. Different from these work, our paper aims to study the effect of frequency bias on corruption robustness of different types that corrupts features of varying frequencies, rather than to propose a new way to better resist these corruptions.

Link between Frequency and Robustness. There is a line of work that seeks to understand how robust models respond to corruption and adversarial perturbations of various frequency profiles [Yin *et al.*, 2019; Vasconcelos *et al.*, 2021]. In contrast to these prior works, our work here takes the frequency analysis in a different direction by studying models through the Fourier spectrum of their Jacobians rather than their test or training data. More concretely, with the original training data, we train models and bias the frequency profile of the model's Jacobians towards low-frequency regions to see its effect on model robustness.

Jacobians of Robust Models. The Jacobian,

$$J:=\nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x},\mathbf{y}) \tag{3}$$

defines how the model's prediction changes with an infinitesimally small change to the input $\mathbf{x}$. For image classification, Jacobians can be loosely interpreted as a map of which pixels affect the model's prediction the most and, hence, give an illustration of important regions in an input image [Smilkov *et al.*, 2017; Etmann *et al.*, 2019; Ilyas *et al.*, 2019]. There is a line of work that seeks to improve the adversarial robustness of models by matching the Jacobians to a target distribution [Chan *et al.*, 2019; Chan *et al.*, 2020] or by constraining their magnitude [Jakubovitz and Giryes, 2018]. Rather than aiming to improve the adversarial robustness, the core aim of our paper here is to investigate the relationship between the Fourier profile of models and robustness against corruptions. Moreover, the regularizing effect of JaFR has a different mechanism that acts directly on the Fourier spectrum of the Jacobians. A detailed comparison is in the Appendix.

3 Jacobian Frequency Bias

Motivation. As mentioned in § 2, it has been shown that there is a link between Fourier profile of input training data and robustness against corruptions with different frequency components. However, there is still no study on how changes to a neural network's Fourier profile would affect its robustness. Since the Jacobian of a model represents a visual map of pixel importance [Smilkov *et al.*, 2017], it offers a medium for us to regularize the Fourier profile of the model. Intuitively, when a model's Jacobians are concentrated with low-frequency components, it places more importance on low-frequency features. Conversely, the model relies more on high-frequency features if its Jacobians have a relatively large proportion of high-frequency components. Here, we aim to study how changing the Fourier spectrum of the Jacobians would affect its robustness.

When analyzing the Fourier profile of Jacobians of adversarially robust models (Table 2), we see that it resembles the profile of the training data much more than the non-robust standard trained models. This raises the question of what would happen to model robustness if we directly train neural networks to have a low-frequency profile, similar to what we see in the images from SVHN, CIFAR-10 and CIFAR-100 (see Figure 1). This motivates a metric to quantitatively measure and control the Fourier profile of the neural network to more directly study the effect of Fourier profile on robustness. In the next sections, we propose the Jacobian frequency bias to achieve this goal and discuss how we can use it to train a neural network.

3.1 Jacobian Frequency Bias

Here, we present the measure of Jacobian frequency bias with an example of single-channel image classification where the input images are denoted as $\mathbf{x} \in \mathbb{R}^{hw}$. Given a training dataset ($\mathcal{D}_{\text{train}}$) where each training sample consists of an input image $\mathbf{x}$ and one-hot label vector of k classes as $\mathbf{y} \in \mathbb{R}^{k}$, we can express $f_{\text{cls}}(\mathbf{x}) \in \mathbb{R}^{k}$ as the prediction of the classifier (f_{cls}), parameterized by θ. Then, the classification cross entropy loss ($\mathcal{L}_{\text{cls}}$) is:

$$\mathcal{L}_{\text{cls}} = -\mathbf{y}^{\top} \log f_{\text{cls}}(\mathbf{x}) \tag{4}$$

where f_{cls} is the classifier model. The Jacobian matrix $J \in \mathbb{R}^{hw}$ of the model's classication loss value with respect to the input layer can be computed through backpropagation:

$$J(\mathbf{x}) \coloneqq \nabla_{\mathbf{x}} \mathcal{L}_{\text{cls}} = \begin{bmatrix} \frac{\partial \mathcal{L}_{\text{cls}}}{\partial \mathbf{x}_{1,1}} & \cdots & \frac{\partial \mathcal{L}_{\text{cls}}}{\partial \mathbf{x}_{w,1}} \\ \vdots & \ddots & \vdots \\ \frac{\partial \mathcal{L}_{\text{cls}}}{\partial \mathbf{x}_{1,h}} & \cdots & \frac{\partial \mathcal{L}_{\text{cls}}}{\partial \mathbf{x}_{w,h}} \end{bmatrix} \tag{5}$$

We can then compute the Jacobian's Fourier spectrum to retrieve a frequency profile of the model. Since the input images are made up of discrete pixels, we can extract this information by applying a discrete Fourier transform ($\mathcal{F}$) to the Jacobian to get a map (M) of its frequency components' magnitude:

$$M_{i,j} = |\mathcal{F}(J)[i,j]| \tag{6}$$

In our experiments where the input images have 3 RGB channels, we compute the Fourier map for each other channel separately (M_r, M_g, M_b) and take the mean across these channels, i.e., $M_{i,j} = \frac{M_r[i,j]+M_g[i,j]+M_b[i,j]}{3}$.

Next, we propose to compute a scalar bias term ($\mathcal{B}_{\text{low}}$) from the 2-D map (M) to measure the relative bias (or ratio) of low-frequency components with respect to high-frequency components. One criterion for $\mathcal{B}_{\text{low}}$ is that the contribution of the frequency magnitude to this term should monotonically decrease as the frequency increase, i.e., larger high-frequency magnitudes results in lower $\mathcal{B}_{\text{low}}$ values. To satisfy this, we monotonically decrease the exponent value on the frequency magnitude as its frequency increases. For a 1-D scenario where l is the dimension of the Fourier spectrum, we can express this bias term $\mathcal{B}_{\text{low}}$ as:

$$\mathcal{B}_{\text{low}} = \Pi_i (M_i)^{\alpha_i}, \quad \alpha_i < \alpha_j, \quad \forall i,j \in [1,l], \quad i < j \tag{7}$$

where M_1 and M_l are the magnitudes of the lowest and highest frequency components respectively. To ensure that $\mathcal{B}_{\text{low}}$ measures the relative ratio between the frequencies rather than absolute values of the frequency components, we use the following constraint on the α so that it is independent of the sum of the components' magnitudes:

$$\alpha_i = -\alpha_{(l-i+1)}, \quad \forall i \in [1,l] \tag{8}$$

In all our experiments, we use an array of values whose values are evenly spaced with distance k from α_1 and α_l for the α values, i.e,

$$\alpha_{i+1} = \alpha_i + k, \ \ \forall i \in [1, l-1] \tag{9}$$

and use $\alpha_1 = 1, \alpha_l = -1$. When generalizing the bias term to two axes, we can compute the sum of all bias terms along each row and column of the 2-D Fourier spectrum to give:

$$\mathcal{B}_{\text{low}} = \left[\sum_j \left[\Pi_i (M_{i,j})^{\alpha_i}\right] + \sum_i \left[\Pi_j (M_{i,j})^{\alpha_j}\right] \right] \tag{10}$$

In the next section, we discuss how $\mathcal{B}_{\text{low}}$ can be used to regularize neural networks' Jacobians to bias them towards low-frequency components.

3.2 Jacobian Frequency Regularization (JaFR)

To recall, the aim here is to control the Fourier spectrum of a model's Jacobians to study how its Fourier profile can affect model robustness. To achieve this, we propose Jacobian Frequency Regularization (JaFR) to bias the Fourier spectrum of a model towards the low-frequency region. Figure 2 shows a summary of how our proposed Jacobian Frequency Regularization (JaFR) trains a model to alter its frequency profile. Since the operations involved in computing $\mathcal{B}_{\text{low}}$ are differentiable, we can simply incorporate it in the following loss function with the aim to maximize the low-frequency bias of a model:

$$\begin{aligned} \mathcal{L}_{\text{freq}} &= -\log \mathcal{B}_{\text{low}} = \\ & -\log \left[\sum_j \left[\Pi_i (M_{i,j})^{\alpha}\right] + \sum_i \left[\Pi_j (M_{i,j})^{\beta}\right] \right] \end{aligned} \tag{11}$$

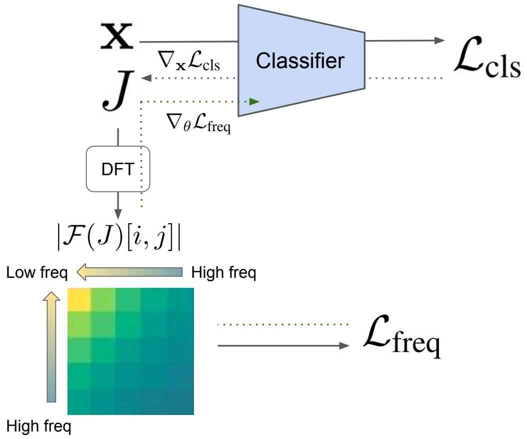

$\mathbf{x}$
J
Classifier
$\nabla_{\mathbf{x}}\mathcal{L}_{\text{cls}}$
$\nabla_{\theta}\mathcal{L}_{\text{freq}}$
$\mathcal{L}_{\text{cls}}$
DFT
$|\mathcal{F}(J)[i,j]|$
Low freq
High freq
High freq
$\mathcal{L}_{\text{freq}}$

Figure 2: Training architecture of Jacobian frequency regularization (JaFR). JaFR trains the model's Jacobian to bias towards low-frequency components by shifting its Fourier spectrum's intensity towards the low-frequency regions, leftwards for the horizontal and upwards for the vertical axis.

Combining with the classification loss in Equation 4, we can optimize through stochastic gradient descent to approximate the optimal parameters for the classifier f_{cls} as follows,

$$\theta^* = \arg\min_{\theta}(\mathcal{L}_{cls} + \lambda_{freq}\mathcal{L}_{freq}) \tag{12}$$

where λ_{freq} determines the weight of JaFR in the model's training. A positive λ_{freq} value regularizes the model to bias towards low-frequency features while a negative value (indicated as JaFR(-)) conversely steers the model towards high-frequency features. The $\mathcal{L}_{freq}$ term can be computed during each standard training iteration with an additional backpropagation step. Algorithm 1 summarizes the training of a classifier with JaFR. In the next sections, we present the results on the effect of JaFR on model robustness under adversarial examples and corruption noises.

Algorithm 1 Jacobian Frequency Regularization Training

```
Input: Train data $\mathcal{D}_{\text{train}}$, learning rate $\gamma$
for each training iteration do
    Sample $(\mathbf{x}, \mathbf{y}) \sim \mathcal{D}_{\text{train}}$
    $\mathcal{L}_{\text{cls}} \leftarrow -\mathbf{y}^\top \log f_{\text{cls}}(\mathbf{x})$    (1) Compute classification cross-entropy loss
    $J \leftarrow \nabla_{\mathbf{x}}\mathcal{L}_{\text{cls}}$    (2) Compute Jacobian matrix
    $M_{i,j} \leftarrow |\mathcal{F}(J)[i,j]|$    (3) Compute frequency magnitudes
    $\mathcal{L}_{\text{freq}} \leftarrow -\log\left[\sum_j \left[\Pi_i(M_{i,j})^\alpha\right] + \sum_i \left[\Pi_j(M_{i,j})^\beta\right]\right]$    (4) Compute frequency bias
    $\theta \leftarrow \theta - \gamma\, \nabla_\theta(\mathcal{L}_{\text{cls}} + \lambda_{\text{freq}}\mathcal{L}_{\text{freq}})$    (5) Update the classifier $f_{cls}$ to minimize $\mathcal{L}_{cls}$ and $\mathcal{L}_{freq}$
end for
```

4 Experiments

We conduct experiments across 4 image datasets (SVHN, CIFAR-10 and CIFAR-100, TinyImageNet) to study the effect of JaFR on model robustness against adversarial examples and common image corruptions. We largely follow the training setting as [Andriushchenko and Flammarion, 2020] where the PreAct ResNet-18 architecture [He *et al.*, 2016] is used for all models. To benchmark JaFR's effect on adversarial robustness, we compare with a relatively weak AT baseline (FGSM AT) and 2 strong AT baselines (FGSM AT + GradAlign [Andriushchenko and Flammarion, 2020] and PGD AT [Madry *et al.*, 2017]). All AT models use $\epsilon = \frac{8}{255}$. The dataset-specific parameters are further detailed in the following sections. Experiments are run on Nvidia V100 GPUs.

SVHN. All models were trained for 15 epochs, with a cyclic learning rate schedule and a max learning rate of 0.05. We use $\lambda_{freq} = 0.001$ and $\lambda_{freq} = 0.05$ for the JaFR and FGSM AT + JaFR model respectively. FGSM + GradAlign model uses the same hyperparameters in [Andriushchenko and Flammarion, 2020] for training.

CIFAR-10. The training setup largely follows the SVHN experiments where the PreAct ResNet-18 architecture is used for all models. All models were trained for 30 epochs, with a cyclic learning schedule. The standard-trained, JaFR and FGSM models use a max learning rate of 0.2 while the FGSM + JaFR, FGSM + GradAlign and PGD AT models used a max learning rate of 0.3. Both JaFR and FGSM AT + JaFR models use $\lambda_{freq} = 0.001$.

CIFAR-100. The training setup is similar to that in CIFAR-10, except for the learning rates and $\lambda_{freq} = 0.001$ values. The max learning rates for all models are 0.3 except for the JaFR (0.1) and standard-trained (0.2) models.

TinyImageNet. The training setup is similar to that in CIFAR-100, except for the $\lambda_{freq} = 0.002$ value.

4.1 Common Corruption Robustness

CIFAR-10-C and CIFAR-100-C are common corruption benchmarks where the CIFAR-10/CIFAR-100 test sets are corrupted with an array of 19 corruption types with 5 levels of severity. The accuracy values are reported as the average over the 5 severities of each corruption.

JaFR improves corruption robustness for standard-trained models. From Table 1 and 8 (Appendix), we observe that the JaFR model achieves the best overall corruption robustness for both CIFAR-10 and -100 images, as indicated by its lowest relative mean corruption errors (mCE). The JaFR outperforms the second-best model (standard-trained) in 14 out of 19 corruptions for CIFAR-10-C and 17 out of 19 corruptions for CIFAR-100-C test samples. This is an encouraging sign for direct frequency regularization of neural networks as an approach for better corruption robustness.

When comparing the spectra of the standard and JaFR model in Table 2, we observe that the intensities shift from the highest-frequency (bottom-right) region to lower frequency regions (left and top) with JaFR. This marks a lower reliance on high-frequency features in the model, which can confer better robustness against high-frequency corruptions. Indeed, with JaFR, we see the biggest improvements for high-frequency corruptions (e.g., Glass, Shot and Gaussian noise). Moreover, we can see that the spectra of our JaFR-only model, unlike the adversarially trained models (e.g.,

	Standard	JaFR	FGSM AT + JaFR(-)	FGSM AT	FGSM AT + JaFR	PGD AT	$\mathcal{B}_{\text{low}}$
Clean	93.11±0.20	93.13±0.11	83.16 ±1.01	84.80±1.37	79.94±0.22	79.31±0.23	-
mCE	100.00	95.84	127.51	124.32	134.37	135.71	-
Fog	86.23±0.45	**86.60±0.28**	**67.15±3.81**	63.00±4.03	56.60±0.72	56.36±1.25	12.85
Saturate	**89.27±0.16**	89.07±0.28	77.55±0.76	**79.54±1.07**	76.61±0.22	76.29±0.21	12.28
Contrast	**75.10±1.08**	73.45±0.95	**50.93±3.10**	44.06±3.97	41.96±0.61	41.19±0.88	12.14
Bright	**91.56±0.15**	91.45±0.18	81.02±0.88	**83.04±1.20**	76.95±0.42	76.01±0.58	12.01
Snow	79.54±0.71	**81.06±0.15**	75.05±1.52	**77.83±1.43**	74.28±0.27	73.73±0.21	11.53
Frost	75.41±0.63	**78.97±0.56**	75.82±1.64	**77.70±1.44**	70.50±0.54	69.06±1.18	10.61
Motion	**75.67±1.35**	75.13±1.21	67.33±2.01	65.15±3.47	**72.65±0.41**	72.15±0.78	10.61
Zoom	76.42±1.65	**76.48±1.20**	71.39±1.87	69.37±3.91	**75.38±0.44**	74.78±0.67	10.37
Elastic	81.62±0.60	**81.98±0.55**	73.88±1.45	73.90±2.27	**74.86±0.37**	74.25±0.58	10.05
Pixel	**72.90±0.70**	72.45±0.89	79.78±0.72	**81.20±0.55**	78.15±0.14	77.47±0.33	8.2
Gauss. B	73.38±1.58	**73.53±1.16**	70.56±1.51	68.49±3.93	**74.39±0.38**	73.71±0.63	7.82
Defocus	81.57±0.83	**81.76±0.71**	74.22±1.37	73.42±3.01	**76.15±0.29**	75.46±0.54	7.49
Glass	50.18±1.75	**55.12±0.58**	66.28±2.89	70.18±1.07	**73.75±0.28**	73.33±0.49	7.01
Spatter	80.77±0.11	**82.29±0.45**	76.46±1.68	**78.49±2.04**	75.83±0.32	75.29±0.32	6.77
JPEG	77.76±0.56	**79.56±0.60**	80.38±0.77	**81.47±1.54**	78.19±0.18	77.58±0.24	6.51
Speckle	62.79±2.44	**66.74±0.51**	68.86±4.83	74.07±4.08	**77.36±0.21**	76.22±0.54	3.76
Shot	59.63±2.75	**64.32±0.76**	68.32±5.06	73.73±4.19	**77.35±0.24**	76.19±0.42	3.68
Impulse	56.50±1.64	**58.75±0.59**	57.73±6.09	64.51±8.91	71.55±0.54	**71.56±0.74**	3.63
Gauss. N	48.70±3.31	**54.34±1.00**	62.91±6.74	69.40±5.27	**76.59±0.33**	75.15±0.49	3.62

Table 1: Accuracy values (↑ better) and mCE (↓ better) for different models under CIFAR-10 corruptions. The corruption types are arranged with descending order of low-frequency bias $\mathcal{B}_{\text{low}}$.

PGD AT), do not concentrate the intensities on the lowest-frequency region (top-left) which may result in an over-reliance on low-frequency features and make the model more susceptible to low-frequency corruptions. During our experiments, larger λ_{freq} values for JaFR do concentrate the intensities even more in the low-frequency regions but result in poorer overall performance. This balance in reliance on both high and low-frequency features may explain the improved overall performance of JaFR against corruption.

Tradeoff between robustness against low- and high-frequency corruption exists. For both CIFAR-10-C (Table 1) and CIFAR-100-C (Table 8), we observe that when JaFR is added to FGSM AT, the performance against low-frequency corruptions (e.g., Fog & Contrast) drops when the accuracy against high-frequency corruptions (e.g., Impulse & Gaussian noise) improves. We also see such behavior when JaFR is added to the standard trained model in CIFAR-10-C. This aligns with previous studies that show such a tradeoff between corruptions with different frequency profiles.

To study the opposite effect of JaFR, a JaFR(-) model variant is added where the λ_{freq} term has a negative value to investigate the effect of biasing the model towards high-frequency components. We can see such tradeoff emerging with an opposite effect when JaFR(-) is combined with FGSM AT to bias the model towards high frequency features, where the performance against high-frequency corruptions (e.g., Impulse & Gaussian noise) drops when the accuracy against low-frequency corruptions (e.g., Fog & Contrast) improves. This suggests that direct frequency regularization of models is a promising way to tailoring robustness against a set of corruptions that is more relevant for their applications.

FGSM AT + JaFR outperforms PGD AT in almost all corruptions. From our experiments, apart from one case of Impulse corruption in CIFAR-10-C, the FGSM AT + JaFR outperforms the PGD AT model in all the other 37 corruptions scenarios. We speculate that the PGD AT model might have overfitted to resist adversarial perturbations, making them excessively reliant on a small subset of low-frequency features that can be easily disrupted by the common corruptions.

4.2 Adversarial Robustness

We evaluate the adversarial robustness of models with FGSM [Goodfellow *et al.*, 2014] and PGD [Madry *et al.*, 2017] attacks. PGD uses 50 gradient iterations and 10 restarts with a step size of $\alpha_{adv} = \epsilon/4$.

Biasing towards low frequency can boost adversarial robustness in weakly adversarially trained models. For all the four image datasets, when JaFR is combined with FGSM, there is a significant boost in both clean and adversarial accuracy values from the FGSM AT model for all attack types (see Table 3, 4, 5 and 6). This observation is similar to what is observed for a recent defense, GradAlign, which also requires combining with FGSM to improve the adversarial robustness of the model. The need of using FGSM training samples to see an improvement in JaFR's adversarial robustness indicates that strong adversarial examples are more than just high-frequency noise and are able to find regions of high error in the uneven loss landscape of a non-adversarially trained model [Liu *et al.*, 2020]. For the CIFAR-10 and -100, the improvement of adding JaFR to FGSM is large enough to achieve a robustness level that is competitive with the strong PGD AT baseline. In Table C, we observe that JaFR scales well to a larger dataset, with a smaller amount of computation than PGD AT. Conversely, combining with JaFR(-) to bias the FGSM AT model towards high frequency with a negative λ_{freq} term was observed in our experiments to worsen

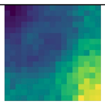
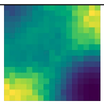
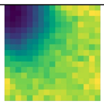
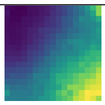
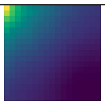
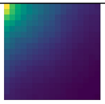
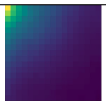
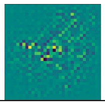
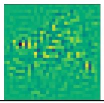
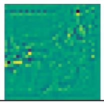
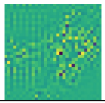
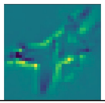
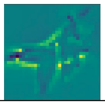
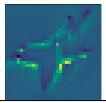
Model	Standard	JaFR	FGSM AT	FGSM AT + JaFR(-)	FGSM AT + JaFR	FGSM AT + GradAlign	PGD AT	Original Image
$\mathbb{E}\left[\lvert\mathcal{F}(J)\rvert\right]$								-
J								
$\mathcal{B}_{\text{low}}$	1.56	16.71	2.68	0.326	15.41	7.88	7.49	-

Table 2: CIFAR-10 models' low-frequency bias ($\mathcal{B}_{\text{low}}$), frequency spectra and Jacobians.

its adversarial robustness against PGD attacks. Experiments (Table 7 in Appendix) on more advanced attacks such as CW [Carlini and Wagner, 2017] and AutoAttack [Croce and Hein, 2020] show that the robustness gain from JaFR can resist even stronger attacks and is not due to gradient masking. Moreover, from Table 6, we observe that JaFR scales well, with a smaller amount of computation than the PGD AT.

Model	Clean	FGSM	PGD
Standard	96.62±0.05	21.86±0.90	0.17±0.02
JaFR	96.58±0.07	21.94±1.16	0.14±0.03
FGSM AT	92.33±0.20	89.16±5.87	0.52±0.90
FGSM + JaFR	87.22±4.63	59.4±12.75	19.64±16.26
FGSM + GA	92.51±0.26	59.52±0.26	46.63±0.15
PGD AT	92.13±0.51	62.38±0.86	56.79±0.29

Table 3: SVHN accuracy (%) on clean and adversarial test samples.

Model	Clean	FGSM	PGD
Standard	93.11±0.20	16.04±0.85	0
JaFR	93.13±0.11	15.86±0.88	0
FGSM AT	84.80±1.37	85.14±4.01	0.01±0.02
FGSM + JaFR	79.94±0.22	53.12±0.25	46.32±0.15
FGSM + GA	80.07±0.21	54.26±0.55	46.97±0.15
PGD AT	79.31±0.23	54.00±0.72	49.63±0.20

Table 4: CIFAR-10 accuracy (%) on clean and adversarial test set.

JaFR removes high-frequency components in models' Jacobian. From Table 2, we can see that models with JaFR have the highest low-frequency bias ($\mathcal{B}_{\text{low}}$) values, indicating that the training regularization is successful in increasing the low-frequency components of the model's Jacobians. Furthermore, from the Jacobian's Fourier spectra ($\mathbb{E}\left[\lvert\mathcal{F}(J)\rvert\right]$), intensities of the spectra shift from the high-frequency regions (right and bottom) towards the left and top regions of the spectra which indicate the low-frequency components along the horizontal and vertical axis respectively. In contrast, using a negative λ_{freq} term (in FGSM AT + JaFR(-)) concentrates the spectra towards highest-frequency (bottom-right) region and drastically lowers the ($\mathcal{B}_{\text{low}}$) value which indicates a strong reliance in high-frequency features.

JaFR improves saliency of Jacobians. When JaFR is added to the FGSM AT model, we observe that its Jacobians (J) become more salient (see Table 2). Together with the boost in adversarial robustness from the FGSM AT model, this corroborates previous results that saliency of Jacobian is correlated with adversarial robustness [Etmann *et al.*, 2019].

Model	Clean	FGSM	PGD
Standard	72.27±0.31	8.04±0.84	0.13±0.03
JaFR	74.59±0.37	8.60±0.81	0.08±0.04
FGSM AT	52.08±2.85	31.43±2.15	0.04±0.04
FGSM + JaFR	50.49±0.21	26.96±0.36	23.42±0.27
FGSM + GA	50.68±0.28	27.44±0.68	23.93±0.19
PGD AT	49.57±0.44	27.68±0.77	25.24±0.25

Table 5: CIFAR-100 accuracy (%) on clean and adversarial test set.

Model	Clean	FGSM	PGD	Complex. (s/iter)
Standard	63.59	2.23	0.05	0.1575
JaFR	63.44	1.87	0.04	0.6015
FGSM AT	45.89	17.54	0	0.4245
FGSM + JaFR	47.43	20.3	18.86	0.6027
FGSM + GA	40.69	19.86	17.81	0.6003
PGD AT	40	17.54	19.91	1.34

Table 6: Model accuracy and computational complexity on the Tiny-ImageNet dataset.

5 Conclusion

Model robustness is growing more important as deep learning models are gaining wider adoption. Here, we delve further into the link between the frequency characteristic of a model and its robustness by proposing Jacobian frequency bias. Through this term, we can control the distribution of high- and low-frequency components in the model's Jacobian and find that it can affect both corruption and adversarial robustness. We hope that our findings here will open an avenue for future work to explore other frequency-focused approaches to improve model robustness.

Acknowledgements

This work was partially supported by the Data Science and Artificial Intelligence Research Centre, School of Computer Science and Engineering at the Nanyang Technological University (NTU), Singapore, and the A*Star Center for Frontier Artificial Intelligence Research.

References

[Andriushchenko and Flammarion, 2020] Maksym Andriushchenko and Nicolas Flammarion. Understanding and improving fast adversarial training. *arXiv preprint arXiv:2007.02617*, 2020.

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

[Chan *et al.*, 2019] Alvin Chan, Yi Tay, Yew Soon Ong, and Jie Fu. Jacobian adversarially regularized networks for robustness. *arXiv preprint arXiv:1912.10185*, 2019.

[Chan *et al.*, 2020] Alvin Chan, Yi Tay, and Yew-Soon Ong. What it thinks is important is important: Robustness transfers through input gradients. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 332–341, 2020.

[Croce and Hein, 2019] Francesco Croce and Matthias Hein. Minimally distorted adversarial examples with a fast adaptive boundary attack. *arXiv preprint arXiv:1907.02044*, 2019.

[Croce and Hein, 2020] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, pages 2206–2216. PMLR, 2020.

[Etmann *et al.*, 2019] Christian Etmann, Sebastian Lunz, Peter Maass, and Carola-Bibiane Schönlieb. On the connection between adversarial robustness and saliency map interpretability. *arXiv preprint arXiv:1905.04172*, 2019.

[Geirhos *et al.*, 2018] 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. *arXiv preprint arXiv:1811.12231*, 2018.

[Goodfellow *et al.*, 2014] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer, 2016.

[Hendrycks and Dietterich, 2019] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

[Hendrycks *et al.*, 2021] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8349, 2021.

[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. *arXiv preprint arXiv:1905.02175*, 2019.

[Jakubovitz and Giryes, 2018] Daniel Jakubovitz and Raja Giryes. Improving dnn robustness to adversarial attacks using jacobian regularization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 514–529, 2018.

[Liu *et al.*, 2020] Chen Liu, Mathieu Salzmann, Tao Lin, Ryota Tomioka, and Sabine Süsstrunk. On the loss landscape of adversarial training: Identifying challenges and how to overcome them. *arXiv preprint arXiv:2006.08403*, 2020.

[Madry *et al.*, 2017] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[Saikia *et al.*, 2021] Tonmoy Saikia, Cordelia Schmid, and Thomas Brox. Improving robustness against common corruptions with frequency biased models. *arXiv preprint arXiv:2103.16241*, 2021.

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

[Szegedy *et al.*, 2013] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[Vasconcelos *et al.*, 2020] Cristina Vasconcelos, Hugo Larochelle, Vincent Dumoulin, Nicolas Le Roux, and Ross Goroshin. An effective anti-aliasing approach for residual networks. *arXiv preprint arXiv:2011.10675*, 2020.

[Vasconcelos *et al.*, 2021] Cristina Vasconcelos, Hugo Larochelle, Vincent Dumoulin, Rob Romijnders, Nicolas Le Roux, and Ross Goroshin. Impact of aliasing on generalization in deep convolutional networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10529–10538, 2021.

[Wu *et al.*, 2020] Dongxian Wu, Shu-Tao Xia, and Yisen Wang. Adversarial weight perturbation helps robust generalization. *Advances in Neural Information Processing Systems*, 33, 2020.

[Yin *et al.*, 2019] Dong Yin, Raphael Gontijo Lopes, Jonathon Shlens, Ekin D Cubuk, and Justin Gilmer. A fourier perspective on model robustness in computer vision. *arXiv preprint arXiv:1906.08988*, 2019.