A Spectral View of Randomized Smoothing under Common Corruptions: Benchmarking and Improving Certified Robustness

Jiachen Sun[1], Akshay Mehra[2], Bhavya Kailkhura[3], Pin-Yu Chen[4], Dan Hendrycks[5], Jihun Hamm[2], and Z. Morley Mao[1]

[1] University of Michigan, Ann Arbor
[2] Tulane University
[3] Lawrence Livermore National Laboratory
[4] IBM Research
[5] University of California, Berkeley

Abstract. Certified robustness guarantee gauges a model's resistance to test-time attacks and can assess the model's readiness for deployment in the real world. In this work, we explore a new problem setting to critically examine how the adversarial robustness guarantees change when state-of-the-art randomized smoothing-based certifications encounter common corruptions of the test data. Our analysis demonstrates a previously unknown vulnerability of these certifiably robust models to low-frequency corruptions such as weather changes, rendering these models unfit for deployment in the wild. To alleviate this issue, we propose a novel data augmentation scheme, *FourierMix*, that produces augmentations to improve the spectral coverage of the training data. Furthermore, we propose a new regularizer that encourages consistent predictions on noise perturbations of the augmented data to improve the quality of the smoothed models. We show that *FourierMix* helps eliminate the spectral bias of certifiably robust models, enabling them to achieve significantly better certified robustness on a range of corruption benchmarks. Our evaluation also uncovers the inability of current corruption benchmarks to highlight the spectral biases of the models. To this end, we propose a comprehensive benchmarking suite that contains corruptions from different regions in the spectral domain. Evaluation of models trained with popular augmentation methods on the proposed suite unveils their spectral biases. It also establishes the superiority of *FourierMix* trained models in achieving stronger certified robustness guarantees under corruptions over the entire frequency spectrum.

Keywords: Certified Robustness; Common Corruption; Benchmark

1 Introduction

Developing machine learning (ML) systems that are robust to adversarial variations in the test data is critical for applied domains that require ML safety [21],

such as autonomous driving and cyber-security. Unfortunately, a large body of work in this direction has fallen into the cycle where new empirical defenses are proposed, followed by new adaptive attacks breaking these defenses [3,55]. Therefore, significant efforts have been dedicated to developing methods that provide provable robustness guarantees [17,42,57]. Most promising among these certified defenses are based on *randomized smoothing (RS)* based certification [9, 32, 33] which are scalable to deep neural networks (DNNs) and high-resolution datasets. Specifically, the RS-based certification procedure relies on a smoothed version of the original classifier, which outputs the class most likely returned by the original classifier under random noise perturbations of the input. Prediction from the RS procedure at the test time is accompanied by a *radius* in which the predictions of the smoothed classifier are guaranteed to remain constant, thereby making them resilient to adversarial attacks within the neighborhood. Training methods such as [9,47,66] have been proposed to maximize the *average certified radius (ACR)*, and models trained using these procedures achieve state-of-the-art (SOTA) adversarial robustness guarantees, all while assuming that the test data is identically distributed to the training data. In this work, we take a critical look at the current status of certifiably robust ML and consider whether these certifiably robust models are ready for deployment in the real world. Our work takes the first steps towards answering this question by evaluating RS-based provably robust ML models under *common corruptions*, as mismatches between the training and deployment distributions are ubiquitous in the wild. Our analysis shows that **common corruptions pose a serious threat to certifiably robust models.** We, therefore, highlight a previously unrecognized threat to certifiably robust models and thereby show that these models are not ready for deployment in the real world. Specifically, we found SOTA certifiably robust models to be surprisingly brittle to low-frequency perturbations, such as weather-related corruptions (*e.g.*, fog and frost). Vulnerability to such corruptions could lead to a detrimental performance of ML models on safety-critical applications. For example, 35%–75% performance drop is observed on low-frequency corruptions rendering RS-based robustness guarantees useless (Fig. 1).

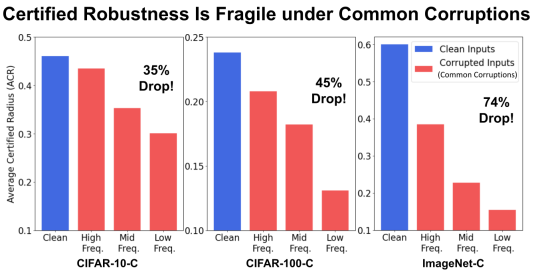


Fig. 1. Robustness guarantees of certified models [9] degrade significantly on corrupted data.

Motivated by our analysis, which shows RS-based smoothed classifiers suffer from low-frequency corruptions, we propose a novel data augmentation method that uses **spectrally diverse yet semantically consistent augmentations** of the training data. Specifically, our proposed *FourierMix* generates augmented data samples by using Fourier-based transformations on the input data to in-

crease the spectral coverage of the training set. *FourierMix* proportionally perturbs the amplitude and phase of the images in the training data and then combines them with the affine transformations of the data, producing spectrally diverse augmentations. To encourage the model to produce consistent predictions on different data augmentations, we propose a *hierarchical consistency regularizer (HCR)*. The use of HCR as the regularizer leads to semantic consistency of representations across random noise perturbations (for RS certification) as well as *FourierMix* generated augmentations (for corruption robustness) of the same input image. *FourierMix* consistently achieves significantly better-certified robustness than existing SOTA data augmentation methods extended to build a smoothed classifier across a range of corruption benchmarks. We further analyze these smoothed models using Fourier sensitivity analysis in the spectral domain. Compared to other methods, models trained on *FourierMix* augmentations coupled with hierarchical consistency regularization are significantly more resilient to perturbations across the entire frequency spectrum.

Our evaluation of certifiably robust models on various corruption benchmark datasets uncovers another peculiar phenomenon–**even popular benchmark datasets may be biased towards certain frequency regions**. Due to the complexity of real-world data, it is extremely challenging and tedious to unveil the spectral biases of the models and identify their failure modes. Because of this, improvements in the performance of the models on these benchmarks may not generalize to other corruption types. Thus, we should be cautious and avoid over-reliance on a specific leaderboard, especially to judge the robustness of models under corruption. To enable the designers to understand the spectral biases of their models and obtain a more comprehensive view of the model robustness to data corruptions, we propose a new benchmark that includes a collection of corruption test sets, each focusing on specific frequency ranges while collectively covering the entire frequency spectrum. Evaluation of the certified robustness of different models on the proposed dataset shows that the smoothed models obtained after training with existing data augmentation schemes are indeed biased towards certain frequency regions. This justifies the observed performance (and ranking) variations across different benchmarks. On the other hand, models trained with our *FourierMix* based data augmentations perform significantly better than the competitors across the entire frequency spectrum, further demonstrating that *FourierMix* helps alleviate the spectral biases.[6]

A detailed discussion on related work is provided in Appendix A, while all the references are included in the main paper.

2 Are Certifiably Robust Models Ready for Deployment in the Wild?

Predictions of certifiably robust ML models are guaranteed to stay constant in a neighborhood of a test point, making them provably resilient to adversaries at the

[6] The codebase and dataset of this work are available at `https://github.com/jiachens/FourierMix`.

test time. This feature of certified defenses makes them an attractive candidate for real-world safety-critical applications. However, progress in this area has been assessed by evaluating these models in idealistic scenarios (*i.e.*, the in-distribution setup), which is not representative of real-world data distributions. To better understand the performance of certified defenses in the real world, in this section, we evaluate SOTA certified defenses under common corruptions.

2.1 Preliminaries on SOTA Certified Defenses

Previous works have proposed different certification methods to obtain provable adversarial robustness guarantees (*e.g.*, convex polytope [57], recursive propagation [17], and linear relaxation [42, 67]). However, their use is limited due to their trivial bounds derived from large-scale datasets and deep models. Recently, randomized smoothing (RS) based certification method was proposed, which is efficient and scalable to large-scale datasets and deep models. Therefore, we use RS-based certification in this study. Let us consider a base classifier $\mathcal{M}$ trained on samples $\boldsymbol{x} \in \mathcal{X} \subset \mathbb{R}^{d \times d \times 3}$ and their corresponding labels $y \in \mathcal{Y} \subset \mathbb{R}^{+}$, obtained from an underlying data distribution $\mathcal{D}$.

Certification. The RS-based certification uses a base classifier $\mathcal{M}$ and provides certified robustness guarantees for its smoothed version defined as $\hat{\mathcal{M}}(\boldsymbol{x}) = \arg\max_{c \in \mathcal{Y}} \mathbb{P}(\mathcal{M}(\boldsymbol{x} + \boldsymbol{\delta}) = c)$ where $\boldsymbol{\delta} \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$. Intuitively, $\hat{\mathcal{M}}$ returns the most probable class evaluated by $\mathcal{M}$ over a number of Gaussian perturbations of the input $\boldsymbol{x}$. The certification guarantees that the prediction of the smoothed classifier $\hat{\mathcal{M}}$ are consistent in the ℓ_2 radius [9] of $\mathrm{CR}(\hat{\mathcal{M}}, \sigma, \boldsymbol{x}; y) = \frac{\sigma}{2}(\Phi^{-1}(p_A) - \Phi^{-1}(p_B))$, where Φ^{-1} is the inverse CDF of the standard Gaussian distribution, $p_A = \mathbb{P}(\mathcal{M}(\boldsymbol{x} + \boldsymbol{\delta}) = c_A)$ is probability of the top class c_A and $p_B = \max_{c \neq c_A} \mathbb{P}(\mathcal{M}(\boldsymbol{x} + \boldsymbol{\delta}) = c)$ is the probability of the runner-up class. Monte Carlo-based sampling [18] is utilized to approximate $\underline{p_A} \leq p_A$ and $\overline{p_B} = 1 - \underline{p_A} \geq p_B$. The certified radius can still be computed using the same formula by replacing p_A and p_B with $\underline{p_A}$ and $\overline{p_B}$.

Improved Training. It has been observed empirically [9] that models trained using the standard procedure do not provide reasonable certified robustness. Therefore, there is an increasing interest in developing improved training techniques to maximize certified robustness. Several works [34] have made significant advances in the training techniques and reported impressive gains in certified radius on in-distribution test data. Specifically, new training methods such as Gaussian augmentation [9], SmoothAdv [47] and MACER [66] have been proposed. Intuitively, Cohen *et al.* [9] propose to leverage Gaussian augmentation with variance σ^2 to train the base classifier. SmoothAdv [47] and MACER [66] both use Gaussian augmentation and further improve Cohen *et al.*'s baseline method by adversarial training and introducing an auxiliary objective to maximize the certified radius, respectively. However, the effect of common corruptions on the robustness guarantees of these models remains unexplored.

Evaluation Metrics. Similar to [37, 47, 66], we use the *average certified radius* (ACR) as our metric to evaluate the robustness:

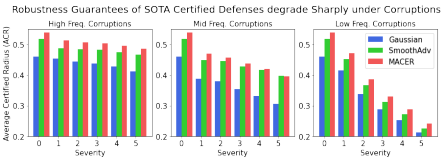
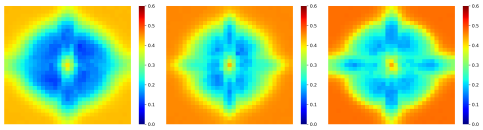


Fig. 2. Randomized smoothing based models [9, 47, 66] suffer up to 54.0% decreases in their certified robustness on **mid-to-low** frequency corruptions from CIFAR-10-C. Severity 0 is in-distribution.



(a) Gaussian (b) SmoothAdv (c) MACER

Fig. 3. Fourier sensitivity analysis on CIFAR-10 shows the ACR of SOTA certified defenses degrade significantly under corruptions from **mid-to-low** frequency region (interpreted in § 2.2).

$$\text{ACR} := \frac{1}{|\mathcal{D}_{test}|} \sum_{(\boldsymbol{x},y)\in\mathcal{D}_{test}} \text{CR}(\hat{\mathcal{M}}, \sigma, \boldsymbol{x}; y) \times \mathbf{1}_{\hat{\mathcal{M}}(\boldsymbol{x},\sigma)=y}$$

which is also equivalent to the area under the certified radius-accuracy curve (AUC). For performance on corruption datasets, we measure the mean ACR (mACR) as an overall metric, $\text{mACR} := \frac{1}{c}\sum_{i=1}^{c} \text{ACR}_i$, where c is the number of corruptions leveraged in a specific test set. For example, $c = 15$ and 10 in CIFAR-C and -$\bar{\text{C}}$ datasets, respectively. Unlike previous studies on empirical defenses, we do not use the *empirical* clean and robust accuracy [9, 47, 66] as a metric in this work since we focus on the *certified* robustness.

2.2 Analyzing Certified Defenses under Common Corruptions.

Real-world test data often do not follow the training data distribution $\mathcal{D}$, although tangible improvements have been made in certifying the robustness of in-distribution data. Therefore, evaluating the performance of $\mathcal{M}$ under distribution shifts (*i.e.*, corrupted data) $\{(\hat{\boldsymbol{x}}, y)_1, ..., (\hat{\boldsymbol{x}}, y)_n\} \sim \hat{\mathcal{D}}$ becomes a major concern. We consider the impact of corrupted data on models trained using SOTA robust training methods [9, 47, 66] and RS-based certified defenses.

Degradation of Certified Robustness Guarantees on Common Corruptions. To measure the performance of certified defenses under data corruptions, we use the prevalent corruptions dataset CIFAR-10-C [22], which contains 15 different corruptions from four categories (with 5 severity levels): noise, blur, weather, and digital corruptions. We re-arrange the corruption dataset into three groups and evaluate the ACR by increasing the severity level of the corruption. Grouping is performed based on the visual similarity of the amplitude spectrum of corrupted images (see Appendix C). Group-H corruptions (roughly categorized as high-frequency corruption type) consist of {`Gaussian noise`, `impulse noise`, `shot noise`, `pixelate`, `JPEG`}; Group-M corruptions (roughly categorized as mid-frequency corruption type) consist of {`defocus blur`, `frosted glass blur`, `motion blur`, `zoom blur`, `elastic`}; and Group-L corruptions (roughly categorized as low-frequency corruption type) consist of {`brightness`, `fog`, `frost`, `snow`, `contrast`}.

The performance of SOTA certified defenses on these groups of corruptions is presented in Fig. 2. SmoothAdv and MACER achieve tangible enhancements in ACR on in-distribution CIFAR-10 data compared to the Gaussian augmentation baseline. However, all methods show a sharp performance drop in ACR as we move from Group-H (high-frequency) to Group-L (low-frequency). We see that these methods are surprisingly brittle in low-frequency corruption regimes, *e.g.*, we see up to 54.0% drop in ACR when moving from severity 0 (*i.e.*, in-distribution) to severity 5. We emphasize that this performance drop points to a methodological shortcoming. The degradation is not due to the corruptions in Group-L being too difficult since the empirical robust accuracy (Fig. 9 in Appendix B) remains consistently high on all the groups and severity levels for empirically robust models [23, 30, 44]. Even though the performance of any ML model is expected to suffer on test data that lies far away from the data used during training, the drastic performance degradation of RS-based certifiably robust models on low-frequency corruptions is particularly concerning. Our findings also generalize to IBP-based certification [59] (Appendix D.1), further demonstrating the vulnerability of certified defenses to low-frequency corruptions.

Validating the Brittleness of Smoothed Models Through a Spectral Lens. To highlight that the vulnerability to low-frequency corruptions is a limitation of provably robust ML models, in this section, we perform a more systematic analysis that corroborates that our finding is not limited to a specific benchmark and holds more broadly. To achieve this, we perform a spectral domain analysis of smoothed models by utilizing the Fourier sensitivity analysis [65].

A Fourier basis image in the pixel space is a real-valued matrix $\boldsymbol{U}_{i,j} \in \mathbb{R}^{d \times d}$ where its $||\boldsymbol{U}_{i,j}||_2 = 1$, and $\text{FFT}(\boldsymbol{U}_{i,j})$ only has two non-zero elements at (i, j) and $(-i, -j)$ in the coordinate that views the image center as the origin. Given a test set and a smoothed model, we evaluate the $\text{CR}(\cdot)$ of $\widetilde{\boldsymbol{x}}_{i,j} = \boldsymbol{x} + r\epsilon\boldsymbol{U}_{i,j}$ for each $\boldsymbol{x}$ in the test set and compute their ACR, where r is randomly sampled in $\{-1, 1\}$, ϵ is the perturbation in ℓ_2 norm, and we treat the RGB channels independently. Each of the evaluated ACR corresponds to a data point in the heat map located at (i, j). Fig. 3 shows the heatmaps of models trained with Gaussian augmentation [9], SmoothAdv [47], and MACER [66] using $\epsilon = 4$ [65]. The center and edges of the heatmap contain the evaluation of the lowest and highest frequency perturbations, respectively. The results in Fig. 3 show that the certifiably robust classifiers achieve small ACR on corrupted data belonging to the low-frequency region (around the center of the image), whereas they achieve a high ACR in the high-frequency region (near the edges). In particular, the ACRs are always less than 0.3 for all three methods in the mid-to-low frequency range, while they perform well in a high-frequency regime. We emphasize that the Fourier sensitivity analysis in Fig. 3 is general and is not specific to corruptions appearing in CIFAR-10-C. Based on our analysis, we find that certifiably robust models are biased towards high-frequency noises and perform surprisingly poor on low-frequency corrupted data. Following this insight, we develop a data augmentation method capable of producing spectrally diverse augmentations to make certifiably robust models perform well on corrupted data across the en-

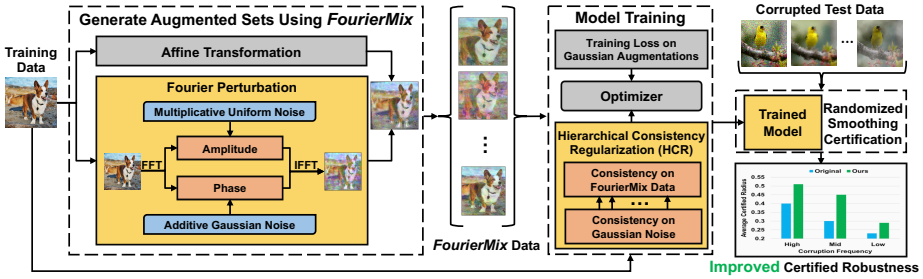


Fig. 4. Overview of Our *FourierMix* Pipeline for Generating Spectrally Diverse Data Augmentations and Training of Certifiably Robust Models with the Proposed Hierarchical Consistency Regularization (HCR).

tire frequency spectrum in § 3. It is also worth noting that although test-time adaptation [56] is another class of methods that improves the *empirical* corruption robustness, we demonstrate that they are ineffective when combined with certified defenses in Appendix G.

3 *FourierMix*: Data Augmentation Strategy with a Broad Spectral Coverage

To improve the certified robustness of RS-based methods under common corruptions, it is intuitively desirable to make the base classifier $\mathcal{M}$ robust against different types of corruptions and their Gaussian perturbations. Motivated by our Fourier sensitivity analysis (§ 2), we propose a novel data augmentation method, *FourierMix*. As opposed to existing data augmentation schemes, *FourierMix* explicitly uses *spectral coverage* as its design objective to boost the certified robustness of corrupted data. To improve the spectral coverage, we introduce Fourier-based operations that manipulate the image in the frequency domain. We also leverage randomly sampled affine transformations to enrich the augmentations in *FourierMix*. We adopt the high-level framework of AugMix [23] for chaining and mixing different augmented images. Figure 4 shows the overall pipeline and Algorithm 1 presents the pseudocode of *FourierMix*.

Fourier Operations. Two-dimensional images can be converted into the frequency domain by applying the Fourier transform and vice versa. Fourier transform has the *duality* property, which provides a unique but equivalent perspective for image analysis. We use fast Fourier transform (FFT) and inverse FFT (IFFT) for the transformation between the pixel and frequency domains. $\mathrm{FFT}(\boldsymbol{x})$ is complex in general, *i.e.*, $\mathrm{FFT}(\boldsymbol{x}) = \mathrm{FFT}_{real}(\boldsymbol{x}) + i\mathrm{FFT}_{imag}(\boldsymbol{x})$, with $\boldsymbol{A} = |\mathrm{FFT}(\boldsymbol{x})|$ as its amplitude and $\boldsymbol{P} = \arctan(\mathrm{FFT}_{imag}(\boldsymbol{x})/\mathrm{FFT}_{real}(\boldsymbol{x}))$ as its phase. The amplitude spectrum of natural images generally follows a power-law distribution, *i.e.*, $\frac{1}{f^\alpha}$, where f is the azimuthal frequency and $\alpha \approx 2$ [5, 54], resulting in extremely small power in the high-frequency areas. However, the amplitude

spectrum of the I.I.D. Gaussian noise is a uniform distribution, so Gaussian augmentation biases the models toward the high-frequency regime relative to the original images. In order to have broad and unbiased spectral coverage, the core of *FourierMix* is to allocate similar proportions of augmentations across all frequencies. We use two spectral operators in *FourierMix* to achieve this goal:

$$\mathbf{A}(u,v) = \boldsymbol{A}^{\text{orig}}_{u,v} \cdot \text{U}(1 - s_{\mathbf{A}}, 1 + s_{\mathbf{A}}) \tag{1}$$

$$\mathbf{P}(u,v) = \boldsymbol{P}^{\text{orig}}_{u,v} + \mathcal{N}^{s_{\mathbf{P}}}_{\text{truncated}}(0, \sigma^2\mathbf{I}) \tag{2}$$

where (u, v) is the coordinate of the 2D frequency in the spectrum, and $s_{\mathbf{A}}$ and $s_{\mathbf{P}}$ control the severity levels of two operators. Formally, the PDF of $\mathcal{N}^{s_{\mathbf{P}}}_{\text{truncated}} = \frac{\phi(x/\sigma)}{\sigma \cdot (2\Phi(s_{\mathbf{P}}/\sigma) - 1)}$, where $\phi(\cdot)$ and $\Phi(\cdot)$ denote the PDF and CDF functions of a standard normal distribution, respectively. On one hand, we apply multiplicative factors sampled from a uniform distribution $\text{U}(\cdot)$ to all frequencies in the amplitude spectrum. Therefore, $\mathbf{A}(u, v)$ ensures that the proportions of augmentation are similar across all frequencies relative to the original spectrum. On the other hand, since the magnitude of the phase spectrum follows a random distribution that is not correlated with the 2D frequency [36], additive phase noises can thus assign similar proportions of augmentations across 2D frequencies. As it is widely acknowledged that the phase component retains most of the high-level semantics [29,61,64], we leverage additive truncated Gaussian to constrain $\mathbf{P}(u, v)$ so that it will not destroy the semantics of the training images. Some sample images generated using *FourierMix* are provided in Appendix E.

Hierarchical Consistency Regularization (HCR). Motivated from [25] that enforces consistency on in-distribution data, we propose *hierarchical consistency regularization* (HCR) to further boost the performance of *FourierMix* in terms of the ACR on corrupted test sets:

$$\mathcal{L}_G = \frac{1}{s} \sum_{i=0}^{s} \text{KL}(\mathcal{M}(\boldsymbol{x}_j + \boldsymbol{\delta}_i) \| \overline{\mathcal{M}}(\boldsymbol{x}_j, \boldsymbol{\delta})) \tag{3}$$

$$\mathcal{L}_{HCR} = \frac{1}{k+1} \sum_{j=0}^{k} \Big[\lambda \cdot \text{KL}(\overline{\mathcal{M}}(\boldsymbol{x}_j, \boldsymbol{\delta}) \| \overline{\mathcal{M}}(\boldsymbol{x}, \boldsymbol{\delta})) + \eta \cdot \mathcal{L}_G \Big] \tag{4}$$

where $\overline{\mathcal{M}}(\boldsymbol{x}, \boldsymbol{\delta}) = \mathbb{E}_{j \in \{0,1,\ldots,k\}}[\overline{\mathcal{M}}(\boldsymbol{x}_j, \boldsymbol{\delta})]$, $\overline{\mathcal{M}}(\boldsymbol{x}_j, \boldsymbol{\delta}) = \mathbb{E}_{i \in \{1,2,\ldots,s\}}[\mathcal{M}(\boldsymbol{x}_j + \boldsymbol{\delta}_i)]$, $\boldsymbol{x}_0$ is the original training image, and $\text{KL}(\cdot\|\cdot)$ denotes the Kullback–Leibler divergence (KLD) [28]. We use $k = 2$ and $s = 2$ for the *FourierMix* and Gaussian augmentation with $\delta_i = \mathcal{N}(0, \sigma^2\mathbf{I})$, respectively. Since Jensen–Shannon divergence (JSD) [15] uses the KLD to calculate a normalized score that is symmetrical, HCR essentially stacks two levels of JSD while training the base classifier to enforce the consistent representations over both augmentations. The first level of consistency $\mathcal{L}_G$ is applied to the Gaussian augmentation, rendering the Gaussian perturbed neighbors of $\boldsymbol{x}_{0,1,2}$ have similar outputs, and the second level of consistency is on the whole $(k + 1)s$ set to enforce *FourierMix* augmented images with consistent outputs. We utilize λ and η as hyper-parameters

to tune the weights of two levels of consistency. The overall training loss is: $\mathcal{L} = \frac{1}{s}\sum_{i=1}^{s}\mathcal{L}(\boldsymbol{x}_0 + \boldsymbol{\delta}_i, y) + \mathcal{L}_{HCR}$.

Algorithm 1: *FourierMix* Pseudocode

Data: Model $\mathcal{M}$, Image $\boldsymbol{x}_{\text{orig}}$, Affine Transformation $\mathbf{T}$, Fourier Amplitude $\mathbf{A}$ and Phase $\mathbf{P}$ Operations

Result: $\boldsymbol{x}_{\text{aug}} = \textit{FourierMix}(\boldsymbol{x}_{\text{orig}},\ k,\ \alpha)$

$\boldsymbol{x}_{\text{aug}} = 0$
Sample mixing weights $(w_1, ..., w_k) \sim \text{Dirichlet}(\alpha, ..., \alpha)$
for $i = 1, 2, ..., k$ **do**
  Sample random affine transformation $\mathbf{T}_i$
  $\boldsymbol{x}_{\text{fourier}} = \text{FFT}(\boldsymbol{x}_{\text{orig}})$
  Sample severity level of operations $s_{\mathbf{A}}, s_{\mathbf{P}}$
  $\boldsymbol{x}_{\text{fourier}} = (\mathbf{A}_{s_{\mathbf{A}}} \circ \mathbf{P}_{s_{\mathbf{P}}})(\boldsymbol{x}_{\text{fourier}})$
  $\boldsymbol{x}_f = \text{IFFT}(\boldsymbol{x}_{\text{fourier}})$
  Sample weight $t \sim \text{Beta}(\alpha, \alpha)$
  $\boldsymbol{x}_{\text{aug}} += w_i \cdot (t\boldsymbol{x}_f + (1-t)\mathbf{T}_i^{\top} \cdot \boldsymbol{x}_{\text{orig}})$
end
Sample weight $m \sim \text{Beta}(\alpha, \alpha)$
$\boldsymbol{x}_{\text{aug}} = m\boldsymbol{x}_{\text{orig}} + (1-m)\boldsymbol{x}_{\text{aug}}$

Comparison with Prior Arts. There are some notable differences between *FourierMix* and prior SOTA in terms of: a) base augmentation operations, and b) data augmentation objective. These differences are later quantitatively highlighted using experimental results.

AugMix leverages the base augmentation operations from AutoAugment [11] that do not overlap with ImageNet-C. In contrast, the augmentations in *FourierMix* utilize a simpler set of generic augmentations. We compare the performance (*i.e.*, ACR) of *FourierMix* with AugMix on multiple corruption datasets in our evaluation (§ 4 and § 5). Another key difference between *FourierMix* and prior arts is that FourierMix explicitly uses spectral coverage as the data augmentation objective. For example, the recently proposed FACT [62] randomly mixes the amplitude spectra of two training samples, which has no control over the spectral coverage (results are presented in Appendix D.1). However, *FourierMix* realizes proportional assignment of augmentation across all frequencies.

4 Experiments on Popular Corruption Benchmarks

Experimental Setup. We use ACR and mACR (see § 2.1) as the main evaluation metrics. We utilize the official implementation from [9] to compute the certified radius CR(·). We use the same base architectures leveraged in prior arts [9, 25, 47, 66], *i.e.*, ResNet-110 and ResNet-50, for experiments on CIFAR-10/100 and ImageNet [19], respectively. We use Gaussian augmentation with $\sigma = 0.25$ and 0.5 for both training and certifying the CIFAR-10/100 and ImageNet models, respectively. Further training details are in Appendix D.

Baselines. We evaluate the certified robustness of models trained with following augmentations schemes on corrupted data: Gaussian [9], AutoAugment [11], and AugMix [23]. We also compare HCR with the baseline JSD regulariza-

Table 1. Models trained with *FourierMix* and HCR achieve significant improvements in the certified robustness (ACR and mACR) guarantees on all popular corruption datasets. **Bold** and underline denote the best and runner-up results, respectively.

	CIFAR-10	CIFAR-10-C				CIFAR-10-$\bar{\text{C}}$
Augmentation	ACR	mACR	-Low	-Mid	-High	mACR
Gaussian	0.461	0.363	0.301	0.353	0.435	0.314
+JSD	**0.535**	0.439	0.346	0.451	0.520	0.393
+AutoAugment	0.411	0.372	0.312	0.364	0.431	0.304
+JSD	0.432	0.400	0.343	0.395	0.464	0.346
+AugMix	0.452	0.385	0.324	0.383	0.449	0.341
+JSD	0.518	0.430	0.357	0.436	0.496	0.382
+**HCR**	0.520	0.437	0.369	0.444	0.497	0.393
+***FourierMix***	0.455	0.388	0.326	0.386	0.453	0.348
+JSD	0.522	0.444	0.375	0.454	0.504	0.397
+**HCR**	**0.535**	**0.460**	**0.384**	**0.473**	**0.521**	**0.419**

	CIFAR-100	CIFAR-100-C				CIFAR-100-$\bar{\text{C}}$
Augmentation	ACR	mACR	-Low	-Mid	-High	mACR
Gaussian	0.238	0.169	0.131	0.182	0.208	0.130
+JSD	0.291	0.232	0.167	0.248	0.280	0.196
+AutoAugment+JSD	0.265	0.225	0.175	0.234	0.265	0.176
+AugMix+JSD	0.286	0.231	0.184	0.240	0.269	0.193
+AugMix+**HCR**	0.296	0.249	0.191	0.263	0.292	0.211
+***FourierMix***+JSD	0.295	0.247	0.190	0.258	0.292	0.207
+***FourierMix***+**HCR**	**0.309**	**0.261**	**0.199**	**0.278**	**0.307**	**0.227**

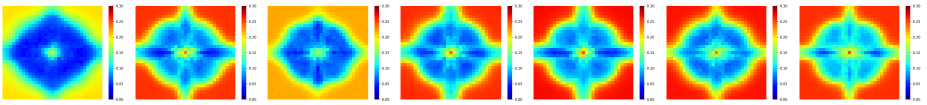
	ImageNet	ImageNet-C				ImageNet-$\bar{\text{C}}$
Augmentation	ACR	mACR	-Low	-Mid	-High	mACR
Gaussian	0.600	0.256	0.155	0.228	0.385	0.266
+JSD	0.736	0.395	0.220	0.382	**0.581**	0.395
+AugMix+JSD	0.717	0.391	0.238	0.387	0.550	0.379
+AugMix+**HCR**	0.727	0.390	0.234	0.383	0.552	0.378
+***FourierMix***+JSD	**0.751**	**0.399**	**0.242**	**0.389**	0.564	**0.413**
+***FourierMix***+**HCR**	0.750	0.397	0.239	0.387	0.567	0.411

tion [25]. We follow Cohen *et al*. [9] and Jeong *et al*. [25] to train the Gaussian and Gaussian+JSD baseline models, respectively. For other augmentation methods, we apply Gaussian noise $\mathcal{N}(0, \sigma^2\mathbf{I})$ to half of the training samples in the mini-batch to ensure good certification performance using RS, and we follow Hendrycks *et al*. to apply JSD to these augmentation methods [23].

Datasets. For the in-distribution evaluation, we use CIFAR-10/100 [31] and ImageNet [12] datasets. CIFAR-10/100 consists of small 32×32 images belonging to 10/100 classes and ImageNet consists of 1.2 million images with 1,000 classes. We crop images in ImageNet into the same size of $224 \times 224 \times 3$ pixels. For the test data, we use the common corruptions datasets [22] (CIFAR-10/100-C and ImageNet-C) and a recently proposed dataset [38] (CIFAR-10/100-$\bar{\text{C}}$ and ImageNet-$\bar{\text{C}}$) which contains human interpretable and perceptually different corruptions as compared to those contained in CIFAR-C/ImageNet-C.

4.1 Results on CIFAR-Based Corruption Benchmarks

The results in Table 1 show the overall mACR of the models trained on CIFAR-10 using different augmentation and regularization methods when evaluated on



(a) Gaussian (b) Gaussian + JSD (c) AutoAug + JSD (d) AugMix + JSD (e) AugMix + **HCR** (f) **F-Mix** + JSD (g) **F-Mix** + **HCR**

Fig. 5. Fourier sensitivity analysis of models trained using different augmentations and regularizers on CIFAR-100 demonstrate their vulnerability to distribution shifts from mid-to-low frequency region (around the center of the plots). (F-Mix: *FourierMix*).

CIFAR-10-C and CIFAR-10-$\bar{\text{C}}$, respectively. The results show that *FourierMix* consistently achieves the highest mACR across different corruption types. *FourierMix*+HCR significantly improves upon the baseline of Gaussian augmented training by 26.7% and 33.4% in terms of the overall mACR on CIFAR-10-C and CIFAR-10-$\bar{\text{C}}$ and also improves upon the stronger baseline, AugMix+HCR, by 5.3% and 6.6% on the two datasets, respectively. We find consistency regularization to be helpful for certified robustness on corruption benchmarks. Especially, adding JSD to Gaussian augmentations significantly improves the robustness on mid- and high-frequency corrupted data. We see that combining HCR with *FourierMix* achieves SOTA ACRs on all corruption types providing significant gains even on low-frequency corruptions. This success is attributed to the spectrally diverse corruptions produced by *FourierMix*. Interestingly, we find AutoAugment overfits to corruptions in CIFAR-10-C since it suffers a major performance degradation on corruptions in CIFAR-10-$\bar{\text{C}}$. We believe the large overlap between the leveraged augmentations and corruptions in CIFAR-10-C and limited spectral diversity are the primary reasons for this performance degradation of AutoAugment. Detailed results for each corruption type in CIFAR-10-C/$\bar{\text{C}}$ are shown in Tables 2 and 3 in Appendix D.1.

Next, we present the mACR (Table 1) of the models trained with CIFAR-100 when evaluated on corrupted data (CIFAR-100-C and CIFAR-100-$\bar{\text{C}}$). Similar to the performance of models trained with CIFAR-10, *FourierMix* achieves the highest overall mACR among all augmentation methods on both corruption datasets. Specifically, *FourierMix*+HCR outperforms the Gaussian baseline by 54.4% and 74.6% on two datasets, respectively. Compared to AugMix+HCR, *FourierMix*+HCR improves the performance by 4.8% and 7.6% on the two datasets, respectively. Detailed results for each corruption type in CIFAR-100-C/$\bar{\text{C}}$ are shown in Tables 4 and 5 in Appendix D.2.

To further corroborate our findings, we carry out the Fourier sensitivity analysis of models trained on CIFAR-100 in Fig. 5. Adding a consistency loss (Gaussian+JSD) improves the ACR of the model in the high-frequency region but is still worse than the ACR achieved by the addition of consistency loss (JSD and HCR) with *FourierMix* augmentations in low-to-mid frequency regions. Similar to our quantitative results, AutoAugment does not improve much over the baseline of Gaussian augmentation which suggests that models trained with AutoAugment may be biased towards high-frequency regions. Heatmaps for CIFAR-10 models report similar findings and are presented in Fig. 7 in Appendix D.1.

4.2 Results on ImageNet-Based Corruption Benchmarks

Table 1 presents the mACR of the models trained on ImageNet when evaluated on ImageNet-C and ImageNet-$\bar{\text{C}}$. We observe that distribution shifts lead to a drastic decline in the certified robustness on ImageNet. The drop between the ACR of clean data and the mACR of corrupted data is ∼57%, whereas it was ∼30% on CIFAR-10/100. Encouragingly, *FourierMix* continues to achieve the highest mACR compared to other baselines. *FourierMix* outperforms the baseline of Gaussian augmented training and AugMix+JSD by 55.9% and 2.1% in terms of the overall mACR, respectively. Detailed results and discussion for ImageNet-C/$\bar{\text{C}}$ can be found in Tables 6 and 7 in Appendix D.3.

Summary. Our results in this section not only highlight the vulnerability of SOTA certified defenses to corrupted data but also uncovers spectral biases in the benchmark datasets that are used to measure corruption robustness. In particular, methods that perform well on one corrupted dataset may not work well on other datasets due to differences in the spectral signatures of the corruptions. This makes it incredibly important to obtain a comprehensive view of the model robustness to avoid issues such as leaderboard bias [39] and model overfitting to a specific benchmark [38]. To help achieve this objective, we propose a new benchmark that has a collection of spectrally diverse corruption datasets.

5 A Spectral Corruption Benchmarking Suite

Although corruptions proposed by [22] can be roughly grouped into different frequency ranges, their spectral diversity is restricted (see Figs. 10 and 11 in Appendix C). This could lead to corruption overfitting for methods that make models robust only on a limited subset of corruption types but fail on others (*e.g.*, Gaussian on ImageNet-C in Table 1). As the nature of test-time corruptions is unknown at train-time, and their form is application-dependent, models must be evaluated under diverse corruption settings. To achieve this, next we discuss the creation and evaluation of models on the proposed corruption benchmarking suite. The goal of this new suite is to complement the existing benchmark datasets and enable researchers to uncover the spectral biases of their models.

Protocol for Dataset Generation. The proposed benchmark is a collection of datasets each focusing on a specific frequency range while collectively covering the entire frequency spectrum. Different from the Fourier sensitivity analysis that only perturbs a single frequency using the Fourier basis, CIFAR-10/100-F leverages power law-based noise [1] to generate complex perturbations in the spectral domain [26]. Note that the power spectrum of several natural data distributions (e.g., natural images) follow power-law distribution [1]. Inspired by this, we model the amplitude perturbation as $\boldsymbol{\delta}_{\text{Fourier}}(f)_{\mathbf{A}} = \frac{P(f)}{(|f-f_c|+1)^{\alpha}} \cdot \text{U}(1-b, 1+b)$, where $P(f)$ approximates the tolerance of corruptions at azimuthal frequency $f = \sqrt{u^2+v^2}$, f_c is the central frequency that the perturbation focuses on, and α denotes the power of the power law distribution. We also use a uniform distribution $\text{U}(1-b, 1+b)$ with b as a hyper-parameter ($b = 0.2$ in our study) to diversify the perturbations.

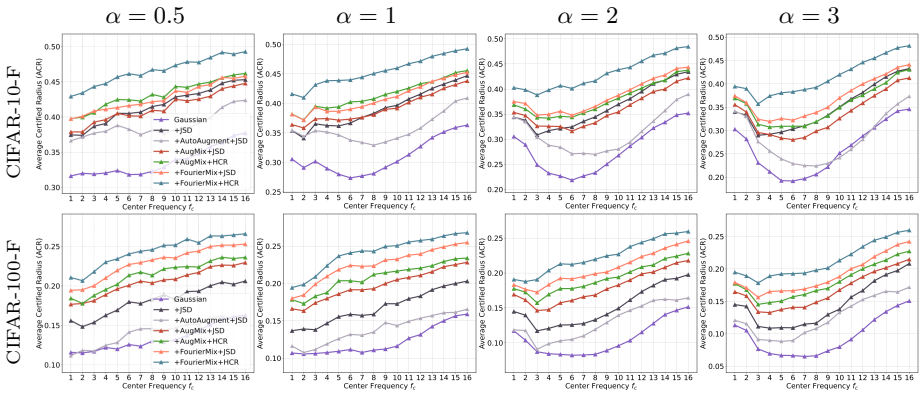


Fig. 6. ACRs on the proposed CIFAR-F dataset averaged over 3 severity levels show that *FourierMix* based models perform consistently better than other baselines across entire spectrum. Increasing α (from left to right), decreases the spread of the frequencies. Dips of ACRs in mid-frequency regions (*e.g.*, $\alpha = 2, 3$) demonstrate the vulnerability of models to low-to-mid frequency corruptions.

We define $P(f) = \text{clip}(\boldsymbol{A}_{\boldsymbol{x}}^{\text{clean}}(f), a_{\text{lower}}, a_{\text{upper}})$ which adds the amount of perturbation based on the power associated with the different frequencies in the clean image [27], *i.e.*, frequencies with higher power have larger perturbations.We leverage the clip$(\cdot)$ function to bound the amount of corruption in each spatial frequency. The maximum and minimum values are chosen to ensure that perturbations do not affect the semantic content of the images. The phase perturbation is formulated as $\boldsymbol{\delta}_{\text{Fourier}}(f)_{\mathbf{P}} = \text{U}(0, 2\pi)$. Given each pair $(\boldsymbol{x}^i, y^i)$ in the original validation set, we synthesize CIFAR-10/100-F images as

$$\boldsymbol{x}_F^i = \boldsymbol{x}^i + \gamma \cdot \text{IFFT}(\boldsymbol{\delta}_{\text{Fourier}}) \tag{5}$$

where $\gamma = \frac{\epsilon}{||\text{IFFT}(\boldsymbol{\delta}_{\text{Fourier}})||_2}$ normalizes the spreading effect of the power-law distribution and, thus, controls the severity level of CIFAR-10/100-F. We create both CIFAR-10/100-F with 3 severity levels with $\epsilon \in \{8, 10, 12\}$. As the images in CIFAR-10/100 are of size 32×32, their FFT spectrum has discrete azimuthal frequencies from 0 to 16. Since zero-frequency noise is a constant in the pixel space, we set the center frequency $f_c \in \{1, 2, ..., 16\}$. We leverage $\alpha \in \{0.5, 1, 2, 3\}$ because power-law noises with $0 < \alpha \leq 3$ arise in both natural signals and in man-made processes [1]. In total, our CIFAR-10/100-F datasets are consisted of $3 \times 4 \times 16 = 192$ test sets from different regions of the frequency spectrum thereby increasing the spectral coverage of the original dataset.

Visual Effect of Varying α and f_c. As shown in Fig. 12 in Appendix F, α controls the frequency dispersion of the corruption at f_c. With a smaller α, *e.g.*, $\alpha = 0.5$, the spreading effect of the power law distribution is more significant. The corrupted images thus contain noises across all azimuthal frequencies. In contrast, for larger α, the corruptions will be focused more on a single frequency *e.g.*, $\alpha = 3$, and higher f_c leads to a higher corruption frequency.

Results on CIFAR-10/100-F. Fig. 6 reports the performance of models used in § 4.1 on CIFAR-10/100-F benchmark. Our results show that both AutoAugment [11] and AugMix [23] based smoothed models are relatively biased toward low-frequency corruptions. The effect of high-frequency corruptions is more pronounced on models trained with AutoAugment which behave similarly to the simple baseline of Gaussian augmentation (Fig. 6). The intersection of the curves of AugMix+JSD and Gaussian+JSD in the mid-frequency region in CIFAR-10-F (Fig. 6), illustrates the different spectral biases introduced by different augmentation methods. Unlike CIFAR-10-F, we find that Gaussian and Gaussian+JSD perform relatively worse on CIFAR-100-F compared to other augmentation methods. In comparison to other methods, we find that models trained with *FourierMix* and HCR do not show significant spectral biases and serve as a strong baseline. Specifically, models trained with *FourierMix*+HCR, on average, outperform AugMix+HCR, by 11.8% and 16.0% on CIFAR-10/100-F, respectively. We emphasize that models trained with *FourierMix* do not overfit to CIFAR-10/100-F datasets since they have different formulations and even visual patterns (see Appendix E and F).

6 Discussion and Conclusion

Our work has shown that certified defenses are surprisingly brittle to distribution shifts such as low-frequency corruptions. To alleviate this issue, we proposed *FourierMix* augmentation to increase the spectral coverage of the training data. We also presented a benchmarking suite to understand the model's corruption robustness comprehensively. Some of our findings are consistent with past results that model evaluation under corruption is a challenging problem, and one should not rely on a single benchmark [20, 38, 43]. However, as opposed to the existing works that focus on *empirical* robustness, we show that these issues persist and may even be more prominent in the problem of certified adversarial defense. Even though evaluation against all possible types of corruptions is infeasible, our results highlighted that eliminating spectral biases of the models improves the certified robustness under common corruptions.

Although we have taken some first steps to address this challenging problem, many questions remain to be answered. First, bridging the gap between robustness guarantees in high-frequency and low-frequency corruption regimes is still an open problem. A deeper theoretical understanding of this phenomenon will likely motivate systematic approaches to overcome this issue. Finally, the analysis done in this work can be explored in the context of certifying other ℓ_p norms [63], spectral deformations [2], and semantic transformations [35].

Acknowledgements. This work was performed under the auspices of the U.S. Department of Energy by the Lawrence Livermore National Laboratory under Contract No. DE-AC52-07NA27344 and LLNL LDRD Program Project No. 20-ER-014. This work was partially supported by NSF under the National AI Institute for Edge Computing Leveraging Next Generation Wireless Networks, Grant # 2112562, in addition to NSF grants CMMI-2038215 and CNS-1930041.

References

1. 1/f noise. http://www.scholarpedia.org/article/1/f_noise (2021)
2. Alfarra, M., Bibi, A., Khan, N., Torr, P.H., Ghanem, B.: Deformrs: Certifying input deformations with randomized smoothing. Proceedings of the AAAI Conference on Artificial Intelligence **36**(6), 6001–6009 (Jun 2022). https://doi.org/10.1609/aaai.v36i6.20546, https://ojs.aaai.org/index.php/AAAI/article/view/20546
3. Athalye, A., Carlini, N., Wagner, D.: Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In: International conference on machine learning. pp. 274–283. PMLR (2018)
4. Bulusu, S., Kailkhura, B., Li, B., Varshney, P.K., Song, D.: Anomalous example detection in deep learning: A survey. IEEE Access **8**, 132330–132347 (2020)
5. Burton, G.J., Moorhead, I.R.: Color and spatial structure in natural scenes. Applied optics **26**(1), 157–170 (1987)
6. Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In: 2017 IEEE Symposium on Security and Privacy (SP). pp. 39–57 (2017). https://doi.org/10.1109/SP.2017.49
7. Chen, P.Y., Sharma, Y., Zhang, H., Yi, J., Hsieh, C.J.: EAD: elastic-net attacks to deep neural networks via adversarial examples. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 10–17 (2018)
8. Chen, P.Y., Zhang, H., Sharma, Y., Yi, J., Hsieh, C.J.: ZOO: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In: ACM Workshop on Artificial Intelligence and Security. pp. 15–26 (2017)
9. Cohen, J., Rosenfeld, E., Kolter, Z.: Certified adversarial robustness via randomized smoothing. In: International Conference on Machine Learning. pp. 1310–1320. PMLR (2019)
10. Croce, F., Andriushchenko, M., Sehwag, V., Debenedetti, E., Flammarion, N., Chiang, M., Mittal, P., Hein, M.: Robustbench: a standardized adversarial robustness benchmark. arXiv preprint arXiv:2010.09670 (2020)
11. Cubuk, E.D., Zoph, B., Mane, D., Vasudevan, V., Le, Q.V.: Autoaugment: Learning augmentation strategies from data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 113–123 (2019)
12. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
13. Dodge, S., Karam, L.: Understanding how image quality affects deep neural networks. In: 2016 eighth international conference on quality of multimedia experience (QoMEX). pp. 1–6. IEEE (2016)
14. Fischer, M., Baader, M., Vechev, M.: Certified defense to image transformations via randomized smoothing. arXiv preprint arXiv:2002.12463 (2020)
15. Fuglede, B., Topsoe, F.: Jensen-shannon divergence and hilbert space embedding. In: International Symposium onInformation Theory, 2004. ISIT 2004. Proceedings. p. 31. IEEE (2004)
16. Gokhale, T., Anirudh, R., Kailkhura, B., Thiagarajan, J.J., Baral, C., Yang, Y.: Attribute-guided adversarial training for robustness to natural perturbations. arXiv preprint arXiv:2012.01806 (2020)
17. Gowal, S., Dvijotham, K., Stanforth, R., Bunel, R., Qin, C., Uesato, J., Arandjelovic, R., Mann, T., Kohli, P.: On the effectiveness of interval bound propagation for training verifiably robust models. arXiv preprint arXiv:1810.12715 (2018)

18. Hammersley, J.: Monte carlo methods. Springer Science & Business Media (2013)
19. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
20. Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., 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. pp. 8340–8349 (2021)
21. Hendrycks, D., Carlini, N., Schulman, J., Steinhardt, J.: Unsolved problems in ml safety. ArXiv **abs/2109.13916** (2021)
22. Hendrycks, D., Dietterich, T.: Benchmarking neural network robustness to common corruptions and perturbations. arXiv preprint arXiv:1903.12261 (2019)
23. Hendrycks, D., Mu, N., Cubuk, E.D., Zoph, B., Gilmer, J., Lakshminarayanan, B.: Augmix: A simple data processing method to improve robustness and uncertainty. arXiv preprint arXiv:1912.02781 (2019)
24. Ilyas, A., Engstrom, L., Athalye, A., Lin, J.: Black-box adversarial attacks with limited queries and information. In: International Conference on Machine Learning. pp. 2137–2146. PMLR (2018)
25. Jeong, J., Shin, J.: Consistency regularization for certified robustness of smoothed classifiers. arXiv preprint arXiv:2006.04062 (2020)
26. Johnson, J.B.: The schottky effect in low frequency circuits. Physical review **26**(1), 71 (1925)
27. Joubert, O.R., Rousselet, G.A., Fabre-Thorpe, M., Fize, D.: Rapid visual categorization of natural scene contexts with equalized amplitude spectrum and increasing phase noise. Journal of Vision **9**(1), 2–2 (2009)
28. Joyce, J.M.: Kullback-Leibler Divergence, pp. 720–722. Springer Berlin Heidelberg, Berlin, Heidelberg (2011)
29. Kermisch, D.: Image reconstruction from phase information only. JOSA **60**(1), 15–17 (1970)
30. Kireev, K., Andriushchenko, M., Flammarion, N.: On the effectiveness of adversarial training against common corruptions. arXiv preprint arXiv:2103.02325 (2021)
31. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
32. Lecuyer, M., Atlidakis, V., Geambasu, R., Hsu, D., Jana, S.: Certified robustness to adversarial examples with differential privacy. In: 2019 IEEE Symposium on Security and Privacy (SP). pp. 656–672. IEEE (2019)
33. Li, B., Chen, C., Wang, W., Carin, L.: Second-order adversarial attack and certifiable robustness (2018)
34. Li, L., Qi, X., Xie, T., Li, B.: Sok: Certified robustness for deep neural networks. arXiv **abs/2009.04131** (2020)
35. Li, L., Weber, M., Xu, X., Rimanic, L., Kailkhura, B., Xie, T., Zhang, C., Li, B.: Tss: Transformation-specific smoothing for robustness certification. In: ACM CCS (2021)
36. Lim, J.S.: Two-dimensional signal and image processing. Englewood Cliffs (1990)
37. Mehra, A., Kailkhura, B., Chen, P.Y., Hamm, J.: How robust are randomized smoothing based defenses to data poisoning? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13244–13253 (2021)
38. Mintun, E., Kirillov, A., Xie, S.: On interaction between augmentations and corruptions in natural corruption robustness. arXiv preprint arXiv:2102.11273 (2021)

39. Mishra, S., Arunkumar, A.: How robust are model rankings: A leaderboard customization approach for equitable evaluation. arXiv preprint arXiv:2106.05532 (2021)
40. Mohapatra, J., Ko, C.Y., Weng, T.W., Chen, P.Y., Liu, S., Daniel, L.: Higher-order certification for randomized smoothing. Neural Information Processing Systems (2020)
41. Mohapatra, J., Weng, T.W., Chen, P.Y., Liu, S., Daniel, L.: Towards verifying robustness of neural networks against a family of semantic perturbations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 244–252 (2020)
42. Raghunathan, A., Steinhardt, J., Liang, P.: Certified defenses against adversarial examples. arXiv preprint arXiv:1801.09344 (2018)
43. Raji, I.D., Bender, E.M., Paullada, A., Denton, E., Hanna, A.: Ai and the everything in the whole wide world benchmark. arXiv preprint arXiv:2111.15366 (2021)
44. Rebuffi, S.A., Gowal, S., Calian, D.A., Stimberg, F., Wiles, O., Mann, T.: Fixing data augmentation to improve adversarial robustness. arXiv preprint arXiv:2103.01946 (2021)
45. Ruder, S.: An overview of gradient descent optimization algorithms. arXiv preprint arXiv:1609.04747 (2016)
46. Saenko, K., Kulis, B., Fritz, M., Darrell, T.: Adapting visual category models to new domains. In: European conference on computer vision. pp. 213–226. Springer (2010)
47. Salman, H., Yang, G., Li, J., Zhang, P., Zhang, H., Razenshteyn, I., Bubeck, S.: Provably robust deep learning via adversarially trained smoothed classifiers. arXiv preprint arXiv:1906.04584 (2019)
48. Schneider, S., Rusak, E., Eck, L., Bringmann, O., Brendel, W., Bethge, M.: Improving robustness against common corruptions by covariate shift adaptation. Advances in Neural Information Processing Systems **33** (2020)
49. Sun, J., Cao, Y., Chen, Q.A., Mao, Z.M.: Towards robust LiDAR-based perception in autonomous driving: General black-box adversarial sensor attack and countermeasures. In: 29th USENIX Security Symposium (USENIX Security 20). pp. 877–894. USENIX Association (Aug 2020), `https://www.usenix.org/conference/usenixsecurity20/presentation/sun`
50. Sun, J., Cao, Y., Choy, C.B., Yu, Z., Anandkumar, A., Mao, Z.M., Xiao, C.: Adversarially robust 3d point cloud recognition using self-supervisions. Advances in Neural Information Processing Systems **34**, 15498–15512 (2021)
51. Sun, J., Koenig, K., Cao, Y., Chen, Q.A., Mao, Z.M.: On adversarial robustness of 3d point cloud classification under adaptive attacks. arXiv preprint arXiv:2011.11922 (2020)
52. Sun, J., Zhang, Q., Kailkhura, B., Yu, Z., Xiao, C., Mao, Z.M.: Benchmarking robustness of 3d point cloud recognition against common corruptions. arXiv preprint arXiv:2201.12296 (2022)
53. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199 (2013)
54. Tolhurst, D., Tadmor, Y., Chao, T.: Amplitude spectra of natural images. Ophthalmic and Physiological Optics **12**(2), 229–232 (1992)
55. Tramer, F., Carlini, N., Brendel, W., Madry, A.: On adaptive attacks to adversarial example defenses. arXiv preprint arXiv:2002.08347 (2020)

56. Wang, D., Shelhamer, E., Liu, S., Olshausen, B., Darrell, T.: Tent: Fully test-time adaptation by entropy minimization. In: International Conference on Learning Representations (2021), https://openreview.net/forum?id=uXl3bZLkr3c
57. Wong, E., Kolter, Z.: Provable defenses against adversarial examples via the convex outer adversarial polytope. In: International Conference on Machine Learning. pp. 5286–5295. PMLR (2018)
58. Xiao, C., Li, B., Zhu, J.Y., He, W., Liu, M., Song, D.: Generating adversarial examples with adversarial networks. arXiv preprint arXiv:1801.02610 (2018)
59. Xu, K., Shi, Z., Zhang, H., Wang, Y., Chang, K.W., Huang, M., Kailkhura, B., Lin, X., Hsieh, C.J.: Provable, scalable and automatic perturbation analysis on general computational graphs. arXiv e-prints pp. arXiv–2002 (2020)
60. Xu, K., Wang, C., Cheng, H., Kailkhura, B., Lin, X., Goldhahn, R.: Mixture of robust experts (more): A robust denoising method towards multiple perturbations. arXiv preprint arXiv:2104.10586 (2021)
61. Xu, Q., Zhang, R., Zhang, Y., Wang, Y., Tian, Q.: A fourier-based framework for domain generalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14383–14392 (2021)
62. Xu, Q., Zhang, R., Zhang, Y., Wang, Y., Tian, Q.: A fourier-based framework for domain generalization. In: IEEE/CVF CVPR. pp. 14383–14392 (June 2021)
63. Yang, G., Duan, T., Hu, J.E., Salman, H., Razenshteyn, I., Li, J.: Randomized smoothing of all shapes and sizes. In: International Conference on Machine Learning. pp. 10693–10705. PMLR (2020)
64. Yang, Y., Lao, D., Sundaramoorthi, G., Soatto, S.: Phase consistent ecological domain adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9011–9020 (2020)
65. Yin, D., Lopes, R.G., Shlens, J., Cubuk, E.D., Gilmer, J.: A fourier perspective on model robustness in computer vision. arXiv preprint arXiv:1906.08988 (2019)
66. Zhai, R., Dan, C., He, D., Zhang, H., Gong, B., Ravikumar, P., Hsieh, C.J., Wang, L.: Macer: Attack-free and scalable robust training via maximizing certified radius. In: International Conference on Learning Representations (2020), https://openreview.net/forum?id=rJx1Na4Fwr
67. Zhang, H., Weng, T.W., Chen, P.Y., Hsieh, C.J., Daniel, L.: Efficient neural network robustness certification with general activation functions. In: Advances in Neural Information Processing Systems. pp. 4944–4953 (2018)