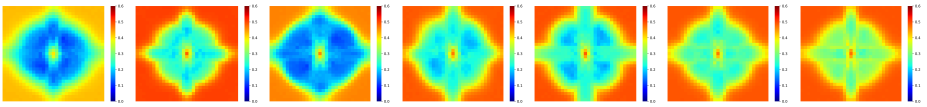
Appendix

A Related Work

Deep neural networks (DNNs) trained using standard gradient descent optimizers [45] have been shown vulnerable to adversarial examples [53]. A number of white- and black-box attacks have been proposed [6–8, 24, 49–51, 58] to construct adversarial examples with small ℓ_p distances to the original data that mislead these DNN models. Besides adversarial attacks, recent studies have devoted efforts to characterizing model performance under common corruptions [4, 22], where natural corruptions lead to a significant impact on the accuracy of SOTA ML models. Thus, it has become imperative to study how ML models can be made robust to test data coming from different distributions when the models are deployed in the real world.

Certified Robustness and Defenses. The authors in [53] have discovered the adversarial examples in DNN models, after which many defenses have been presented to mitigate such vulnerability [3]. However, many of the proposed countermeasures have been shown to rely on gradient obfuscation, limiting malicious agents from accessing the accurate gradients. Such defenses are vulnerable to adaptive attacks, which give a false sense of security [3] of the models. Certified defenses are thus highly desirable. Along with a prediction on the test point, these defenses output a certified radius r such that for any $||\boldsymbol{\delta}||_2 < r$, the model continues to have the same prediction. Such techniques include convex polytope [57], recursive propagation [17], and linear relaxation [42, 67]. These methods provide a lower bound on the perturbation required to change the model's prediction on a target point. However, such methods can merely be applied to shallow models, which limits their practicality. Recently, [9, 32, 33, 40] have proposed randomized smoothing (RS)-based certified defenses that produce better lower bounds and are scalable to large networks. In this paper, we study the corruption robustness of such certified defenses. Unlike a recent work [37], which uses data poisoning attacks to hurt the robustness guarantees of the RS-based models, our work demonstrates the failure of these models on test-time corruptions, which might be encountered by the model deployed in the real world.

Robustness against Common Corruptions – Benchmarks and Defenses. Pioneering studies have identified vulnerabilities of deep learning models to common corruptions. Dodge *et al.* find that standard trained DNNs are vulnerable to blur and Gaussian noise [13]. Hendrycks *et al.* [22] present CIFAR-10/100-C and ImageNet-C, consisting of fifteen different common corruptions with five severity levels to facilitate robustness evaluations of CIFAR [31] and ImageNet [12] models. Sun *et al.* [52] present common corruption benchmarks for 3D point cloud data. Recently, Mintum *et al.* further propose CIFAR-10/100-$\bar{\text{C}}$ and ImageNet-$\bar{\text{C}}$ to provide new corruptions [38]. There are two popular lines of work on improving the robustness against common corruptions: *test-time adaptation* [48] and *data augmentation* [11, 23]. The authors in [46] propose a method to update the batch normalization (BN) statistics for improving domain adaptation. Another



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

Fig. 7. Average Certified Radius (ACR) of Fourier Basis Analysis on CIFAR-100 with $\epsilon = 4$ (AutoAug: AutoAugment, F-Mix: *FourierMix*).

Table 2. Average Certified Radius (ACR) of Models Trained with Different Methods on CIFAR-10-C. Models trained with *FourierMix* and HCR achieve significant improvements in the certified robustness (ACR) guarantees on all corruption types from the CIFAR-10-C dataset.

Augmentation	CIFAR-10	mACR	-low	-mid	-high	Gauss.	Shot	Impulse	Defocus	Glass	Motion	Zoom	Snow	Frost	Fog	Bright	Contrast	Elastic	Pixel	JPEG
Gaussian	0.461	0.363	0.301	0.353	0.435	0.448	0.448	0.421	0.380	0.346	0.338	0.357	0.394	0.347	0.187	0.439	0.137	0.342	0.420	0.440
+JSD	**0.535**	0.439	0.346	0.451	0.520	**0.529**	0.514	**0.528**	0.471	0.445	0.443	0.453	0.449	0.378	0.235	0.485	0.185	0.444	0.506	0.521
+AutoAugment	0.411	0.372	0.312	0.364	0.431	0.451	0.452	0.419	0.411	0.356	0.342	0.360	0.403	0.354	0.201	0.446	0.158	0.352	0.429	0.445
+JSD	0.432	0.400	0.343	0.395	0.464	0.473	0.476	0.443	0.423	0.385	0.394	0.390	0.427	0.403	0.212	0.483	0.189	0.382	0.453	0.473
+AugMix	0.452	0.385	0.324	0.383	0.449	0.459	0.460	0.436	0.412	0.369	0.372	0.391	0.413	0.374	0.216	0.457	0.159	0.371	0.439	0.453
+JSD	0.518	0.430	0.357	0.436	0.496	0.504	0.507	0.481	0.461	0.426	0.429	0.441	0.452	0.408	0.240	0.501	0.185	0.425	0.485	0.502
+**HCR**	0.520	0.437	0.369	0.444	0.497	0.505	0.506	0.484	0.464	0.438	0.435	0.447	0.460	0.426	0.252	0.505	0.200	0.437	0.487	0.501
+***FourierMix***	0.455	0.388	0.326	0.386	0.453	0.461	0.462	0.446	0.417	0.369	0.378	0.393	0.415	0.376	0.220	0.457	0.160	0.373	0.439	0.456
+JSD	0.522	0.444	0.375	0.454	0.504	0.512	0.513	0.491	0.474	0.448	0.446	0.456	0.464	0.432	0.257	**0.519**	0.201	0.445	0.495	0.508
+**HCR**	**0.535**	**0.460**	**0.384**	**0.473**	**0.521**	0.528	**0.530**	0.513	**0.492**	**0.470**	**0.464**	**0.477**	**0.477**	**0.432**	**0.275**	0.517	**0.220**	**0.462**	**0.511**	**0.524**

recent method, TENT [56] updates both the affine transformation and statistics of BN by using self-entropy minimization. On the other hand, methods such as AutoAugment [11] leverages reinforcement learning to learn an augmentation policy that produces a diverse set of augmentations to help make the models robust to corrupted data. Another popular method, AugMix [23] achieves impressive performance improvement on corrupted data using augmentations generated by mixing up images obtained from applying randomly sampled operations along with using a Jenson-Shannon-based consistency loss during training. The authors in [16, 60] leveraged adversarial training schemes to improve the corruption robustness. Unlike existing data augmentation schemes which intend to improve the empirical robust accuracy of the models, the data augmentation schemes of interest to this paper aim to improve the adversarial robustness guarantees under common corruptions.

Certified Semantic Robustness. Recent work [14, 35, 41] have also focused on developing techniques to provide performance guarantees to seen (or known) common corruption types (such as rotation or brightness changes). However, in this work, we are interested in more realistic scenarios with unseen (or unknown) test-time corruptions. It is worth noting that the susceptibility analysis and defense techniques developed in this work can be extended to SOTA semantic robustness techniques.

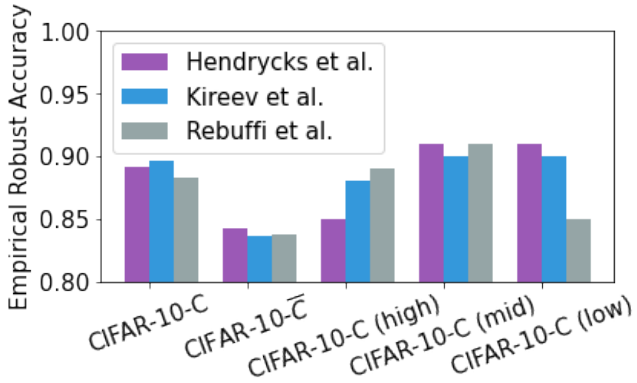


Fig. 8. The ranking of SOTA models [23, 30, 44] (based on empirical robust accuracy) changes across datasets and corruption types, suggesting there is no single model which performs the best on different corruption benchmarks.

Table 3. Average Certified Radius (ACR) of Models Trained with Different Methods on CIFAR-10-C̄. Models trained with *FourierMix* and HCR achieve significant improvements in the certified robustness (ACR) guarantees on corruptions from the CIFAR-10-C̄ dataset.

Augmentation	mACR	Blue	Brown	Checkerboard	Circular	Inv. Sparkle	Lines	Pinch	Ripple	Sparkles	Trans. Chromatic
Gaussian	0.314	0.351	0.255	0.310	0.386	0.222	0.336	0.398	0.365	0.251	0.269
+JSD	0.393	0.458	0.303	0.395	0.452	0.252	0.430	0.492	0.463	0.306	0.376
+AutoAugment	0.304	0.351	0.263	0.312	0.395	0.223	0.348	0.406	0.248	0.235	0.256
+JSD	0.346	0.354	0.297	0.335	0.445	0.238	0.374	0.436	0.402	0.269	0.308
+AugMix	0.341	0.389	0.269	0.334	0.439	0.233	0.358	0.416	0.397	0.272	0.307
+JSD	0.382	0.429	0.303	0.372	0.483	0.255	0.404	0.467	0.450	0.306	0.350
+HCR	0.393	0.442	0.309	0.384	0.486	0.268	0.419	0.471	0.464	0.320	0.368
+FourierMix	0.348	0.391	0.269	0.331	0.441	0.237	0.368	0.432	0.401	0.280	0.325
+JSD	0.397	0.445	0.307	0.395	0.482	0.265	0.430	0.490	0.463	0.320	0.377
+HCR	**0.419**	**0.474**	**0.317**	**0.418**	**0.504**	**0.289**	**0.459**	**0.501**	**0.486**	**0.339**	**0.406**

B Empirical Robust Accuracy of SOTA Models on Corrupted Data

The results in Fig. 8 show empirical robust accuracy of state-of-the-art models on existing corruption benchmarks. We use the recently proposed RobustBench [10] benchmark and selected the top-performing models on CIFAR-10-C for this experiment [23, 30, 44]. As evident from the figure, the performance of the models varies across datasets and corruption types showing that a single model is not able to achieve the best performance on all types of corruptions. Evaluating the models on a single benchmark is not enough to obtain the true picture of the corruption robustness of a model. Thus to eliminate the biases present in corruption benchmarks, one should gauge the corruption robustness of a model by evaluating it on a variety of datasets. Our proposed CIFAR-10/100-F benchmark can be used by designers to probe the spectral biases of the models.

Table 4. Average Certified Radius (ACR) of Models Trained with Different Methods on CIFAR-100-C. Models trained with *FourierMix* and HCR achieve significant improvements in the certified robustness (ACR) guarantees on corruptions from the CIFAR-100-C dataset.

Augmentation	CIFAR-100	mACR	-low	-mid	-high	Gauss.	Shot	Impulse	Defocus	Glass	Motion	Zoom	Snow	Frost	Fog	Bright	Contrast	Elastic	Pixel	JPEG
Gaussian	0.238	0.169	0.131	0.182	0.208	0.214	0.218	0.193	0.181	0.170	0.157	0.169	0.177	0.153	0.069	0.207	0.051	0.159	0.206	0.209
+JSD	0.291	0.232	0.167	0.248	0.280	0.283	0.285	0.273	0.261	0.252	0.240	0.250	0.226	0.188	0.104	0.242	0.079	0.235	0.278	0.281
+AutoAugment + JSD	0.265	0.225	0.175	0.234	0.265	0.275	0.273	0.252	0.248	0.230	0.230	0.238	0.232	0.202	0.104	0.257	0.082	0.225	0.261	0.266
+AugMix + JSD	0.286	0.231	0.184	0.240	0.269	0.274	0.278	0.256	0.255	0.236	0.233	0.243	0.239	0.211	0.111	0.267	0.092	0.232	0.267	0.270
+AugMix + **HCR**	0.296	0.249	0.191	0.263	0.292	0.296	0.301	0.282	0.278	0.264	0.255	0.263	0.249	0.215	0.118	0.274	0.097	0.253	0.291	0.292
+*FourierMix* + JSD	0.295	0.247	0.190	0.258	0.292	0.295	0.300	0.283	0.273	0.257	0.249	0.260	0.251	0.217	0.115	0.275	0.092	0.250	0.288	0.292
+*FourierMix* + **HCR**	**0.309**	**0.261**	**0.199**	**0.278**	**0.307**	**0.310**	**0.313**	**0.302**	**0.291**	**0.283**	**0.270**	**0.277**	**0.260**	**0.221**	**0.128**	**0.284**	**0.102**	**0.267**	**0.303**	**0.307**

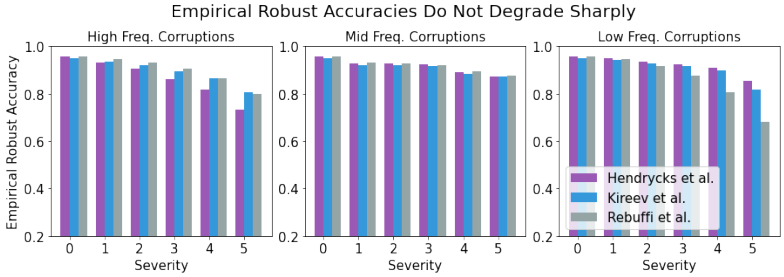


Fig. 9. The performance gaps (*i.e.*, the robust accuracy) are remained small/reasonable in state-of-the-art empirically robust models [23, 30, 44]. Severity 0 denotes the in-distribution data.

C Amplitude Spectrum of CIFAR-10-C/$\overline{\text{C}}$

As introduced in § 2, we arrange the amplitude specturm of corruptions from CIFAR-10-C into three groups, roughly categorized as high/mid/low-frequency corruptions. Specifically, we compute the the $\mathbb{E}[\text{FFT}(\boldsymbol{x})]$ and $\mathbb{E}[\text{FFT}(C(\boldsymbol{x}) = \boldsymbol{x})]$ by averaging over all the validation images [65] for CIFAR-10 and each corruption in CIFAR-10-C, respectively, where $C(\cdot)$ denotes the corruption function. As Fig. 10 shows, CIFAR-10 (clean) images follow a distribution of $\frac{1}{f^\alpha}$, where $f = \sqrt{u^2 + v^2}$ is the azimuthal frequency and $\alpha \approx 2$. Therefore, clean images have extremely low power in the high-frequency regions (the edges and corner). Due to this, all the noise perturbations corresponding to `JPEG` and `pixelate` can be considered as high-frequency corruptions, relative to the clean images' distribution. On the other hand, weather-related and `contrast` corruptions are all centered in the low-frequency region. We categorize remaining perturbations as mid-frequency corruptions.

We also visualize the amplitude spectrum of corruptions from CIFAR-10-$\bar{\text{C}}$ in Fig. 11. We find that most of the corruptions from CIFAR-10-$\bar{\text{C}}$ are centered in the low/mid-frequency ranges, explaining why *FourierMix* achieves lager improvements on CIFAR-10-$\bar{\text{C}}$ than CIFAR-10-C compared to spectrally-biased baselines.

Table 5. Average Certified Radius (ACR) of Models Trained with Different Methods on CIFAR-100-$\bar{\text{C}}$. Models trained with *FourierMix* and HCR achieve significant improvements in the certified robustness (ACR) guarantees on corruptions from the CIFAR-100-$\bar{\text{C}}$ dataset.

Augmentation	mACR	Blue	Brown	Checkerboard	Circular	Inv. Sparkle	Lines	Pinch	Ripple	Sparkles	Trans. Chromatic
Gaussian	0.130	0.151	0.070	0.114	0.159	0.097	0.137	0.199	0.160	0.097	0.116
+JSD	0.196	0.228	0.106	0.186	0.233	0.124	0.221	0.274	0.242	0.151	0.193
+AutoAugment + JSD	0.176	0.211	0.087	0.152	0.229	0.119	0.184	0.236	0.217	0.140	0.185
+AugMix + JSD	0.193	0.227	0.107	0.176	0.259	0.131	0.206	0.260	0.244	0.153	0.191
+AugMix + **HCR**	0.211	0.253	0.120	0.199	0.276	0.136	0.224	0.283	0.263	0.156	0.203
+***FourierMix*** + JSD	0.207	0.243	0.106	0.194	0.262	0.136	0.226	0.281	0.258	0.154	0.205
+***FourierMix*** + **HCR**	**0.227**	**0.260**	**0.129**	**0.219**	**0.281**	**0.151**	**0.247**	**0.300**	**0.278**	**0.172**	**0.228**

D Training and Evaluation Details

Training. We train CIFAR-10/100 and ImageNet models for 200 and 90 epochs for all methods with an SGD optimizer, respectively [45]. We exclude the input normalization layer as it will degrade the certification performance on corrupted data. We use different σ to train CIFAR-10/100 and ImageNet models, as specified in § 4.

Evaluation. Recall from the theorem derived in § 2 of Cohen *et al.*, CR($\cdot$) approaches ∞ when $\underline{p_A}$ approaches the value 1 [9]. However, this will also require the Gaussian perturbed samples $n \approx \infty$. Consider that the base classifier $\mathcal{M}(\boldsymbol{x}+\boldsymbol{\delta})$ has observed n samples that all equal to c_A, $p_A \geq \alpha^{(1/n)}$ has a probability $1-\alpha$ [9]. To both constrain the computational complexity and achieve a tight bound, we use $n = 100,000$, $n_0 = 100$, and $\alpha = 0.001$ as the hyper-parameters to get high confidence of the computed radius, following prior arts [9, 25, 47, 66]. Since we need to evaluate corruption datasets with 125$\times$ larger sizes than the original test sets, we certify 500 and 350 examples from each corruption and each severity level of the CIFAR-10/100 and ImageNet corruption datasets (*i.e.*, -C/$\bar{\text{C}}$). For the Fourier sensitivity analysis of CIFAR-10/100, each data point in the heat map is the corresponding ACR of 200 examples.

D.1 Detailed Results on CIFAR-10-Based Corruption Benchmarks

In this section, we present detailed results for our evaluation on CIFAR-10-C/$\bar{\text{C}}$. We fix $\eta = 10$ and use $\lambda = 40$ for HCR (Equation 4) in our experiments on CIFAR-10. Tables 2 and 3 present the ACR on individual corruption types from CIFAR-10-C/$\bar{\text{C}}$, respectively. *FourierMix* consistently achieves the highest ACR on most of the corruption types in both corruption datasets. Especially, we find *FourierMix* helps achieve larger improvements on weather-related corruptions, which have real-world implications (*e.g.*, safety of autonomous driving). We also perform Fourier sensitivity analysis to confirm our findings. Fig. 5 shows the heat maps, which also corroborate our insights in § 4.1.

We opted for RS-based certification due to its scalability to large datasets and models. Our findings and claims, however, are general. To show this, we choose the next best baseline using improved CROWN-IBP [59]. Unfortunately, this method cannot scale to ImageNet due to the large image size. Even on CIFAR-10, it provides trivially loose bounds, *i.e.*, ACR $\approx$ 0, for ResNet-110

Table 6. Average Certified Radius (ACR) of Models Trained with Different Methods on ImageNet-C. Spectrally diverse augmentations from *FourierMix* brings significant gains to certified robustness of the models trained on ImageNet against corruptions from ImageNet-C.

Augmentation	ImageNet	mACR	-low	-mid	-high	Gauss.	Shot	Impulse	Defocus	Glass	Motion	Zoom	Snow	Frost	Fog	Bright	Contrast	Elastic	Pixel	JPEG
Gaussian	0.600	0.256	0.155	0.228	0.385	0.342	0.324	0.310	0.174	0.227	0.212	0.201	0.148	0.170	0.013	0.419	0.027	0.325	0.440	0.507
+JSD	0.736	0.395	0.220	0.382	**0.581**	**0.537**	**0.519**	**0.508**	0.289	0.378	0.351	**0.374**	0.254	0.245	0.013	**0.551**	0.039	**0.518**	**0.640**	**0.702**
+AugMix + JSD	0.717	0.391	0.238	0.387	0.550	0.496	0.489	0.473	**0.329**	**0.395**	0.376	0.352	**0.255**	0.286	**0.041**	0.542	0.064	0.481	0.622	0.668
+AugMix + **HCR**	0.727	0.390	0.234	0.383	0.552	0.500	0.494	0.480	0.320	0.391	0.374	0.349	0.249	0.283	0.040	0.539	0.061	0.481	0.624	0.662
+***FourierMix*** + JSD	**0.751**	**0.399**	**0.242**	**0.389**	0.564	0.515	0.493	0.483	0.315	0.384	**0.380**	0.370	0.254	**0.300**	**0.041**	0.544	**0.073**	0.497	0.637	0.694
+***FourierMix*** + **HCR**	0.750	0.397	0.239	0.387	0.567	0.518	0.499	0.492	0.312	0.382	0.377	0.370	0.249	0.295	0.039	0.544	0.069	0.494	0.637	0.689

Table 7. Average Certified Radius (ACR) of Models Trained with Different Methods on ImageNet-$\bar{\text{C}}$. Spectrally diverse augmentations from *FourierMix* brings significant gains to certified robustness of the models trained on ImageNet against corruptions from ImageNet-$\bar{\text{C}}$.

Augmentation	mACR	Blue	Brown	Caustic	Checkboard	Cocentric	Inv. Sparkle	Perlin	Plasma	Single Freq.	Sparkle
Gaussian	0.266	0.394	0.284	0.325	0.250	0.235	0.152	0.274	0.065	0.284	0.400
+JSD	0.395	**0.579**	0.395	**0.512**	**0.370**	**0.374**	**0.224**	0.404	0.113	0.408	**0.567**
+AugMix + JSD	0.379	0.560	0.381	0.461	0.365	0.342	0.212	**0.413**	0.121	0.397	0.538
+AugMix + **HCR**	0.378	0.563	0.377	0.464	0.361	0.342	0.210	0.410	0.115	0.396	0.539
+***FourierMix*** + JSD	**0.413**	0.562	**0.544**	0.479	**0.370**	0.366	0.215	**0.413**	**0.227**	**0.417**	0.547
+***FourierMix*** + **HCR**	0.411	0.565	0.535	0.481	0.365	0.367	0.210	0.408	0.215	0.415	0.550

(due to its depth) used in our paper. Thus, we use ResNet-18 in Table 9 which shows that our low-freq brittleness finding also extends to these methods.

To distinguish *FourierMix*, which directly uses spectral diversity objective, from other Fourier augmentation methods, we select a recent method FACT. which mixes the spectra of different clean data samples. As can be seen from Table 10, *FourierMix* outperforms FACT by a significant margin, which can be attributed to its better spectral diversity.

D.2 Detailed Results on CIFAR-100-Based Corruption Benchmarks

In this section, we present detailed results for our evaluation on CIFAR-100-C/$\bar{\text{C}}$ and CIFAR-100-F. We fix $\eta = 10$ and use $\lambda = 20$ for HCR in our experiments on CIFAR-100. Tables 4 and 5 present the ACR on individual corruption types from CIFAR-100-C/$\bar{\text{C}}$, respectively. CIFAR-100 is more difficult for RS-based certification compared to CIFAR-10. We find that *FourierMix*+HCR helps achieve the highest ACR on *all* corruption types in both datasets with significant enhancements compared to existing augmentation methods.

D.3 Detailed Results on ImageNet-Based Corruption Benchmarks

ImageNet appears to be the most challenging dataset for certified defenses, to which only RS-based techniques can be applied [9]. We select representative combinations of augmentations and regularization schemes that perform well on CIFAR-10/100 for our experiments on ImageNet. We exclude the input normalization layer, which trades off the ACR on clean data for the ACR on corrupted data. We use $\eta = 5$ and $\lambda = 5$ for our experiments with HCR. Tables 6 and 7 present the detailed results on our evaluation on ImageNet-C/$\bar{\text{C}}$. Note the the corruption types in ImageNet-$\bar{\text{C}}$ are different from the ones in CIFAR-10/100-$\bar{\text{C}}$.

Table 8. Average Certified Radius (ACR) of Clean (CIFAR-10) and Corrupted (CIFAR-10-C) Data with $\sigma = 0.25$ Using SOTA Certified Defense Methods.

Method	clean	mACR	Gauss.	Shot	Impulse	Defocus	Glass	Motion	Zoom	Snow	Frost	Fog	Bright	Contrast	Elastic	Pixel	JPEG
Gaussian	0.461	0.363	0.448	0.448	0.421	0.380	0.346	0.338	0.357	0.394	0.347	0.187	0.439	0.137	0.342	0.420	0.440
MACER	**0.539**	**0.426**	**0.509**	**0.509**	**0.492**	**0.460**	**0.436**	0.422	**0.433**	**0.443**	**0.381**	**0.232**	**0.477**	**0.185**	**0.428**	**0.490**	**0.503**
SmoothAdv	0.519	0.411	0.483	0.485	0.471	0.448	0.426	**0.423**	0.425	0.418	0.361	0.222	0.451	0.175	0.415	0.472	0.483

Table 9. RACC of RS and CROWN-IBP on CIFAR-10-C at $\epsilon = 0.25$.

	ResNet-18					ResNet-110				
RACC (%)↑	CIFAR-10	CIFAR-10-C	High	Mid	Low	CIFAR-10	CIFAR-10-C	High	Mid	Low
IBP	39.1	32.0	37.0	34.2	24.9	N/A	N/A	N/A	N/A	N/A
RS-Gaussian	65.0	52.4	61.6	52.9	42.8	65.4	53.4	61.9	53.7	44.6
RS-*FourierMix*	**67.8**	**55.7**	**62.5**	**58.5**	**46.2**	**69.2**	**60.3**	**66.8**	**61.9**	**52.2**

We find that the spectral biases of other baselines become much more noticeable on ImageNet-based corruption benchmarks. Gaussian+JSD accomplishes the highest ACR on high-frequency corruptions, while AugMix+JSD performs the best on several low-frequency corruptions in ImageNet-C. As RS-based models generally suffer performance degradation on low-frequency corruptions, Gaussain+JSD beats AugMix+JSD in terms of overall mACR. However, *FourierMix* performs well across the spectrum, reaching the highest mACR on both datasets.

However, HCR does not play an essential role in ImageNet. We find this might also related to the difficulty of certification on ImageNet. HCR as a strict regularization term will trade off certified radius for accuracy, resulting in similar ACR, *i.e.*, the area under the radius-accuracy curve. This observation is consistent with prior studies on in-distribution data certification [25]. Despite HCR not making a significant difference over JSD regularization, it is worth noting that substantial improvements can still be gained by *FourierMix* on ImageNet due to its broad spectral coverage. Although tangible improvements have been realized by *FourierMix* on ImageNet-based corruption benchmarks, we want to highlight that there is still large room for future research to improve over our baselines. We hope this work will motivate more studies on certified defenses for ImageNet under common corruptions, as discussed in § 6.

E *FourierMix* Details

Hyper-parameter Settings. We detail the chosen hyper-parameters used in the experiments with *FourierMix*. As illustrated in Algorithm 1 and Equations 1 and 2, we leverage 5 different severity levels and truncated Gaussian distribution. We use a large $\sigma = 5$ for the truncated Gaussian distribution to make *FourierMix* render more diverse augmentation. For CIFAR-10/100, we set $s_{\mathbf{A}} \in [0.2, 0.3, 0.4, 0.5, 0.6]$ and $s_{\mathbf{P}} \in [\frac{\pi}{12}, \frac{\pi}{10}, \frac{\pi}{8}, \frac{\pi}{6}, \frac{\pi}{4}]$ as the 5 severity levels in Equations 1 and 2, respectively. For ImageNet, we use the same set of $s_{\mathbf{A}}$ and set $s_{\mathbf{P}} \in [\frac{\pi}{4}, \frac{3\pi}{10}, \frac{3\pi}{8}, \frac{\pi}{2}, \frac{3\pi}{4}]$ since high-resolution images can tolerate more perturbations in the phase spectrum.

Sample Images from *FourierMix*. We visualize randomly sampled images from CIFAR-10/100 and ImageNet in Figs. 13, 14, and 15, respectively.

Table 10. ACR Comparison of FACT and our *FourierMix*.

ACR ↑	CIFAR-10	CIFAR-10-C	High	Mid	Low
FourierMix+JSD	**0.522**	**0.444**	**0.504**	**0.454**	**0.375**
FACT [62]+JSD	0.503	0.410	0.478	0.406	0.345

Table 11. Failure of Existing Methods: Average Certified Radius (ACR) of CIFAR10-C with $\sigma = 0.25$ Using Test-Time Adaptation.

Adaptation	mACR	Gauss.	Shot	Impulse	Defocus	Glass	Motion	Zoom	Snow	Frost	Fog	Bright	Contrast	Elastic	Pixel	JPEG
Gaussian	**0.363**	**0.448**	**0.448**	**0.421**	**0.380**	**0.346**	**0.338**	**0.357**	**0.394**	**0.347**	**0.187**	**0.439**	**0.137**	**0.342**	**0.420**	**0.440**
+BN	0.356	0.441	0.442	0.417	0.369	0.338	0.326	0.345	0.392	**0.347**	0.181	0.436	0.133	0.332	0.411	0.432
+TENT	0.357	0.442	0.442	0.419	0.369	0.337	0.328	0.346	**0.394**	0.345	0.182	0.436	0.132	0.330	0.412	0.434

F Sample Images from CIFAR-10/100-F

We visualize more sample images from our created datasets in Fig. 12 using different classes. It is also worth noting that *FourierMix* augmented images (Figs. 13 and 14) have different patterns with CIFAR-10/100-F.

It is worth noting that the generation protocol of CIFAR-10/100-F is general and we plan to construct ImageNet-F from a representative subset of ImageNet [12] as a future study.

G Discussion on Test-Time Adaptation

As discussed in Appendix A, another widely acknowledged approach to counter distribution shifts is test-time adaptation. We thus perform a preliminary study on how test-time adaptation will affect the certified robustness. Specifically, we use BN [46] and TENT [56] as representative methods. Since the theorem derived by Cohen *et al.* [9] requires the base classifier $\mathcal{M}$ to be deterministic, we cannot apply BN and TENT in an online manner. To deal with such a problem, while evaluating the ACR of corrupted data from a specific corruption type, we randomly sample 500 (out of 10,000) images from the corruption test set for the adaptation. We follow other settings specified in [46,56] for our experimentation. Table 11 presents the detailed results on CIFAR-10-C. We find that test-time adaptations fail to improve the ACR in under common corruptions. The reason is that *one-shot* adaptation relies upon a small amount of data which is not sufficient to correct the distribution shift caused by corruptions. In contrast, it may cause the base classifier $\mathcal{M}$ to become biased towards the small subset of test data used for adaptation. We highlight that certification of adaptive models is also a potential direction that can help with certified robustness under common corruptions. More theoretical support is needed in this direction, and we leave it as a promising future work.

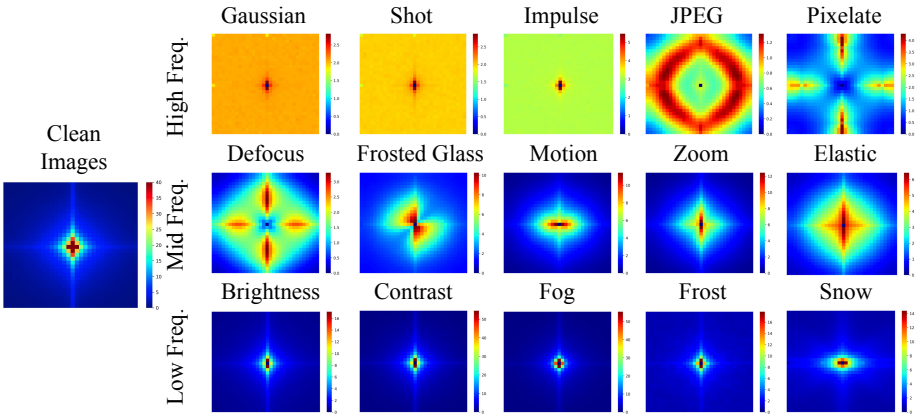


Fig. 10. Amplitude Spectrum $\boldsymbol{A}$ of Different Corruptions in CIFAR-10/100-C with severity 3.

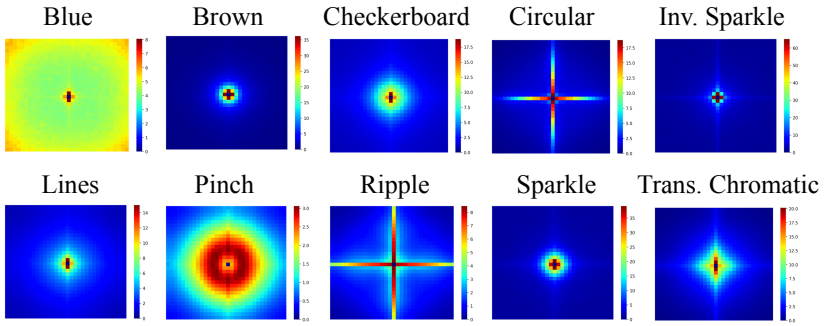
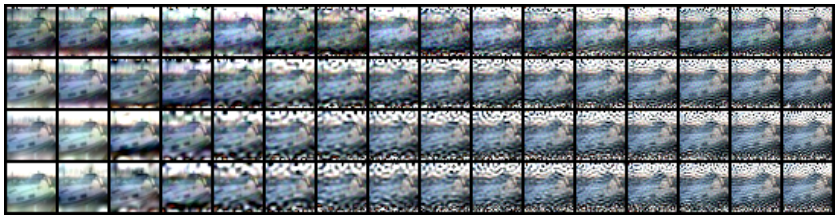
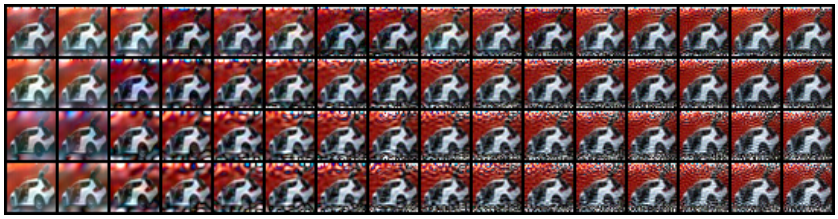


Fig. 11. Amplitude Spectrum $\boldsymbol{A}$ of Different Corruptions in CIFAR-10/100-$\bar{\text{C}}$ with severity 3.



(a) Examples of the Ship Class

(b) Examples of the Airplane Class



(c) Examples of the Car Class

(d) Examples of the Bird Class

(e) Examples of the Horse Class

Fig. 12. Sample Images from CIFAR-10/100-F with $\epsilon = 12$. From top down row-wise, the images are from $\alpha \in \{0.5, 1, 2, 3\}$ and from left to right column-wise, the images are from $f_c \in \{1, 2, ..., 16\}$

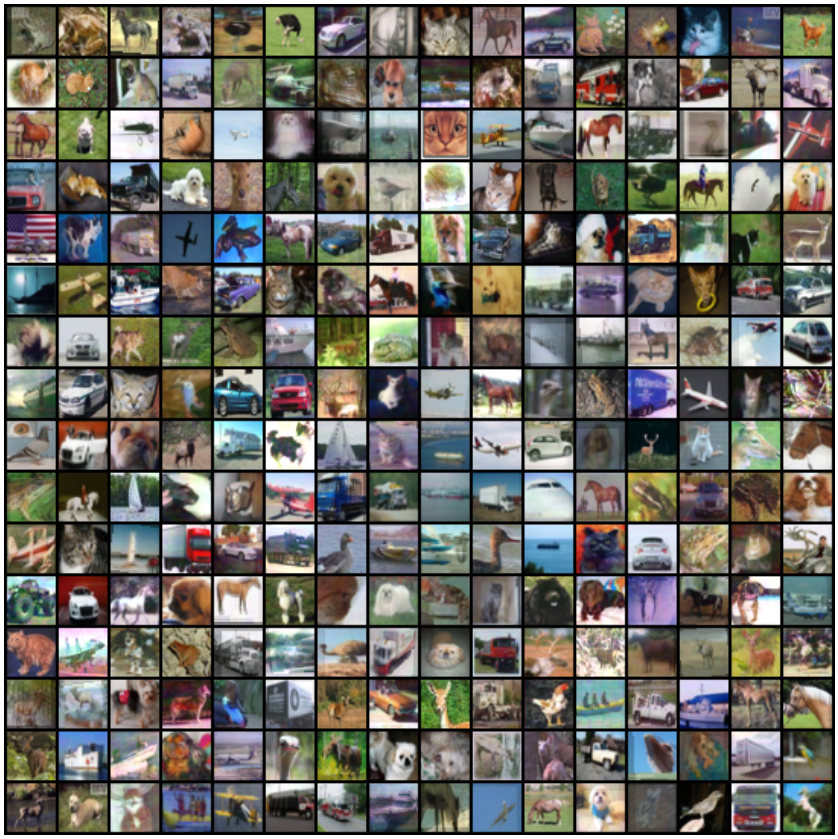


Fig. 13. Sample Images from *FourierMix* Data Augmentation on CIFAR-10. To better highlight the visual patterns of *FourierMix*, we utilize the highest severity level for $\mathbf{A}(\cdot)$ and $\mathbf{P}(\cdot)$ in this figure.

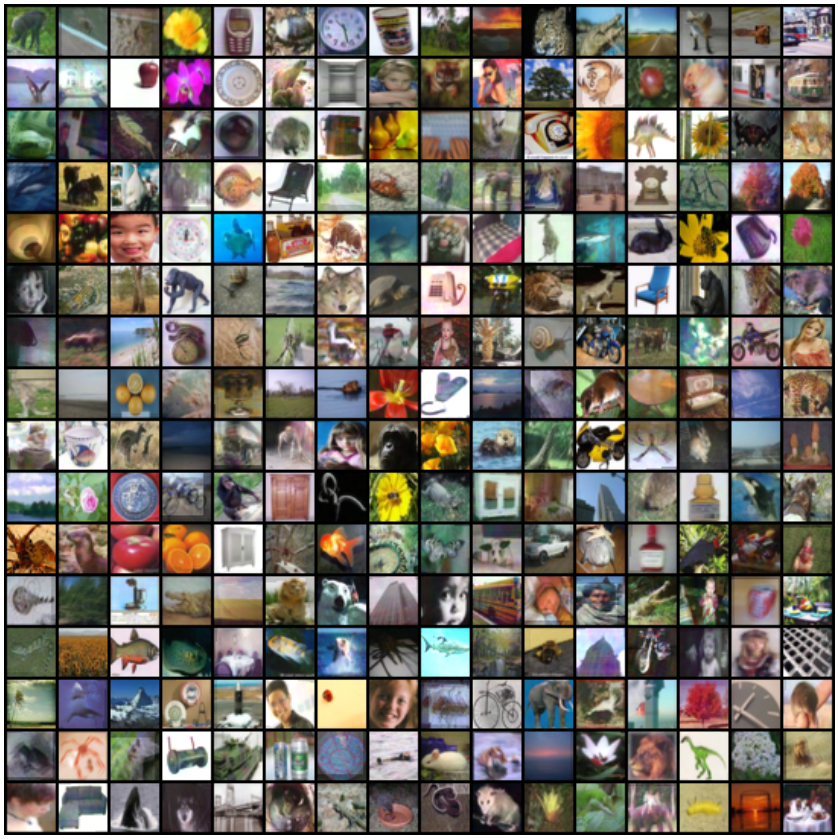


Fig. 14. Sample Images from *FourierMix* Data Augmentation on CIFAR-100. To better highlight the visual patterns of *FourierMix*, we utilize the highest severity level for $\mathbf{A}(\cdot)$ and $\mathbf{P}(\cdot)$ in this figure.

Fig. 15. Sample Images from *FourierMix* Data Augmentation on ImageNet. To better highlight the visual patterns of *FourierMix*, we utilize the highest severity level for $\mathbf{A}(\cdot)$ and $\mathbf{P}(\cdot)$ in this figure.