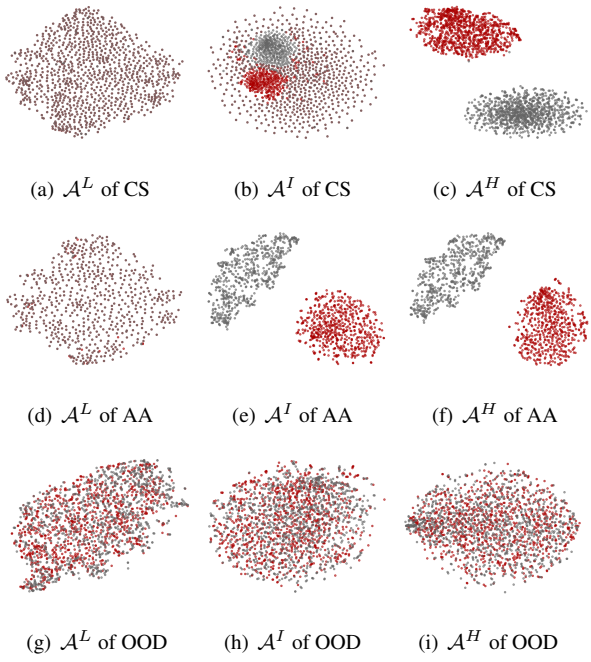
Amplitude-Phase Recombination: Rethinking Robustness of Convolutional Neural Networks in Frequency Domain

Supplemental Material I

Appendix

In the supplemental material, we firstly visualize the distributions of the corrupted samples, adversarial samples, and OOD samples in the frequency domain to validate the **Assumption** 2 in the main text in Section A. Then, several typical templates of the phase spectrum are shown in Section B to intuitionally explain the rationality of the proposed APR method, and the implementation details of the data augmentation of APR-S are listed in Section C. In Section D, the Fourier analysis is provided to demonstrate the various gains from APR-P and APR-S, and the clean error analysis and OOD detection on ImageNet are also listed to clarify the excellent scalability of our method.



(a) $\mathcal{A}^L$ of CS (b) $\mathcal{A}^I$ of CS (c) $\mathcal{A}^H$ of CS

(d) $\mathcal{A}^L$ of AA (e) $\mathcal{A}^I$ of AA (f) $\mathcal{A}^H$ of AA

(g) $\mathcal{A}^L$ of OOD (h) $\mathcal{A}^I$ of OOD (i) $\mathcal{A}^H$ of OOD

Figure 1. The T-SNE [11] visualization of the different types of the amplitude spectrum. Red represents the original image or in-distribution (ID) samples in CIFAR-10, and gray represents Corrupted Samples (CS), the samples generated by Adversarial Attacks (AA) from CIFAR-10, or OOD samples from CIFAR-100.

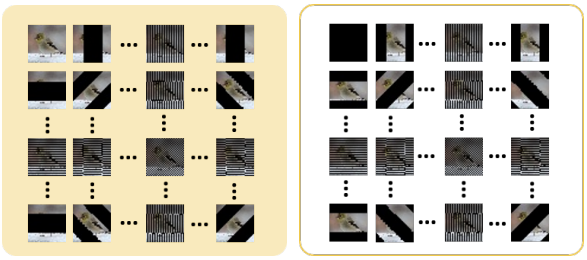
A. More Studies on the Frequency Domain.

Here, we show more studies of the different types of the amplitude spectrum. For all samples in CIFAR, we generate the low-frequency, intermediate-frequency and high-frequency counterparts with r for non-zero parts set to $[0, 8]$, $[8, 16]$ and $[16, 16\sqrt{2}]$, respectively. In Figure 1, we show the amplitude spectrum distributions of low-frequency, intermediate-frequency, and high-frequency from original samples, their corrupted samples, adversarial samples, and OOD samples respectively.

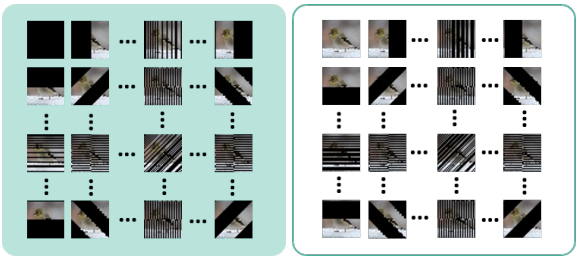
Firstly, for *corrupted samples* and *adversarial samples* from a single category, we could observe that their amplitude spectrum in high-frequency and intermediate-frequency has different distribution with the original samples even if only invisible noises are introduced. Moreover, the amplitude spectrums in low-frequency of corrupted samples and adversarial samples are indistinguishable from the original images. It also explains that CNN captures the high-frequency image components for classification [12]. Hence, CNN would make a wrong prediction for 'similar' images (corruption and adversarial samples) when the parts of the amplitude spectrum are changed.

Then, for the *OOD samples* from CIFAR-100, it is evident that any type of the amplitude spectrum of in-distribution and out-of-distribution could be not able to distinguish. CNN focusing on the amplitude spectrum has a huge risk that any OOD sample with a similar amplitude part of in-distribution samples would be classified as an in-distribution sample. Hence, CNN would be overconfident for some out-of-distributions when similar amplitude information appears.

Overall, the above analyses explain our **Assumption** 2 in the main text, that the counter-intuitive behaviors of the sensitivity to common perturbations and the overconfidence of OOD maybe both be related to CNN's over-dependence on the amplitude spectrum. We do not focus on the high frequency only, because the OOD samples may come from the similarity of any amplitude part as shown in Figure 1. As a result, the focusing of some parts of the amplitude spectrum may create an invisible way to attack CNN, such as the adversarial attack and various corruptions (Corollary 1), and the amplitude attack or OOD attack (Corollary 2).



(a) Positive Real Part $\mathcal{T}^{\mathcal{R}+}$ (b) Negative Real Part $\mathcal{T}^{\mathcal{R}-}$



(c) Positive Imaginary Part $\mathcal{T}^{\mathcal{I}+}$ (d) Negative Imaginary Part $\mathcal{T}^{\mathcal{I}-}$

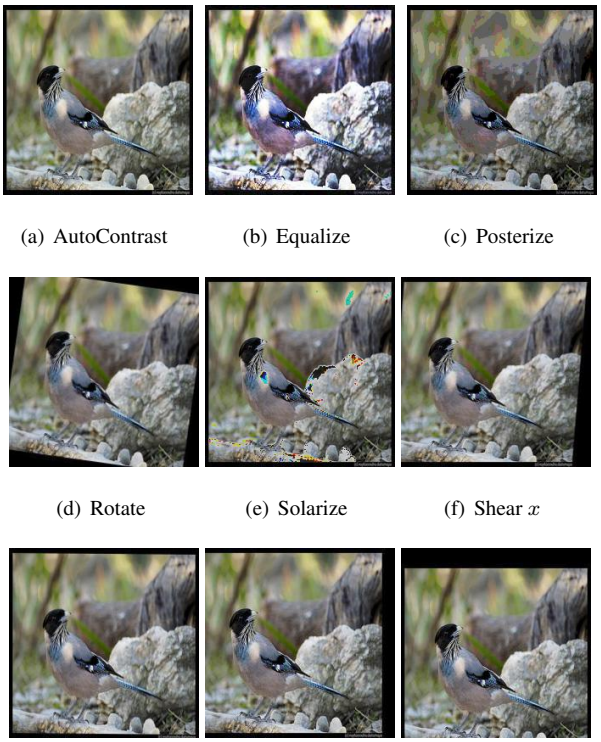
Figure 2. Fourier transform can be interpreted as dividing an image with $4 \times N^2$ templates for contrast computation.

B. Templates from the Phase Spectrum

To better reveal the role of the phase spectrum, we analyze the phase spectrum in the different frequency domains. When DFT is applied on a gray image, there are totally $4 \times N^2$ templates which are used to compute $2 \times N^2$ contrast scores (as shown in Figure 2). Consequently, the frequency spectrum stores contrast values obtained at multiple scales and directions. The classification or other visual tasks could benefit from capturing the difference between targets and distractors by these templates. The templates of the lowest frequencies divide images into large regions, which are "coarse" partitions. Then, the templates of the highest frequencies provide "fine" partitions that achieve only high responses to noises and textures. In addition, the templates of the intermediate frequencies provide "moderate" partitions which may include the target object, and it has also been proved that it's beneficial for fixation prediction in [6]. These templates in the phase spectrum could help to recover the structural information of the original image even without the original amplitude spectrum [8]. The robustness human visual system can also rely on this visible structured information for recognition [8, 7].

C. Augmentation Operations

The augmentation operations used in APR-S are same with [4] as shown in Figure 3. We do not use *contrast*, *color*, *brightness*, *sharpness* and *Cutout* as they may overlap with the corruptions of CIFAR-C and ImageNet-C.



(a) AutoContrast (b) Equalize (c) Posterize

(d) Rotate (e) Solarize (f) Shear x

(g) Shear y (h) Translate x (i) Translate y

Figure 3. Illustration of augmentation operations applied to the same image.

D. Additional Results

D.1. Fourier Analysis

In order to better understand the reliance of our methods on different frequencies, here we measure the model sensitivity to the additive noise at differing frequencies. We add a total of 33×33 Fourier basis vector to the CIFAR-10 test set, one at a time, and record the resulting error rate after adding each Fourier basis vector. Each point in the 33×33 sensitivity heatmap shows the error rate on the CIFAR-10 test set after it has been perturbed by a single Fourier basis vector. Points corresponding to the low-frequency vectors are shown in the center of the heatmap, whereas the high-frequency vectors are farther than the center.

In Figure 4, we observe that the standard model is robust to the low-frequency perturbations but severely lacks robustness to the high-frequency perturbations, where the error rates exceed 80%. Then, the model trained by APR-P is more robust to all frequencies, especially to the low and intermediate frequencies. Moreover, the model trained by APR-S maintains robustness to low-frequency perturbations and improves robustness to the high-frequency perturbations, but is still sensitive to the additive noise in the intermediate frequencies. This further explains the exper-

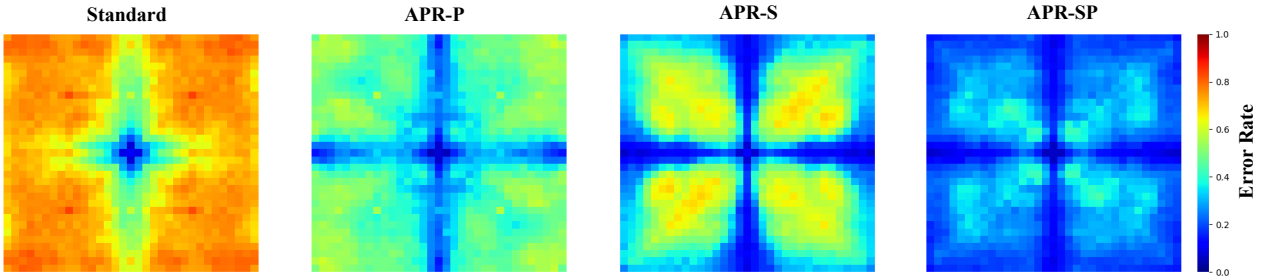


Figure 4. Model (Wide ResNet) sensitivity to the additive noise aligned with different Fourier basis vectors on CIFAR-10 validation images. We fix the additive noise to have L_2 norm 15 and evaluate four methods: a standard trained model, APR-P, APR-S, APR-SP. Error rates are averaged over 1000 randomly sampled images from the test set. The standard trained model is highly sensitive to the additive noise in all but the lowest frequencies. APR-SP could substantially improve robustness to most frequency perturbations.

Table 1. CIFAR-10 Clean Error. All values are percentages and the best results are indicated in bold.

		Standard	Cutout	Mixup	CutMix	AutoAugment	Adv Training	AugMix	APR-P	APR-S	APR-SP
CIFAR-10-C	AllConvNet	6.1	6.1	6.3	6.4	6.6	18.9	6.5	**5.5**	6.5	5.7
	DenseNet	5.8	**4.8**	5.5	5.3	**4.8**	17.9	4.9	5.0	5.1	**4.8**
	WideResNet	5.2	4.4	4.9	4.6	4.8	17.1	4.9	4.8	5.0	**4.3**
	ResNeXt	4.3	4.4	4.2	**3.9**	3.8	15.4	4.2	4.5	4.5	**3.9**
Mean		5.4	4.9	5.2	5.0	5.0	17.3	5.1	5.0	5.2	**4.7**

Table 2. OOD performance of different methods on the larger and more difficult datasets, where ImageNet-1K is the in-distribution dataset and ImageNet-O is the OOD dataset.

Method	AUROC	OSCR
Standard	40.9	36.8
APR-SP	**62.3**	**53.2**

iments in Section 5.1.2 (main text) that APR-P improves performances of both OOD detection and defense adversarial attacks tasks. From Appendix A, the adversarial samples are more different from original samples in intermediate and high frequencies, while the OOD samples may share similarities with the original samples in any frequencies. The gains of APR-P and APR-S to different frequency domains bring the gains to different tasks.

Furthermore, APR-SP (combining APR-S and APR-P) could substantially improve robustness to most frequency perturbations. The weak sensitivity to the intermediate frequencies is reasonable because of the gains for target prediction from intermediate frequencies in Appendix B.

D.2. Clean Error

Table 1 reports clean error [4] of CIFAR-10 by different methods, and the proposed method achieves the best performances on various backbone networks. APR-SP not only improves the model adaptability to the common corruptions, surface variations and OOD detection, but also improves the classification accuracy of the clean images.

D.3. OOD Detection on ImageNet.

We conduct OOD experiments on the larger and more difficult ImageNet-1K dataset [9]. ImageNet-O [5] is adopted as the out-of-distribution dataset of ImageNet-1K. ImageNet-O includes 2K examples from ImageNet-22K [9] excluding ImageNet-1K. The ResNet 50 [3] is trained on ImageNet-1K and tested on both ImageNet-1K and ImageNet-O.

In order to evaluate the accuracy of in-distribution and the ability of OOD detection simultaneously, we introduce *Open Set Classification Rate* (OSCR) [2, 1] as an evaluation metric. Let δ is a score threshold. The *Correct Classification Rate* (CCR) is the fraction of the samples where the correct class k has maximum probability and has a probability greater than δ:

$$CCR(\delta) = \frac{|\{x \in \mathcal{D}_{\mathcal{I}}^k \wedge argmax_k P(k|x) = \hat{k} \wedge P(\hat{k}|x) \geq \delta\}}{|\mathcal{D}_{\mathcal{I}}^k|}. \tag{A.1}$$

where $\mathcal{D}_{\mathcal{I}}^k$ is the interest in-distribution classes that the neural network shall identify. The *False Positive Rate* (FPR) is the fraction of samples from OOD data $\mathcal{D}_{\mathcal{O}}$ that are classified as *any* in-distribution class k with a probability greater than δ:

$$FPR(\delta) = \frac{|\{x|x \in \mathcal{D}_{\mathcal{O}} \wedge max_k P(k|x) \geq \delta\}|}{|\mathcal{D}_{\mathcal{O}}|}. \tag{A.2}$$

A larger value of OSCR indicates a better detection performance. As shown in Table 2, APR-SP performs better than the standard augmentations even on the large and difficult

dataset. Especially, APR-SP achieves about 22% improvement on AUROC. From the OSCR, APR-SP improves the performances of the OOD detection while maintaining test accuracy. These results indicate the excellent scalability of APR in larger-scale datasets.

D.4. More CAM Visualization Examples

We also list more visualization examples with various corruptions in Figure 5 and 6, the CNN trained by APR-SP is able to focus on the target objects for classification even with different common corruptions and surface variations.

References

[1] Guangyao Chen, Peixi Peng, Xiangqian Wang, and Yonghong Tian. Adversarial reciprocal points learning for open set recognition. *arXiv preprint arXiv:2103.00953*, 2021. 3

[2] Akshay Raj Dhamija, Manuel Günther, and Terrance Boult. Reducing network agnostophobia. In *Advances in Neural Information Processing Systems*, pages 9157–9168, 2018. 3

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[4] Dan Hendrycks, Norman Mu, Ekin Dogus Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. In *International Conference on Learning Representations*, 2019. 2, 3

[5] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. *arXiv preprint arXiv:1907.07174*, 2019. 3

[6] Jia Li, Ling-Yu Duan, Xiaowu Chen, Tiejun Huang, and Yonghong Tian. Finding the secret of image saliency in the frequency domain. *IEEE transactions on pattern analysis and machine intelligence*, 37(12):2428–2440, 2015. 2

[7] Honggu Liu, Xiaodan Li, Wenbo Zhou, Yuefeng Chen, Yuan He, Hui Xue, Weiming Zhang, and Nenghai Yu. Spatial-phase shallow learning: Rethinking face forgery detection in frequency domain. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 2

[8] Alan V Oppenheim and Jae S Lim. The importance of phase in signals. *Proceedings of the IEEE*, 69(5):529–541, 1981. 2

[9] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. 3

[10] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 5, 6

[11] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 1

[12] Haohan Wang, Xindi Wu, Zeyi Huang, and Eric P Xing. High-frequency component helps explain the generalization of convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8684–8694, 2020. 1

(a) Noise: Gaussion

(b) Noise: Shot

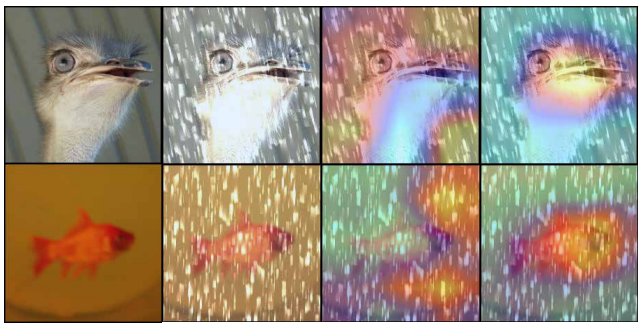
(c) Noise: Impulse

(d) Blur: Defocus

(e) Blur: Glass

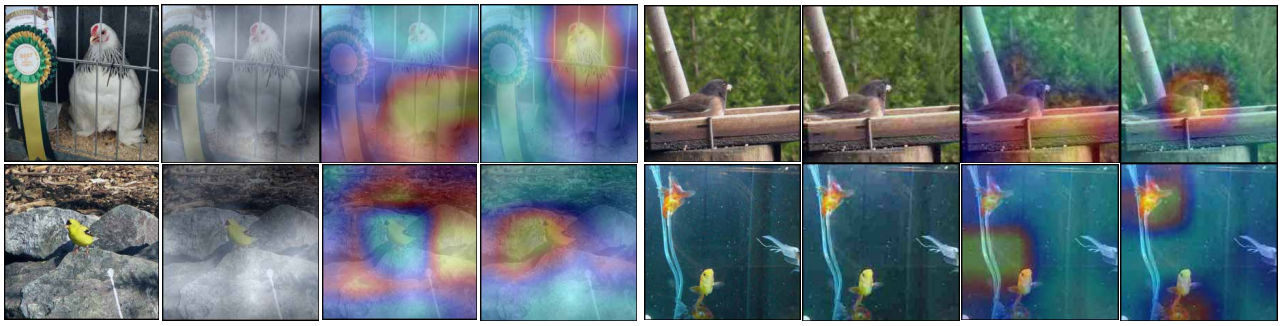
(f) Blur: Motion

(g) Blur: Zoom



(h) Weather: Snow

Figure 5. The Gradient-weighted Class Activation Mapping [10] of the baseline (the third column in each panel) and the proposed APR-SP (the fourth column in each panel) for images with different common corruptions and surface variations (the second column in each panel). The original images are in the first column in each panel. Best viewed in color. APR-SP still is robust even in various corruptions.



(a) Weather: Fog

(b) Digital: JPEG Compression

(c) Digital: Elastic Transform

(d) Digital: Pixelate

Figure 6. The Gradient-weighted Class Activation Mapping [10] of the baseline (the third column in each panel) and the proposed APR-SP (the fourth column in each panel) for images with different common corruptions and surface variations (the second column in each panel). The original images are in the first column in each panel. Best viewed in color. APR-SP still is robust even in various corruptions.