Exploring connections of spectral analysis and transfer learning in medical imaging

Yucheng Lu[a], Dovile Juodelyte[a], Jonathan D. Victor[b], and Veronika Cheplygina[a]

[a]IT University of Copenhagen, Rued Langgaards Vej 7, Copenhagen, Denmark
[b]Weill Cornell Medicine, 1300 York Avenue, New York, USA

ABSTRACT

In this paper, we use spectral analysis to investigate transfer learning and study model sensitivity to frequency shortcuts in medical imaging. By analyzing the power spectrum density of both pre-trained and fine-tuned model gradients, as well as artificially generated frequency shortcuts, we observe notable differences in learning priorities between models pre-trained on natural vs medical images, which generally persist during fine-tuning. We find that when a model's learning priority aligns with the power spectrum density of an artifact, it results in overfitting to that artifact. Based on these observations, we show that source data editing can alter the model's resistance to shortcut learning. Code available at: https://github.com/YCL92/Shortcut-PSD

Keywords: Shortcut learning, Transfer learning, Image statistics, Medical imaging

1. INTRODUCTION

Deep learning has achieved many advances in medical image classification, even showing performances on par with medical experts. However, convolutional neural networks (CNNs) may be prone to shortcut learning,[1] such as surgical markers.[2] As a consequence, instead of capturing the semantic contents, the model makes predictions based on the shortcuts, which, in the worst case, leads to unreliable results if their association with semantics differs between the training dataset and the images used in real-world applications.

Most studies investigate shortcut learning in the context of training from scratch. However, little is understood about the importance of shortcuts in transfer learning, which is crucial in the medical domain for two reasons. First, transfer learning is often involved in medical image analysis due to the limited amount of labeled data.[3–5] Second, next to obvious shortcuts like pen markings, CT and MR scans, in particular, can have subtle shortcuts in the spectrum domain that may not be noticed by the human eye. This prompts us to explore the sensitivity of transfer learning to spectral shortcuts in medical image classification tasks and how to mitigate the negative impacts it brings about. To this end, we use spectral analysis to investigate the role of power spectrum density (PSD) in pre-training and fine-tuning and observe distinct differences in their learning priorities, which are related to shortcut learning. Based on these observations we show through experiments that resistance to common detrimental frequency shortcuts could be altered via source data editing.

2. RELATED WORK

The standard transfer learning protocol involves fine-tuning a pre-trained model on a small target dataset, where the pre-trained weights are determined by training on a large-scale source dataset, with ImageNet[6] being a popular choice. However, due to the domain gap between ImageNet and target medical datasets, several studies showed that pre-training on medical data can improve performance,[4,7,8] leading to a large-scale radiological image dataset RadImageNet.[9] Another factor related to model performance in transfer learning is shortcut learning. For example,[10] found that RadImageNet is more robust to shortcuts, such as those relying on noise and denoising. Since we notice that these shortcuts are *spectrum-related*, and the spectra of natural and medical images have distinct differences,[11] we classify related work into three categories: frequency bias, frequency shortcuts, and spectrum augmentation.

Contact details: {yucl,vech}@itu.dk

Medical Imaging 2025: Image Processing, edited by Olivier Colliot, Jhimli Mitra, Proc. of SPIE Vol. 13406, 1340631 · © 2025 SPIE
1605-7422 · doi: 10.1117/12.3047670

2.1 Frequency Bias

After[12] showed that CNNs often learn a stronger bias towards a frequency band that is highly correlated with the characteristics of image degradation, many researchers further investigated this phenomenon.[13] found that while humans make use of semantic content, powerful CNNs tend to rely on high-frequency components, leading to lower robustness. Their experiments also revealed a trade-off between the model's robustness and performance. Similarly,[14] pointed out that this counter-intuitive behavior is related to an over-emphasis on amplitude, compared to relative phase.[15] examined the importance of each frequency band based on an energy distribution model to control the ratio between classification performance deterioration and image quality degradation for each band. They suggested that the frequency bias is not tightly related to CNN architectures or model depth but to discriminative features in the mid-frequency band.[16] further observed that the preference for low-to-mid-frequency is due to the considerable suppression of high-frequency bands in feature extraction.

While these studies showed the presence of frequency bias, they only considered natural images. In comparison, we analyze frequency bias in both natural image and medical image domains and observe distinct differences.

2.2 Frequency Shortcut

Previous work showed that CNNs are often biased towards specific bands, learning to recognize signature patterns in the spectrum, especially when these patterns are confounders.[17,18] offered a deep understanding of these frequency shortcuts by selectively extracting spectral patterns while disregarding irrelevant frequencies. Their experiments showed that during initial training, CNNs seek to find simple solutions, such as the most distinct spectrum characteristics. These shortcuts are not limited to low-frequency bands but can also be high-frequency ones. Building on these findings,[19] proposed a method to mitigate shortcut learning by filtering samples of each class based on the frequency content of the other classes.

Here, we focus on the model's robustness against frequency shortcuts in transfer learning. Particularly, we are interested in exploring the link between the kernels' frequency response and their initial weights as determined from different source datasets. Our study reveals that the fine-tuned model robustness relates to the pre-trained model's learning priority.

2.3 Spectrum Augmentation

It is difficult to improve the learning of domain-invariant features via data augmentation in the spatial domain alone.[12] An alternative way to mitigate frequency bias is to apply data augmentation in the frequency domain. Several variations on this theme have been pursued.[20] studied the frequency decomposition of learned functions by adding noise to various frequencies via label smoothing. They observed that a high-accuracy model responds well to high frequencies across classes but focuses more on low frequencies within each class.[21] exploited the impact of frequency components in few-shot learning, where a class-discriminative filtering scheme based on Grad-CAM[22] was applied to the training samples to enhance the model's ability to capture task-relevant frequencies.[23] blended the amplitude components from two images while keeping the phase components unaltered based on their semantic-preserving property; while[24] fused the spectra of two random classes with independent focuses on low and high frequencies and trained the model to predict the weighted probabilities for the two classes, retaining a controllable sensitivity across various frequency bands.[25] swapped the frequency bands between two randomly selected images or augmented variants of a single image and further applied phase perturbation from a third image to enrich the data augmentation.

Together, these studies provided new approaches to data augmentation but did not answer the question of how transfer learning might benefit – which we address here. Our experiments show that data editing in the source domain affects the model's robustness against shortcut learning in the target domain.

3. METHODOLOGY

3.1 Datasets and Models

As *sources* we use ImageNet[6] and RadImageNet.[9] ImageNet has 1.2M training and 50K validation images in 1K classes, while RadImageNet has 1M training and 112K validation images in 165 classes. We pre-train a ResNet-50[26] (implementation details in Supplementary) as it is a common choice for medical images. As *targets* we

select two small medical datasets: LoDoPaB-CT[27] – a subset of LIDC-IDRI,[28] and KneeMRI.[29] We chose these datasets as both of their imaging pipelines involve frequency-domain reconstruction. To simplify the analysis of frequency shortcuts, we binarize the tasks to benign (malignancy score < 3) vs malignant for LoDoPaB-CT, and healthy vs injured ligament for KneeMRI. This results in the following train/validation/test partitions: 375/125/1033 samples (198/66/548 studies) for LoDoPaB-CT, and 375/125/871 samples (375/125/582 studies) for KneeMRI. As for pre-training, we followed the official data splits. We used the Adam optimizer with a batch size of 32 and a learning rate of 0.01 and 0.001 for pre-training and fine-tuning, respectively.

3.2 Frequency Shortcuts

We introduce shortcuts by altering the images in two frequency-related ways: noise and denoising – here denoted as "artifacts". The noise level in CT images varies because of automatic exposure control and the choice of reconstruction filters. Denoising is commonly applied after reconstruction as a spatial filtering operation, but the extent of denoising can vary from image to image. Thus, both result in alterations of the frequency content of the image and could lead to frequency shortcuts.

We select projection-domain Photon noise in CT[27] and non-local means (NLM) denoising in MRI[30] because they have distinct spectral statistics. To create a spurious correlation between the artifacts and the labels, we add the artifacts to all negative samples in the test set and a certain amount (e.g. 50%) of positive samples in the training set. This design ensures that if the model detects the shortcut, its out-of-distribution (O.O.D.) performance will decrease, while the independent-and-identically-distributed (I.I.D.) performance will improve.

3.3 Power Spectrum Density

To characterize the statistics of datasets and model weights in the frequency domain, we convert the standard 2D spatial power spectrum into a 1D PSD by integrating the spectrum values over all angles. The resulting quantity provides a comprehensive measure of power distribution across frequencies and is especially useful when an artifact or a feature lies in a specific frequency band. The PSD is computed as follows:

$$PSD\left(\omega_k\right) = \int_0^{2\pi} \left\| \mathcal{F}\left(X\right)\left(\omega_k \cos\phi, \omega_k \sin\phi\right) \right\| d\phi, \tag{1}$$

where ω_k is a radial frequency, $k \in \left\{0, 1, \cdots, \frac{1}{2}M - 1\right\}$, M is the input size (assuming square shape). ϕ is a polar angle, and $\mathcal{F}$ is the Fourier transform. An example of PSD is presented in Fig. 1.

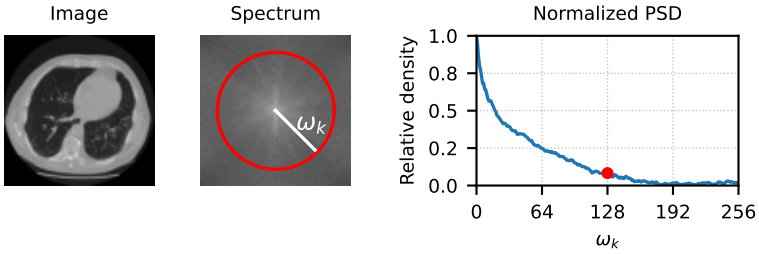

Figure 1. Example of a PSD. From left to right: original image, its spectrum with a selected frequency ω_{128}, and the PSD with the highlighted frequency ω_{128}.

It is worth noting that PSD is versatile. When the input X is image data, the PSD shows the overall spectral distribution. To analyze a trained model, one can compute the gradient map back-propagated from the prediction loss to the input image as X to analyze the model's spectral *learning priority*.[16]

4. EXPERIMENT RESULTS

4.1 ImageNet is Prone to Shortcut Learning

We pre-trained the model on the original ImageNet and RadImageNet and fine-tuned it on the target datasets as the baseline. The 5-fold cross-validation results are shown in Fig. 2. RadImageNet has higher robustness against

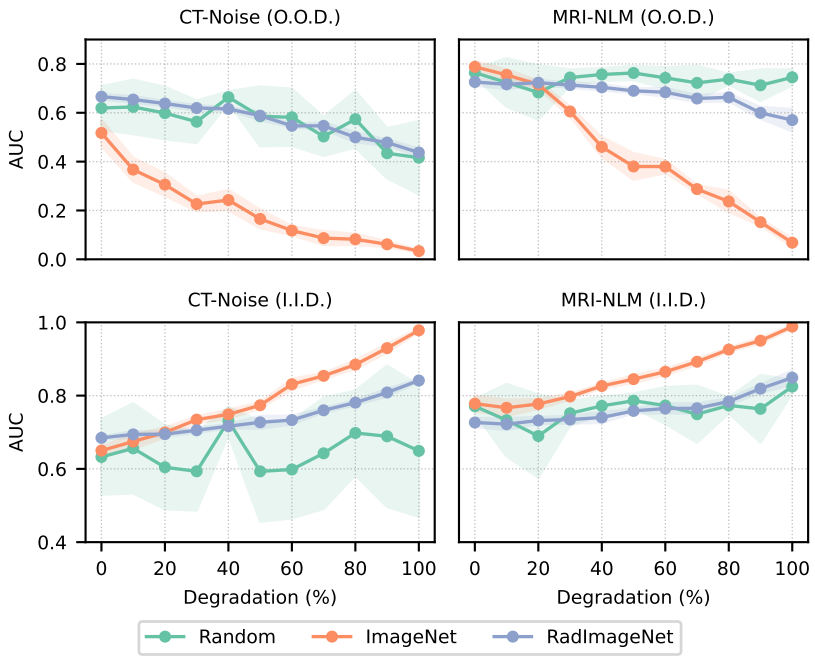

Figure 2. Baseline results (mean and standard deviation of AUC across 5-folds) as a function of degradation (amount of artifacts in the training set), performance on O.O.D. (top) and I.I.D. (bottom) test sets.

frequency shortcuts, whereas ImageNet exhibits poor generalization ability when tested on O.O.D. images. In comparison, random initialization (i.e. training from scratch, dubbed "random") shows dramatic fluctuation in performance across folds, indicating its instability on small datasets. However, both ImageNet and RadImageNet pre-trained models have competitive performance on I.I.D. data, which reveals that the source dataset plays an important role in shortcut learning.

4.2 Learning Priority is Stable During Transfer

We computed the PSDs of models (i.e. learning priorities) pre-trained on ImageNet and RadImageNet. The results are plotted in Fig. 3 (top row). We notice that ImageNet pre-trained model has higher gradients in the mid-to-high frequency bands, indicating that it focuses on extracting features from these bands;[16] while RadImageNet pre-trained model responds more actively to low-frequency features. Similar trends are observed in the learning priorities after fine-tuning, as shown in Fig. 3 (second row). Although the peaks eventually shift to higher frequencies, the overall PSDs still resemble their pre-trained counterparts. This is unsurprising, considering that kernels in early layers show minimal change during fine-tuning,[5] thereby inheriting the predominant spectral response from pre-training.

4.3 PSD is Related to Shortcut Learning

We computed the average PSDs of artificially generated artifacts by extracting the residual between the original and modified images, plotted in Fig. 3 (green solid lines). We observe that the spectral distribution of the artifacts mainly falls in the mid-to-high frequencies. Interestingly, the learning priority of ImageNet pre-trained model shows a higher level of overlap with the PSD of the artifact, while the results in Fig. 2 indicate that ImageNet is prone to shortcut learning. As gradients reflect how much the loss is affected by changes in the input, higher density indicates that kernels are more sensitive to corresponding frequency perturbations.[31] Therefore, it is reasonable to believe that the learning priority of a pre-trained model and its robustness to frequency shortcuts are related: kernels pre-trained on ImageNet have stronger response to mid-to-high frequencies and thus can quickly detect shortcuts with similar spectral distributions.

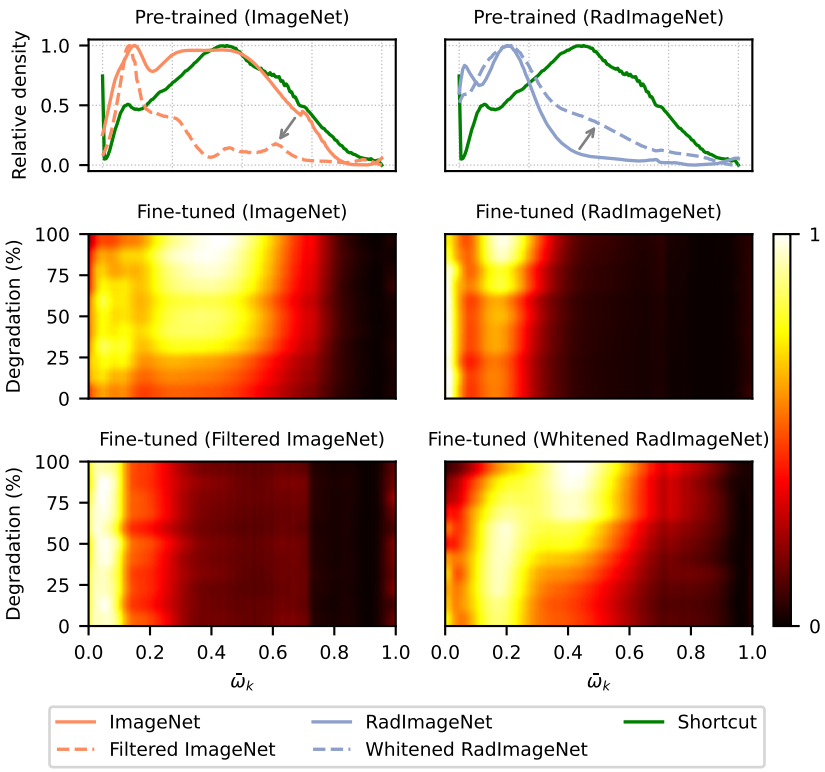

Figure 3. Normalized learning priorities of pre-trained and fine-tuned models. $\bar{\omega}_k$ is the normalized radical frequency with respect to the highest value (x-axis shared between rows). **Top**: normalized PSDs from Eq. 1. The arrows show how the pre-trained model PSDs change before and after source data editing. **Middle**: PSD as a heat map, after different degrees of degradation (amount of artifacts in the training set) for the original datasets. **Bottom**: Same as above but for the edited datasets.

4.4 Source Data Affects Robustness

Previous experiments show that the frequency response of the early layers remains largely unchanged in transfer learning, thus it is possible to enhance or reduce shortcut learning by modifying the model's learning priority via source data editing. Specifically, we altered the model's response to mid-to-high frequencies during pre-training. We encouraged RadImageNet model to focus more on learning high frequencies by normalizing the spectrum of images in RadImageNet. Additionally, whitening was applied to ensure that the normalized images maintain the same mean and standard deviation as the originals:

$$I_n = \mathcal{F}^{-1}\left(\frac{\mathcal{F}(I)}{\|\mathcal{F}(I)\|}\right), \quad I_w = (I_n - \mu_n)\,\frac{\sigma_o}{\sigma_n} + \mu_o, \tag{2}$$

where I, I_n, and I_w represent the original, normalized, and whitened images, respectively. μ_o, μ_n and σ_o, σ_n are the mean and standard deviation of the original image and the normalized image, respectively. $\mathcal{F}^{-1}$ denotes the inverse Fourier transform.

On the contrary, we constrained ImageNet model to exclusively learn low-frequency patterns by eliminating high-frequency details from the ImageNet images. Due to the missing fine details between sub-classes, we merged similar classes based on hierarchy, reducing the number of classes to three: *living thing*, *artifact*, and *miscellaneous*, to guarantee convergence. The performance of models pre-trained on modified datasets is illustrated in Fig. 4, with their learning priorities in Fig. 3 (third row).

As expected, the model pre-trained on whitened RadImageNet no longer shows low learning priority in high frequencies and picks up the shortcut during fine-tuning. In contrast, the model pre-trained on filtered ImageNet

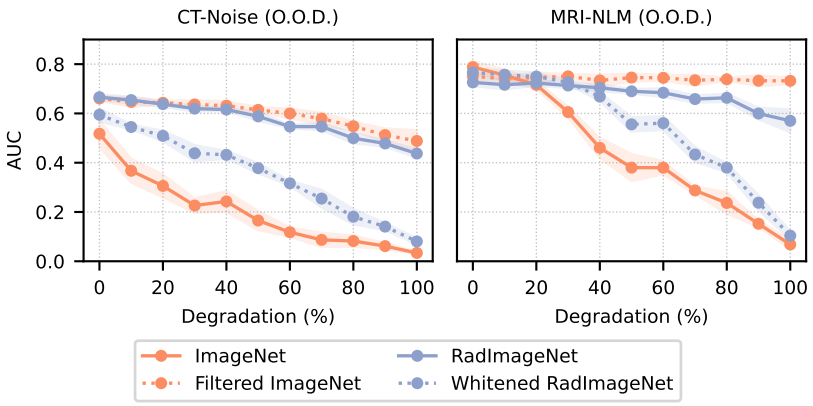

Figure 4. O.O.D. performance (mean and standard deviation of AUC across 5-folds) with (doted lines) and without (solid lines) source data editing.

has limited capability to learn high-frequency features, resulting in a learning priority similar to that of the model pre-trained on original RadImageNet and thereby achieving comparable or even improved robustness.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we discovered that a model's response to frequency shortcuts in transfer learning is influenced by the similarity between the spectral distribution of the shortcut and the learning priority of the pre-trained model. By modifying source data, we showed that it is possible to alter the fine-tuned model robustness against frequency shortcuts. Although frequency analysis is a promising technique for understanding model robustness in transfer learning, several questions remain. First, it is unclear how the statistics of the untouched source data may affect the model's learning priority during pre-training. Second, although we showed that fine-tuned model robustness can be altered, a fine-grained method to manipulate the model's PSD is preferred. Lastly, it would also be interesting to investigate other types of non-frequency confounders, such as patient gender, medical equipment, or markers, from the perspective of the frequency domain.

ACKNOWLEDGMENTS

This research was supported by National Institutes of Health under Grant NIH EY07977, and Novo Nordisk Foundation under Grant NNF21OC0068816.

REFERENCES

[1] Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F. A., "Shortcut learning in deep neural networks," *Nature Machine Intelligence* **2**(11), 665–673 (2020).

[2] Winkler, J. K., Fink, C., Toberer, F., Enk, A., Deinlein, T., Hofmann-Wellenhof, R., Thomas, L., Lallas, A., Blum, A., Stolz, W., et al., "Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition," *JAMA dermatology* **155**(10), 1135–1141 (2019).

[3] Pan, S. J. and Yang, Q., "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering* **22**(10), 1345–1359 (2010).

[4] Cheplygina, V., "Cats or cat scans: Transfer learning from natural or medical image source data sets?," *Current Opinion in Biomedical Engineering* **9**, 21–27 (2019).

[5] Raghu, M., Zhang, C., Kleinberg, J., and Bengio, S., "Transfusion: Understanding transfer learning for medical imaging," *Advances in neural information processing systems* **32** (2019).

[6] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L., "Imagenet: A large-scale hierarchical image database," in [*2009 IEEE conference on computer vision and pattern recognition*], 248–255 (2009).

[7] Xie, Y. and Richmond, D., "Pre-training on grayscale imagenet improves medical image classification," in [*Proceedings of the European conference on computer vision (ECCV) workshops*], 0–0 (2018).

[8] Ghesu, F., Georgescu, B., Mansoor, A., Yoo, Y., Neumann, D., Patel, P., Vishwanath, R., Balter, J., Cao, Y., Grbic, S., et al., "Self-supervised learning from 100 million medical images," *arXiv preprint arXiv:2201.01283* (2022).

[9] Mei, X., Liu, Z., Robson, P. M., Marinelli, B., Huang, M., Doshi, A., Jacobi, A., Cao, C., Link, K. E., Yang, T., et al., "Radimagenet: an open radiologic deep learning research dataset for effective transfer learning," *Radiology: Artificial Intelligence* **4**(5), e210315 (2022).

[10] Juodelyte, D., Lu, Y., Jiménez-Sánchez, A., Bottazzi, S., Ferrante, E., and Cheplygina, V., "Source matters: Source dataset impact on model robustness in medical imaging," *arXiv preprint arXiv:2403.04484* (2024).

[11] Xu, Y., Raj, A., and Victor, J. D., "Systematic differences between perceptually relevant image statistics of brain mri and natural images," *Frontiers in neuroinformatics* **13**, 46 (2019).

[12] Yin, D., Gontijo Lopes, R., Shlens, J., Cubuk, E. D., and Gilmer, J., "A fourier perspective on model robustness in computer vision," *Advances in Neural Information Processing Systems* **32** (2019).

[13] Wang, H., Wu, X., Huang, Z., and Xing, E. P., "High-frequency component helps explain the generalization of convolutional neural networks," in [*Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*], 8684–8694 (2020).

[14] Chen, G., Peng, P., Ma, L., Li, J., Du, L., and Tian, Y., "Amplitude-phase recombination: Rethinking robustness of convolutional neural networks in frequency domain," in [*Proceedings of the IEEE/CVF International Conference on Computer Vision*], 458–467 (2021).

[15] Abello, A. A., Hirata, R., and Wang, Z., "Dissecting the high-frequency bias in convolutional neural networks," in [*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*], 863–871 (2021).

[16] Lin, Z., Gao, Y., and Sang, J., "Investigating and explaining the frequency bias in image classification," *arXiv preprint arXiv:2205.03154* (2022).

[17] Wang, S., Veldhuis, R., Brune, C., and Strisciuglio, N., "Frequency shortcut learning in neural networks," in [*NeurIPS 2022 Workshop on Distribution Shifts: Connecting Methods and Applications*], (2022).

[18] Wang, S., Veldhuis, R., Brune, C., and Strisciuglio, N., "What do neural networks learn in image classification? a frequency shortcut perspective," in [*Proceedings of the IEEE/CVF International Conference on Computer Vision*], 1433–1442 (2023).

[19] Wang, S., Brune, C., Veldhuis, R., and Strisciuglio, N., "Dfm-x: Augmentation by leveraging prior knowledge of shortcut learning," in [*Proceedings of the IEEE/CVF International Conference on Computer Vision*], 129–138 (2023).

[20] Fridovich-Keil, S., Gontijo Lopes, R., and Roelofs, R., "Spectral bias in practice: The role of function frequency in generalization," *Advances in Neural Information Processing Systems* **35**, 7368–7382 (2022).

[21] Cheng, H., Yang, S., Zhou, J. T., Guo, L., and Wen, B., "Frequency guidance matters in few-shot learning," in [*Proceedings of the IEEE/CVF International Conference on Computer Vision*], 11814–11824 (2023).

[22] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D., "Grad-cam: Visual explanations from deep networks via gradient-based localization," in [*Proceedings of the IEEE international conference on computer vision*], 618–626 (2017).

[23] Xu, Q., Zhang, R., Zhang, Y., Wang, Y., and Tian, Q., "A fourier-based framework for domain generalization," in [*Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*], 14383–14392 (2021).

[24] Ngnawe, J., NJIFON, M. A., Heek, J., and Dauphin, Y., "Robustmix: Improving robustness by regularizing the frequency bias of deep nets," *arXiv preprint arXiv:2304.02847* (2023).

[25] Yucel, M. K., Cinbis, R. G., and Duygulu, P., "Hybridaugment++: Unified frequency spectra perturbations for model robustness," in [*Proceedings of the IEEE/CVF International Conference on Computer Vision*], 5718–5728 (2023).

[26] He, K., Zhang, X., Ren, S., and Sun, J., "Deep residual learning for image recognition," in [*Proceedings of the IEEE conference on computer vision and pattern recognition*], 770–778 (2016).

[27] Leuschner, J., Schmidt, M., Baguer, D. O., and Maass, P., "Lodopab-ct, a benchmark dataset for low-dose computed tomography reconstruction," *Scientific Data* **8**(1), 109 (2021).

[28] Armato III, S. G., McLennan, G., Bidaut, L., McNitt-Gray, M. F., Meyer, C. R., Reeves, A. P., Zhao, B., Aberle, D. R., Henschke, C. I., Hoffman, E. A., et al., "The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans," *Medical physics* **38**(2), 915–931 (2011).

[29] Štajduhar, I., Mamula, M., Miletić, D., and Uenal, G., "Semi-automated detection of anterior cruciate ligament injury from mri," *Computer methods and programs in biomedicine* **140**, 151–164 (2017).

[30] Manjón, J. V., Carbonell-Caballero, J., Lull, J. J., García-Martí, G., Martí-Bonmatí, L., and Robles, M., "Mri denoising using non-local means," *Medical image analysis* **12**(4), 514–523 (2008).

[31] Rahaman, N., Baratin, A., Arpit, D., Draxler, F., Lin, M., Hamprecht, F., Bengio, Y., and Courville, A., "On the spectral bias of neural networks," in [*International conference on machine learning*], 5301–5310 (2019).