Can Biases in ImageNet Models Explain Generalization?

INSTITUTE FOR MACHINE LEARNING AND ANALYTICS
UNIVERSITÄT MANNHEIM

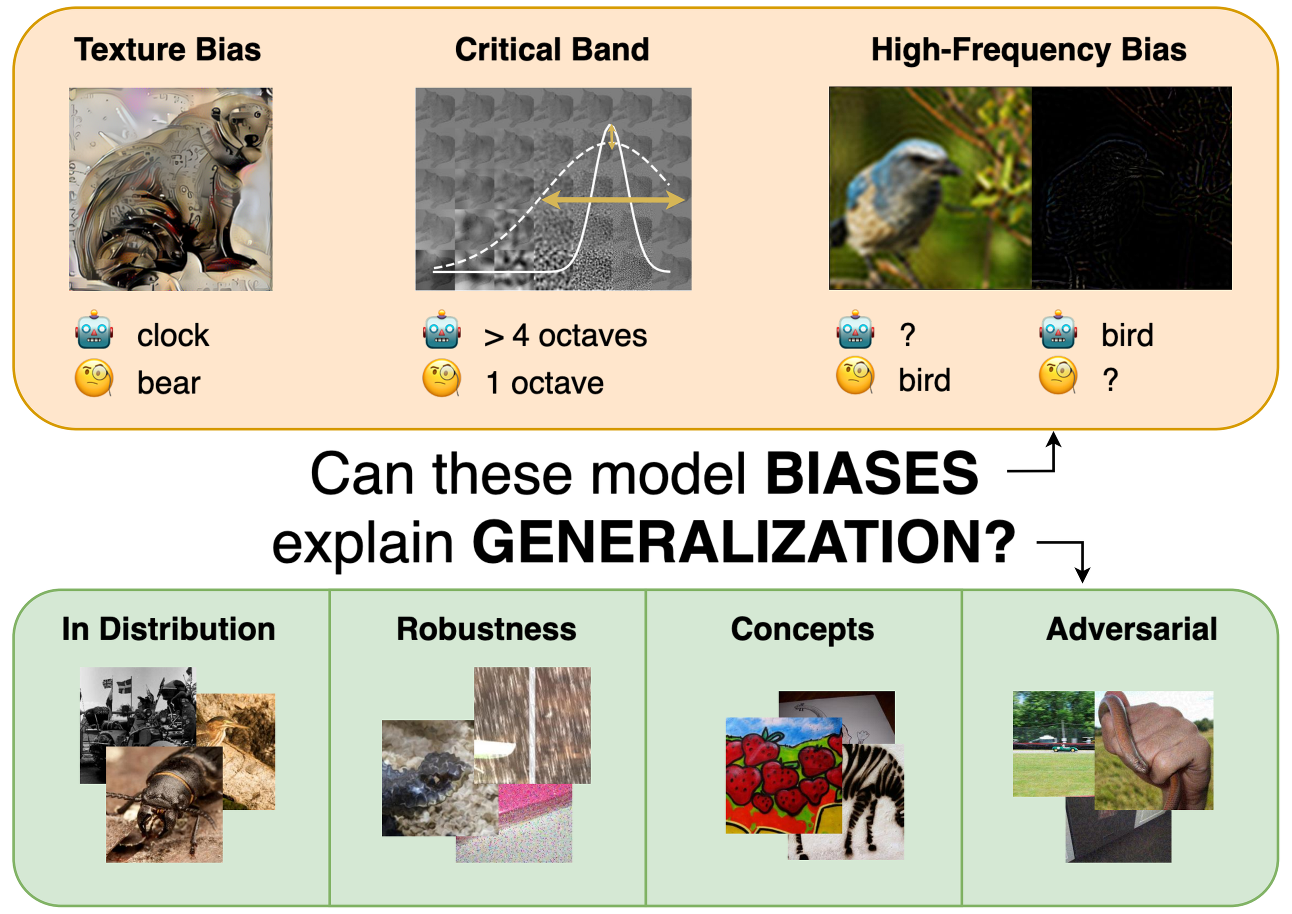
Paul Gavrikov [1] Janis Keuper [1,2]

[1]IMLA, Offenburg University [2]University of Mannheim

GitHub

CVPR SEATTLE, WA JUNE 17-21, 2024

Study Overview



ImageNet-trained models struggle to generalize beyond the training data and are often misaligned with human vision (biased). Aligning these biases was often suggested to improve robustness [1,2,3].

But is perception alignment really the missing key to generalization?

We study 3 recent biases on a diverse zoo of pretrained models for deeper insights.

3 Biases

- **Texture/Shape Bias** [1]
- **Critical Band** [2]
- **Spectral Bias** (low/high-frequency biases) [3]

4 Dimensions of Generalization

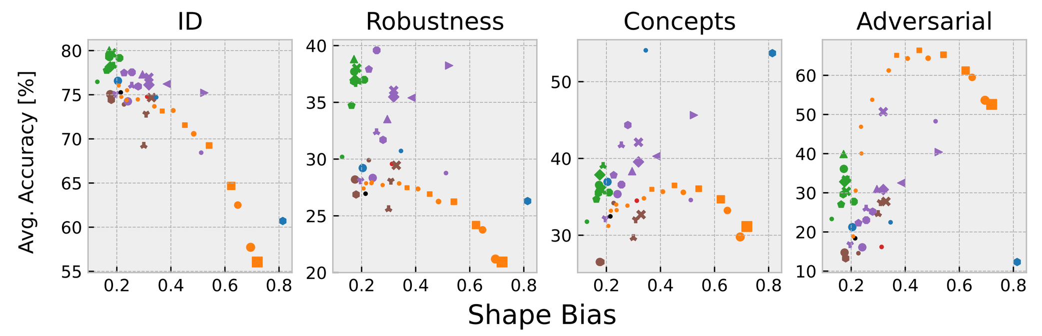
- **In Distribution:** ImageNet validation/v2/ReaL
- **Robustness:** ImageNet-C/Cbar/A
- **Concepts:** ImageNet-Renditions/Sketch and Stylized ImageNet
- **Adversarial:** (low-budget) PGD attack

48 Models

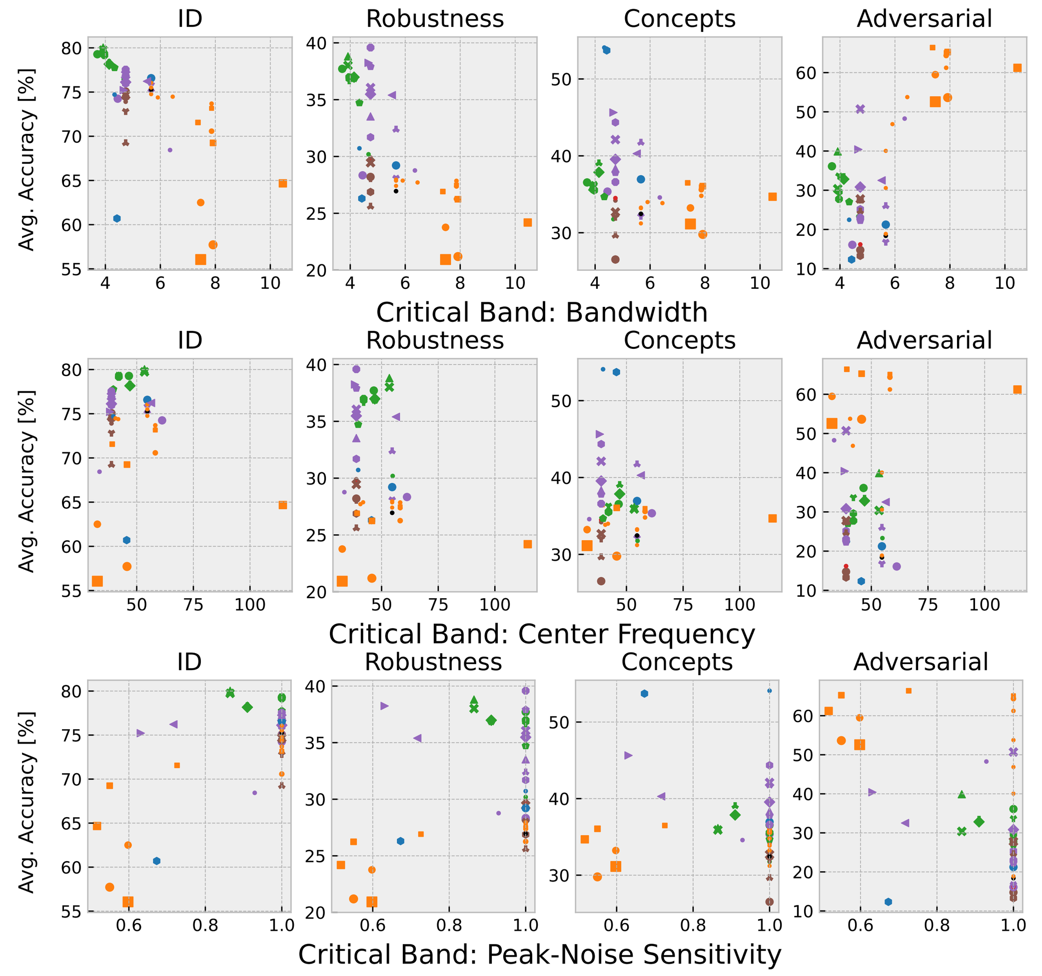
Identical architecture (ResNet-50) & training data (IN-1k) but different training methods:

- baseline of He et al.
- augmentation techniques
- (pre)training on Stylized ImageNet
- adversarial training (AT)
- contrastive learners
- improved training recipes
- randomly weighted convolutions

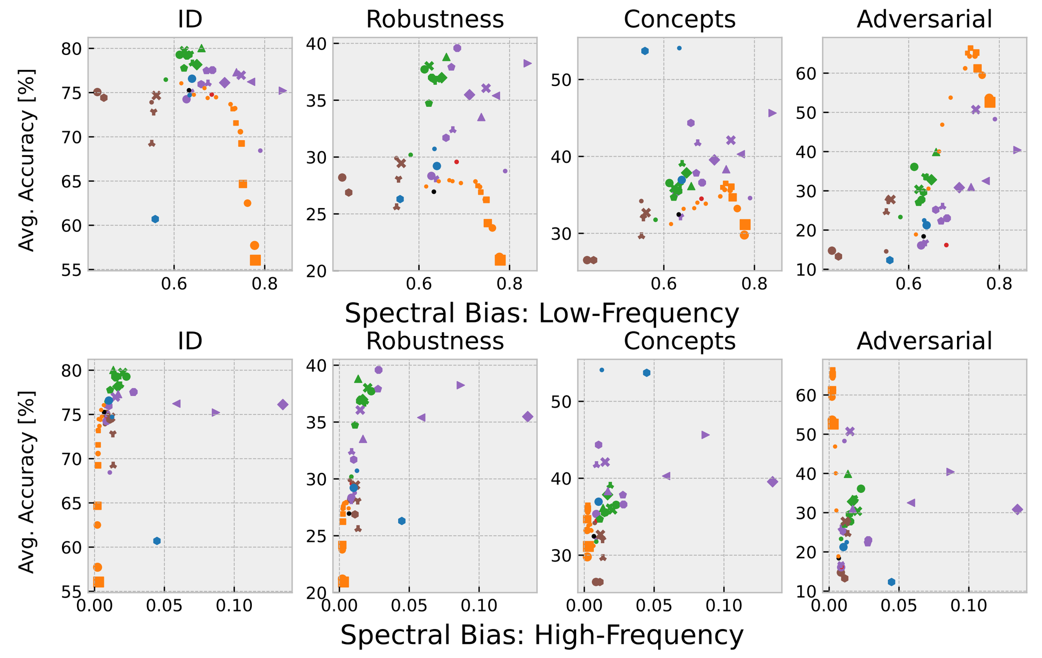
Texture/Shape Bias



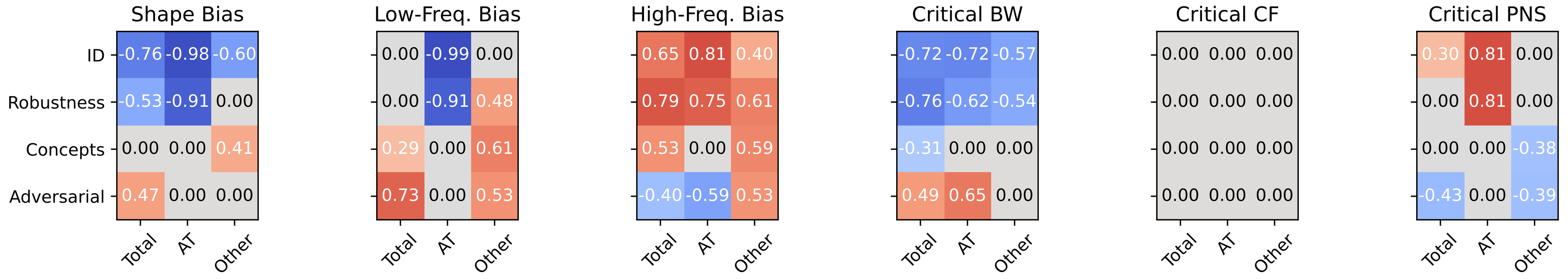
Critical Band



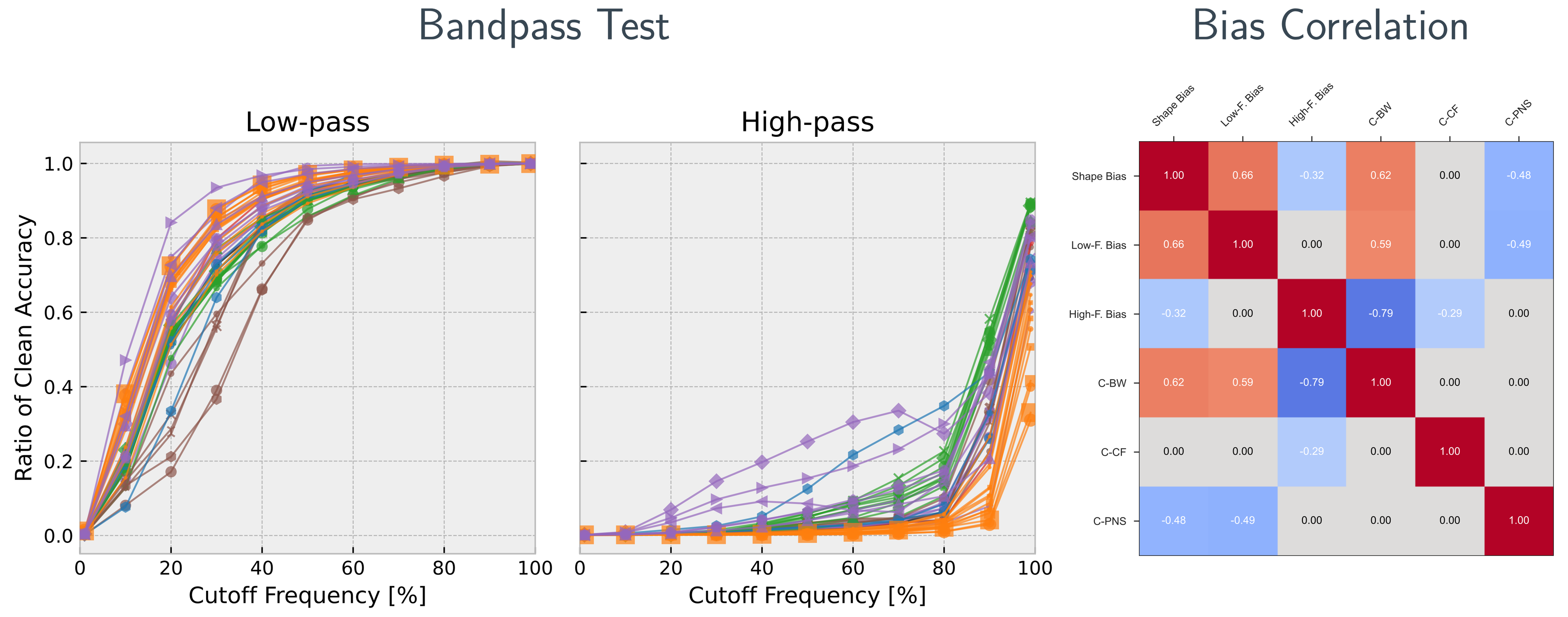
Spectral Bias



Key Findings & Suggestions



- **No single bias could predict generalization performance *holistically*.**
- At best, some biases correlate if we single out specific training or benchmarks.
- Stronger misalignment to human perception can improve performance (e.g., stronger texture bias improves in-distribution accuracy).
- High-frequency is not your enemy! Some level of HF detection seems helpful.



- We need better tests for alignment. E.g., shape bias as per [1] does not consider the accuracy; critical band [2] only uses 30 *random* samples per condition and does not work well for AT-models.
- Future studies should ablate *inductive biases* and *external data*.
- Pay attention to *system noise* and *consistent test transformations*!
- Evaluate models across a wide range of benchmarks that test different aspects of generalization.
- Limitation: this may only apply to ImageNet!

References

[1] R. Geirhos et al., “ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness”, ICLR, 2019.
[2] A. Subramanian et al., “Spatial-frequency channels, shape bias, and adversarial robustness”, NeurIPS, 2023.
[3] H. Wang et al., “High-frequency Component Helps Explain the Generalization of Convolutional Neural Networks”, CVPR, 2020.

Code: `https://github.com/paulgavrikov/biases_vs_generalization`