Email: as15003@nyu.edu
Twitter/X: ajaysub110
Website: ajaysubramanian.com

Spatial-frequency channels, shape bias, and adversarial robustness

What spatial frequencies do humans and neural networks use to recognize natural objects?

Ajay Subramanian, Elena Sizikova, Najib J. Majaj, Denis G. Pelli

NEURAL INFORMATION PROCESSING SYSTEMS **Oral**

NYU

Critical band masking of object recognition

The critical-band masking paradigm *(Fletcher, 1940; Solomon & Pelli, 1994)* characterizes the spatial frequency channel used for object recognition by measuring its sensitivity to frequency-filtered noise. Frequencies that are key to object recognition will be more affected by noise.

We measured thresholds of 14 human observers and 76 neural networks on 16-way object recognition of 1100 ImageNet images in the presence of frequency-filtered noise.

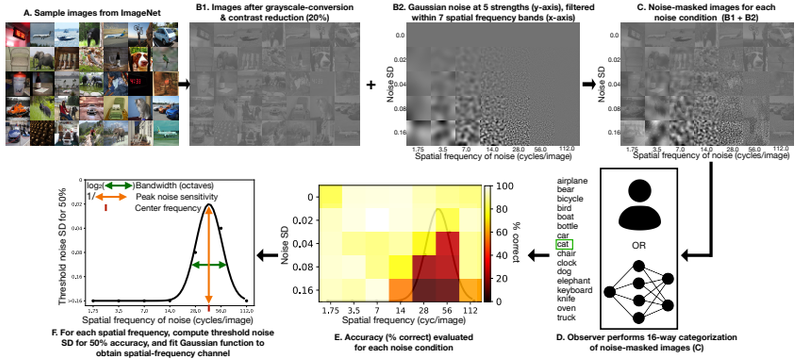
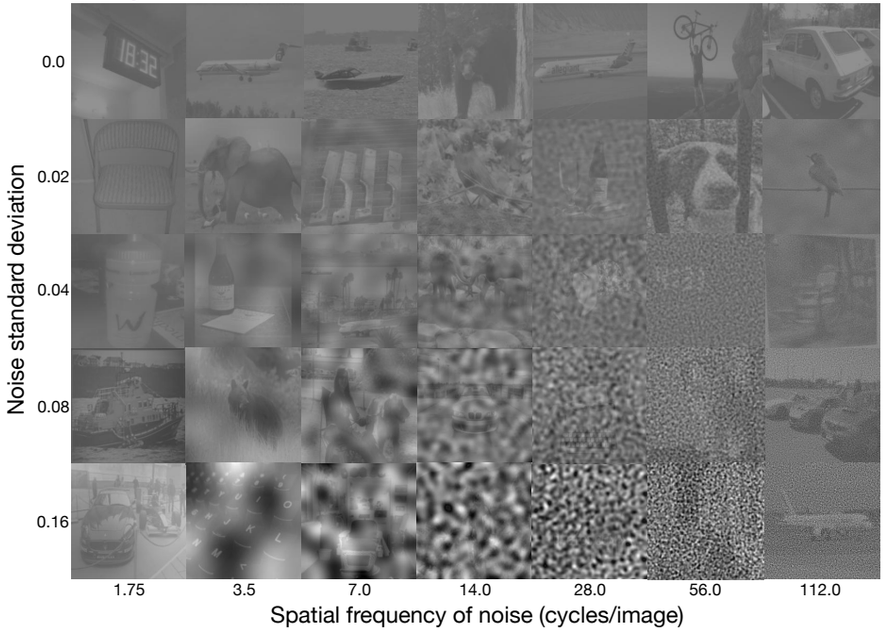


Figure 2. *Critical band masking stimuli, task, and analysis.*
A. 224 x 224 RGB images from ImageNet. **B1.** Images after grayscale-conversion and contrast reduction to 20%. **B2.** 224 x 224 Gaussian noise of 5 strengths filtered within 7 octave-wide spatial-frequency bands. **C.** Sample noise masked images used in our experiment: C = images in B1 + noise in B2.
D. Human or neural network observer performs 16-way categorization of noise-masked images.
E. Heat map showing % correct of a sample observer (human average) separately for each noise condition. **F.** Threshold noise SD for 50% accuracy is computed and Gaussian function is fit to obtain the observer's channel.

Figure 1. ***Demo:*** **Note how far down each column you can recognize objects. The edge of visibility (with an inverted-U shape) reveals the spatial frequencies that you use for recognition i.e., your** ***spatial frequency channel.***



Metrics

a. *Channel properties: center frequency, bandwidth, peak noise sensitivity.* Figure 2F shows the three properties that characterize the channel - **peak noise sensitivity** (1 / channel height), **center frequency** (frequency for peak noise sensitivity), and **bandwidth** in octaves ($\log_2$ full-width half-height). An octave is a doubling of frequency.

b. *Shape bias.* How strongly observers rely on shape features for categorization. We use the metric proposed by *Geirhos et al. (2019)* — % of shape-texture cue-conflict images classified by shape (0 - texture bias, 1 - shape bias)

c. *Adversarial robustness.* How susceptible neural networks are to adversarial perturbations, targeted noise that is known to severely impair network performance but is often imperceptible to human observers. We measure this using whitebox accuracy (low - not robust, high - robust)

Results & Conclusion

1. Humans recognize natural objects using the same 1-octave-wide spatial frequency channel that they use for letters, gratings, and faces, making it a canonical feature of human object recognition.
2. The neural network channel is 2-4 times wider than the human channel.

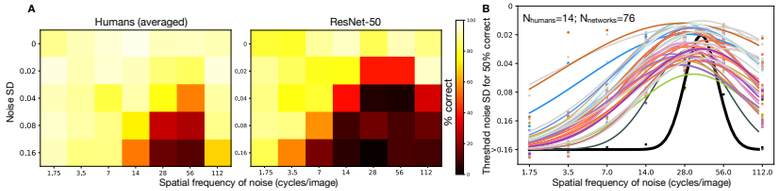


Figure 3. Accuracy heat maps and spatial frequency channels for humans (average) and 76 networks.

3. Channel properties correlate strongly with shape bias and with robustness of adversarially-trained networks.
4. Adversarial training further widens the already-too-wide network channel.

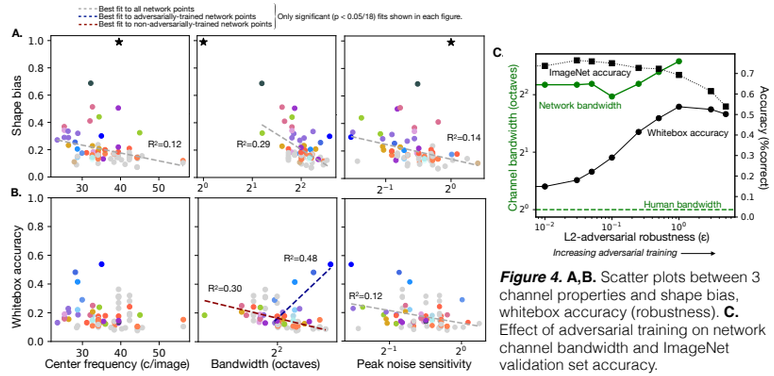


Figure 4. **A,B.** Scatter plots between 3 channel properties and shape bias, whitebox accuracy (robustness). **C.** Effect of adversarial training on network channel bandwidth and ImageNet validation set accuracy.

References

1. Fletcher, H. (1940). Auditory patterns. *Reviews of modern physics.*
2. Solomon, J. A., & Pelli, D. G. (1994). The visual filter mediating letter identification. *Nature.*
3. Geirhos, R. et al. (2019). ImageNet-trained CNNs … shape bias improves accuracy and robustness. *ICLR.*

Citation: Subramanian, A., Sizikova, E., Majaj, N., Pelli, D. G. (2023). Spatial-frequency channels, shape bias, and adversarial robustness. *Advances in neural information processing systems, 37.*