VFA: Vision Frequency Analysis of Foundation Models and Human

Mohammad-Javad Darvishi-Bayazi [1 2] **Md Rifat Arefin** [1 2] **Jocelyn Faubert** [2] **Irina Rish** [1 2]

Abstract

Machine learning models often struggle with distribution shifts in real-world scenarios, whereas humans exhibit robust adaptation. Models that better align with human perception may achieve higher out-of-distribution generalization. In this study, we investigate how various characteristics of large-scale computer vision models influence their alignment with human capabilities and robustness. Our findings indicate that increasing model and data size, along with incorporating rich semantic information and multiple modalities, significantly enhances models' alignment with human perception and their overall robustness.

1. Introduction

The deployment of machine learning models in real-world scenarios is challenging due to distribution shifts (Koh et al., 2021). Several methods have attempted to improve the out-of-distribution (OOD) generalization of models by learning robust representations (Gulrajani & Lopez-Paz, 2020). Humans, on the other hand, exhibit remarkable robustness to distribution shifts. It is argued that aligning models with human perception can enhance their robustness (Geirhos et al., 2019a; Fel et al., 2022).

To compare these two systems, we need a method to assess not only their performance but also their underlying mechanisms. Frequency analysis is a promising approach to studying human vision (Campbell & Robson, 1968), and recently, critical frequency band masking has been used to study Artificial Neural Networks (ANNs) (Subramanian et al., 2024). It has been shown that **humans** recognize objects in natural images using a narrow, one-octave-wide channel, which is consistent across various stimuli such as letters and gratings (Solomon & Pelli, 1994; Majaj et al., 2002), establishing it as a canonical feature of human object recognition. In contrast, **ANNs** utilize frequency channels that are $2-4$ times wider than those of humans (see Figure 1), making them sensitive to a broader range of frequencies (Subramanian et al., 2024) and therefore prone to failure in real-world applications.

In this study, we conduct an extensive exploration of numerous computer vision models to answer these questions: 1) *Are ANNs similar to humans in object recognition tasks?* 2) *Can modern computer vision models match or outperform humans amid frequency noise?* 3) *What factors contribute to their proximity to human performance?* 4) *Humans rely more on the shape of objects than their texture and models are texture-biased (Geirhos et al., 2018a). Therefore can Bandwidth (BW) predict shape bias?* 5) *Would decreasing the bandwidth to be closer to that of humans improve the robustness?*

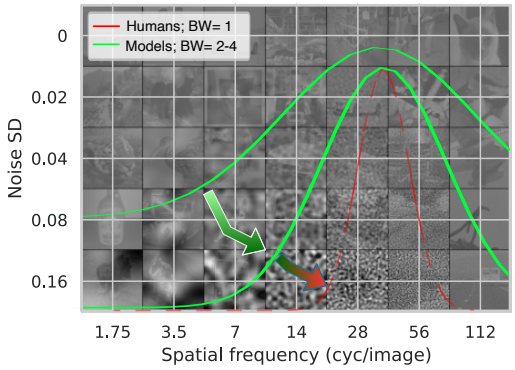

Figure 1. Frequency Bandwidths of Humans and Models. Humans are sensitive to a narrow frequency band, and adding noise within this band (under the red curve) degrades their performance. In contrast, models exhibit a wider frequency band (green curves), making them more vulnerable to noise across a broader range of frequencies. Narrowing the band might improve robustness.

2. Related work

Over the years, different methods have been proposed to address poor generalization under distribution shifts (Zhang et al., 2022; Du et al., 2020; Li et al., 2018; Sun & Saenko, 2016; Sagawa et al., 2019; Shi et al., 2021). However, the underlying principles for better generalization remain un-

[1]Mila, Qubec AI Institute, Montral, QC, Canada [2]Universit de Montral, Montral, QC, Canada. Correspondence to: Mohammad-Javad Darvishi-Bayazi <mj.darvishi92@gmail.com>.

Published at ICML 2024 Workshop on Foundation Models in the Wild. Copyright 2024 by the author(s).

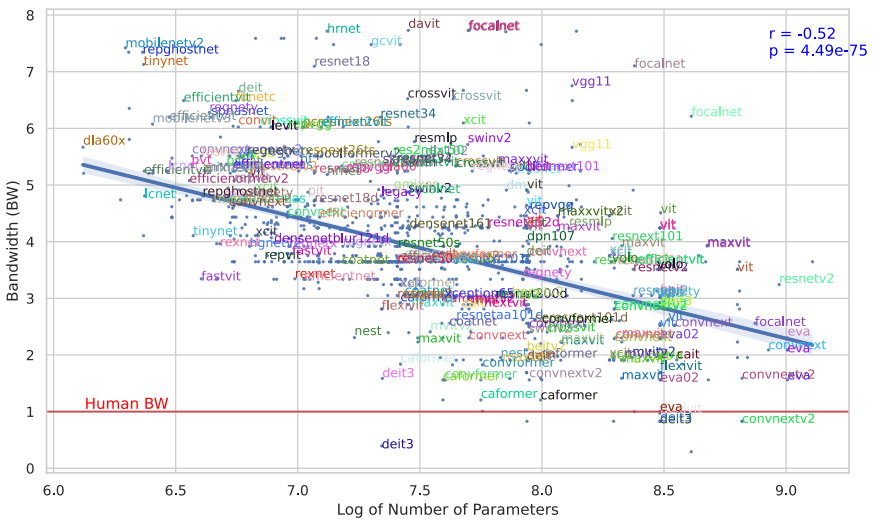

Figure 2. Correlation between Bandwidth (BW) and Model Size in Logarithmic Scale. The regression line represents that as the model size scales, the bandwidth decreases, converging towards human levels.

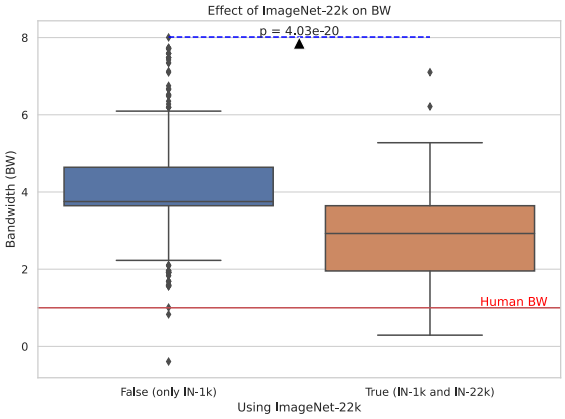

Figure 3. Effect of Data on Bandwidth: Comparing models trained with ImageNet-22K. Models that benefit from ImageNet-22K training exhibit significantly smaller bandwidths.

known. Comparison of the robustness of models to humans has also been studied, as humans show more robustness to distribution changes (Geirhos et al., 2021). Inspired by human robustness, Fel et al. (2022) propose a strategy to align models with human behavior. Recently, based on frequency analysis(Campbell & Robson, 1968) critical frequency band masking has been applied to models (Subramanian et al., 2024). Humans recognize objects in natural images using a narrow one-octave-wide channel, consistent with stimuli such as letters and gratings (Solomon & Pelli, 1994; Majaj et al., 2002), establishing it as a canonical feature of human object recognition. Our study is inspired by this frequency analysis to understand the behavior of the models and their robustness analysis based on different metrics such as OOD accuracy and shape bias as introduced in (Geirhos et al., 2021).

3. Methodology

We explore what characteristics of ANNs can close their gap with human performance. We follow the same procedure as (Subramanian et al., 2024), adding different spatial noise in various frequency bands and different noise standard deviations (SD) as shown in Figure 1. Then we evaluate the systems using these distorted images and fit a Gaussian curve to the point where they reach the 50% accuracy threshold. We calculate the bandwidth as the logarithm of the width at half-maximum in octaves. In this work, we tested **more than 1200** discriminative models available on HuggingFace timm (Wightman, 2019), multimodal zero-shot and fine-tuned CLIP (Radford et al., 2021; Ilharco et al., 2021; Cherti et al., 2023) models to analyze their bandwidth and robustness.

To examine the OOD accuracy and shape bias of models and humans, we utilized a comprehensive collection of 17 OOD datasets curated by Geirhos et al. (2021). These datasets contain 16 superclasses of ImageNet categories and include human responses. This benchmark also allows us to compare models with humans in object recognition in challenging OOD scenarios. The collection includes sketches, edge-filtered images, silhouettes, images with texture-shape cue conflicts, and stylized images with textures replaced by painting styles. Additionally, twelve datasets involve

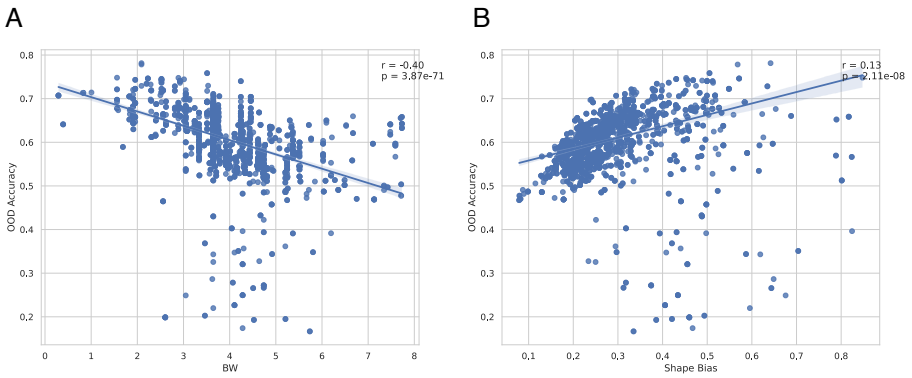

Figure 4. BW and Shape bias comparison. A) OOD versus Bandwidth. B) OOD versus Shape bias. Both BW and Shape bias are predictive of OOD performance and are correlated, with BW showing a higher correlation with OOD generalization.

parametric image degradation and varying factors such as noise and blur. Images within these OOD datasets were sourced from various datasets (Wang et al., 2019; Geirhos et al., 2019b; Wichmann et al., 2017; Geirhos et al., 2018b; 2019b). OOD accuracy is defined as the mean accuracy of a model across these 17 datasets, providing a comprehensive measure of OOD performance. Shape bias is the ratio of model accuracy on shapes to the sum of accuracy on shapes and textures in the cue-conflict dataset.

4. Results

In this work, we study the characteristics of different models regarding human alignment based on their parameter size, the dataset they were trained on, and their methods of learning. We also examine the relationship between human alignment metrics and the robustness of models.

4.1. Impact of Scaling on Frequency Bandwidth

Model Scaling. We conduct a comparative analysis of the frequency bandwidth of various models relative to humans by increasing the model sizes, irrespective of the underlying architecture, learning objective, or data augmentation methodologies. Figure 2 demonstrates that with the increase of model size (X-axis), there is a reduction in bandwidth (Y-axis), signifying a progression towards human-level performance. By extrapolating this trend line, we can predict that models with approximately 31 billion parameters have the potential to achieve a human-level one-octave bandwidth.

Data Scaling. We examine the effect of data scaling and training on a larger number of categories on the model's bandwidth. Figure 3 shows that models trained in ImageNet-22K, a data set with 22K labels, exhibit a bandwidth closer to that of humans. This highlights the importance of data scaling to achieve human-like capabilities in visual tasks.

Additionally, In Table 1 we observe that models trained on the LAION-2B (Schuhmann et al., 2022) dataset initially show a smaller bandwidth. However, fine-tuning these models on ImageNet-1K increases the bandwidth, and further fine-tuning on ImageNet-22K (with 22K classes) before final tuning on a subset of ImageNet-22K significantly improves the bandwidth, bringing it even closer to human levels (see Table 1).

4.2. Relationship of BW to OOD Accuracy

We also examine different metrics (frequency bandwidth, shape bias) that more accurately predict OOD accuracy. Figure 4 demonstrates that bandwidth serves as a superior predictor of OOD accuracy when considering all networks in the regression analysis. As the BW decreases (approach towards humans), OOD accuracy increases. This inverse relationship (with a strong negative correlation of $r = -0.4$) underscores the importance of bandwidth as a predictive metric for the generalization of OOD, compared to shape bias, which has a weaker positive correlation ($r = 0.13$).

4.3. Language Guidance leads to Human-like Bandwidth

We investigated models that show the most human-like performance as case studies. In Figure 5, we show BEiT families with different setups. BEiTv2 (Peng et al., 2022) uses a semantically rich visual tokenizer (distilling knowledge from multimodal pre-trained CLIP model) compared to the original BEiT (Bao et al., 2021). Integration of **semantic tokenization**, **fine-tuning on ImageNet-22K with**, and **model size** all contribute to aligning the model with the human frequency bandwidth.

In Figure 5B, several models **outperform humans** or have bandwidths close to human levels, warranting further exploration. Models using both convolution and attention

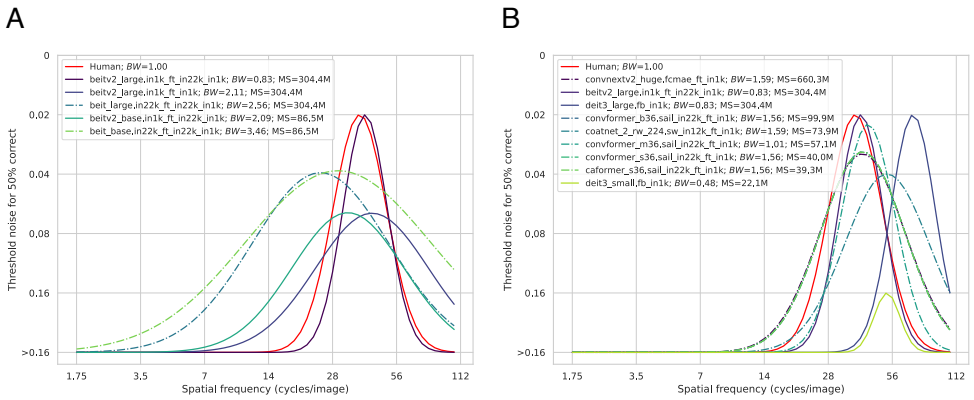

Figure 5. Case studies. (A) Comparing Bandwidth of BEiT Models. (B) Human-like models. Many models exhibit human-like bandwidth, with a version of the BEiT-V2 model perfectly matching the human curve.

mechanisms (Vaswani et al., 2017), such as *CoAtNet* (Dai et al., 2021) and *ConvFormer* (Yu et al., 2023) and *DeiT-III* (Touvron et al., 2022) with frequency-based data augmentation, and *ConvNeXt-V2* (Woo et al., 2023) with masked autoencoders and global response normalization show small bandwidths. More research is needed to understand these models, suggesting a direction for future studies.

5. Discussion

In this paper, our goal was to investigate various factors affecting the robustness of computer vision models and their alignment with human capabilities. We aimed to understand how model scaling, data scaling, semantic richness, data augmentation, large language model supervision, and multi-modality contribute to the performance of these models.

Our study revealed several key findings. Firstly, we observe that increasing the number of parameters through model scaling brings models closer to human performance. Furthermore, we find scaling up the training data through data scaling results in a decrease in bandwidth. Moreover, providing more detailed information about the data helps models learn better representations which echo the findings of (Hong et al., 2023). Furthermore, data augmentation with noise improves the robustness of models. Methods that use CLIP instead of supervised learning with labels, called large language model supervision, preserve more information and are robust against noise distortion. Finally, incorporating multiple modalities helps foundation models to learn semantics.

There are many illusions where human vision does not perceive the actual facts about an image (Anderson, 1997), suggesting that humans are prone to mistakes. This observation might indicate that a model capable of surpassing human performance could achieve superhuman vision. Additionally, it might signify that humans utilize a wealth of contextual information to perceive images. For instance, in the checker shadow illusion, humans interpret two squares with the same shade of gray as differently coloured white and black squares. This phenomenon highlights the complex and often non-literal nature of human visual perception, which incorporates contextual cues and prior knowledge to construct a coherent understanding of visual stimuli.

These findings contribute to our understanding of the factors that influence the performance of computer vision models and provide insight into improving their alignment with human capabilities.

6. Conclusion

Our results lead us to think that scaling foundation models might be the path to more robust machine learning models. However, several questions need to be answered: 1) While models are closing the gap with humans, what would be the next benchmark? Would new benchmarks such as OpenEQA (Majumdar et al., 2024) that evaluate models' capability performance in different aspects beyond object recognition be necessary? 2) In the future, can we simulate human vision with a foundation model? Can we use these models to cure and study the brain?

Impact Statement

In practical applications where computer vision models are utilized as human assistants, these systems must mimic human behaviours and show robustness. Our work compares these systems by exploring their characteristics. We aim to contribute to the development of AI systems that prioritize safety, reliability, and ethical considerations in real-world scenarios.

Acknowledgements

This work was funded by the Canada CIFAR AI Chair Program from the Canada Excellence Research Chairs (CERC) program and NSERC discovery grant RGPIN-2022-05122. We thank the Digital Research Alliance of Canada and Mila for providing computational resources.

References

Adelson, E. Checker-shadow illusion. retrieved august 27 2018, 1995.

Anderson, B. L. A theory of illusory lightness and transparency in monocular and binocular images: The role of contour junctions. *Perception*, 26(4):419–453, 1997.

Bao, H., Dong, L., Piao, S., and Wei, F. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

Campbell, F. W. and Robson, J. G. Application of fourier analysis to the visibility of gratings. *The Journal of physiology*, 197(3):551, 1968.

Cherti, M., Beaumont, R., Wightman, R., Wortsman, M., Ilharco, G., Gordon, C., Schuhmann, C., Schmidt, L., and Jitsev, J. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2818–2829, 2023.

Dai, Z., Liu, H., Le, Q. V., and Tan, M. Coatnet: Marrying convolution and attention for all data sizes. *Advances in neural information processing systems*, 34:3965–3977, 2021.

Du, Y., Zhen, X., Shao, L., and Snoek, C. G. Metanorm: Learning to normalize few-shot batches across domains. In *International Conference on Learning Representations*, 2020.

Fel, T., Rodriguez Rodriguez, I. F., Linsley, D., and Serre, T. Harmonizing the object recognition strategies of deep neural networks with humans. *Advances in neural information processing systems*, 35:9432–9446, 2022.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., and Brendel, W. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018a.

Geirhos, R., Temme, C. R., Rauber, J., Schütt, H. H., Bethge, M., and Wichmann, F. A. Generalisation in humans and deep neural networks. *Advances in neural information processing systems*, 31, 2018b.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., and Brendel, W. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019a. URL https://openreview.net/forum?id=Bygh9j09KX.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., and Brendel, W. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019b.

Geirhos, R., Narayanappa, K., Mitzkus, B., Thieringer, T., Bethge, M., Wichmann, F. A., and Brendel, W. Partial success in closing the gap between human and machine vision. *Advances in Neural Information Processing Systems*, 34:23885–23899, 2021.

Gulrajani, I. and Lopez-Paz, D. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020.

Hong, G. Z., Cui, Y., Fuxman, A., Chan, S. H., and Luo, E. Towards understanding the effect of pretraining label granularity. *arXiv preprint arXiv:2303.16887*, 2023.

Ilharco, G., Wortsman, M., Wightman, R., Gordon, C., Carlini, N., Taori, R., Dave, A., Shankar, V., Namkoong, H., Miller, J., Hajishirzi, H., Farhadi, A., and Schmidt, L. Openclip, 2021. URL https://doi.org/10.5281/zenodo.5143773.

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International conference on machine learning*, pp. 5637–5664. PMLR, 2021.

Li, D., Yang, Y., Song, Y.-Z., and Hospedales, T. Learning to generalize: Meta-learning for domain generalization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Majaj, N. J., Pelli, D. G., Kurshan, P., and Palomares, M. The role of spatial frequency channels in letter identification. *Vision research*, 42(9):1165–1184, 2002.

Majumdar, A., Ajay, A., Zhang, X., Putta, P., Yenamandra, S., Henaff, M., Silwal, S., Mcvay, P., Maksymets, O., Arnaud, S., et al. Openeqa: Embodied question answering in the era of foundation models. In *2nd Workshop on Mobile Manipulation and Embodied Intelligence at ICRA 2024*, 2024.

Peng, Z., Dong, L., Bao, H., Ye, Q., and Wei, F. Beit v2: Masked image modeling with vector-quantized visual tokenizers. *ArXiv*, abs/2208.06366, 2022. URL https://api.semanticscholar.org/CorpusID:251554649.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35: 25278–25294, 2022.

Shi, Y., Seely, J., Torr, P. H., Siddharth, N., Hannun, A., Usunier, N., and Synnaeve, G. Gradient matching for domain generalization. *arXiv preprint arXiv:2104.09937*, 2021.

Solomon, J. A. and Pelli, D. G. The visual filter mediating letter identification. *Nature*, 369(6479):395–397, 1994.

Subramanian, A., Sizikova, E., Majaj, N., and Pelli, D. Spatial-frequency channels, shape bias, and adversarial robustness. *Advances in Neural Information Processing Systems*, 36, 2024.

Sun, B. and Saenko, K. Deep coral: Correlation alignment for deep domain adaptation. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*, pp. 443–450. Springer, 2016.

Touvron, H., Cord, M., and Jegou, H. Deit iii: Revenge of the vit. *arXiv preprint arXiv:2204.07118*, 2022.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Wang, H., Ge, S., Lipton, Z., and Xing, E. P. Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems*, 32, 2019.

Wichmann, F. A., Janssen, D. H., Geirhos, R., Aguilar, G., Schütt, H. H., Maertens, M., and Bethge, M. Methods and measurements to compare men against machines. *Electronic Imaging*, 29:36–45, 2017.

Wightman, R. Pytorch image models. https://github.com/rwightman/pytorch-image-models, 2019.

Woo, S., Debnath, S., Hu, R., Chen, X., Liu, Z., Kweon, I. S., and Xie, S. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16133–16142, 2023.

Yu, W., Si, C., Zhou, P., Luo, M., Zhou, Y., Feng, J., Yan, S., and Wang, X. Metaformer baselines for vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

Zhang, X., Zhou, L., Xu, R., Cui, P., Shen, Z., and Liu, H. Towards unsupervised domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4910–4920, 2022.

A. Checker shadow Illusion

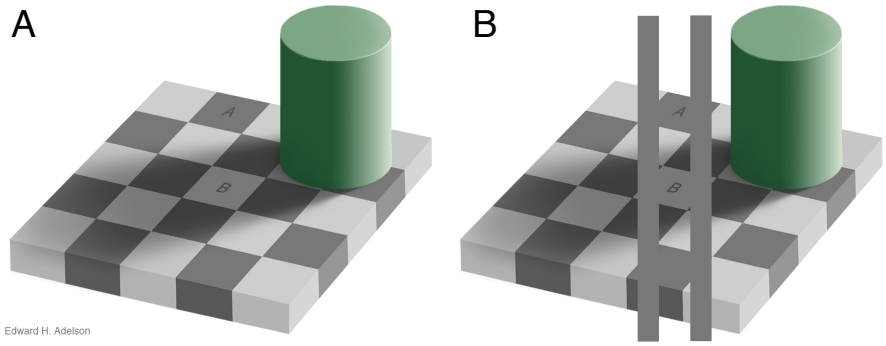

Figure 6. Checker Shadow Illusion (from (Adelson, 1995)). A) The squares marked A and B are actually the same shade of gray, yet they appear different due to the surrounding context. B) By connecting the squares marked A and B with vertical stripes of the same shade of gray, it becomes evident that both squares are indeed the same shade.

The checkerboard shadow illusion shows how context affects how we perceive brightness and colour. Due to the checkerboard pattern and shadows around them, two identically coloured squares in this illusion appear to be different hues. In particular, one perceives a square in shadow as lighter (white), whereas one perceives the same square in well-lit areas as darker (black). This illusion draws attention to the intricate processes underlying human vision, whereby the brain interprets visual stimuli using contextual knowledge, frequently resulting in incorrect perceptions.

B. Tables

Table 1 represents various setups of ViT-L/16 to elucidate why BEiT-V2 training achieves human-like behaviour and its connection to other metrics such as OOD accuracy, and shape bias. BEiT-V2 utilizes CLIP ViT-B/16 as a teacher for tokenization, resulting in superior accuracy and shape bias compared to BeiT and ViT-L/16. Analysis reveals that both CLIP supervision and training on ImageNet-22K contribute to bandwidth, with ImageNet-22K having a more pronounced effect. In the final section of the table, we compare OpenClip, trained on Laion-2B and then fine-tuned on subsets of ImageNet-22K (ImageNet-12K) and ImageNet-1K, and only fine-tuned on ImageNet-1K. The results demonstrate that fine-tuning on ImageNet-1K alone adversely affects bandwidth. Moreover, an increase in the number of labels improves shape bias, but the trend in OOD accuracy differs.

Table 1. Factors Contributing to Model Alignment with Human Bandwidth. The best performing configurations are highlighted in green and the second best in yellow. The results indicate that using CLIP ViT-B/16 for tokenization and training on ImageNet-22K enhances the performance of BEiT-V2 ViT-L/16, bringing it closer to the one-octave bandwidth characteristic of human vision.

Model	Z-Shot	CLIP	IN-1k	IN-22k	BW	OOD	Shape Bias
Humans					1.0000	0.7304	0.9600
CLIP ViT-B/16 [1]	✓	✓			2.7556	0.6950	0.4731
Original ViT-L/16			✓	✓	3.5121	0.7200	0.5381
BEiT ViT-L/16			✓	✓	2.5577	0.4898	0.4411
BEiT-V2 ViT-L/16		✓	✓		2.1084	0.7332	0.5364
BEiT-V2 ViT-L/16		✓	✓	✓	0.8285	0.7560	0.5610
OpenCLIP ViT-L/14	✓	✓			2.8895	0.6931	0.5665
OpenCLIP ViT-L/14 ft-IN1k		✓	✓		3.7526	0.7184	0.4738
OpenCLIP ViT-L/14 ft-IN12k-IN1k		✓	✓	✓	2.5012	0.7401	0.5121

[1] Teacher for BEiT-V2 tokenizer.