# Configural-Shape Representation in Deep Neural Networks

**Fenil R. Doshi (fenil_doshi@fas.harvard.edu)**
Department of Psychology, Harvard University
33 Kirkland Street, Cambridge, Massachusetts 02138

**Talia Konkle (talia_konkle@harvard.edu)**
Department of Psychology and Center for Brain Sciences, Harvard University
33 Kirkland Street, Cambridge, Massachusetts 02138

**George A. Alvarez (alvarez@wjh.harvard.edu)**
Department of Psychology, Harvard University
33 Kirkland Street, Cambridge, Massachusetts 02138

## Abstract

**We introduce a 'configural shape index' to quantify the quality of configural shape information in deep neural networks used to model human visual processing. Unlike shape-vs-texture bias measures (Geirhos et al. 2018), which capture the relative importance of shape in making classification decisions, our index captures the quality of shape representations in absolute terms (not relative to texture), and can be applied to any layer of any DNN model, regardless of model objective. Over a set of 92 models (including CNNs and transformers trained on a variety of tasks), we find low to modest sensitivity to configural shape, even in models with near human levels of shape-bias. These results suggest that there remains significant room for improving the quality of configural shape representations in DNN models of object recognition.**



**Keywords:** Shape bias; Holistic processing; Deep Neural Networks; Mid-level vision; Visual Perception


## Introduction

There is significant interest in the extent to which humans and deep neural network models (DNNs) rely on object shape when performing object-recognition tasks (e.g., Baker et al. 2018, Herman et al., 2020), with an emphasis on the degree of shape-vs-texture bias in humans versus DNNs (Geirhos et al. 2018) — human classification decisions are biased toward shape information, whereas most models are biased towards texture information. Bias, however, is orthogonal to the nature and quality of representations — for instance, it is possible to be 100% shape-biased when making only 1 correct shape-based decision and no other correct responses; (Doshi et al., 2024). Moreover, an emphasis on bias suggests that systems must be good at either shape or texture representation, but not both, when in fact the human perceptual system excels at both (e.g. Biederman, 1987; Julesz, 1975). Thus, it seems worthwhile to focus not only on whether decisions are dominated by shape or texture, but to focus on the nature and quality of shape (and texture) representations. In the present study, we attempt to answer the question — what qualities should a "good" shape representation possess, and how can we measure the **strength** of shape representations in DNNs used to model human visual processing?

Although there is no widely agreed upon formal definition of shape, information regarding object-shape should be preserved by affine transformations (translation, rotation, scale) since the exact same shape-defining features are depicted under each view. In contrast, object-identity (and presumably shape) are destroyed by "scrambling" transforms which alter the configuration of features in a way that is inconsistent with a "different view" of the same object. In the current study, we created a *Configural Shape Benchmark* that embodies these properties, quantifying the degree to which DNN model representations are tolerant to affine transformations while being intolerant to scrambling transformations (*configural-shape-index* = affine-transform tolerance minus scramble tolerance). This measure differs from shape-bias, which measures the relative importance of shape versus texture in making classification decisions. In contrast our *configural-shape-index* measures the strength of configural shape representations in absolute terms (not relative to texture), which allows for the possibility that a system simultaneously has strong shape and texture representations (Herman et al., 2020; Jagadeesh & Gardner, 2022; Long et al., 2018). Moreover, the configural-shape-index is computed using the intermediate activations of the model rather than at the output stage, and thus can be used on any layer of any model regardless of objective.

## Method

**Image Dataset.** We used the Imagenette validation set (Howard, 2019), which is a subset of the official Imagenet1k validation set, limited to 3,925 images sourced from 10 easily distinguished categories: Tench, English Springer, Cassette player, Chain saw, Church, French horn, Garbage truck, Gas pump, Golf ball, Parachute.

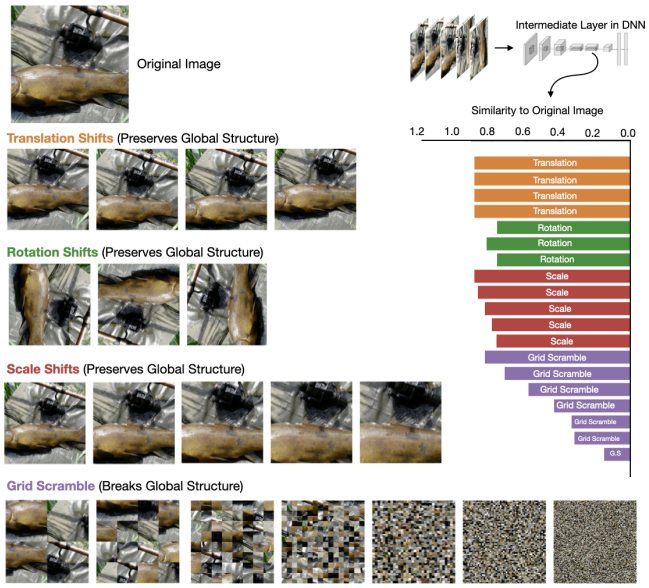


**Fig. 1**. Example translation, rotation and scale shifts in top three rows. Bottom row shows Grid Scrambling which breaks an image in a grid of different patch sizes and then shuffles the patches. On the right is the similarity between each transformed image and the original image within the DNN representations from an intermediate layer of Alexnet model.

**Critical Stimuli Manipulation:** As shown in **Fig. 1**, we systematically modify an image to either maintain or disrupt its global structure, while retaining local features at different scales (depending on the scramble grid size). The top three rows depict modifications through translation, rotation, and scaling, respectively, which preserve both global and local characteristics of the image. Conversely, the final row illustrates a scrambling process that parametrically disrupts the global structure: the image is segmented into grids of patches of varying sizes, with the ensuing patches then shuffled. This progressively degrades spatial coherence, with

the leftmost images retaining more of the local context due to larger patch sizes.

**Configural-Shape-Index:** To compute the *tolerance* to any given transformation within an intermediate model layer, we compute the Cosine Similarity between activations to the original image, and the transformed image (see **Fig. 1**). The goal is to give high scores to models that have high tolerance to affine transformations (translation, rotation, scale), but to penalize high tolerance to scrambling. A mean tolerance score is computed for translation (orange bars), rotation (green bars), and scaling (red bars) individually and the average aggregate of these scores represents the tolerance for shape-preserving transformations. For grid scrambles that disrupt configural shape (purple bars), we compute tolerance by deriving a Normalized Area-Under-the-Curve, giving greater weight to images with larger patches (i.e. more pixels in each patch). The difference between tolerance for shape-preserving transforms (average of translation, rotation, and scale tolerance) and shape-disrupting transforms is then taken as the configural-shape-index — a measure of the strength/quality of configural shape representation in a given model layer. The scores range from -1.0 to 1.0, 1.0 being a perfect configural shape representations (perfectly invariant to affine transforms, perfectly intolerant to any level of scrambling). Although scores of -1 are mathematically possible, they would indicate perfect tolerance to scrambling and zero tolerance to affine-transformation, and thus the effective range is between 0 and 1.

**Models:** We tested 92 models spanning a range of factors putatively impacting shape-representation in DNNs, including 6 standard feedforward object-recognition CNNs, 5 CNNs trained on Stylized ImageNet (Geirhos et al., 2018), 34 CNNs designed for robustness (Salman et al., 2020), 8 networks with constrained receptive fields (Doshi et al., 2023; Brendel & Bethge, 2019), 3 self-supervised CNNs (Chen et al., 2020), 9 trained with semi-supervised or semi-weakly supervised networks on expansive datasets (Yalniz et al., 2019), 10 Sparse Top-K Networks (Li et al., 2024), and 7 Vision Transformers (Dosovitskiy et al., 2020), which include a Masked Autoencoder variant (He et al., 2022), and lastly 10 AlexNet variations with distinct architectural or training modifications. For Vision Transformers, we assessed the configural-shape-index over the class token representation extracted from the last embedding layer, and for the remaining models, across all ReLU activation layers (focusing on the last ReLU layer for model comparisons).

## Results

Across the full set of models tested, untrained models have no sensitivity to configural shape, whereas trained models have at best a modest configural shape index score, with sensitivity increasing for deeper layers within trained models (see **Fig. 2a** and **Fig. 2b**). In general, we find that failure towards configural shape sensitivity arise because models are not tolerant enough to rotation, and are too tolerant to scrambling (**Fig. 2c**), particularly for large patch sizes.

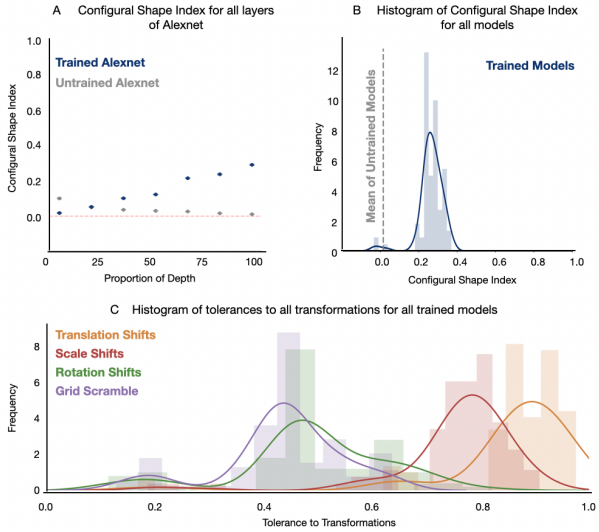


Fig. 2. (A) Configural Shape Indexes for Alexnet layers. (B) Histogram of Configural Shape Indexes for all models. Gray dashed line is the mean index for the 6 untrained models. (C) Tolerance to Affine Transformations (Translation, Rotation and Scale Shifts) and Scrambling.

**Relationship between configural-shape-index and shape-bias scores:** Our findings suggest that models with higher shape bias (Geirhos et al., 2018) are not necessarily also better at our configural shape processing (r=0.628). For example, the most shape-biased model has a shape-bias score of 0.814, but a configural-shape-index of only 0.249. While training Resnet50s with stylization dramatically improves their shape bias (from 0.214 to 0.814), it leads to only modest improvements on our configural shape index (0.2419 to 0.2497) Similarly, other approaches that increase shape bias exhibit similar effects: adversarial robust training increases shape bias by 0.511 but the configural shape index by only 0.08 in Resnet50; diffusion-guided training increases shape bias by 0.382 but configural shape by just 0.045 in Alexnet; and sparsity raises shape bias by 0.468 while actually reducing configural shape by 0.05 in Sparse Top-K networks. Thus, models that presumably emphasize shape over textural features on the shape bias metric, show no or only marginal advancements on the configural-shape-index, highlighting that configural shape-index serves as a distinct (and perhaps stricter) measure of shape-representation.

## Conclusion

We find that the proposed configural-shape-index of configural shape-quality is low across a wide range of DNN models trained with different architectures (convolutional neural networks, vision transformers), and training objectives (category supervised, self-supervised), even for models with substantially heightened shape-bias scores. Taken together, our findings reveal a general insensitivity to configural shape across models, even for models that show near-human levels of shape-bias (Geirhos et al., 2018; Salman et al., 2020; Jaini et al., 2023; Li et al., 2024), indicating that the lack of shape-based representation in DNNs (Baker et al., 2018) remains an important challenge for DNN models of object recognition.

## Acknowledgments

This work was supported by Kempner Graduate Fellowship to FRD, NSF CAREER BCS-1942438 to TK and NSF PAC COMP-COG 1946308 to GAA. Thanks to Spandan Madan for helpful feedback and insightful comments on the project.

## References

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

Baker, N., Lu, H., Erlikhman, G., & Kellman, P. J. (2018). Deep convolutional networks do not classify based on global object shape. PLoS computational biology, 14(12), e1006613.

Hermann, K., Chen, T., & Kornblith, S. (2020). The origins and prevalence of texture bias in convolutional neural networks. *Advances in Neural Information Processing Systems*, *33*, 19000-19015.

Doshi, F. R., Konkle, T., Alvarez G. A. (2024). Quantifying the Quality of Shape and Texture Representations in Deep Neural Network Models. *In Vision Science Society, 2024.*

Biederman, I. (1987). Recognition-by-components: a theory of human image understanding. Psychological review, 94(2), 115.

Julesz, B. (1975). Experiments in the visual perception of texture. Scientific American, 232(4), 34-43.

Jagadeesh, A. V., & Gardner, J. L. (2022). Texture-like representation of objects in human visual cortex. Proceedings of the National Academy of Sciences, 119(17), e2115302119.

Long, B., Yu, C. P., & Konkle, T. (2018). Mid-level visual features underlie the high-level categorical organization of the ventral stream. *Proceedings of the National Academy of Sciences*, *115*(38), E9015-E9024.

Howard, J. (2019). Imagenette. Github repository with links to dataset. https://github.com/fastai/imagenette

Salman, H., Ilyas, A., Engstrom, L., Kapoor, A., & Madry, A. (2020). Do adversarially robust imagenet models transfer better?. Advances in Neural Information Processing Systems, 33, 3533-3545.

Doshi, F., Konkle, T., Alvarez, G.A. (2023). Feedforward Neural Networks can capture Humanlike Perceptual and Behavioral Signatures of Contour Integration. In Cognitive Computational Neuroscience (CCN), 2023.

Brendel, W., & Bethge, M. (2019). Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. *arXiv preprint arXiv:1904.00760*.

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020, November). A simple framework for contrastive learning of visual representations. In *International conference on machine learning* (pp. 1597-1607). PMLR.

Yalniz, I. Z., Jégou, H., Chen, K., Paluri, M., & Mahajan, D. (2019). Billion-scale semi-supervised learning for image classification. *arXiv preprint arXiv:1905.00546*.

Li, T., Wen, Z., Li, Y., & Lee, T. S. (2024). Emergence of Shape Bias in Convolutional Neural Networks through Activation Sparsity. *Advances in Neural Information Processing Systems*, *36*.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., & Girshick, R. (2022). Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 16000-16009).

Jaini, P., Clark, K., & Geirhos, R. (2023). Intriguing properties of generative classifiers. *arXiv preprint arXiv:2309.16779*.