A Developmentally-Inspired Examination of Shape versus Texture Bias in Machines

Alexa R. Tartaglini, Wai Keen Vong, and Brenden M. Lake

{art481, waikeen.vong, brenden}@nyu.edu

New York University

Abstract

Early in development, children learn to extend novel category labels to objects with the same shape, a phenomenon known as the shape bias. Inspired by these findings, Geirhos et al. (2019) examined whether deep neural networks show a shape or texture bias by constructing images with conflicting shape and texture cues. They found that convolutional neural networks strongly preferred to classify familiar objects based on texture as opposed to shape, suggesting a texture bias. However, there are a number of differences between how the networks were tested in this study versus how children are typically tested. In this work, we re-examine the inductive biases of neural networks by adapting the stimuli and procedure from Geirhos et al. (2019) to more closely follow the developmental paradigm and test on a wide range of pre-trained neural networks. Across three experiments, we find that deep neural networks exhibit a preference for shape rather than texture when tested under conditions that more closely replicate the developmental procedure.

Keywords: shape bias, inductive bias, neural networks, word learning

Figure 1: **Two different approaches to measuring shape bias**. In the standard developmental procedure (left), an anchor stimulus (top) is presented and assigned a novel label. The participant is then asked whether a shape match (bottom left) or texture match (bottom right) has the same novel label as the anchor stimulus, and shape bias is calculated as the proportion of participants who prefer the shape match. In the shape bias procedure from Geirhos et al. (2019) (right), a cue-conflict stimulus with a cat shape and elephant texture is presented to a model, and a subset of the 1000 resulting ImageNet class probabilities (small circles) are averaged to attain scores for 16 basic categories. For example, probabilities for the ImageNet categories "Indian elephant" and "African elephant" are averaged to obtain a score for the basic category "elephant". All other ImageNet probabilities (dark gray circles) are ignored. The model's decision is considered shape biased if it predicts the shape category (cat), texture biased if it predicts the texture category (elephant), or is discarded if it predicts any other category.

Introduction

When presented with a new object and its label (e.g. "dax"), how do humans determine how to generalize the label to other objects? Starting around the age of two, children preferentially generalize novel category labels to solid objects of the same shape rather than the same size, color, or texture, a phenomenon known as the **shape bias** (Landau, Smith, & Jones, 1988). The standard experimental approach to measuring the shape bias in developmental psychology involves first presenting the participant with an anchor stimulus consisting of a novel shape and texture (see Figure 1(a) for an illustration). This is followed by the presentation of additional stimuli that match the anchor stimulus in either shape, texture, size, or color (but not in any other dimension). Participants are asked which of the additional stimuli are the same category as the anchor; choosing the stimulus that matches along the shape dimension at levels above chance indicates a shape bias. The emergence of the shape bias indicates a crucial developmental shift in which children begin to recognize that shape information is a reliable indicator of object names, facilitating the acceleration of noun learning (Smith, Jones, Landau, Gershkoff-Stowe, & Samuelson, 2002; Diesendruck & Bloom, 2003; Gershkoff-Stowe & Smith, 2004). The shape bias strengthens as we grow older and gain more visual and linguistic experience; in fact, by the time we are adults, shape appears to serve as the primary predictor of our recognition of familiar categories (Biederman, 1995).

Due to its importance in understanding generalization and learning in humans, multiple computational accounts of the shape bias have been proposed. These include models based on associative learning (Samuelson, 2002; Colunga & Smith, 2005) or overhypotheses in hierarchical Bayesian models (Kemp, Perfors, & Tenenbaum, 2007), although one limitation is that these models used simplified feature representations for capturing shape rather than naturalistic images. In the past decade, following the impressive achievements of deep neural networks (DNNs) on tasks such as image classification (Krizhevsky, Sutskever, & Hinton, 2012), there has been a push to better understand how machines encode and process images and whether they can be used as computational accounts of human vision (Yamins et al., 2014; Geirhos et al., 2021). In particular, this has led to a renewed interest in the shape bias, with initial accounts demonstrating that DNNs did have a preference for shape. Ritter, Barrett, Santoro, and Botvinick (2017) showed that pre-trained convolutional neural networks (CNNs) preferred to categorize novel objects on the basis of shape rather than color, and Feinman and Lake (2018) found that simple neural networks could learn a shape bias from very few examples.

However, while this earlier work showed that DNNs displayed a shape bias, this may have been the result of limiting the comparison to shape versus color. More recent work has argued that deep neural networks actually rely on local texture
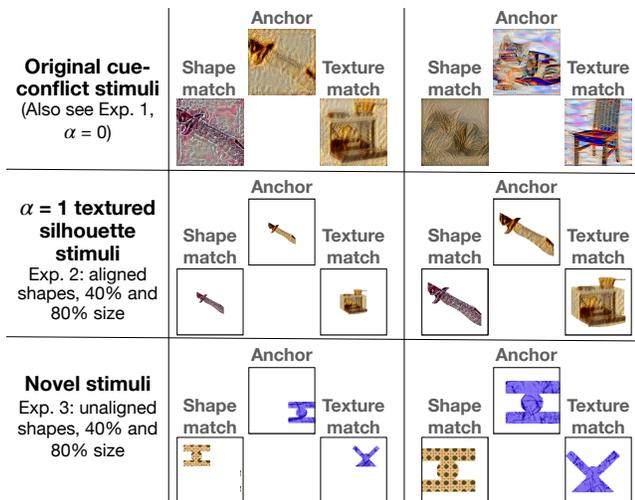
Figure 2: **Example triplets from several experimental conditions**. Each row consists of two triplet-based trials, each with an anchor stimulus, a shape match, and a texture match. The first row demonstrates two triplets created from the original cue-conflict stimuli. For the second and third rows, the triplets on the left demonstrate the 40% size condition, while the triplets on the right demonstrate the 80% size condition. The textured silhouette stimuli in the second row also demonstrate the aligned shape condition, while the novel stimuli in the third row demonstrate the unaligned shape condition.

information rather than global shape information for classification (Baker, Lu, Erlikhman, & Kellman, 2018; Brendel & Bethge, 2019; Geirhos et al., 2019). In particular, Geirhos et al. (2019) conducted a deeper examination by measuring shape versus texture bias in ImageNet pre-trained deep neural networks. They created a set of novel cue-conflict stimuli by overlaying the shape information of one category with the texture information of another category via style transfer (Gatys, Ecker, & Bethge, 2016). Shape bias was then calculated as the proportion of trials in which a model classified a cue-conflict stimulus as its shape category divided by the total number of classifications to either its shape or texture category (and similarly for texture bias; see Figure 1(b) for procedure details). Using this procedure, they found that ImageNet pre-trained CNNs were actually strongly biased towards texture, while humans tested using the same stimuli showed a strong preference for shape, highlighting a large discrepancy between human and machine behavior. In light of this finding, additional research has aimed to find ways of reducing the degree of texture bias in deep neural networks, such as work showing that shape bias can be increased via different kinds of data augmentations (Hermann, Chen, & Kornblith, 2020), or that vision transformers show less of a texture bias than CNNs (Tuli, Dasgupta, Grant, & Griffiths, 2021).

Despite the significant attention this finding has received, there are a number of important differences between the procedure used by Geirhos et al. (2019) and the procedure used in developmental psychology for assessing the shape bias, making it difficult to directly compare these recent findings with traditional findings. First, as shown in Figure 1, the style transfer method for creating the cue-conflict stimuli from Geirhos et al. (2019) produces stimuli where both the texture information may be inadvertently emphasized and the shape information de-emphasized, by covering the entire image—both the shape itself and the background—with the texture pattern. On the other hand, in developmental studies, the stimulus texture is always contained within the shape boundary, more closely reflecting how texture applies to objects in the real world; in fact, physical objects are often used rather than images of stimuli on a computer screen. Second, the procedure is limited to evaluating shape bias for 16 highly *familiar* categories (due to the use of a model's output layer from the set of 1000 ImageNet categories), whereas developmental studies focus on testing shape bias via *novel* shapes. Finally, the use of model classification outputs to determine a shape or texture decision for each individual stimulus differs from the typical procedure showing an anchor against shape or texture matches. In fact, the procedure from Geirhos et al. (2019) requires discarding a large proportion of trials that do not result in a correct shape or texture classification in their measure of shape bias.

Motivated by these differences, we re-examine the texture bias claim and outline an alternative procedure for measuring shape bias in artificial neural networks that is more closely aligned with the developmental procedure. In Experiment 1, we adapt the cue-conflict stimuli from Geirhos et al. (2019) to parametrically reduce the saliency of the background texture. Our main result shows that across all models tested, removing the influence of background texture results in a preference for shape over texture. In Experiment 2, we conduct additional tests to check the robustness of this finding by varying the size and positioning of the adapted stimuli, showing that this preference for shape over texture holds across different presentation conditions. Finally, in Experiment 3, we test the shape bias using a set of stimuli with novel shapes and textures, which most closely replicates the developmental procedure. We find that the overall degree of shape bias across models is lower for the novel stimuli compared to the adapted stimuli from Experiment 2. Furthermore, pre-trained models exhibit a higher rate of shape-based responses than their untrained counterparts, indicating that training increases shape bias. The shape bias testing code and datasets are available at: `https://github.com/alexatartaglini/developmental-shape-bias`.

Methods

Measuring Shape Bias. In this section, we describe the procedure used in the following three experiments to measure shape bias in the models. Following the developmental procedure from Landau et al. (1988), a trial consists of a triplet of image stimuli: an anchor stimulus with a given shape and texture, a shape match stimulus that shares only its shape with the anchor, and a texture match stimulus that shares only its texture with the anchor. We assembled a large number of

unique triplets for a given dataset by first considering each stimulus in the dataset as an anchor, then selected shape and texture matches for the anchor from other stimuli that have the same shape or texture class respectively.[1]

To process a trial, all three images in a triplet were passed individually to a given pre-trained model up to the penultimate layer to extract three embeddings of visual features. We then determined whether the model considered the shape or texture match to be more similar to the anchor by computing the cosine similarity between the anchor and the two matches.[2] If the cosine similarity between the anchor and shape match is higher, the trial is considered a shape decision. Otherwise, it is considered a texture decision. The resulting shape bias of a model is computed as the proportion of the number of its shape decisions to the total number of trials. Because no trials are discarded with this method, texture bias of the models equals the remaining number of texture decisions over the total number of trials.

Models. We measured shape bias for a variety of models with different architectures, types of supervision, and training data. This included a **ResNet-50** convolutional network (He, Zhang, Ren, & Sun, 2016) and a **ViT-B/16** vision transformer (Dosovitskiy et al., 2021). As a baseline, we first ran all experiments with 10 randomly-initialized ResNet-50s and ViT-B/16s then computed the average shape bias for each across all 10 models. These models are referred to as **random ResNet-50** and **random ViT-B/16** respectively. We also tested supervised variants of ResNet-50 and ViT-B/16 that were pre-trained on ImageNet (Deng et al., 2009), as well as a self-supervised ResNet-50 trained via **DINO** (Caron et al., 2021), that did not require labels during pre-training.

We also included **CLIP ViT-B/16** (Radford et al., 2021), which was pre-trained on a dataset of 400 million image-caption pairs via contrastive learning, and was recently been shown to be most comparable to human vision on a range of benchmarks (Geirhos et al., 2021). Finally, we tested **SAYCam-S** model (Orhan, Gupta, & Lake, 2020), which uses a different convolutional neural network (ResNeXt-50, Xie et al., 2017) and was trained using a self-supervised objective on a longitudinal egocentric dataset of headcam footage filmed from the perspective of one child sampled regularly from the age of 6 months to 32 months. We included this final model because the footage is recorded during the developmental period and in the typical environment that English-speaking children tend to acquire a shape bias (although this model does not receive the kind of labeled supervision thought to be important in acquiring the shape bias, Smith et al., 2002; Gershkoff-Stowe & Smith, 2004).

---

[1] Only shapes and textures from the same source image are considered to be a match; for example, if two stimuli both have the shape of a cat but use two different pictures of cats, they are not considered shape matches, and similarly for texture matches.

[2] We also ran all experiments using the dot product and Euclidean distance as alternative measures of perceptual similarity. We found that the results were very similar across all distance metrics used, so only results using cosine similarity are reported.

Experiment 1

In this experiment, we examined the possibility that the texture bias observed in convolutional neural networks may be due to an over-emphasis of texture in the cue-conflict stimuli used in Geirhos et al. (2019). Specifically, the texture information in these test stimuli is highly salient, covering both the shape and background and obscuring some of the underlying shape information (see the top row of Figure 2). This is in contrast to the stimuli used in the developmental setup, in which the texture is contained within the shape of a stimulus and presented on a white or neutral background.

Dataset. In order to bridge these visual differences, we modified the original cue-conflict stimuli created by Geirhos et al. (2019) by parametrically decreasing the opacity of the background texture in varying degrees. This was achieved by making use of the separate "filled silhouette" dataset (Geirhos et al., 2019), which contained a silhouette of each shape instance used in the original cue-conflict stimuli. The silhouette of each shape instance was used to create a mask of the background for each matching cue-conflict stimulus. These masks were then superimposed onto the corresponding cue-conflict stimuli, obscuring the texture background with white pixels and highlighting the shape of the stimulus. We refer to these modified stimuli as the "textured silhouette" stimuli. The opacity of the white background mask was controlled by a variable $\alpha \in [0, 1]$, where $\alpha = 0$ is equivalent to the original cue-conflict stimuli, and $\alpha = 1$ removes all background texture and replaces it with a white background. We created distinct datasets for 6 equally spaced α values, as shown in Figure 3.

All possible (anchor, shape match, texture match) triplets of stimuli were generated as described in the **Measuring Shape Bias** section for each α-valued textured silhouette dataset. In the original dataset used by Geirhos et al. (2019), stimuli with certain shapes and textures appeared more frequently than others, which resulted in a larger number of possible triplets with these anchor stimuli. To ensure that all shape and texture classes were equally represented in the final shape bias computation, we randomly selected 28 unique triplets for each of the $1,200$ anchor stimuli, producing a total of $33,600$ triplets used for evaluation. We repeated this procedure a total of 3 times, reporting the average measurements across replications.

Results and Discussion. When $\alpha = 0$ and the background texture is fully salient, we predictably find a pattern of results that mirrors the observations made by Geirhos et al. (2019). Across the board, the models are highly texture biased, with the notable exception of the random ResNet-50 and DINO ResNet-50, which are weakly shape biased (see Figure 3). Surprisingly, we observed an even stronger texture bias than Geirhos et al. (2019); for example, they observed about a 20% shape bias in a pre-trained ResNet-50 compared to the mere 2% shape bias we observed using the triplet-based procedure. As α increases and background texture salience decreases, all models displayed a monotonic increase in shape bias, al-
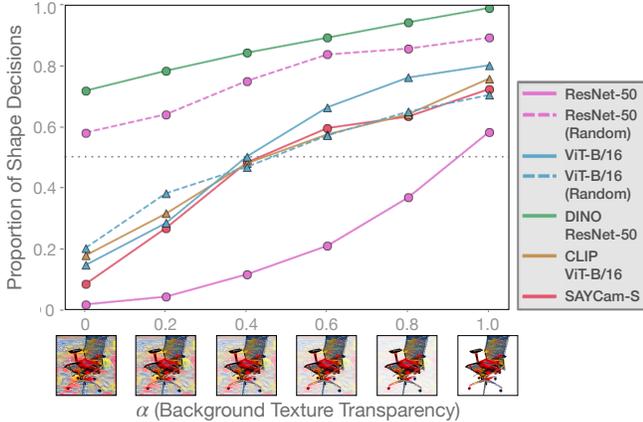
Figure 3: **Shape bias vs.** α. Under each α value on the x-axis, there is an example of a stimulus demonstrating the degree of background transparency. Each line represents a different model (see the legend to the right); CNN architectures are represented with circular markers, while vision transformers are represented with triangular markers. Results for randomly-initialized models are plotted with a dashed line. The horizontal, dotted gray line represents chance levels of shape bias; any model above this line is shape biased, while any model below is texture biased. All models show an increase in shape bias as the texture in the background becomes more transparent.

though at varying rates. The vision transformer models including the CLIP model and the SAYCam-S model showed a substantial increase in shape bias, and were between 70-80% shape biased when the background is completely white. On the other hand, the supervised ResNet-50 remained relatively texture biased throughout, only attaining a slight shape bias when α = 1. Still, this increase is surprising given the extremely strong texture bias observed for ResNet-50 when α = 0. DINO ResNet-50 is the most shape biased for all α values and attained nearly 100% shape bias when the background texture is completely removed. This may be the result of the data augmentations used when training DINO such as Gaussian blur and color augmentation, which have been shown to increase shape bias in CNNs (Hermann et al., 2020). Finally, random ResNet-50 also displayed a rather strong shape bias with low variance across seeds throughout, which is curious given its lack of training.[3]

These results demonstrate that the degree of shape bias using the cue-conflict stimuli from Geirhos et al. (2019) is variable across architectures and can be modulated by altering the background of the stimuli. In the developmental test, the background is irrelevant, so shape bias as measured in artificial neural networks should also reflect this in their set-up. However, there may be alternative explanations for the pattern of results we observed in this experiment. One possibility is that the higher degree of shape bias is an artifact of the perfect alignment in position and size of the shape information between the anchor stimulus and shape match stimulus.

---

[3]We also ran Experiment 1 for all supervised models using the original classification decision procedure from Geirhos et al. (2019) and found a very similar pattern of results.

In particular, this could explain the high degree of shape bias for the randomly-initialized ResNet-50, which may be especially sensitive to pixel overlap.

Experiment 2

In this experiment, we expand upon the results from Experiment 1 and examined whether the high level of shape bias observed in all models when α = 1 is robust to the size and position of the shape information in the stimuli. Real world objects vary in these factors, and thus, it is useful to understand how these properties influence model behavior.

Dataset. Starting with the subset of the textured silhouette stimuli with a white background (α = 1), we created five different stimulus **size** conditions ranging from 20% to 100% of their original size, with each triplet being uniform in size. We also created two positional **alignment** conditions. In the aligned shape condition, all stimuli occupied the center of an image, resulting in perfect overlap between the anchor and shape match stimuli. In the unaligned shape condition, all stimuli were individually placed in random locations of the image; thus, in a given triplet, the three stimuli occupy different parts of the image. Figure 2 provides examples of both the size and alignment variations.[4]

Results and Discussion. When the shapes of the stimuli are aligned, all models were highly shape biased across all stimulus sizes, although they were relatively more shape biased for smaller stimuli (Figure 4, top left). When the shapes were 20% of their original size, all models demonstrated a shape bias between 90% and 100% percent; for the largest stimuli, shape bias for the models ranged between 60% and 100%, with the ImageNet-trained ResNet-50 demonstrating the most significant decrease in shape bias. One possibility for this decrease is that shape bias decreased for larger shapes because a greater surface area allowed for a larger, more detailed patch of texture to be visible, thereby increasing the salience of the texture information. Another possibility is that the receptive field size or patch size in different models is smaller relative to the larger size conditions. When the shapes were randomly positioned in the image and no longer overlap in the shape unaligned condition, the models displayed more variation in their degree of shape bias but were still above the chance level for all stimulus sizes (Figure 4, top right).

These results show that a robust shape bias can be observed in all models by removing the background texture from the stimuli used by Geirhos et al. (2019) despite the different variations in size and positioning we explored in this simulation. This strengthens the findings in Experiment 1 by ruling out the hypothesis that pixel-overlap between the anchor and shape match are driving the preference for shape; rather, the sensitivity to shape is preserved despite the lack of pixel-level overlap.

---

[4]Note that the 100% size condition in Experiment 2 is technically always aligned and is equivalent to the α = 1 textured silhouette stimuli from Experiment 1, so the rightmost data points of Figure 3 are the same as the rightmost data points of the top two plots in Figure 4.
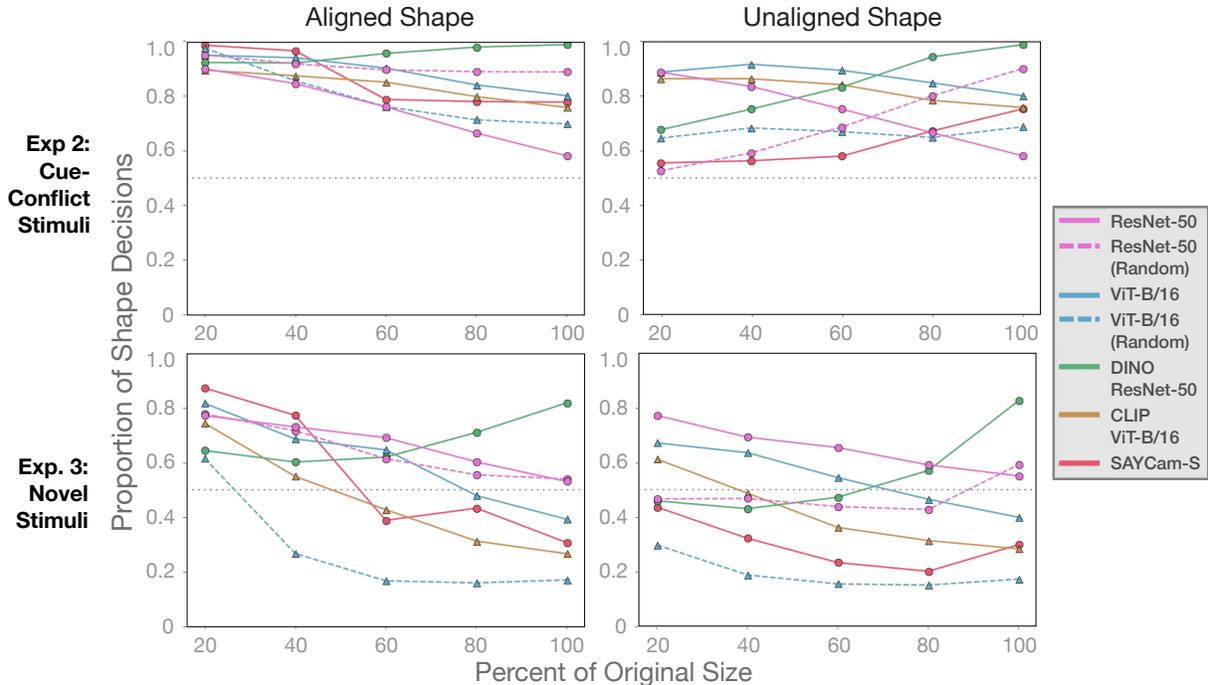
Figure 4: **Results from Experiments 2 and 3**. The rows indicate the experiment, while the columns represent the shape alignment condition with aligned shape results on the left and unaligned results on the right. **Top row**: using textured silhouette stimuli in Experiment 2, we find that all models are shape biased but respond differently to variations in shape size and alignment. **Bottom row**: using novel stimuli in Experiment 3, we find that all models are less shape-biased overall than Experiment 2. Unlike most of the models, DINO ResNet-50 increases in shape bias as shape size increases; also, for the unaligned shape condition, the random ResNet-50 hovers around chance levels of shape bias for all stimulus sizes with low variance.

Experiment 3

The previous two experiments demonstrated that we could produce a robust shape bias response in a variety of artificial neural networks using cue-conflict stimuli based on familiar categories. In this final experiment, we take one more step towards matching the developmental paradigm and measure shape bias using novel shapes and textures, differing from the categories the networks were trained on. This procedure is similar to that taken by Ritter et al. (2017) in which they also measured shape bias in DNNs using triplets of novel stimuli. However, while their work focused on testing shape versus color bias, our aim is to test shape versus texture bias.

Dataset. We created 256 unique novel stimuli using 16 high-quality textures (Brodatz, 1966) and 16 simple shapes (Parks, Griffith, Armstrong, & Stevenson, 2020) by overlaying random patches of each texture with a mask of each shape, resulting in misaligned textures between texture matches. As in Experiment 2, we varied the size and alignment of these novel stimuli, and generated all possible triplets as described in the **Measuring Shape Bias** section. Unlike Experiments 1 and 2, all anchor stimulus classes were equally represented in these datasets, so shape bias is measured once for a given condition using all 57,600 possible triplets. See Figure 1(a) and the bottom row of Figure 2 for examples of these stimuli.

Results and Discussion. When the novel stimuli are aligned in shape, all models were shape biased for the smallest size (see Figure 4, bottom-left). As shape size increased, most models showed a decrease in shape bias (whether aligned or not), leading some models to display a texture bias in the largest size conditions. To our surprise, DINO ResNet-50 was an exception, showing an increase in shape bias as the shapes become larger. It is not entirely clear why DINO behaves qualitatively differently to the other models we tested, and we leave this as an open question for future research.

When the stimuli were unaligned in shape, all models made fewer shape-based decisions when compared to the aligned stimuli (see Figure 4, bottom-right); the one exception was ResNet-50, which maintained the same pattern of behavior regardless of shape alignment. This drop may be due to the impoverished nature of the silhouettes compared to natural images, or the fact these stimuli are further out-of-distribution compared to pre-training. An advantage of this test setting is the neutral performance of the untrained random ResNet-50, which hovers around chance. Notably, ResNet-50, ViT-B/16, and CLIP ViT-B/16 all exhibited a shape bias relative to their untrained counterparts, indicating that they acquired more sensitivity to shape for novel stimuli during their training. This is consistent with the developmental picture, in which children acquire a shape bias as they learn more words (Smith et al., 2002; Gershkoff-Stowe & Smith, 2004).

Discussion

We outlined a more developmentally consistent procedure for testing shape bias in artificial neural networks. One advan-

tage of using embedding similarity instead of model output is that it allows for the measurement of shape bias for both pre-trained and randomly-initialized models, and for both familiar and novel stimulus categories. We measured shape bias using this procedure and three different types of cue-conflict test stimuli for a range of architectures with varying pre-training data and learning objectives. In Experiment 1, we augmented the cue-conflict stimuli used by Geirhos et al. (2019) by removing the textured background in varying degrees. We found that shape bias in all models increased as background texture salience decreased, demonstrating that the sensitivity of shape bias measurements could be explained due to characteristics of the cue-conflict stimuli. In Experiment 2, we tested whether the shape-based responses in Experiment 1 are robust to variations in shape size and position. We found that all models showed more shape-based than texture-based responses regardless of stimulus size or alignment. Finally, in Experiment 3, we generated triplets of stimuli with novel shapes and textures that varied in size and position over a white background. We found that the ImageNet pre-trained ResNet-50 and ViT-B/16 models exhibited a marked increase in shape bias over their randomly initialized variants, especially when the stimuli were randomly positioned, suggesting that these models had learned an inductive bias for shape that generalized to novel stimuli. Furthermore, shape bias across models was generally lower for the novel stimuli compared to the adapted stimuli containing familiar categories in Experiment 2.

Our results suggest that the previously observed texture bias in DNNs arises, in part, from how texture is emphasized in both the foreground and background (Geirhos et al., 2019). In fact, our results from Experiments 1 and 2 show that, despite variability in architecture and training, most models demonstrate a robust shape bias given stimuli with clean backgrounds when tested using a procedure that more closely aligns with developmental tests. However, there are still discrepancies between human and machine behavior that deserve more exploration. Geirhos et al. (2019) observed that humans tested using the original cue-conflict stimuli still showed a strong shape bias despite the exaggerated texture present in the stimuli. One possible explanation is that humans' shape decisions were influenced by the specific testing procedure used. In their experiments, despite the neutral instructions to not bias judgments towards shape or texture, participants selected their classification decisions by clicking on one of 16 shape-based icons (Geirhos et al., 2018). It is possible that the participants identified both the shape and texture classes in many stimuli, and since the task was somewhat ambiguous, the icons could have encouraged them to respond based on shape. Alternatively, models may be responding differently than people due to difficulties in segmenting the foreground of the images (i.e. the shape) from the background. Further work is needed to understand precisely why neural networks struggle with exaggerated background textures and if there are other ways for DNNs to match human behavior

on the original cue-conflict stimuli.

It is worth noting that both Baker et al. (2018) and Geirhos et al. (2019) performed other experiments in which they fill a silhouette (shape) from one class with a texture from another, similar to our $\alpha = 1$ condition in Experiment 1. Unlike our results, they find that texture is emphasized over shape. Their stimuli, however, are notably different. Filling a silhouette with a cutout from another image can not only erase internal features of the first category but can also introduce recognizable features from a different familiar category. These differences further highlight that the degree of shape bias in current models is highly dependent on task and stimulus details. Of the evaluations we consider, our Experiment 3 most closely mirrors the developmental procedure as run with children, although no single test seems to tell the whole story.

The developmental account of the interrelationship between the shape bias and word learning in children seems to align with our observations that models with pre-training on labels or linguistic data exhibit a stronger shape bias than their untrained counterparts. However, the strong shape preference exhibited by the self-supervised models in some conditions is somewhat puzzling from the developmental point of view given their complete lack of language exposure. It remains unclear why a robust shape bias is observed in all models regardless of their pre-training objective. Moreover, it is unclear why models display substantially more variation than people with regards to the shape bias.

As deep neural networks have continued to demonstrate impressive performance on visual tasks and have thus seen an expanding range of applications, it has become increasingly important to understand the nature of the biases they employ to classify objects and how these biases compare to those learned by humans. The recent explorations into the existence of a shape bias in DNNs have yielded a number of unexpected and significant results. In this paper, we add to this line of research by measuring shape versus texture bias in a range of DNN architectures using a procedure that more closely matches the developmental setup, highlighting that close attention needs to be paid to all aspects of the evaluation in order to conduct a "species-fair comparison" (Firestone, 2020). Ultimately, the shape bias is only one of many inductive biases employed by the human mind to understand and make sense of the world. Both machine learning and cognitive science stand to gain from further investigation into models that capture a range of inductive biases as well as methods for most fruitfully comparing human and machine intelligence.

Acknowledgments

References

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.

Biederman, I. (1995). Visual object recognition. In S. M. Kosslyn & D. N. Osherson (Eds.), *Visual cognition: An invitation to cognitive science* (Vol. 2, p. 121–165). MIT press Cambridge, MA, USA.

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

Brodatz, P. (1966). *Textures: a photographic album for artists and designers*. Dover publications.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., & Joulin, A. (2021). Emerging properties in self-supervised vision transformers. *arXiv preprint arXiv:2104.14294*.

Colunga, E., & Smith, L. B. (2005). From the lexicon to expectations about kinds: A role for associative learning. *Psychological Review*, *112*(2), 347.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *IEEE conference on Computer Vision and Pattern Recognition* (pp. 248–255).

Diesendruck, G., & Bloom, P. (2003). How specific is the shape bias? *Child Development*, *74*(1), 168–178.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... others (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Feinman, R., & Lake, B. M. (2018). Learning inductive biases with simple neural networks. In *Proceedings of the 40th Annual Conference of the Cognitive Science Society*.

Firestone, C. (2020). Performance vs. competence in human–machine comparisons. *Proceedings of the National Academy of Sciences*, *117*(43), 26562–26571.

Gatys, L. A., Ecker, A. S., & Bethge, M. (2016). Image style transfer using convolutional neural networks. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 2414–2423).

Geirhos, R., Narayanappa, K., Mitzkus, B., Thieringer, T., Bethge, M., Wichmann, F. A., & Brendel, W. (2021). Partial success in closing the gap between human and machine vision. In *Advances in Neural Information Processing Systems*.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2019). ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International conference on learning representations*.

Geirhos, R., Temme, C. R. M., Rauber, J., Schütt, H. H., Bethge, M., & Wichmann, F. A. (2018). Generalisation in humans and deep neural networks. In *Advances in Neural Information Processing Systems* (pp. 7549–7561).

Gershkoff-Stowe, L., & Smith, L. B. (2004). Shape and the first hundred nouns. *Child development*, *75*(4), 1098–1114.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition* (pp. 770–778).

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*.

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental Science*, *10*(3), 307–321.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems* (Vol. 25, pp. 1097–1105).

Landau, B., Smith, L. B., & Jones, S. S. (1988). The importance of shape in early lexical learning. *Cognitive Development*, *3*(3), 299–321.

Orhan, E., Gupta, V., & Lake, B. M. (2020). Self-supervised learning through the eyes of a child. In *Advances in Neural Information Processing Systems* (Vol. 33).

Parks, K. M., Griffith, L. A., Armstrong, N. B., & Stevenson, R. A. (2020). Statistical learning and social competency: The mediating role of language. *Scientific Reports*, *10*(1), 1–15.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... others (2021). Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.

Ritter, S., Barrett, D. G., Santoro, A., & Botvinick, M. M. (2017). Cognitive psychology for deep neural networks: A shape bias case study. In *International Conference on Machine Learning* (pp. 2940–2949).

Samuelson, L. K. (2002). Statistical regularities in vocabulary guide language acquisition in connectionist models and 15-20-month-olds. *Developmental Psychology*, *38*(6), 1016.

Smith, L. B., Jones, S. S., Landau, B., Gershkoff-Stowe, L., & Samuelson, L. (2002). Object name learning provides on-the-job training for attention. *Psychological Science*, *13*(1), 13–19.

Tuli, S., Dasgupta, I., Grant, E., & Griffiths, T. L. (2021). Are convolutional neural networks or transformers more like human vision? In *Proceedings of the 43rd Annual Conference of the Cognitive Science Society*.

Xie, S., Girshick, R., Dollár, P., Tu, Z., & He, K. (2017). Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition* (pp. 1492–1500).

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619–8624.