Journal of Experimental Psychology: General


ISSN: 0096-3445

2023, Vol. 152, No. 12, 3380–3402
https://doi.org/10.1037/xge0001440

# Human Shape Representations Are Not an Emergent Property of Learning to Classify Objects

Gaurav Malhotra[1], Marin Dujmović[1], John Hummel[2], and Jeffrey S. Bowers[1]
[1] School of Psychological Sciences, University of Bristol
[2] Department of Psychology, University of Illinois Urbana-Champaign

Humans are particularly sensitive to relationships between parts of objects. It remains unclear why this is. One hypothesis is that relational features are highly diagnostic of object categories and emerge as a result of learning to classify objects. We tested this by analyzing the internal representations of supervised convolutional neural networks (CNNs) trained to classify large sets of objects. We found that CNNs do not show the same sensitivity to relational changes as previously observed for human participants. Furthermore, when we precisely controlled the deformations to objects, human behavior was best predicted by the number of relational changes while CNNs were equally sensitive to all changes. Even changing the statistics of the learning environment by making relations uniquely diagnostic did not make networks more sensitive to relations in general. Our results show that learning to classify objects is not sufficient for the emergence of human shape representations. Instead, these results suggest that humans are selectively sensitive to relational changes because they build representations of distal objects from their retinal images and interpret relational changes as changes to these distal objects. This inferential process makes human shape representations qualitatively different from those of artificial neural networks optimized to perform image classification.

***Public Significance Statement***
This study shows significant differences in how the human visual system and recent artificial intelligence (AI) models represent objects. This difference between the two systems likely stems from their different goals—while AI models are built to classify objects, humans must additionally reason and interact with them. Our results suggest that the human visual system represents objects in a manner that enables us to perform these additional tasks.

*Keywords:* vision, convolutional neural networks, object recognition, shape representation, relational representation



A great deal of research into human vision is driven by the observation that visual perception is biased. For example, we prefer to group objects in a scene based on certain Gestalt principles—a bias to look for proximity, similarity, closure, and continuity (Ellis, 2013). We also prefer to view objects from certain viewpoints—a bias for canonical-perspectives (Palmer et al., 1981). This article is focused on one such bias—the *shape-bias*—the observation that humans, from a very young age, prefer to categorize objects based on their shape, rather than other prominent features such as color, size, or texture (Biederman & Ju, 1988; Landau et al., 1988). One manifestation of this bias is that we can identify most objects from line drawings as quickly and accurately as we can identify them from full-color photographs (Biederman & Ju, 1988) and we can do this even if we have no previous experience with line drawings (Hochberg & Brooks, 1962).

Two different explanations have been proposed regarding the origin of these biases. The first view, which we call the *heuristic approach*, proposes that biases originate because the visual system needs to transform the *proximal stimulus*—that is, the retinal image—into a representation of the *distal stimulus*—that is, a

This article was published Online First September 11, 2023.
Gaurav Malhotra https://orcid.org/0000-0002-6868-9655
This research was supported by the European Research Council Grant Generalization in Mind and Machine (741134). A preprint of this article has been made available at https://doi.org/10.1101/2021.12.14.472546 and a version of code for running the simulations reported in the manuscript as well as participant data from Experiment 4 is available at https://github.com/gammagit/distal. All code for generating the data sets, simulating the model as well as participant data from Experiment 4, can be downloaded from https://github.com/gammagit/distal.

Gaurav Malhotra served as lead for conceptualization, data curation, formal analysis, investigation, methodology, software, validation, and visualization. Jeffrey S. Bowers served as lead for funding acquisition, project administration, resources, and supervision and served in a supporting role for writing–review and editing. Gaurav Malhotra and Jeffrey S. Bowers contributed equally to writing–original draft and writing–review and editing. Marin Dujmović contributed equally to data curation and formal analysis. John Hummel and Jeffrey S. Bowers contributed equally to conceptualization.

Correspondence concerning this article should be addressed to Gaurav Malhotra, School of Psychological Sciences, University of Bristol, 12a Priory Road, Bristol, BS8 1TU, United Kingdom. Email: gaurav.malhotra@bristol.ac.uk

veridical representation of the cause of the stimulus. Of course, the simple act of transforming one representation to another should not necessarily lead to biases. But, in this case, mapping the retinal image to the distal stimulus is an ill-posed problem: there is not enough information in the proximal stimulus to unambiguously recover the properties of the distal stimulus (Nakayama et al., 1995; Pizlo, 2001). To overcome this problem, the visual system makes assumptions (i.e., employs heuristics) to determine which properties of the proximal stimulus are used to build distal representations (Knill, 1992; Mamassian & Landy, 1998; Pizlo & Stevenson, 1999; Stevens, 1981). A striking example of such assumptions is the Kanizsa triangle (Kanizsa, 1979), where the visual system encodes the multiple collinearities of edges present in the proximal image and uses these to build contours of a triangle even though these contours do not exist in the retinal image. The advantage of distal representations is that they are relevant for a broad range of tasks—the same representation of an object can be used for recognition and visual reasoning (Hummel & Biederman, 1992) among other visual skills.

A second view proposes that these biases can emerge as a result of internalization of the biases present in the environment relevant to classifying objects. According to this view, humans prefer to view objects from a canonical perspective because these perspectives most frequently enable us to tell objects apart, and they prefer to classify objects based on shape because the shape is more diagnostic during object classification. In other words, biases are a consequence of performing statistical learning on a large set of objects, with the goal of optimizing behavior on object classification. Under this hypothesis, human object representations are shaped by recognition alone, rather than being determined by a broad range of tasks, such as visual reasoning, object manipulation, and temporal prediction. We will call this the optimization-for-classification approach or, more briefly, the *optimization approach*.

The goal of this study was to test the second view—whether inferences about distal stimuli can emerge as a result of learning to classify a large set of objects. We tested this by focusing on supervised convolutional neural networks (CNNs)—which are machine learning models that recognize objects by learning statistical features of their proximal stimuli that can be used to optimally classify each stimulus, given some training data. Some recent studies suggest that CNNs can indeed show a shape bias provided they are trained to classify images using ecologically plausible training sets (Hermann et al., 2020). The learned representations that support object recognition in a supervised CNN are specialized for image classification. There is no pressure to learn distal representations of objects. As such, CNNs trained using supervised learning to classify objects provide a concrete model to test the optimization view. If human perceptual biases are acquired purely through internalizing the statistics of the environment in order to classify objects, then training CNNs to perform classification on ecologically realistic data sets should lead to perceptual shape-biases similar to the ones observed for humans.

Initial studies testing shape-bias in CNNs showed that CNNs trained in a supervised setting on large data sets of naturalistic images (e.g., `ImageNet`) frequently lacked a shape-bias, instead preferring to classify images based on texture (Geirhos et al., 2018) or other local features (Baker et al., 2018; Malhotra et al., 2022). However, it has been argued that CNNs can also be trained to infer an object's shape given the right type of training. For example, Geirhos et al. (2018) trained standard CNNs on a new image data set that mixes the shape of images from one class with a set of randomly chosen textures so that shape becomes more diagnostic of category. This new data set was constructed using the style-transfer algorithm (Gatys et al., 2016), where a subset of images from `ImageNet` were modified to match the style of 48 texture images. Geirhos et al. (2018) showed that CNNs trained on this data set (called "`Stylized-ImageNet`") learned to classify objects by shape. In another study, Feinman and Lake (2018) found CNNs were capable of learning a shape-bias based on a small set of images, as long as the training data was carefully controlled.

Similarly, Hermann et al. (2020) argued that texture bias reported for CNNs trained on `ImageNet` may be a consequence of how data are frequently augmented during training—input images are randomly cropped, which may remove shape information but preserve texture. They showed that instead of randomly cropping images, introducing more psychologically plausible forms of data augmentation (e.g., introduction of color distortion, noise, and blur to input images) make standard CNNs rely more on shape when classifying images. Indeed, the authors found that data augmentation was more effective in inducing a shape-bias than modifying the learning algorithms or architectures of networks, and concluded: "Our results indicate that apparent differences in the way humans and ImageNet-trained CNNs process images may arise not primarily from differences in their internal workings, but from differences in the data that they see" (Hermann et al., 2020, Abstract).

These results raise the possibility that human biases are indeed a consequence of internalizing the statistical properties of the environment relevant to classifying objects rather than the product of heuristics involved in building distal representations of objects. But studies so far have focused on judging whether or not CNNs are able to develop a shape-bias, rather than examining the type of shape representations they acquire. If humans and CNNs indeed acquire a shape-bias through a similar process of statistical optimization, then CNNs should not only show a shape-bias, but also develop shape representations that are similar to human shape representations.

A key finding about human shape representations is that humans do not give equal weight to all shape-related features. For example, it has been shown that human participants are more sensitive to distortions of shape that change relations between parts of objects than distortions that preserve these relations (Biederman, 1987; Hummel & Stankiewicz, 1996). These observations have typically been taken to support a heuristic view according to which relations present in the proximal images are used to build distal representations of objects (Hummel, 1994). The question we ask is whether CNNs trained to classify objects learn to encode these relational features of shape. If they do, it would suggest that the relational sensitivity of human shape representations can emerge as a consequence of learning to classify large sets of objects and that shape biases in object recognition are the product of optimizing performance on object classification alone. But if not, it would suggest that these biases are best characterized as heuristics designed to build distal representations of shape and that learning to classify objects is not sufficient for the emergence of such distal representations.

In the rest of the article, we discuss a series of experiments (simulation studies with CNNs as well as behavioral experiments with human participants) that show that the shape representations that

emerge as a result of classifying images in CNNs are qualitatively different from human shape representations. In the first two experiments, we examine objects that consist of multiple parts, while the following experiments examine objects that consist of a single part. The deformations required to infer the shape representations of these two types of objects are different, but related. Therefore, we begin each section by describing these deformations and how these deformations are predicted to affect shape representations under the two (optimization and heuristic) views. We then present the results of experiments where humans and CNNs were trained on the same set of shapes and then presented these deformations. In the final section, we discuss how our findings pose a challenge to developing models of human vision.

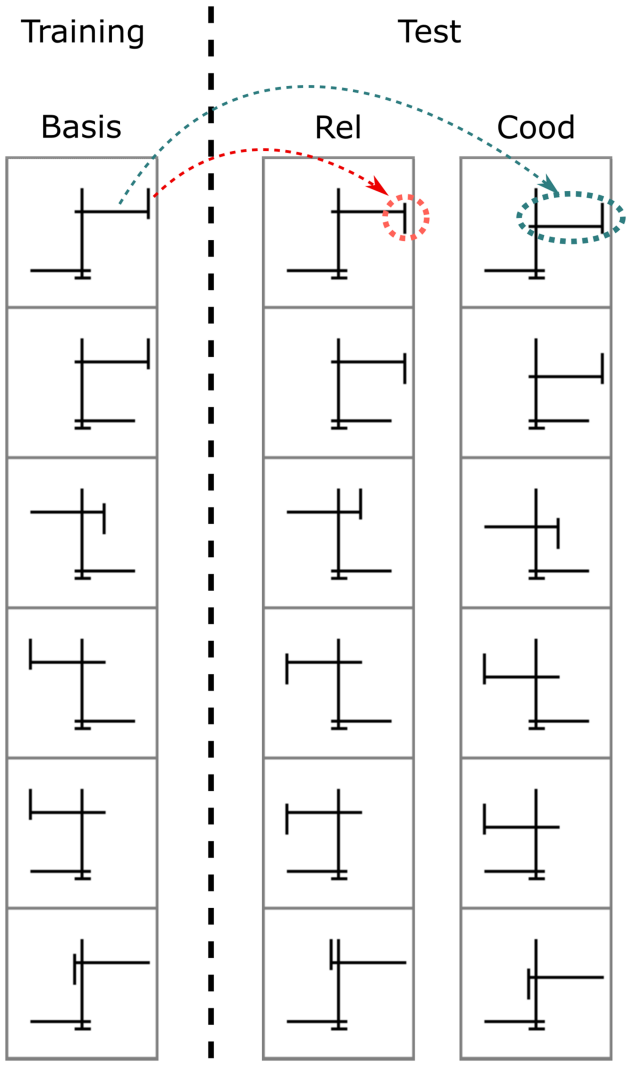
## Experiment 1

In our first experiment, we asked whether models that learn to optimize their performance by classifying large sets of objects develop a key property of human shape representations—its sensitivity to a subset of object deformations. According to the structural description theory (Biederman, 1987), humans represent objects as collections of convex parts in specific categorical spatial relations. For example, consider two objects—a bucket and a mug—both of which consist of the same parts: a curved cylinder (the handle) and a truncated cone (the body). The encoding of objects through parts and relations between parts makes it possible to support a range of visual skills. For example, it is possible to appreciate the similarity between a mug and a bucket because they both contain the same parts (curved cylinder and truncated cone) as well as their differences (the different relations between the object parts). That is, the representational scheme supports visual reasoning. In addition, the parts themselves are coded so that they can be identified from a wide range of viewing conditions (e.g., invariance to scale, translation, and viewing angle, as well as robustness to occlusion), allowing objects to be classified from novel poses and under degraded conditions.

Note that the reliance on categorical relations to build up distal representations of multipart objects is a built-in assumption of the model (one of the model's heuristics), and it leads to the first hypothesis we test, namely that image deformation that changes a categorical relation between an object's parts should have a larger impact on the object's representation than metrically equivalent deformations that leave the categorical relations intact (as might be produced by viewing a given object from different angles). By contrast, any model that relies only on the properties of the proximal stimulus might be expected to treat all metrically equivalent deformations as equivalent. Such a model may learn that some distortions are more important—that is, diagnostic—than others in the context of specific objects, but it is unclear why they would show a general tendency to treat relational deformations as different than metric ones since there is no heuristic that assumes that categorical relations between parts are a central feature of object shape representations. (Indeed, it may have no explicit encoding of parts at all.) Instead, all deformations are simply changes in the locations of features in the image.

Hummel and Stankiewicz (1996) designed an experiment to test this prediction of structural description theory and compare it to the prediction of view-based models (Poggio & Edelman, 1990) of human vision. They created a collection of shapes modeled on Tarr and Pinker's (1989) simple "objects." Each object consisted of a collection of lines connected at right angles (Figure 1). Hummel and Stankiewicz (1996) then created two deformations of each of these *Basis* objects. One deformation, the *relational* deformation (`Rel`), was identical to the Basis object from which it was created except that one line was moved so that its "above/below" relation to the line to which it was connected changed (from above to below or vice-versa). This deformation differed from the Basis object in the coordinates of one part and in the categorical relation of one part to another. The other deformation, the *coordinates* deformation (`Cood`), moved two lines in the Basis object in a way that preserved the categorical spatial relations between all the lines

**Figure 1**
*Stimuli Used by Hummel and Stankiewicz (1996)*




*Note*. The first column shows a set of six (`Basis`) shapes that participants were trained to recognize. Participants were then tested on shapes in the second and third columns, which were generated by deforming the Basis shape in the corresponding row. In the second column (`Rel` deformation) a shape is generated by changing one categorical relation (highlighted in red circle). In the third column (`Cood` deformation) all categorical relations are preserved but the coordinates of some elements are shifted (highlighted in blue ellipse). Rel = relational; Cood = coordinate. See the online article for the color version of this figure.

composing the object, but changed the coordinates of two lines. Note that both types of deformations can, in principle, indicate a change in distal stimulus. But, a system that uses relational changes as a heuristic for changes to distal stimuli will be more sensitive to Rel changes than Cood changes.

Across five experiments participants first learned to classify a set of base objects and then tested their ability to distinguish them from their relational (Rel) and coordinate (Cood) deformations. The experiments differed in the specific set of images used, the specific tasks, and the duration of the stimuli, but across all experiments, participants found it easy to discriminate the Rel deformations from their corresponding basis object and difficult to distinguish the Cood deformations. The effects were not subtle. In Experiment 1 (that used the stimuli from Figure 1) participants mistook the Rel and Cood images as the base approximately 10% and 90%, respectively, with similar findings observed across experiments. Hummel and Stankiewicz (1996) took these findings to support the claim that humans encode objects in terms of the categorical relations between their parts, consistent with the predictions of the structural description theories that propose a heuristic approach to human shape representation (Hummel, 1994).

However, an optimization approach may also be able to explain the findings of Hummel and Stankiewicz (1996)—a bias for perceiving objects in terms of parts and relations may simply *emerge* as a result of learning to classify objects. In Experiment 1, we tested this hypothesis by replicating the experimental setup of Hummel and Stankiewicz (1996), replacing human participants with two well-known CNNs—VGG-16 and AlexNet—that have been previously argued to capture human-like representations (Kriegeskorte, 2015; Yamins & DiCarlo, 2016) and an ability to develop a shape-bias (Geirhos et al., 2018; Hermann et al., 2020).

## Method

### Training Stimuli

We constructed six Basis shapes that were identical to the shapes used by Hummel and Stankiewicz (1996) in their Experiments 1–3. Each image was sized 196 × 196 pixels and consisted of five black line segments on a white background organized into different shapes. All images had one short (horizontal) segment at the bottom and one long (vertical) segment in the middle. This left three segments, two long, which were always horizontal, and one short, which was always vertical. The two horizontal segments could be either left-of or right-of the central vertical segment. Additionally, the short vertical segment could be attached to the left of or the right of the upper horizontal segment. This means that there were a total of eight ($2 \times 2 \times 2$) possible Basis shapes. We selected six out of these to match the six shapes used by Hummel and Stankiewicz (1996). Each training set contained 5,000 images in each category constructed using data augmentation, where the basis image was translated to a random location (in the range $[-50, +50]$ pixels) on the canvas and randomly scaled ([0.5, 1]) or rotated ($[-20°, +20°]$).

### Test Stimuli

Following Hummel and Stankiewicz (1996), we constructed Rel (relational) deformations of each Basis shape by shifting the location of the top vertical segment, so that its categorical relation to the upper horizontal segment changed from "above" to "below." Similarly, we constructed Cood (coordinate) deformations by shifting the location of *both* the top horizontal line and the short vertical segments together, so that the categorical relations between all the segments remained the same but the pixel distance (e.g., cosine distance) was at least as large as the pixel distance for the corresponding Rel deformation. The test set consisted of 1,000 triplets of basis, Rel and Cood images for each category, which were again generated using the same data augmentation method.

### Model Architecture and Pretraining

We evaluated two deep CNNs, VGG-16 (Simonyan & Zisserman, 2014) and AlexNet (Krizhevsky et al., 2017) on the image classification tasks described in the Results section. We obtained qualitatively similar results for both architectures. Therefore, we focus on the results of VGG-16 in the main text and describe the results of AlexNet in Appendix C. Since human participants had a lifetime experience of classifying naturalistic objects prior to the experiment, we used network implementations that had been pretrained on a set of naturalistic images. Two types of pretraining were used: networks were either pretrained in the standard manner on ImageNet (a large database of naturalistic images), or pretrained on a set of images where shape was made more predictive than texture by using style-transfer (Gatys et al., 2016). We used networks pretrained by Geirhos et al. (2018), who have shown that networks trained in this manner have a greater shape-bias than networks trained on ImageNet. For example, in their cue-conflict paradigm, Geirhos et al. (2018) observed that shape-bias for ResNet-50 increased from 22% for the network trained on ImageNet to 81% for the corresponding network trained on Stylized-ImageNet.

### Further Training

Networks were either tested in a *Zero-shot* condition, where no further training was given on any of our data sets and we recorded the response of the pretrained networks to the test images, or in a *fine-tuned* condition, where the pretrained network was fine-tuned to classify the 5,000 basis images of each category described in the stimuli above. In both the zero-shot and fine-tuned conditions, we replaced the last layer of the classifier to reflect the number of target classes in each data set. In the fine-tuned condition, the models learned to minimize the cross-entropy error by using the Adam optimizer (Kingma & Ba, 2014) with a small learning rate of $10^{-5}$ and a weight-decay of $10^{-3}$. In all simulations, learning continued until the loss function converged. To check for overfitting, we created cross-validation sets and ensured performance on the training set was not higher than on the cross-validation sets. We also trained networks using standard regularization methods such as batch normalization and dropout and obtained qualitatively similar results. In most cases, the networks achieved nearly perfect classification on the training set. All simulations were performed using the Pytorch framework (Paszke et al., 2017) and we used torchvision implementation of all models.

### Analysis of Internal Representations

To test the similarity of internal representations of basis images and their Rel and Cood deformations, we obtained the embedding of each image at each convolution and fully connected layer of the CNN. We used the same version of the network (i.e., the same set

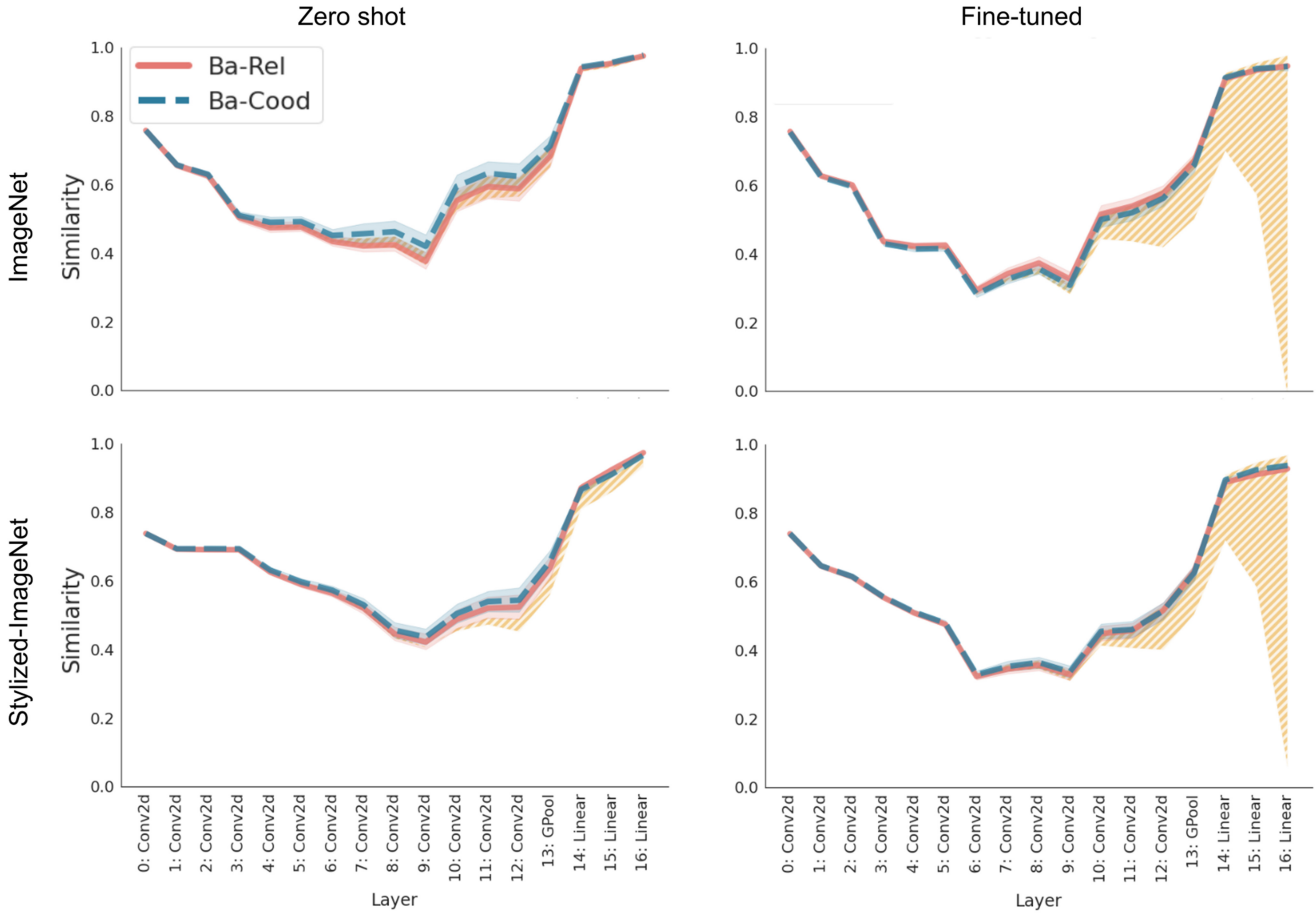
of learned weights) to obtain the embedding for each image. For a given category, we randomly sampled 100 pairs of images from the basis and `Rel` test sets and computed the cosine similarity between embeddings of each pair. We then averaged this distance across the six categories. This gave us the estimated average distance in the `Ba-Rel` condition. Similarly, the average cosine similarity across the six categories, obtained by computing the cosine similarity between 100 pairs of basis and `Cood` test images for each category gave us the `Ba-Cood` distance. These distances were compared against two baseline conditions. The upper limit of similarity was given by the average similarity of 100 pairs of basis images from the same category. The lower limit was given by the average similarity of 100 pairs of basis images from different categories (in each pair, one of the images was from one category and the other from one of the other six categories).

## Results and Discussion

An analysis of the internal representations of `VGG-16` is shown in Figure 2 and its classification performance is shown in Figure A1 in Appendix A. (Results for `AlexNet` followed the same qualitative pattern and are shown in Appendix C). The classification accuracy for all types of images in the *zero-shot* condition was at chance. This is not surprising since the new output layer (matching the number of target classes in our data set) had a set of random weights that had never been trained on our data set. But the internal representations in the zero-shot condition provided insight into how novel stimuli (from our study) were represented in networks pretrained on naturalistic images. An inspection of these internal representations showed that (left-hand column in Figure 2) the similarity between a basis image and its relational variant was the same on average (computed

**Figure 2**

*Cosine Similarities in Internal Representations for a `VGG-16` Network*




*Note.* In each panel, the solid (red) line plots the cosine similarity between the internal representations of a Basis shape and its `Rel` deformation, while the dashed (blue) line plots the cosine similarity between the internal representations of a Basis shape and its `Cood` deformation. Layers of the network are along the *x*-axis. Networks were either pretrained on `ImageNet` (first row) or on `Stylized-ImageNet` (second row). Their internal representations were then probed either without any further training (zero-shot, first column) or fine-tuning on the Hummel and Stankiewicz data set (second column). The hatched (yellow) area shows the upper and lower bounds on cosine similarity (obtained by computing the cosine similarity of images from the same and different categories, respectively). Shaded regions around each line show a 95% confidence interval. Based on the results of Hummel and Stankiewicz (1996), we would expect the solid (red) line (`Ba-Rel`) to be closer to the lower, rather than upper bound. Ba = Basis; Rel = relational; Cood = coordinate; Conv2d = two-dimensional convolution layer; GPool = global pooling. See the online article for the color version of this figure.

across the six categories) as the similarity between the basis image and its coordinate variant throughout the networks. That is, the networks failed to distinguish between the basis images and their relational and coordinate variants. In fact, networks also failed to distinguish between basis images from different categories (note the narrow hatched [yellow] region in the *zero-shot* condition in Figure 2). Unless the CNNs were trained to categorize the six basis images, the networks trained on data sets of naturalistic images failed to distinguish between the kinds of relations that differentiate the six categories—that is, to the model all line drawings in our stimuli set look approximately the same. This was true even for networks trained on `Stylized-ImageNet` (second row in Figure 2) that showed a strong shape-bias when tested on naturalistic images. This behavior contrasts with that of humans—an adult who has never seen these figures before in their lives would still see them as noticeably different shapes—and is an early existence proof of the contrasting shape representations that underlie humans and supervised CNNs.

Nevertheless, when the networks were trained on the six Basis shapes (the *fine-tuned* condition), both types of networks successfully learned to distinguish between stimuli from different categories with classification accuracy more than 95% for novel basis images in the test set (see Figure A1). It is again not surprising that the networks learn to classify the six Basis shapes even though they failed to tell them apart before training—previous studies have shown that CNNs have a large capacity to learn diagnostic features present in images. For example, CNNs can learn to classify large data sets of images consisting of random pixels (Tsvetkov et al., 2023) and even images where single pixels are diagnostic of category (Malhotra et al., 2020). To examine whether these networks learn to distinguish shapes based on relational features rather than some local features diagnostic of a category (e.g., local contours as shown by Baker et al., 2018), we examined the internal representations of relational and coordinate deformations in these fine-tuned networks. Examining these representations (right-hand column in Figure 2) showed that the networks represented all types of images in a similar manner in the early convolution layers (there is no difference between similarities within or between categories in the early layers) but representations begin to separate in the deeper convolution and fully connected layers (the hatched [yellow] region increases in size as we move left to right because images from different categories have lower similarity than images from the same category). However, for both types of pretrained networks, the basis images were equally distant to their relational and coordinate deformations (see the overlapping `Ba-Rel` and `Ba-Cood` lines in Figure 2). In keeping with this result, we also observed that the networks were able to classify images with both types of distortions nearly perfectly (see Figure A1). These results suggest that the networks learn to distinguish between the six Basis shapes based on some local features of these shapes which are preserved during the relational and coordinate deformations, rather than learning to distinguish shapes based on global relations between object parts. In summary, we did not find any evidence that suggests that the CNN represents a relational change to an image in any privileged manner compared to a coordinate change.

## Experiment 2

The results of Experiment 1 suggested that learning to classify the naturalistic images in `ImageNet` or even `Stylized-ImageNet` is not sufficient for CNNs to perceive the objects in terms of their categorical relations. But it could be argued that this is not because of a limitation of the optimization approach, but due to the limitation of data sets that the model was trained on. It is possible that, if the classification model was trained on data sets where relational differences were diagnostic of object categories, it may have internalized this statistic and started perceiving objects in terms of their categorical relations, just like humans. We tested this hypothesis in the next set of simulations, where we created a training environment with a "relational bias." We show next that when we do this, the network can learn specific changes to relations but it does not generalize this knowledge to novel (but highly similar) relational changes.

### Method

Experiment 2 used the same model architectures, pretraining, and analysis methods as Experiment 1. However, instead of using the training data set based on Hummel and Stankiewicz (1996), we created three new data sets where relational changes were diagnostic of image categories.

#### *Training and Test Stimuli*

We generated three data sets—shown as Set 1, Set 2, and Set 3 in Figure 3—for teaching CNNs to recognize relational deformations on Hummel and Stankiewicz's stimuli. Each data set again contained the six Basis shapes (and their translation, rotation, and scale transformations) from the training set in Experiment 1. Additionally, they also contained five new Basis shapes. These five shapes were the relational (`Rel`) deformations of the first five Basis shapes. In other words, the training set assigned different categories to a shape and its `Rel` deformation for five out of six figures. The test set consisted of `Rel` and `Cood` deformations of the final (unpaired) shape. Each training set again consisted of 5,000 images per category, where each image was constructed by translating, scaling, and rotating the Basis shape for that category. The test set consisted of 1,000 images per category where each image was constructed by randomly translating, scaling, and rotating the `Rel` and `Cood` deformations of the unpaired Basis shape.

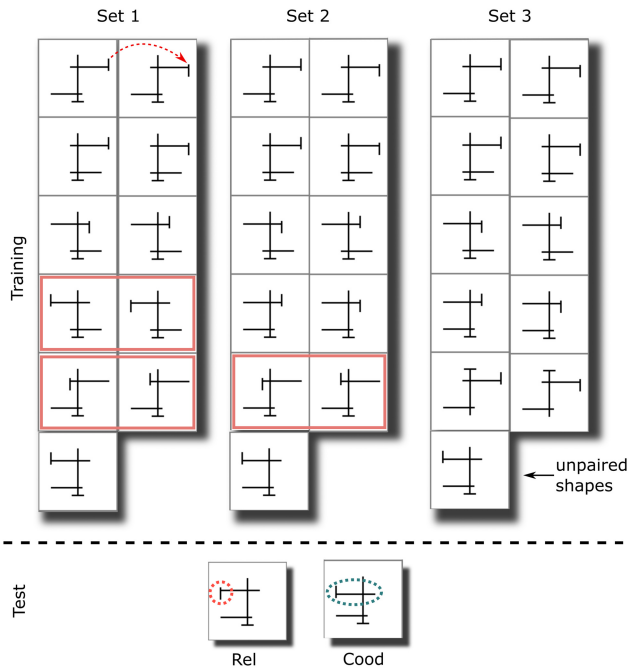
The difference between Sets 1, 2, and 3 lay in the degree of novelty of test images. In all three data sets the same relation (dashed red circle in Figure 3) was changed between the unpaired Basis shape and its `Rel` deformation. However, in the first set, there were four other categories (two pairs, highlighted in red rectangles) in the training set where a similar change in relation occurred—that is, for all highlighted categories, there existed another category where the short red segment at the left end of the top bar flipped from "above" to "below" or vice versa. In the second training set (Set 2 in Figure 3), there were two categories in the training set where the tested relation changed. However, in this case, this relational change occurred in a different location (closer to the central vertical line). In the third training set (Set 3 in Figure 3), the tested relational change was the least similar to training (relational changes only occurred to the right of the central vertical line for all other trained images).

#### *Further Training*

As in Experiment 1, the CNNs were fine-tuned on the training stimuli, which here consisted of 11 (5 + 5 + 1) Basis shapes. All

**Figure 3**

*Three Training Sets That Try to Teach the Network to Recognize Relational Changes*




*Note.* In each set, the first column shows a set of six unique Basis shapes, while the second column shows Rel deformations of the first five shapes (see red arrow). At the bottom are the two test shapes. These test shapes are identical to the 11th (unpaired) training shape, except for one relational (dashed red circle) or coordinate (dashed blue ellipse) deformation. In Set 1, the difference between the untrained shape and the tested Rel deformation is exactly the same as the relational change distinguishing one pair of shapes and similar to another pair in the training set (both highlighted in solid red rectangles). In Set 2, the exact relational change is not trained, however, there is a similar relational change at a close location (pair again highlighted in solid red rectangle). Set 3 is the most challenging, where none of the diagnostic relational changes in the training set occur at similar locations to the tested relational deformation. Rel = relational; Cood = coordinate. See the online article for the color version of this figure.

other details of training, including hyperparameters were kept the same as in Experiment 1.

## Results and Discussion

Figure 4 shows the cosine similarity in internal representations for VGG-16 trained on these three modified data sets (we obtained a similar pattern of results for AlexNet—see Figure C3). We observed that when networks were trained on Set 1 (left column in Figure 4), the cosine similarity Ba-Rel was lower than Ba-Cood in deeper layers of the CNN. That is, the networks treated the relational deformation as *less* similar to basis figures than the coordinate deformations. This looks much more like the behavior of human participants in Hummel and Stankiewicz (1996). But note that Set 1 contained two pairs of categories with the same relational change that distinguishes the tested Rel deformation from the corresponding basis figure. A stronger test is provided by Set 2 that excludes the pair of categories distinguished by the critical relational change from the training set. Here, we observed that this effect was significantly reduced (middle column in Figure 4, also compare results in Figure C3 in the Appendix for AlexNet, where this effect is slightly more pronounced but qualitatively similar). The strongest test for whether the network learns relational representations is provided by Set 3, where none of the categories in the training set changed the exact relation that distinguishes the Rel deformation from the basis image in the test set. Here, we observed (Figure 4, right-hand column) that the effect disappeared completely—the cosine similarity Ba-Rel was indistinguishable from Ba-Cood and both similarities were at the upper bound. All networks failed to learn that novel relational changes are more important for classification than coordinate changes even when the learning environment contained a "relational bias"—that is, changing relations led to a change in an image's category mapping. We also observed that when the networks were more sensitive to relational changes than coordinate changes (left-hand column of Figure 4), the difference in similarities between basis and relational deformations was greater for the network pretrained on ImageNet than the network pretrained on Stylized-ImageNet. This suggests that the type of shape bias induced by training the networks on stylized images does not help in selecting the features that distinguish the shapes used in this stimuli. That is, even though the network trained on stylized images shows a shape bias, these results suggest that the shape representations used by this network are qualitatively different from the shape representations used by humans.

## Experiment 3

Experiments 1 and 2 used stimuli that consisted of multiple parts and relational deformations involved changing the relationships between these parts. But of course, in order to build distal representations of complex objects, it is also necessary to build distal representations of the parts themselves. While Experiments 1 and 2 show that CNNs and humans differ in their representation of multipart objects, the stimuli used in these experiments did not allow us to compare the representations of the parts themselves, or indeed single-part objects. Another limitation of the stimuli in Experiments 1 and 2 was that they used discrete relations ("up" vs. "down," "left of" vs. "right of"), which do not permit manipulation of the degree of relational or coordinate change. This meant that we could not match the extent of relational change with *an equivalent* coordinate change and compare the sensitivity to each of these changes.

What sorts of deformations of the proximal stimulus should allow us to contrast optimization and heuristic approaches for identifying the component parts of complex objects or single-part objects? According to the structural description theory (Biederman, 1987), certain shape properties of the proximal image are taken by the visual system as strong evidence that individual parts have those properties. For example, if there is a straight or parallel line in the image, the visual system infers that the part contains a straight edge or parallel edges. If the proximal stimulus is symmetrical, it is assumed that the part is symmetrical (see, e.g., Pizlo et al., 2010). These (and other) shape features used to build a distal representation of the object part are called nonaccidental because they would only rarely be produced by accidental alignments of viewpoint. The visual system ignores the possibility that a given nonaccidental feature in the proximal stimulus (e.g., a straight line) is

**Figure 4**

*Cosine Similarity for VGG-16 Networks That Have Been Trained on Diagnostic Relations*


*Note.* Each panel shows cosine similarity in internal representations between basis images and `Rel` (solid, red) or `Cood` (dashed, blue) deformations of those images. The network was either pretrained on `ImageNet` (first row) or `Stylized-ImageNet` (second row) and fine-tuned on Set 1 (left), Set 2 (middle), or Set 3 (right) shown in Figure 3. Like Figure 2, the hatched (yellow) region shows the upper and lower bound on similarity. We can see that the network fine-tuned on Set 1 represents relational deformations as significantly different from basis images as well as coordinate deformations (the solid red line is much lower than the upper bound and dashed blue line for deeper layers in the network). However, this is not the case for networks fine-tuned on Set 2 or Set 3. Ba = Basis; Rel = relational; Cod = coordinate; Conv2d = two-dimensional convolution layer; GPool = global pooling. See the online article for the color version of this figure.

the product of an accidental alignment of the eye and distal stimulus (e.g., a curved edge). That is, the human visual system uses nonaccidental proximal features as a heuristic to infer distal representations of object parts. Critical for our purpose, many of the nonaccidental features described by Biederman (1987) are relational features, and indeed, many of the features are associated with Gestalt rules of perceptual organization, such as good continuation, symmetry, and Pragnanz (simplicity). Accordingly, any deformations of the proximal stimulus that alter these nonaccidental features (such as disrupting symmetry) should have a larger impact on classifications than deformations that do not.

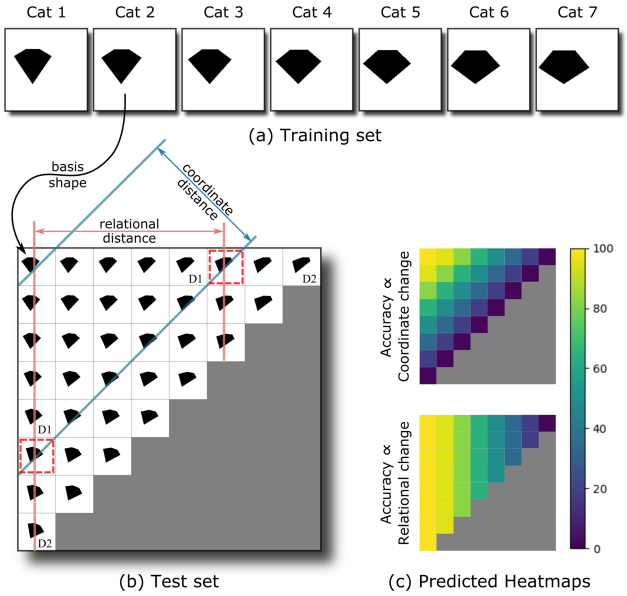
With this in mind, we designed a new stimuli set that allowed us to precisely manipulate the relational and coordinate deformations of single-part objects. The stimuli set consisted of seven symmetrical polygons (see Figure 5A), and we deformed these polygons by altering the locations of the vertices composing the polygons in a way that precisely controlled the metric change in the vertices' locations (in the retinal image). Like Experiment 1, we created two types of deformations: (a) a coordinate deformation that parametrically varied the degree to which a polygon rotated in the visual image versus (b) a relational change that had an equivalent impact as the corresponding rotation, but instead introduced a shear that changed the relative location of the polygon's vertices. Note that rotating an object preserves all nonaccidental features (Biederman, 1987), while shearing it changes its symmetry—a nonaccidental property of the object. To a model that looks only at the proximal stimulus, both deformations lead to an equivalent pixel-by-pixel change, while to a model that infers properties such as symmetry and solidity of the distal stimulus, the coordinate deformation preserves these properties while the relational deformation changes them.

Figure 5B shows some examples of test images for one of the trained shapes. These test shapes are organized based on the degree and type of deformation. The degree of relational deformation (shear) of a test image increases as we move from left to right, while the degree of coordinate deformation (rotation) increases as we move from top to bottom. We can also construct test shapes that are a combination of these relational and coordinate deformations. Every shape in Figure 5B is a combination of a rotation and a shear of the Basis shape in the top-left corner. We have organized these test shapes based on their distance to the basis figure: all shapes along each diagonal have the same cosine distance to the Basis shape,[1] and diagonals farther from the Basis shape are at a larger distance. Thus, this method gives us a set of test shapes organized according to increasing relational and coordinate changes and matched based on the distance to the Basis shape. We could now ask how accuracy degrades on this landscape of test shapes. If the visual system encodes shape as a set of diagnostic features of the proximal (retinal) image, accuracy should fall as one moves across

[1] We obtained qualitatively similar results when deformations were organized based on their Euclidean distance to the Basis shape.

**Figure 5**

*Stimuli Used to Test Shape Representations in Single-Part Objects*




*Note.* (A) The shapes in the basis set used for training. Each shape is presented at various translations and scales. (B) The test set for one of the categories (Cat 2) is obtained by deforming the Basis shape (in the top-left corner) through a combination of rotation and shear operations. Here we have organized these deformations in a matrix based on their coordinate distance (measured as cosine distance) and relational distance (measured as change in the relative location of vertices) from the Basis shape. All deformations on a diagonal of this matrix are at the same coordinate distance from the Basis shape and all deformations in a column are at the same relational distance from the Basis shape. Highlighted (red) squares show stimuli for computing cosine distance in Figure 7 below. Deformations marked D1 and D2 are used for testing human participants. (C) The predicted accuracy on the test set is presented as heat maps, assuming that accuracy is a function of coordinate distance (top), or relational distance (bottom). Cat = category; D = deformation. See the online article for the color version of this figure.

(perpendicular to) the diagonals on the landscape. On the other hand, if the visual system encodes shape as a property of the distal stimulus, then changing internal relations should lead to a larger change in classification accuracy than an equivalent coordinate change—that is, the accuracy should fall sharply as one moves left to right along each diagonal. Figure 5C shows predicted accuracy on this landscape for the two types of shape representations.

## Method

### Training Stimuli

The training set for Experiment 2 consisted of seven symmetric filled pentagons, presented on a white canvas. Each category contained 5,000 training images. The training set presented these polygons at different translations and scales, so it was not possible to classify them based on the position of a local feature or the area of the polygon. The difference between the Basis shapes for the two categories was the angles between the edges. Note that all polygons in the training set were presented in the upright orientation since rotation is one of the transforms (the coordinate transform) that the model is tested on.

### Test Stimuli

The test set consisted of a grid of shapes that were obtained by deforming the Basis shape of the corresponding category. We used two deformations: rotation, which preserved the internal angles between edges, and shear, which changed internal angles. To shear a shape, its vertices were horizontally moved by a distance that depended on the vertical distance to the apex. For a vertex with coordinates ($x_{old}$, $y_{old}$), we obtained a new set of vertices, ($x_{new}$, $y_{new}$) = ($x_{old} + \lambda(\Delta y)^2$, $y_{old}$), where $\lambda$ was the degree of shear and $\Delta y$ was the distance between $y_{old}$ and $y_{apex}$, the $y$-coordinate of the vertex at the apex. Images could also be a combination of rotations and shears. To do this, the basis image was first sheared, then rotated. We measured the (cosine or Euclidean) distance of a deformed image from the basis image and used this distance to organize the test images on a grid (see Figure 5). We then obtained 100 exemplars of each deformed image on the grid by randomly translating and scaling the image.

### Model Architecture and Pretraining

We used the same set of models as Experiments 1 and 2 (`VGG-16` and `AlexNet`) pretrained in the same manner (either on `ImageNet` or on `Stylized-ImageNet`).

### Further Training

Like Experiment 1, models were tested in either the *zero-shot* condition, where we did not train the model on our training set and simply examined the internal representations in response to test images, or in the *fine-tuned* condition, where the pretrained model underwent further training (with a reduced learning rate) on the training stimuli. We again observed that the models failed to distinguish any shape in the zero-shot condition, therefore, we restrict the presentation of our results to the fine-tuned condition.

### Analysis of Internal Representations

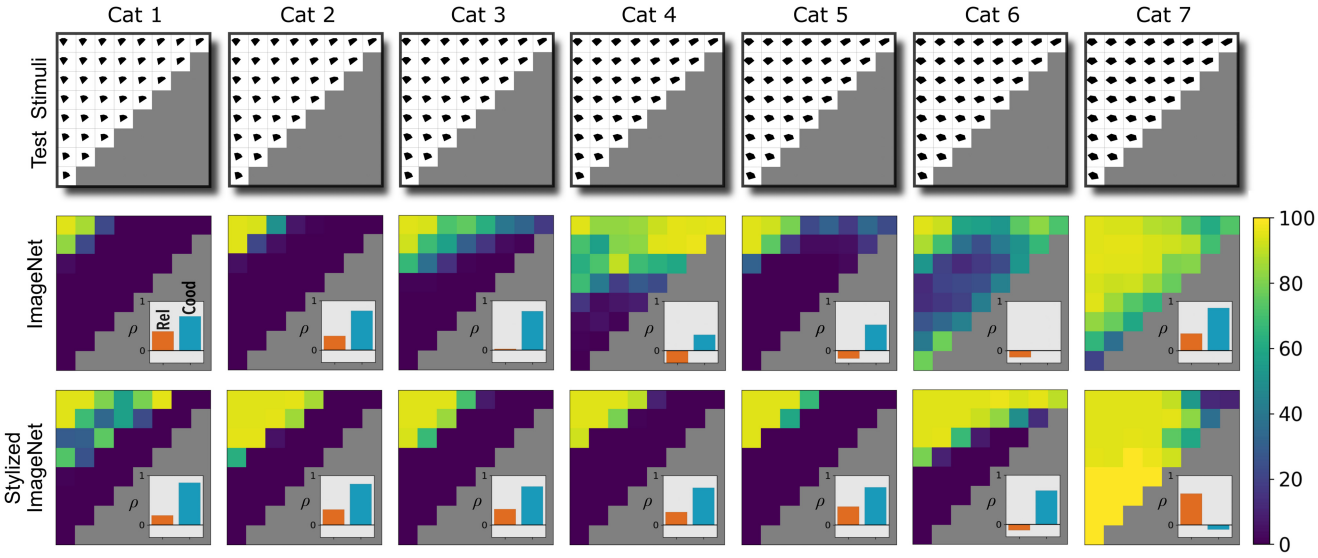
We obtained an estimate of how each deformation on the grid affects performance by computing the average accuracy across the 100 test images for that location. This gave us an empirical heatmap for the model's sensitivity to relational and coordinate deformations. We compared this empirical heatmap to the predicted heatmaps (Figure 5C) by computing Spearman's rank correlation coefficients between the observed heatmap and each of the predicted heatmaps. The similarity of internal representations for the polygons stimuli was obtained in a similar manner to the Hummel and Stankiewicz's stimuli. The similarity of a shear transformation to the corresponding basis image (`Ba-Sh`) was estimated by measuring the average cosine similarity between embeddings (at all convolutional and fully connected layers of the network) of 100 pairs of images from the basis and sheared sets of the same category. Similarly, the similarity between a basis image and its rotational transformation (`Ba-Rot`) was estimated by measuring the average cosine similarity between embeddings of 100 pairs of images from basis and rotated sets of the same category.

## Results and Discussion

The classification performance of `VGG-16` for images from the test set is shown in Figure 6 (we obtained a qualitatively similar pattern of results for `AlexNet`, see Appendix C). For all networks, we observed

**Figure 6**

*Performance of* `VGG-16` *on Deformations of Single-Part Objects*




*Note.* Test stimuli for each category are shown in the top row. Middle and bottom rows show accuracy on the landscape of relational and coordinate deformations for the network pretrained on `ImageNet` (middle row) or `Stylized-ImageNet` (bottom row). In each case, the network was fine-tuned on the set of seven polygons shown in Figure 5A. Each heatmap (in middle and bottom rows) corresponds to a category and shows the percent of shapes (with a relational and coordinate deformation given by the position on the landscape) accurately classified as the category from which the stimulus was derived. Insets show correlation coefficients (Spearman's ρ) of the correlation between the observed heatmap and predicted heatmaps (Figure 5C) for which accuracy decreases as a function of relational distance (red) and coordinate distance (blue). Cat=category; Rel = relational; Cood = coordinate. See the online article for the color version of this figure.

that test accuracy was highest at the top-left corner (i.e., for the Basis shape) and reduced as the degree of relational and coordinate change was increased. Crucially, we observed that for most categories, accuracy decreased as a function of distance to the Basis shape (perpendicular to the diagonals), rather than relational change (left to right). Consistent with this qualitative observation, we observed a large (and significant) correlation between most of the observed heatmaps and the predicted heatmap for coordinate change but a small (and nonsignificant) correlation between most of the observed heatmaps and the predicted heatmap for relational change (Figure 6, insets). In fact, for some categories, accuracy *improved* as one moved from left to right along the diagonals. Occasionally, we observed high accuracy for large rotations in one category. This was generally due to false positives, where large rotations for all categories were classified as the same category by the network (see Figure B1 in Appendix B for details). Overall, these results suggest that the network does not represent the shapes in this task in a relational manner. If it did, its performance on relational changes should have been a lot worse than its performance on relation-preserving rotations.

In order to get more insight into the network's internal representations for relational and coordinate deformations, we examined the similarity between representations of basis images and their relational and coordinate deformations. An example of these images is highlighted (dashed red squares) in Figure 5B. We took 100 such triplets that varied in location and scale of these exemplars and computed the average cosine similarity between representations of basis and shear (`Ba-Sh`) deformations as well as basis and rotation (`Ba-Rot`) deformations for all categories and at all layers within the network. These cosine similarities for two `VGG-16` networks are plotted in Figure 7 (we again obtained qualitatively similar results for `AlexNet`—see Figure C6 in Appendix). At all internal layers, we observed that the average similarity between a basis image and its relational (shear) deformation was equal to or higher than the average similarity between the basis image and its coordinate (rotation) deformation (compare solid [red] and dashed [blue] lines in Figure 7). In other words, the relational deformation of an image was closer to the basis image than its coordinate deformation, and pretraining on the `Stylized-ImageNet` data set to give the network a shape-bias did not change this pattern. This is the opposite of what one would expect if the network represented the stimuli in a relational manner.

## Experiment 4

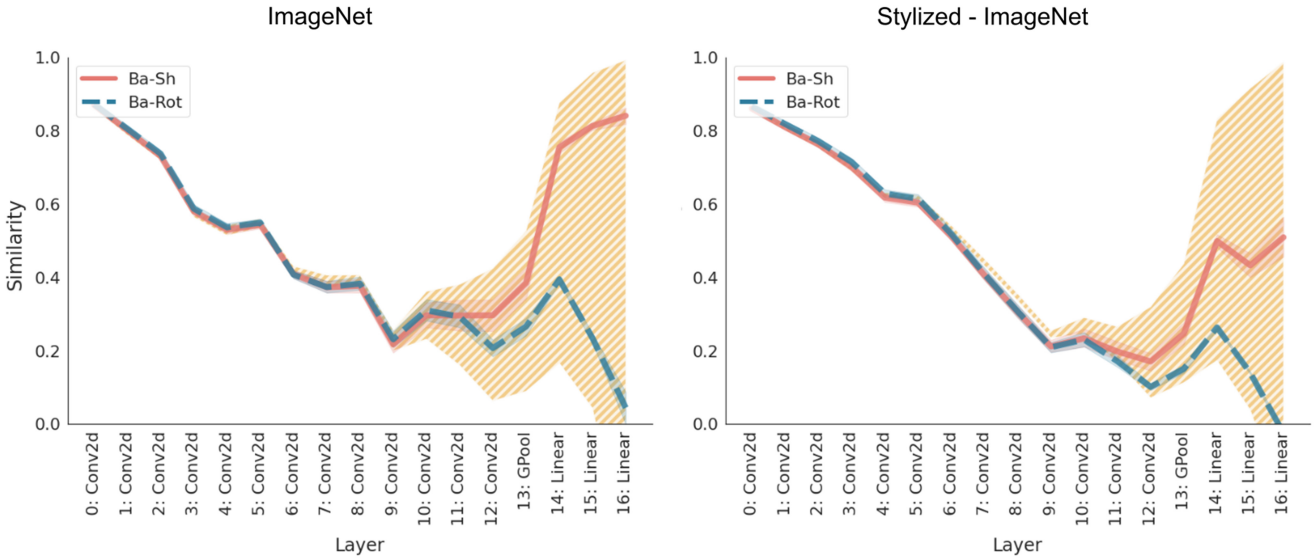
Results of Experiment 3 showed that CNNs trained to classify objects do not show any enhanced sensitivity to deformations of relations between features of single-part objects. In other words, we did not observe any evidence suggesting that the CNNs infer properties of distal stimuli based on the proximal input image. In our next experiment, we examined how humans trained on the exact same stimuli responded to the two types of deformations.

### Participants

Participants ($N = 37$, $M_{age} = 33$, 70% women, 30% men) with normal or corrected-to-normal vision were recruited via prolific for an online study and the experiment was conducted on the Pavlovia platform. We did not elicit gender, sex, or age information from participants during the experiment and no participant was excluded based

**Figure 7**

*Cosine Similarity in Internal Representations of VGG-16 Fine-Tuned on Stimuli in Experiment 3*




*Note.* The solid (red) and dashed (blue) lines show the average cosine similarity between basis images and relational (shear) and coordinate (rotation) deformations, respectively. The hatched (yellow) region shows the bounds on this similarity, with the upper bound determined by the average similarity between basis images from the same category and lower bound determined by the average similarity between basis images of different categories. If relational (shear) deformation has a larger effect on internal representations than a coordinate (rotation) deformation, one would expect the solid (red) line to be below the dashed (blue) line. Ba = Basis; Sh = shear; Rot = rotational; Conv2d = two-dimensional convolution layer; GPool = global pooling. See the online article for the color version of this figure.

on their gender. The proportion of male and female participants reported here is based on the demographic information collected by prolific when participants register on the platform. Participants were reimbursed a fixed 2 GBP and participants who proceeded to the testing phase ($N = 23$) had a chance to earn a bonus of up to another 2 GBP depending on their performance during testing. The average payment was 8 GBP/hr. A written ethics approval for the study was obtained from the University of Bristol Ethics Board.

## Stimuli

Four categories (out of seven) were chosen from the data set in Experiment 3 to train participants. These were Cat 1, Cat 3, Cat 5, and Cat 7 from Figure 5A. For the test data, we selected two deformations of each type that were matched according to the cosine distance from the basis (trained) image. For the relational deformation, these were the fifth (Deformation D1) and final (Deformation D2) shears in the top row of Figure 5B. For the coordinate deformation, these were the fifth (D1) and final (D2) rotations in the leftmost column of Figure 5B. This made up the five conditions in the experiment: basis, D1 (shear), D2 (shear), D1 (rotation), and D2 (rotation). The original stimuli were 224 × 224 pixels but were rescaled for each participant to 50% of the vertical resolution of the participant's screen to account for the variability in screen size and resolution when running the study online. Just like in Experiment 3, training stimuli for each category were obtained by modifying the basis image for various scales and translations but were always shown in an upright orientation. Similarly, test stimuli also varied in scale and location, but were additionally sheared and rotated according to the condition.

## Procedure

Participants completed a supervised training phase in which they learned to categorize basis versions of the four categories. Each training block consisted of 40 stimuli for a total of 200 training trials (50 per category). Feedback on overall accuracy was given at the end of each block. Participants completed up to a maximum of five training blocks, or until they reached 85% categorization accuracy in a block. Participants who managed to reach 85% accuracy continued to the test blocks. The order of trials was randomized for each participant. Each trial started with a fixation cross (750 ms), then the stimulus was presented (500 ms) followed by four on-screen response buttons labeled 1 …4 corresponding to the four categories (until the participant responded by clicking on the chosen category button). After participants responded, feedback was given—CORRECT (1 s) if the response was correct and INCORRECT with additional information about what the correct response should have been (1.5 s) if the response was incorrect. The training phase was followed by a test phase consisting of five test blocks. Each block consisted of 20 trials for a total of 100 test trials (25 per condition). Like the training phase, the order of test trials was randomized for each participant. The procedure for each test trial was the same as in the training phase except that participants were not given any feedback during testing.

## Analysis

Four planned comparisons (*t* tests) were conducted in order to test whether accuracy rates in each of the shear and rotation conditions differed from accuracy in the basis condition.

## Transparency and Openness

We report all data exclusions (if any), all manipulations, and all measures in the study. The number of participants was chosen such that at least 20 participants got to the test phase. With this in mind, we advertised the task on prolific and stopped recruiting when $N = 37$ participants had completed the study, which gave us a sample size of $N = 23$ participants who completed the test phase. All data, analysis code, and research materials are available at https://github.com/gammagit/distal. Data were analyzed using Python, Version 3.8.3, and visualized using Matplotlib, Version 3.3.2. This study's design and its analysis were not preregistered.

## Results and Discussion

The average CNN and human accuracy of classification on each of these deformations are shown in Figure 8. We can see that irrespective of training, `VGG-16` was more sensitive to rotation than to shear (see Figure C7 for `AlexNet`). While performance decreases for both deformations, it decreases more rapidly for rotations. Human participants showed the opposite pattern (Figure 8, right-hand panel). There was no significant difference in performance between the basis image and the two rotation deformations, both $t(22) < 1.11$, $p > .28$, while performance decreased significantly for each of the shear deformations, both $t(22) > 3.99$, $p < .001$, $d_z > 1.65$. The largest shear resulted in the largest decrease in performance ($M_{\text{difference}} = 25.87\%$). Thus, the behavior of participants was in line with the prediction of structural description theories, where a shape is encoded based on relations between features, and in the opposite direction to the performance of the CNNs trained to classify objects.

## Experiment 5

One response to the difference between CNNs and humans in Experiments 3 and 4 is that it arises due to the difference in experience between the two systems. Humans experience objects in a variety of rotations and consequently represent a novel object in a rotation-invariant manner. CNNs, on the other hand, have not been explicitly trained on objects in different orientations (although `ImageNet` includes some objects in various poses). It could therefore be argued that CNNs do not learn relational representations in Experiment 3 because the training set did not provide an incentive for learning such a representation. Indeed, the optimization view argues that a bias must be present in the training environment for the visual system to internalize it.

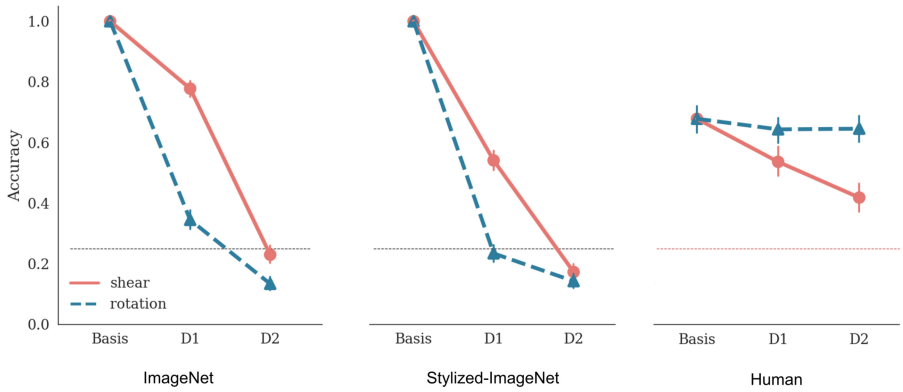
To give the network a better chance of learning to classify based on internal relations, we conducted two further simulations. In the first simulation, we trained the networks on some rotations for all Basis shapes and tested them on unseen rotations. This simulation emulates generalizing the concept of rotation for each object after observing some of the rotations for that object. In the second simulation, the networks were shown *all* rotations of some Basis shapes and tested on unseen rotations of the left-out Basis shapes. This simulation emulates generalizing the concept of rotation from one object to another.

## Method

All methods in Experiment 5 remained the same as in Experiment 3, except for the images in the training sets. In the first simulation, the training set now consisted of basis (polygon) shapes presented at random translations and scales (just like Experiment 3) but additionally, also at rotations in the range $[-45°, 0°]$ for all polygons. We then

**Figure 8**

*Comparison of Humans and `VGG-16` on How Classification Accuracy Changes With Deformations*




*Note*. Left and middle panels show classification accuracy for `VGG-16` trained on `ImageNet` and `Stylized-ImageNet`, respectively. Both types of networks were fine-tuned on the polygons stimuli shown in Figure 5. Right panel shows the performance of human participants on the same stimuli. Each panel shows performance under three conditions: basis image, deformation D1, and deformation D2. For the shear deformation (solid, red line), D1 and D2 consist of images in the top row in the fourth and eighth column in Figure 5. For the rotation deformation (dashed, blue line), D1 and D2 consist of images in the first column and fourth and eighth rows. Error bars show a 95% confidence interval and the dotted black line shows chance performance. D = deformation. See the online article for the color version of this figure.

tested the networks on rotations in the range [0°, +45°]. In the second simulation, we selected six (out of seven) categories and trained the network on random translations, scales, and *all* rotations [0°, 360°] for these categories. For the seventh category (`Cat 3`), images were still randomly translated and scaled, but always presented in the upright orientation. We then tested how the network generalized to the two types of deformations for this critical category. We obtained qualitatively similar results for networks pretrained on `ImageNet` and `Stylized-ImageNet`. Since the network trained on `Stylized-ImageNet` has the best chance of capturing human data, here we present the results of this network for both simulations.

## Results and Discussion

The network performance for the first simulation is shown in Figure 9. For most categories, performance degraded equally or more with a change in rotation than with an equivalent change in shear. That is, the network was *better* at generalizing to large relational deformations (shears) than large relation-preserving deformations (rotation). The pattern was different for Cat 6, where the network showed good performance on large rotations. But examining the confusion matrix again revealed that the high accuracy at large rotations for this category was misleading as it was accompanied by large Type I errors: large rotations for shapes of any category were misclassified as belonging to Cat 6. Overall, we did not find any evidence for the network learning shapes based on their internal relations.

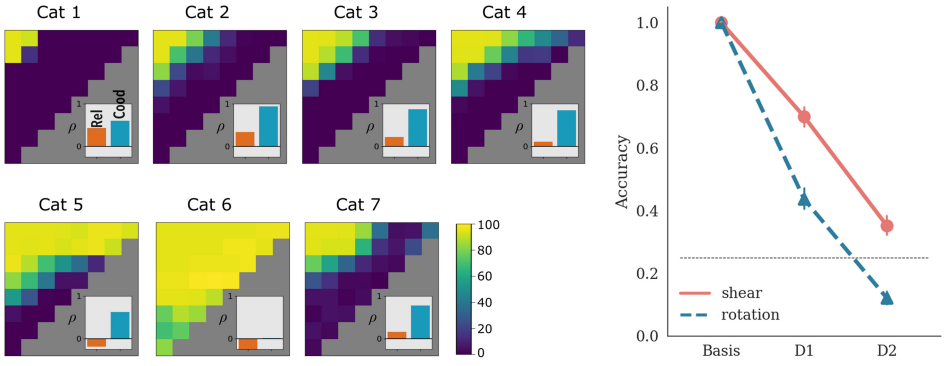
The results of the second simulation are shown in Figure 10. Figure 10A shows the heat map of accuracy on the test grid for the left-out category. This heat map showed that the network continued showing the pattern observed above—its performance decreases across (perpendicular to) the diagonals, but increases as one moves from left to right along these diagonals. Figure 10B shows the performance of VGG-16 on the same conditions as the human experiment (see Figure 8). We observed that performance for the two shear deformations (dashed line) dropped at the same rate or faster (compare with results for `AlexNet` in Figure C9B in the Appendix) than the two rotation deformations (solid line). This figure makes it clear that training other orientations on all rotations does not help the network generalize better to novel orientations for the left-out category. In fact, the performance drops more quickly than when none of the categories were rotated in the training set (compare with Figure 8). This is because the network starts classifying novel orientations of the left-out shape as the shapes that it had seen being rotated in the training set.

It may be tempting to think that the differences between humans and CNNs can be reconciled by training CNNs that learn rotation-invariant shapes. However, consider how a CNN achieves rotation invariance. Figure 11, based on Goodfellow et al. (2016, Chapter 9), illustrates how a network consisting of convolution and pooling layers may learn to recognize digits in different orientations. As a result of training on digits (here, the digit 5) oriented in three different directions, the convolution layer develops three different filters, one for each orientation. A downstream pooling unit then amalgamates this knowledge and fires when any one of the convolution filters is activated. Therefore, this pooling unit can be considered as representing the rotation-invariant digit 5. During testing, when the network is presented with the digit 5 in any orientation, the corresponding convolution filter gets activated, resulting in a large response in the pooling unit, and the network successfully recognizes the digit 5, irrespective of its orientation.

In contrast, a relational account of shape representation does not rely on developing filters for each orientation of a shape. Indeed, it is not even necessary to observe a shape in all orientations to get, at least some degree of, rotation invariance. Instead, a relational account proposes that the visual system starts by identifying internal parts of an object, and binds these parts to roles in a relational structure (e.g., *above*; $x$ = curved cylinder; $y$ = truncated cone) and checks whether this relational structure and bindings match those of the learned

**Figure 9**

*Performance of `VGG-16` Trained on Some Rotations of All Categories*

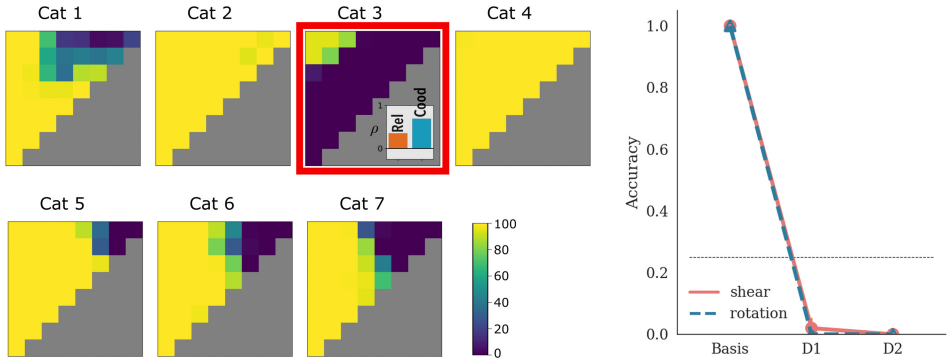



(a) *Accuracy landscape* (b) *Accuracy for conditions in experiment*

*Note.* (A) The accuracy of the network plotted as percent of correct classifications for each rotation and shear deformation of each category. Insets show correlation coefficients (Spearman's ρ) for correlation with heatmap predicted as a function of relational change (red) and coordinate change (blue). (B) Accuracy for shear (solid, red) and rotation (dashed, blue) as a function of deformations used in Experiment 4 with human participants (compare with Figure 8, right-hand panel). Cat = category; Rel = relational; Cood = coordinate. See the online article for the color version of this figure.

**Figure 10**

*Performance of `VGG-16` Trained on All Rotations of Some Categories*




(a) *Accuracy landscape for left-out Category* (b) *Accuracy for conditions in experiment*

*Note*. (A) The accuracy of the network plotted as percent of correct classifications for rotation and shear deformations for all categories. Note the high performance for all rotations of most categories is expected as the network is trained on these rotations. The critical (out-of-training-distribution) test is the network's performance on the left-out category—Cat 3 (highlighted using a red rectangle). Inset shows correlation coefficients (Spearman's ρ) for correlation with heatmap predicted as a function of relational change (red) and coordinate change (blue). (B) The accuracy of the network for the set of Deformations D1 and D2 for Cat 3, tested in Experiment 4 with human participants (compare with Figure 8, right-hand panel). Cat = category; Rel = relational; Cood = coordinate; D = deformation. See the online article for the color version of this figure.

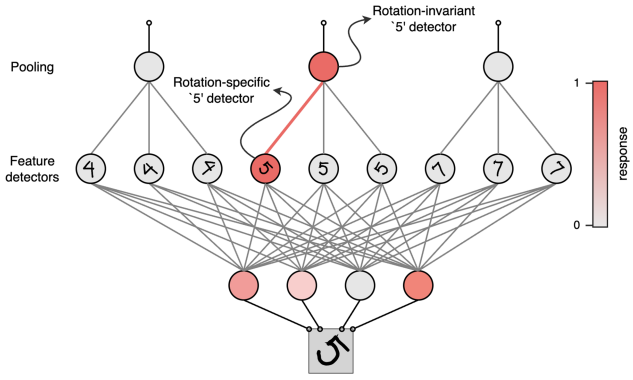
shape. (We elaborate in the General Discussion why CNNs may be ill-equipped to represent and compare relations in this fashion.) Support for this relational account comes from many psychological studies have shown that invariance, such as rotation invariance, precedes recognition (Biederman & Cooper, 1991, 1992; Biederman & Gerhardstein, 1995; Hummel, 2013).

## General Discussion

In a series of experiments, we have shown that humans represent shape in qualitatively different ways to CNNs that learn to classify large data sets of objects using supervised learning. In Experiment 1, we found that CNNs trained to classify objects were entirely insensitive to deformations in categorical relations between object parts. Furthermore, we could not train CNNs to be sensitive to relational changes in general even when we made relational changes diagnostic of category classification (Experiment 2). In Experiments 3 and 4, where we precisely matched the extent of relational and coordinate deformations, we found that humans were highly sensitive to relational deformations of single-part objects, whereas CNNs were only sensitive to coordinate distance, and once again, CNNs could not learn to be sensitive to relational manipulations (Experiment 5).

**Figure 11**

*A Proposal for Achieving Rotation Invariance in CNNs (Based on Goodfellow et al., 2016, Chapter 9)*




*Note*. A network that learns to detect digits at various rotations. It does this by learning a large set of rotation-specific filters, one matching each rotation of each digit. A downstream unit pools across all rotation filters for a given digit, essentially performing a disjunction over all filter activations. CNN = convolutional neural network. See the online article for the color version of this figure.

These findings challenge the hypothesis that humans perceive objects based on similar principles as CNNs trained to classify large sets of objects and that apparent differences arise due to "differences in the data that they see" (Hermann et al., 2020). These results show that even CNNs that have been trained to classify objects on the basis of shape (trained on the `Stylized-ImageNet`) learn the wrong sort of shape representation. These findings add to other studies that also highlight the different types of shape representation used by CNNs and the human visual system. For example, Baker and Elder (2022) show that, unlike humans, CNNs are insensitive to configural relations between local shape features and training innovations (such as training on `Stylized-ImageNet`) do not lead to configural processing in CNNs. Similarly, Puebla and Bowers (2021) have found that CNNs fail to support a simple relational judgment with shapes, namely, whether two shapes are the same or different. Again, this highlights how CNNs trained to process shape ignore relational information. In addition, Baker et al. (2018) have shown that CNNs that classify objects based on shape focus on local features and ignore how local features relate to one another in order to encode the global structure of objects.

More generally, our results speak to the debate between heuristic and optimization views sketched out in the Introduction. Simulations with

CNNs suggest that a human-like sensitivity to relational properties (Experiment 1) and symmetry (Experiment 3) does *not* emerge by optimizing performance on image classification. We also found that simply training a CNN on some relational changes (Experiments 2, 4, and 5) is *not* enough for the network to generalize this knowledge to novel relational changes. These results suggest that human sensitivity to relational properties and symmetry of objects are inductive biases of our visual systems—that is, they arise because our visual system uses heuristics such as constancy of relations and preservation of symmetry to solve the ill-posed problem of inferring distal representations of three-dimensional objects from their two-dimensional retinal images. Such an inference mechanism is neither built into the architecture of supervised CNNs nor required in the task of classifying images.

The difficulty of learning relations such as *above* and *larger-than* in a CNN is exacerbated by two additional factors. The first is that, by adulthood, most such relations are invariant with their arguments (Doumas et al., 2008): an adult understands that *larger-than* means the same thing in *larger-than* (star, planet), *larger-than* (elephant, mouse), and *larger-than* (nucleus, electron) even though the relation's arguments are extremely different across these cases. The cognitive architecture somehow learns an invariant representation despite the fact the *larger-than* relation is never experienced in its fully abstract, argument-free form. (One never sees an example of disembodied *larger-than*-ness; instead, one only sees examples of specific things that are larger than other specific things.) Learning and generalizing such an invariant relation is a real challenge for CNNs, particularly in an object classification task where CNNs are prone to learning alternative features ("shortcuts") that allow them to perform well (Geirhos et al., 2020). But even if such an invariant relation could somehow be learned and generalized by using a different task, CNNs face a second hurdle. During a trial, adults not only identify these invariant relations but also understand that these relations accept arguments (such as the *larger-than* relation has *larger* and *smaller* arguments) and that these arguments bind to instances (e.g., *larger* binds to elephant). These bindings need to be flexible and vary from trial to trial, based on the compositional structure of the input stimulus. There is no mechanism in a CNN to perform this dynamic binding (Doumas et al., 2008, 2022). The only mechanism of binding available in CNNs is conjunctive coding—that is, a static binding where otherwise independent properties (such as *larger* and elephant) are "entangled." But learning entangled representations is mutually inconsistent with learning invariant representations. Thus CNNs learning object classification may be fundamentally ill-equipped to represent relations both because of the nature of the task and because of architectural limitations.

Of course, it is possible that other deep neural network architectures, such as transformers (Vaswani et al., 2017), capsule networks (Sabour et al., 2017) or neural turning machines (Graves et al., 2014), trained on a range of different tasks (especially tasks where the objective is to approximate the distal representation) or on tasks with different objectives rather than classification (e.g., unsupervised learning of image sequences; Parker & Serre, 2015, or generative modeling; Kingma & Welling, 2013, or on a "self-supervised" task; Grill et al., 2020) may lead to shape representations that are more similar to those formed in human visual cortex. However, here we wanted to focus on CNNs trained on recognizing objects through supervised learning because of two reasons. Firstly, it has been argued that CNNs trained under these settings learn to classify objects based on human-like shape representations (Geirhos et al., 2018; Hermann et al., 2020; Kubilius et al., 2016). Secondly, these models have had the largest success in predicting neural representations in human and primate visual system (Cadieu et al., 2014; Schrimpf et al., 2020; Yamins & DiCarlo, 2016) and it has been argued that there is a strong correlation between a CNN's categorization performance and its ability to predict individual-level IT neural unit response data (Yamins et al., 2014). Our findings challenge the view that optimizing performance in a classification task can explain shape representations used during human shape perception. Instead, these findings are well predicted by the classic structural description theory of object recognition that builds a distal representation of objects using heuristics (e.g., Biederman, 1987).

## References


Baker, N., & Elder, J. H. (2022). Deep learning models fail to capture the configural nature of human shape perception. *iScience*, *25*(9), Article 104913. https://doi.org/10.1016/j.isci.2022.104913

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), Article e1006613. https://doi.org/10.1371/journal.pcbi.1006613

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review*, *94*(2), 115–147. https://doi.org/10.1037/0033-295X.94.2.115

Biederman, I., & Cooper, E. E. (1991). Evidence for complete translational and reflectional invariance in visual object priming. *Perception*, *20*(5), 585–593. https://doi.org/10.1068/p200585

Biederman, I., & Cooper, E. E. (1992). Size invariance in visual object priming. *Journal of Experimental Psychology: Human Perception and Performance*, *18*(1), 121–133. https://doi.org/10.1037/0096-1523.18.1.121

Biederman, I., & Gerhardstein, P. C. (1995). Viewpoint-dependent mechanisms in visual object recognition: Reply to Tarr and Bülthoff (1995). *Journal of Experimental Psychology: Human Perception and Performance*, *21*(6), 1506–1514. https://doi.org/10.1037/0096-1523.21.6.1506

Biederman, I., & Ju, G. (1988). Surface versus edge-based determinants of visual recognition. *Cognitive Psychology*, *20*(1), 38–64. https://doi.org/10.1016/0010-0285(88)90024-2

Cadieu, C. F., Hong, H., Yamins, D. L., Pinto, N., Ardila, D., Solomon, E. A., Majaj, N. J., & DiCarlo, J. J. (2014). Deep neural networks rival the representation of primate IT cortex for core visual object recognition. *PLoS Computational Biology*, *10*(12), Article e1003963. https://doi.org/10.1371/journal.pcbi.1003963

Doumas, L. A., Hummel, J. E., & Sandhofer, C. M. (2008). A theory of the discovery and prediction of relational concepts. *Psychological Review*, *115*(1), 1–43. https://doi.org/10.1037/0033-295X.115.1.1

Doumas, L. A., Puebla, G., Martin, A. E., & Hummel, J. E. (2022). A theory of relation learning and cross-domain generalization. *Psychological Review*, *129*(5), 999–1041. https://doi.org/10.1037/rev0000346

Ellis, W. D. (2013). *A source book of gestalt psychology*. Routledge.

Feinman, R., & Lake, B. M. (2018). *Learning inductive biases with simple neural networks*. Preprint. arXiv:1802.02745

Gatys, L. A., Ecker, A. S., & Bethge, M. (2016, 26th June–1st July). Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition, Las Vegas, Nevada* (pp. 2414–2423). IEEE Computer Society.

Geirhos, R., Jacobsen, J. H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., & Wichmann, F. A. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence*, *2*(11), 665–673. https://doi.org/10.1038/s42256-020-00257-z

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*. Preprint. arXiv:1811.12231

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Graves, A., Wayne, G., & Danihelka, I. (2014). *Neural turing machines*. Preprint. arXiv:1410.5401

Grill, J. B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Pires, B. A., Guo, Z. D., Azar, M. G., Piot, B., Kavukcuoglu, K., Munos, R., & Valko, M. (2020). Bootstrap your own latent—a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, *33*, 21271–21284.

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.

Hochberg, J., & Brooks, V. (1962). Pictorial recognition as an unlearned ability: A study of one child's performance. *The American Journal of Psychology*, *75*(4), 624–628. https://doi.org/10.2307/1420286

Hummel, J. E. (1994). Reference frames and relations in computational models of object recognition. *Current Directions in Psychological Science*, *3*(4), 111–116. https://doi.org/10.1111/1467-8721.ep10770560

Hummel, J. E. (2013). Object recognition. In *Oxford handbook of cognitive psychology* (pp. 32–46). Oxford University Press.

Hummel, J. E., & Biederman, I. (1992). Dynamic binding in a neural network for shape recognition. *Psychological Review*, *99*(3), 480–517. https://doi.org/10.1037/0033-295X.99.3.480

Hummel, J. E., & Stankiewicz, B. J. (1996). Categorical relations in shape perception. *Spatial Vision*, *10*(3), 201–236. https://doi.org/10.1163/156856896X00141

Kanizsa, G. (1979). *Organization in vision: Essays on gestalt perception*. Praeger Publishers.

Kingma, D. P., & Ba, J. (2014). *Adam: A method for stochastic optimization*. Preprint. arXiv:1412.6980

Kingma, D. P., & Welling, M. (2013). *Auto-encoding variational Bayes*. Preprint. arXiv:1312.6114

Knill, D. C. (1992). Perception of surface contours and surface shape: From computation to psychophysics. *Journal of the Optical Society of America A*, *9*(9), 1449–1464. https://doi.org/10.1364/JOSAA.9.001449

Kriegeskorte, N. (2015). Deep neural networks: A new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science*, *1*(1), 417–446. https://doi.org/10.1146/vision.2015.1.issue-1

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84–90. https://doi.org/10.1145/3065386

Kubilius, J., Bracci, S., & Op de Beeck, H. P. (2016). Deep neural networks as a computational model for human shape sensitivity. *PLoS Computational Biology*, *12*(4), Article e1004896. https://doi.org/10.1371/journal.pcbi.1004896

Landau, B., Smith, L. B., & Jones, S. S. (1988). The importance of shape in early lexical learning. *Cognitive Development*, *3*(3), 299–321. https://doi.org/10.1016/0885-2014(88)90014-7

Malhotra, G., Dujmovic, M., & Bowers, J. S. (2022). Feature blindness: A challenge for understanding and modelling visual object recognition. *PLOS Computational Biology*, *18*(5). Article e1009572. https://doi.org/10.1371/journal.pcbi.1009572

Malhotra, G., Evans, B. D., & Bowers, J. S. (2020). Hiding a plane with a pixel: Examining shape-bias in CNNs and the benefit of building in biological constraints. *Vision Research*, *174*, 57–68. https://doi.org/10.1016/j.visres.2020.04.013

Mamassian, P., & Landy, M. S. (1998). Observer biases in the 3D interpretation of line drawings. *Vision Research*, *38*(18), 2817–2832. https://doi.org/10.1016/S0042-6989(97)00438-0

Nakayama, K., He, Z. J., & Shimojo, S. (1995). Visual surface representation: A critical link between lower-level and higher-level vision. In S. M. Kosslyn, & D. N. Osherson (Eds.), *Visual cognition: An invitation to cognitive science* (pp. 1–70). The MIT Press.

Palmer, S., Rosch, E., & Chase, P. (1981). Canonical perspective and the perception of objects. In J. Long, & A. Baddeley (Eds.), *International symposium on attention and performance (Attention and performance IX)* (pp. 135–151). Lawrence Erlbaum Associates.

Parker, S. M., & Serre, T. (2015). Unsupervised invariance learning of transformation sequences in a model of object recognition yields selectivity for non-accidental properties. *Frontiers in Computational Neuroscience*, *9*, Article 115. https://doi.org/10.3389/fncom.2015.00115

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., & Lerer, A. (2017, December 4–9). *Automatic differentiation in PyTorch*. 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, United States.

Pizlo, Z. (2001). Perception viewed as an inverse problem. *Vision Research*, *41*(24), 3145–3161. https://doi.org/10.1016/S0042-6989(01)00173-0

Pizlo, Z., Sawada, T., Li, Y., Kropatsch, W. G., & Steinman, R. M. (2010). New approach to the perception of 3D shape based on veridicality, complexity, symmetry and volume. *Vision Research*, *50*(1), 1–11. https://doi.org/10.1016/j.visres.2009.09.024

Pizlo, Z., & Stevenson, A. K. (1999). Shape constancy from novel views. *Perception & Psychophysics*, *61*(7), 1299–1307. https://doi.org/10.3758/BF03206181

Poggio, T., & Edelman, S. (1990). A network that learns to recognize three-dimensional objects. *Nature*, *343*(6255), 263–266. https://doi.org/10.1038/343263a0

Puebla, G., & Bowers, J. (2021). Can deep convolutional neural networks support relational reasoning in the same-different task?. *Journal of Vision*, *22*(10), 1–18. https://doi.org/10.1167/jov.22.10.11

Sabour, S., Frosst, N., & Hinton, G. E. (2017). *Dynamic routing between capsules*. Preprint. arXiv:1710.09829

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., Kar, K., Bashivan, P., Prescott-Roy, J., Schmidt, K., Yamins, D. L. K., & DiCarlo, J. J. (2020). *Brain-score: Which artificial neural network for object recognition is most brain-like?* BioRxiv. https://doi.org/10.1101/407007

Simonyan, K., & Zisserman, A. (2014). *Very deep convolutional networks for large-scale image recognition*. Preprint. arXiv:1409.1556

Stevens, K. A. (1981). The visual interpretation of surface contours. *Artificial Intelligence*, *17*(1–3), 47–73. https://doi.org/10.1016/0004-3702(81)90020-5

Tarr, M. J., & Pinker, S. (1989). Mental rotation and orientation-dependence in shape recognition. *Cognitive Psychology*, *21*(2), 233–282. https://doi.org/10.1016/0010-0285(89)90009-1

Tsvetkov, C., Malhotra, G., Evans, B. D., & Bowers, J. S. (2023). The role of capacity constraints in Convolutional Neural Networks for learning random versus natural data. *Neural Networks*, *161*, 515–524. https://doi.org/10.1016/j.neunet.2023.01.011

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention is all you need*. Preprint. arXiv:1706.03762.

Yamins, D. L., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, *19*(3), 356–365. https://doi.org/10.1038/nn.4244

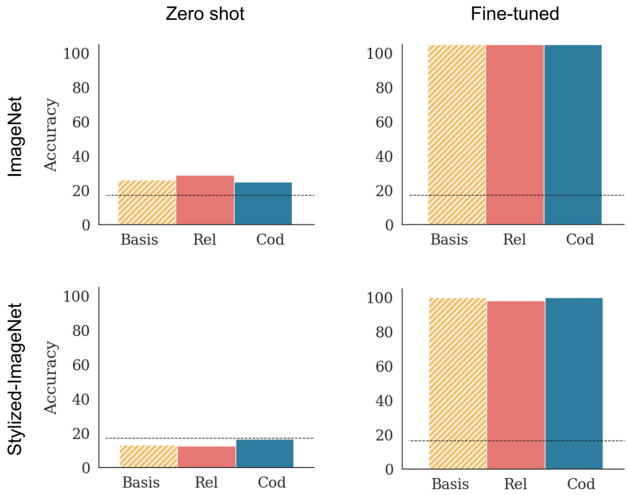
Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(23), 8619–8624. https://doi.org/10.1073/pnas.1403112111

(*Appendices follow*)

# Appendix A

## Classification Performance

**Figure A1**
*Classification Accuracy for* `VGG-16` *in Experiment 1*




*Note.* Panels in the first column show classification accuracy on the pretrained network without any further training, while panels in the second column show test performance for a model that was fine-tuned on the set of Basis shapes. The dashed black line shows chance performance. We observed that models in the *zero-shot* condition failed to classify the Basis shapes or their deformations (accuracy was statistically at chance across models) and models in the *fine-tuned* condition learned to perfectly classify basis images, but failed to distinguish them from relational or coordinate deformations. Rel = relational deformation; Cod = coordinate deformation. See the online article for the color version of this figure.

# Appendix B

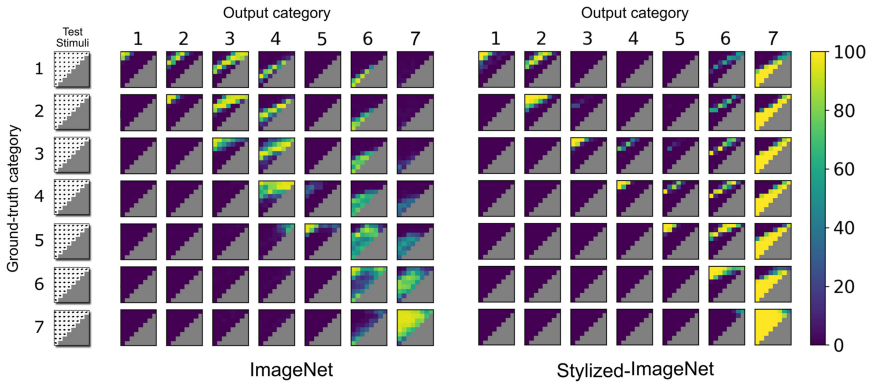
## Examining Errors in Experiment 3

In Experiment 3, we observed that performance decreased as a function of coordinate distance for most categories. However, in most simulations, we also observed that there was one category where performance was really high for most deformations, including large rotations. For example, in Figure 6, most categories show a large decrease in performance with an increase in rotation of test images, except for Cat 7 (both middle and bottom rows). To understand why this was the case, it is useful to look at the errors made by the network. Figure B1 shows confusion matrices for two models (`VGG-16` pretrained on `ImageNet` and `Stylized-ImageNet` respectively). Each heat map shows the number of times an output category was chosen for all deformations of a given input category. This confusion matrix shows that both networks were prone to misclassify large rotations from any category as belonging to Cat 7 (note a large number of classifications in the final column of each matrix for large rotations). These false positives (Type I errors) create a bias in the accuracy results for Cat 7 in Figure 6—that is, the high accuracy for large rotations for Cat 7 are, in fact, misleading as the networks classify large rotations for any category as Cat 7. These confusion matrices also show that the networks showed a "rightward" bias—there are more Type I errors in the upper triangle of each matrix than in the lower triangle. In other words, the network was more likely to misclassify images from each category as the category above rather than the category below.

(*Appendices continue*)

**Figure B1**

*Confusion Matrices for* `VGG-16` *in Experiment 3*




*Note*. For any heat map, the category label along each row shows the ground truth—that is, all test shapes used to obtain the heat map were obtained by distorting the Basis shape from that category. The category label along the column shows the output category label assigned by the network. Therefore, in each row, the diagonal heat map shows the correct classifications, while the off-diagonal heat maps show how each deformation was misclassified. See the online article for the color version of this figure.



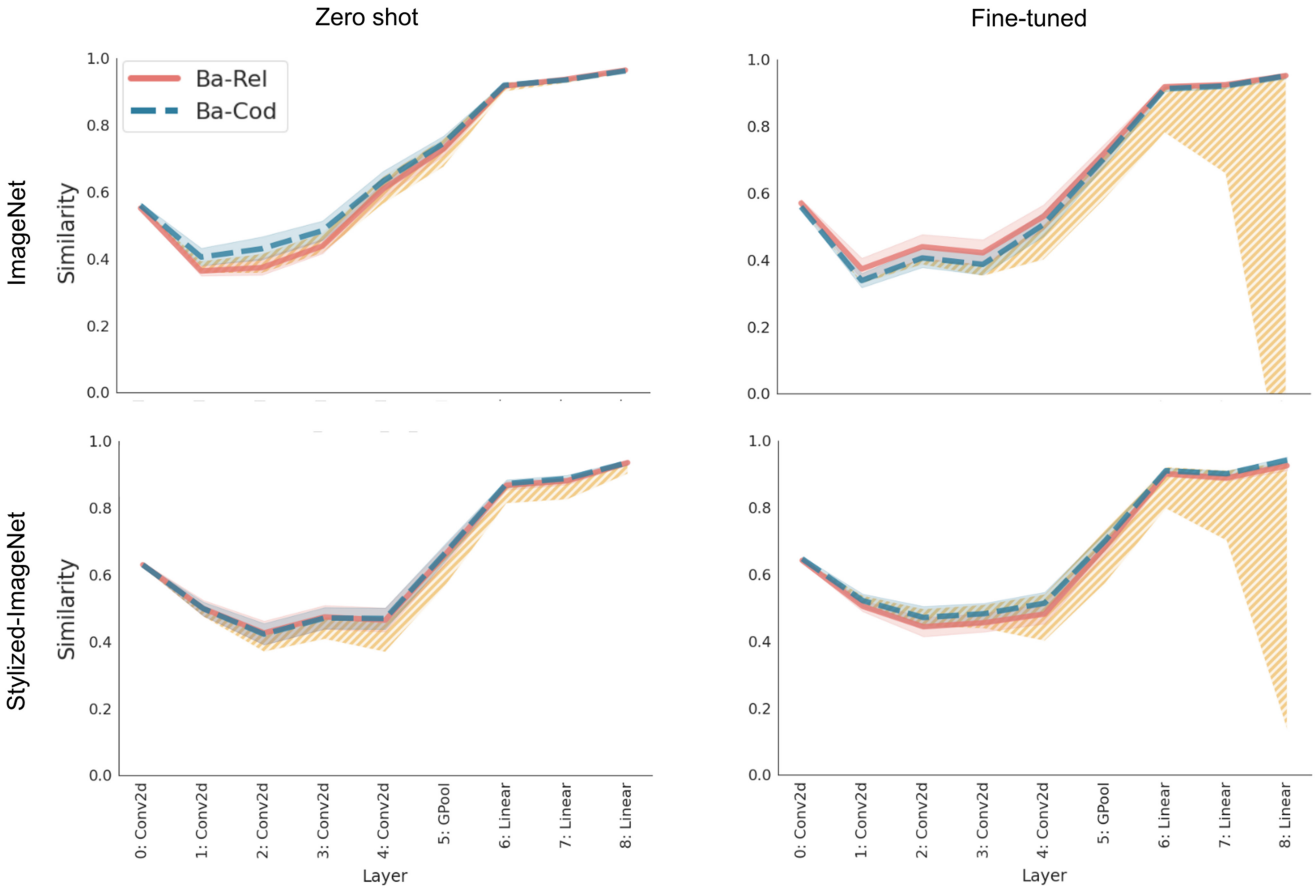
(*Appendices continue*)

# Appendix C

## Results for AlexNet

**Figure C1**

*Cosine Similarity for Internal Representations for* `AlexNet` *in Experiment 1*

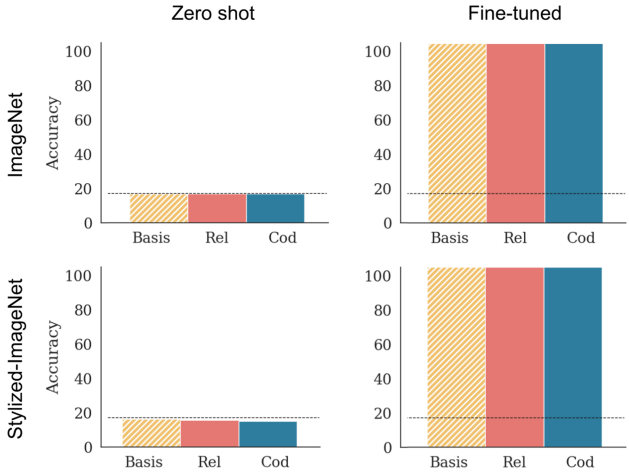



*Note.* Like the results for `VGG-16` (compare with Figure 2 in the main text), the similarity between basis images and both types of deformations is at the upper bound throughout the network, showing that the network does not distinguish the trained (basis) image from its `Rel` and `Cood` deformations. Ba = Basis; Rel = relational; Cod = coordinate; Conv2d = two-dimensional convolution layer; GPool = global pooling. See the online article for the color version of this figure.

(*Appendices continue*)

**Figure C2**

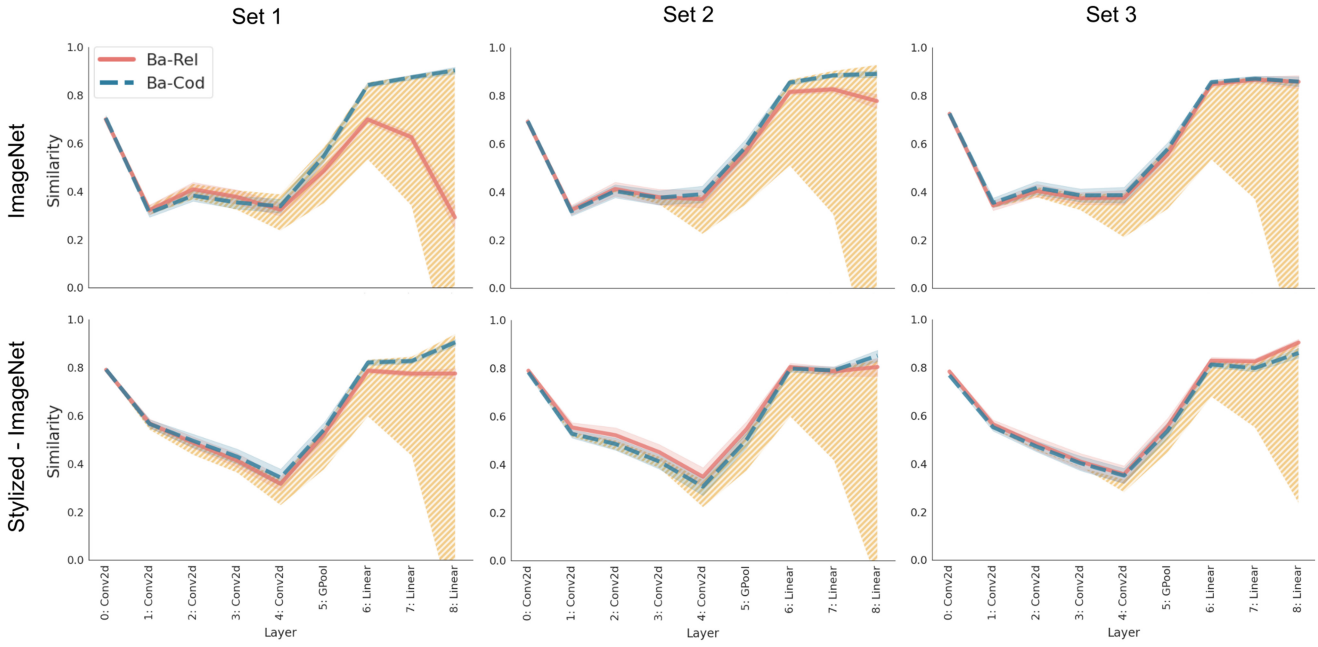
*Performance of `AlexNet` in the Test Set for Experiment 1*




*Note.* Each panel shows accuracy on the basis of shapes as well as the two types of deformations: relational (`Rel`) which changes a categorical relation and coordinate (`Cood`), which preserves all categorical relations. Compare with performance of `VGG-16` in Figure A1. Rel = relational; Cod=coordinate. See the online article for the color version of this figure.

**Figure C3**

*Cosine Similarity for `AlexNet`, Trained on Diagnostic Relations in Experiment 2*

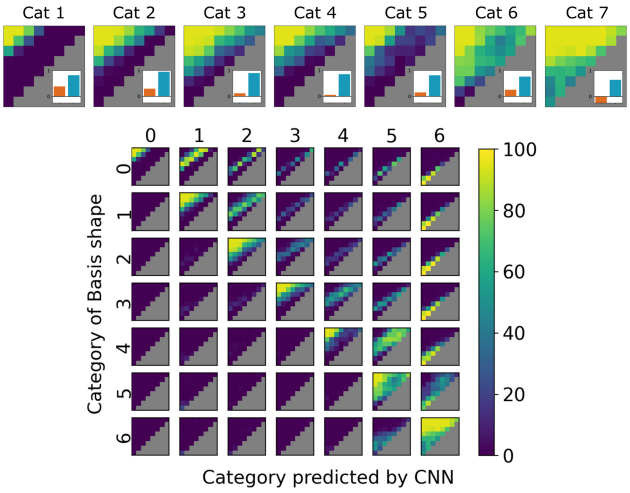



*Note.* Like the results for `VGG-16` (compare with Figure 4), we see that networks learn to distinguish the `Rel` deformation from the basis image for Set 1 (left column) when it has seen the specific deformation in the training set. But this sensitivity to `Rel` deformation diminishes in Set 2 (middle column) when only one pair of trained shapes have a similar deformation and is completely lost for Set 3 (right column) when the network has been trained on the `Rel` deformations, but the specific deformation tested is novel. Ba = Basis; Rel = relational; Cod = coordinate; Conv2d = two-dimensional convolution layer; GPool = global pooling. See the online article for the color version of this figure.

(*Appendices continue*)

**Figure C4**

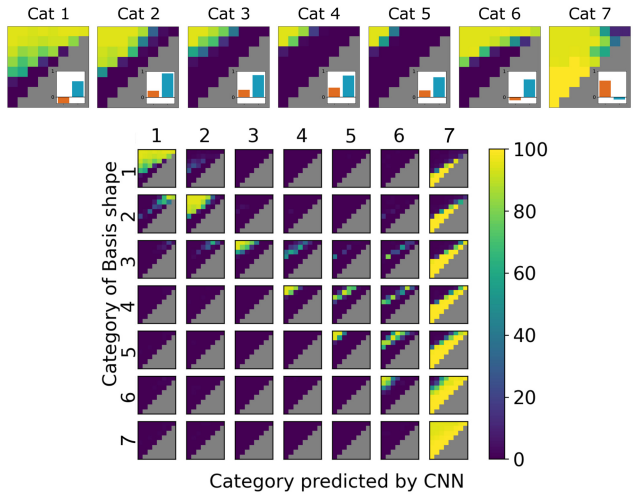
*Classification Performance for* `AlexNet` *Trained on* `ImageNet` *in Experiment 3*




*Note.* Each heatmap shows accuracy on test items for a particular category for `AlexNet` pretrained on `ImageNet` and fine-tuned on the data set in Figure 5. Each cell in the heatmap corresponds to a deformation that is a combination of relational (shear) and coordinate (rotation) transformations of the trained Basis shapes (see Figure 5A). The grid at the bottom shows the "confusion matrix"—each heatmap in the grid shows the proportion of responses predicted as the category along the column for a deformation with Basis shape taken from the category along the row. Like the results for `VGG-16` (compare with Figure 6), we see that accuracy decreases as a function of coordinate distance from the Basis shape, rather than the relational distance. Cat = category; CNN = convolutional neural network. See the online article for the color version of this figure.

**Figure C5**

*Classification Performance for* `AlexNet` *Trained on* `Stylized-ImageNet` *in Experiment 3*



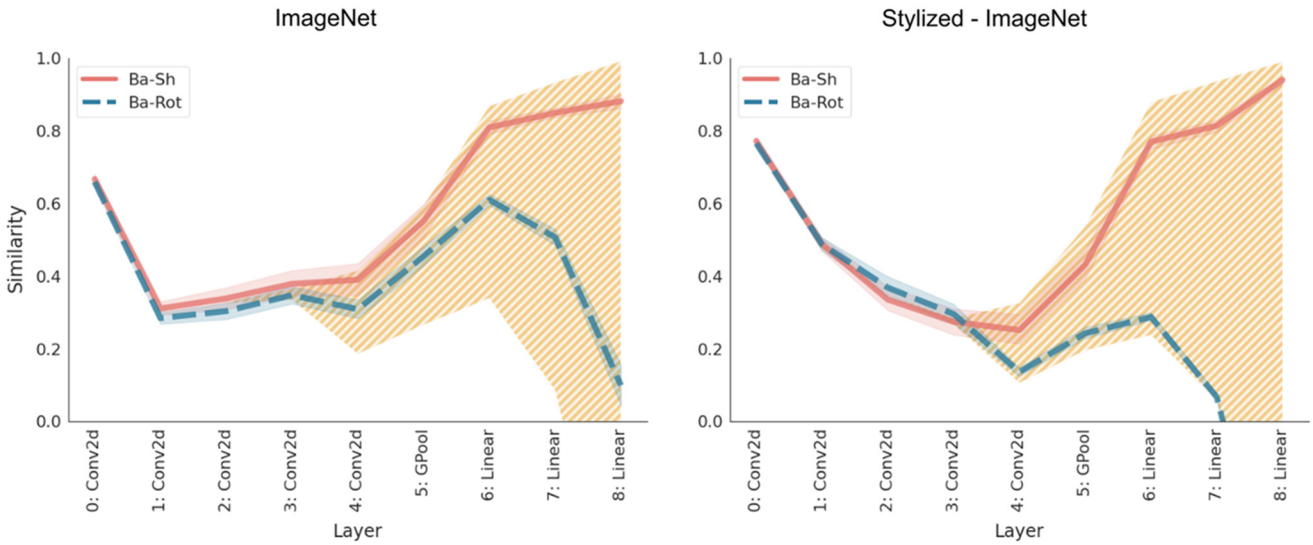

*Note.* Each heatmap in the top row shows accuracy on test items for a particular category for `AlexNet` pretrained on `Stylized-ImageNet` and fine-tuned on the data set in Figure 5. The bottom panel shows the confusion matrix. See Figure C4 for explanation. Cat = category; CNN = convolutional neural network. See the online article for the color version of this figure.



(*Appendices continue*)

**Figure C6**

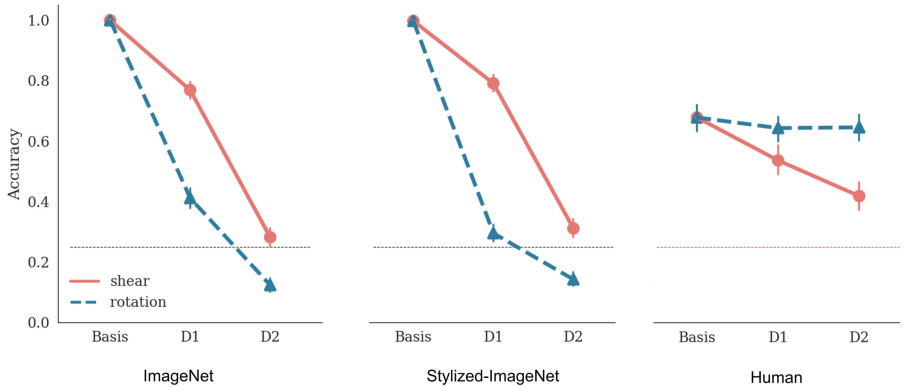
*Cosine Similarity for `AlexNet` in Experiment 3*




*Note.* Cosine similarity between internal representations for the Basis shapes and two deformations of the Basis shape (dashed red squares in Figure 5B) from the polygons data set at each convolution and fully connected layer of `AlexNet`. The solid (red) line shows the average similarity between representations for a Basis shape and its relational (shear) deformation, while the dashed (blue) line shows the average similarity between a Basis shape and its coordinate (rotation) transformation. The hatched area shows the bounds on similarity, with the upper bound determined by the average similarity between two Basis shapes from the same category and the lower bounds determined by the average similarity between two Basis shapes of different categories. Like the results for `VGG-16` (compare with Figure 7), we observed that the network treated the relational (shear) deformation as being more similar to the Basis shape than the coordinate (rotation) deformation. Ba=Basis; Sh=shear; Rot=rotational; Conv2d = two-dimensional convolution layer; GPool = global pooling. See the online article for the color version of this figure.

**Figure C7**

*Comparison of Classification Accuracy for `AlexNet` (Experiment 3) and Human Participants (Experiment 4)*

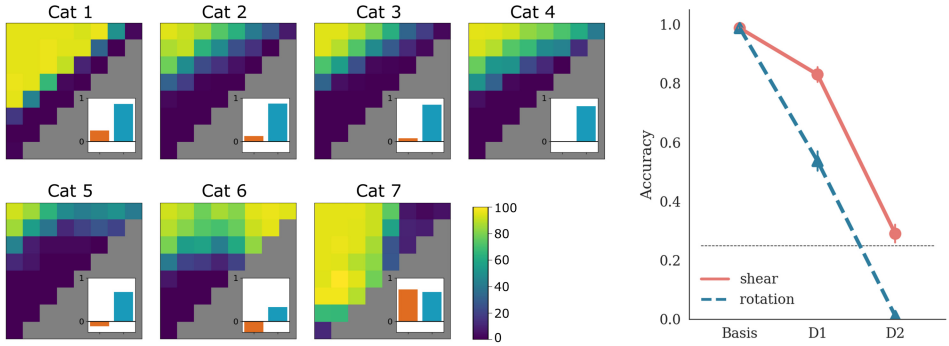



*Note.* Each panel shows performance under three conditions: basis image, deformation D1, and deformation D2. For the shear deformation (solid, red line), D1 and D2 consist of images in the top row in the fourth and eighth column in Figure 5. For the rotation deformation (dashed, blue line), D1 and D2 consist of images in the first column and fourth and eighth rows. Error bars show a 95% confidence interval and the dashed red line shows chance performance. Note that the results in the right-hand panel are reproduced here for convenience but are the results of the same experiment reported in Figure 8, right-hand panel. D = deformation. See the online article for the color version of this figure.

(*Appendices continue*)

**Figure C8**

*Performance of* `AlexNet` *in Experiment 5, Simulation 1*

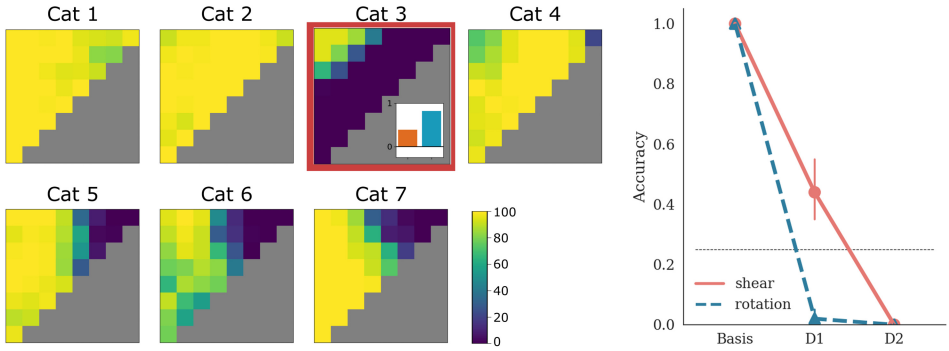



(a) *Accuracy landscape* (b) *Accuracy for conditions in experiment*

*Note*. Performance of `AlexNet` fine-tuned on an augmented data set where the Basis shapes are not only translated and scaled but also randomly rotated in the range [ − 45°, 0°]. The network is then tested on shear and rotation deformations in the range [0°, + 45°]. Like the results for `VGG-16` (compare with Figure 9), we observed that even when the network was trained on some rotations, its performance on untrained rotations (a coordinate transformation) was still worse than shears (a relational transformation). (b) shows accuracy for deformation levels D1 and D2 used for testing human participants. Cat = category; D = deformation. See the online article for the color version of this figure.

**Figure C9**

*Performance of* `AlexNet` *in Experiment 5, Simulation 2*




(a) *Accuracy landscape for left-out Category* (b) *Accuracy for conditions in experiment*

*Note*. Performance of `AlexNet` fine-tuned on an augmented data set where the Basis shapes are not only randomly translated and scaled but also rotated. For six out of seven categories, the network is trained on *all* rotations [0°, 360°]. We then tested the network on the left-out category (Cat 3, highlighted with a red square in Panel A) on untrained rotations and shears. However, we observed that despite being trained in this manner, the accuracy degraded as a function of the coordinate deformation, rather than the relation deformation. (B) shows the performance of this network for deformations D1 and D2 used to test human participants. Cat = category; D = deformation. See the online article for the color version of this figure.

Received August 9, 2022
Revision received February 17, 2023
Accepted April 22, 2023 ■