The contrasting shape representations that support object recognition in humans and CNNs

Gaurav Malhotra,[1*] Marin Dujmovic[1], John Hummel,[2] Jeff Bowers[1]
[1]School of Psychological Sciences, University of Bristol, Bristol, UK
[2]Department of Psychology, University of Illinois Urbana-Champaign, Champaign, USA
* gaurav.malhotra@bristol.ac.uk.

1 Abstract

Recently it has been argued that CNNs trained to perform object classification through supervised learning can emulate a key property of human vision, namely, they can classify objects based on their shape. However, it is unclear whether their learned shape representations are human-like. We explored this question in the context of a well-known observation from psychology showing that human shape representations encode the relations between object features. We ran a series of simulations where we trained CNNs on datasets of novel shapes and tested them on a set of controlled deformations of these shapes. We found that CNNs do not show any enhanced sensitivity to deformations which alter relations between features, even when explicitly trained on such deformations. This behaviour contrasted with human participants in previous studies as well as in a new experiment. We argue that these results are a consequence of a fundamental difference between how humans and CNNs learn to recognise objects: while CNNs select features that allow them to optimally classify the proximal stimulus, humans select features that they infer to be properties of the distal stimulus. This makes human representations more generalisable to novel contexts and tasks.

Introduction

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 [10]. We also prefer to view objects from certain viewpoints – a bias for canonical-perspectives [39]. This paper is focused on one such bias – the *shape-bias* – the observation that humans, from a very young age, prefer to categorise objects based on their shape, rather than other prominent features such as colour, size or texture [34, 6]. 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 [6] and we can do this even if we have no previous experience with line drawings [18].

Two different explanations have been proposed regarding the origin of these biases. The first view proposes that these biases are an internalisation of the biases present in the environment relevant for classifying objects. According to this view, humans prefer to view objects from a canonical perspective because these perspectives are more frequent in the visual environment and they prefer to classify objects based on shape because shape is more diagnostic during object classification. In other words, biases are a consequence of performing statistical learning with the goal of optimising behaviour on a particular task. We will call this the optimisation-for-classification approach or, more briefly, the *optimisation approach*.

The second view, which we call the *heuristic approach*, proposes that biases originate because the visual system needs to transform the *proximal stimulus* – i.e., the retinal image – into a representation of the *distal stimulus* – i.e., a 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 [41, 38]. 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 [48, 30, 36, 42]. The advantage of distal representations is that they are relevant for broad range of tasks – the same representation of an object can be used for recognition and visual reasoning [23] amongst other visual skills.

The goal of this study was to adjudicate which of these approaches provides a better explanation for shape-bias in human object recognition. We tested this by focusing on supervised Convolutional Neural Networks (CNNs) – which are machine learning models that recognise objects by learning statistical features of their proximal stimuli that can be used to optimally classify a each stimulus, given some training data. The learned representations that support object recognition are specialized for image classification. There is no pressure to learn representations of objects than can perform a range of tasks, let alone learn distal representations of objects. As such, CNNs trained using supervised learning to classify objects provide a concrete model to test the optimisation view. If human perceptual biases are acquired through internalising the statistics of the environment in order to classify objects, then training CNNs to perform classification on ecologically realistic datasets 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 datasets of naturalistic images (e.g. `ImageNet`) frequently lacked a shape-bias, instead preferring to classify images based on texture [13] or other local features [1, 35]. 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. [13] trained standard CNNs on Style-Transfer image dataset that mixes the shape of images from one class with the texture from other classes so that only shape was diagnostic of category. CNNs trained on this dataset learned to classify objects by shape. In another study, Feinman and Lake [11] 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. [17] showed that more psychologically plausible forms of data augmentation, namely the 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".

These results raise the possibility that human biases are indeed a consequence of internalising 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 optimisation, 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 [2, 26]. 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 [20]. The question we ask is whether CNNs trained to classify objects learn to encode these relational features of shape. Across two sets of experiments we show that even though CNNs are able to categorise objects based on shape, they do this on the basis of qualitatively different shape representations to humans. In both experiments human behaviour is better explained by the heuristic approach than an optimisation approach.

The rest of the paper is divided into three sections. In the first section, we focus on objects that consist of multiple parts and, in the second section, on 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 (optimisation and heuristic) views. We then present 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 for developing models of human vision.

Experiment 1: multi-part objects

Proximal and distal encodings of multi-part objects

What sort of deformations of the proximal stimulus should allow us to contrast the optimisation and heuristic approaches? Specific hypotheses can be derived from the structural description theory [2], which assumes we build representations of distal stimuli during the process of identifying objects. On this theory, objects are represented 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 code) 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 multi-part 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 deformations that change 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 than others in the context of specific objects, but it is unclear why they would show a general tendency to treat categorical deformations as different than metric ones since there is no heuristic that assumes that categorical relations between parts is central feature of object shape representations. (Indeed, there is no explicit encoding of parts at all.)

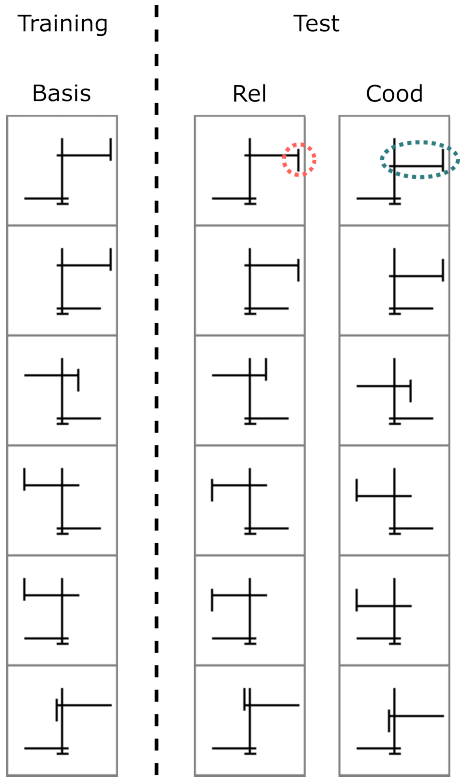


Fig 1. Stimuli used by Hummel and Stankiewicz [26]. The first column shows a set of six (Basis) shapes that participants were trained to recognise. Participants were then tested on shapes in the second and third columns, which were generated by deforming the Basis shape in the corresponding row. In the 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 coordinates of some elements are shifted (highlighted in blue ellipse).

Instead, all deformations are simply changes in the locations of features in the image.

Hummel and Stankiewicz [26] explored this question in the context of comparing structural description and view based models 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 then created two deformations of each of these *Basis* object. 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 composing the object, but changed the coordinates of two lines. Note that both variants deformed the objects (proximal stimulus) but the relational variant changes categorical relations between parts of the object.

Across five experiments participants first learned to classify a set of base objects and then tested on 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, 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 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 [20].

Testing CNNs on the Hummel and Stankiewicz [26] stimuli

The findings of Hummel and Stankiewicz [26] provide a critical test for any model that claims to be a theory of human vision: if shape-bias in humans is a consequence of optimising over an object recognition task, then it should also lead to shape representations that are more sensitive to relational deformations than coordinate deformations. To test this hypothesis, we replicated the experimental setup of Hummel and Stankiewicz, replacing human participants with two well-known CNNs – `VGG-16` and `AlexNet` – that have been previously argued to capture human-like representations [51, 31] and an ability to develop a shape-bias [17, 13].

We took CNNs that were either pre-trained on `ImageNet` (a large database of naturalistic images) or `Stylized-ImageNet` (a database specifically constructed to induce a shape-bias, see [13]). We then tested these networks under either a *Zero-shot* learning condition, where we observed the response of the pre-trained networks to the Basis, Rel and Cood images, or a *Fine-tuned* learning condition, where we further trained these pre-trained networks to categorise the `Basis` images (at variout rotations, translations and scales) and observed their response to the Rel and Cood variants (see Methods for details).

We determined each network's ability to distinguish between the basis shapes and their relational and coordinate deformations by computing the similarity between the internal representation for a basis shape and each type of deformation (see Appendix 1.1 for the networks' classification performance). Results for `VGG-16` are shown in Figure 2 and for `AlexNet` in Appendix 1.4. Each panel in Figure 2 corresponds to a combination of pre-training and test conditions and shows the average cosine similarity between internal representations for a Basis image and it's relational (Ba-Rel, solid red line) and coordinate (Ba-Cood, dashed blue line) deformations. The internal representations are computed at all convolutional and fully connected layers within the network. We compared these similarities to two baselines: the average similarity between two Basis images that belong to the same category and the average similarity between two Basis images that belong to different categories. These two baselines provide the upper and lower bounds on similarities (hatched yellow region).

We observed that in the *Zero-shot* condition (left-hand column in Figure 2), the similarity between a basis image and it's relational variant was statistically the same as the similarity between the basis image and it's 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 and the classification performance in Figure S1). Thus pre-training on `ImageNet` or `Stylized-ImageNet` was not sufficient for networks to distinguish between the stimuli or their deformations used by Hummel and Stankiewicz – to these models, all line drawings are alike. In contrast, the networks successfully learned to distinguish between stimuli from different categories in the *Fine-tuned* condition (see classification performance in Figure S1). Examining the internal representations showed that the networks represented all types of images in a similar

Fig 2. Cosine similarity between internal representations at convolutional (Conv2d) and fully connected (Linear) layers of `VGG-16` network that is either pre-trained on `ImageNet` or on `Stylized-ImageNet` – a dataset developed to induce a shape-bias in CNNs. 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. 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 95% confidence interval. Based on the results of Hummel and Stankiewicz [26], we would expect the solid (red) line (Ba-Rel) to be closer to the lower, rather than upper bound. Instead we observe that it stays at the upper bound throughout the network and is statistically indistinguishable from the dashed (blue) line, showing that there is no significant difference between the Basis shape and either (relational or coordinate) deformation at any layer of the network.

manner in the early convolution layers (there is no statistical 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, both types of pre-treained 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). Note that this is not because the networks overfit to the training data. In fact, networks showed very good generalisation to both novel Basis images (in unseen combinations of rotations, translation and scale) and the two types of deformations, with the cosine distance between a basis shape and either deformation close to the upper bound of similarity (also see the high classification performance for both deformations in Figure S1). 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.

Teaching relational representations of multi-part objects

The above results suggest that training CNNs on `ImageNet` or even `Stylized-ImageNet` is not sufficient for these models 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 optimisation approach, but due to the limitation of these datasets. It is possible that if the classification model is trained on a dataset where relational differences are important for classification, it may internalise this statistic and start 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 generalise this knowledge to novel (but highly similar) relational changes.

Consider the three augmented training sets shown in Figure 3. In each set the pre-trained networks are trained on the six Basis shapes (and their translation, rotation and scale transformations) just like in the *Fine-tuned* condition in the experiments above. In addition, they are also trained on five new shapes. These five shapes are the Rel deformations of the first five Basis shapes. In other words, the training set assigns different categories to a shape and it's Rel deformation for five out of six figures. After training the networks on these eleven $(5 + 5 + 1)$ shapes, we tested them on the Rel and Cood deformations of the final (unpaired) Basis shape.

The difference between the three datasets lies in the degree of novelty of test images. In all three datasets in Figure 3, the same relation (dashed red circle) is changed between the unpaired Basis shape and it's 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 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 central vertical line for all other trained images).

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 S8). As in previous simulations, we tested networks that were either pre-trained on `ImageNet` (first row) or on `Stylized-ImageNet` (second row) and fine-tuned to each training set. 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 behaviour of human participants in Hummel and Stankiewicz [26]. 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 the this effect was significantly reduced (middle column in Figure 4) – the cosine similarity Ba-Rel was slightly lower than Ba-Cood but by a much smaller degree and the difference only existed only for the networks pre-trained on `ImageNet` and only in the fully connected layers (also compare results in Figure S8 in Appendix for `AlexNet`, where this effect is slightly more

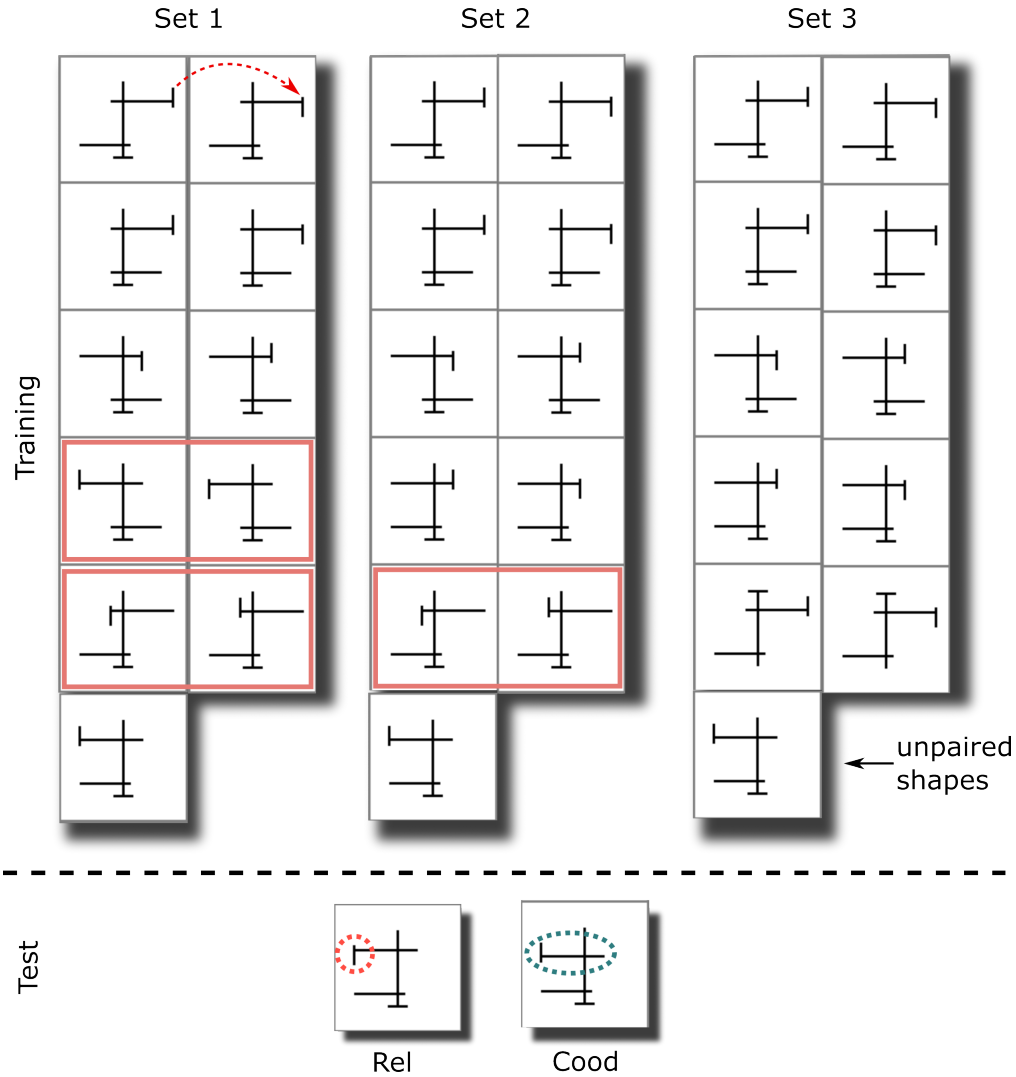


Fig 3. Three training sets that try to teach the network to recognise relational changes. In each set, the first column shows a set of six unique Basis shapes, while the second column shows Rel deformations of the first five shapes (see red arrow). At the bottom are the two test shapes. These test shapes are identical to the eleventh (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 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.

Fig 4. Cosine similarity between Basis image and two types of deformations for a `VGG-16` network pre-trained on either `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, the solid (red) line shows similarity between Basis and the relational deformation while the dashed (blue) lines shows similarity between Basis and the coordinate deformation. 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 (solid red line is much lower than 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.

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" – i.e., changing relations led to a change in an image's category mapping.

Experiment 2: single-part objects

Deformations for testing single-part objects

As detailed above, structural description theories claim that the categorical relations between object parts are encoded in order to build distal representations of multi-part objects. But of course, in order to build distal representations of complex objects, it is also necessary to build distal representations of the parts themselves. This raises the question of what sorts of deformations of the proximal stimulus should allow us to contrast optimisation and heuristic approaches for identifying the component parts of complex objects or single-part objects? According to the structural description theory [2], 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, for example, 43]. 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 nonaccidential feature in the proximal stimulus (e.g., a straight line) is the product of an accidental alignment of 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 [2] 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. By contrast, it is not clear that CNNs optimized to classify objects will encode symmetry or other relational features used to build distal representations. Accordingly, CNNs may be insensitive to deformations of symmetry or other relations present in the proximal stimulus.

With this in mind, we created set of seven symmetrical pentagons (Figure 5(a)), and made deformations of 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, vs. (b) a relational change that had an equivalent impact as the corresponding rotation, but instead introduced a shear that changed relative location of the polygon's vertices. Although the specific manipulation is different from that used in Experiment 1, the general logic is the same: one of the deformations preserves the relations between object features while the other changes them. To a model that looks only at 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 5(b) shows some examples of test images for one of the trained shapes. These test shapes are organised 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 5(b) is a combination of a rotation and a shear of the basis shape in the top-left corner. We have organised 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 organised 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 (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 5(c) shows predicted accuracy on this landscape for the two types of shape representations.

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

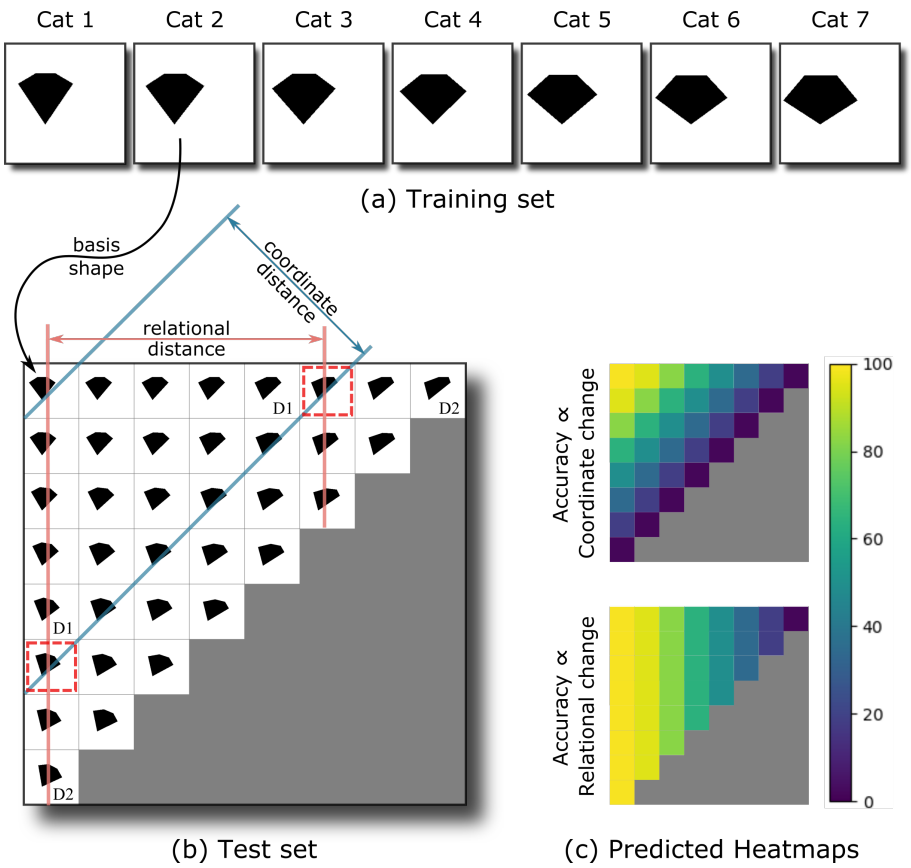


Fig 5. Stimuli used to test shape representations in single-part objects. (a) The shapes in the Basis set used for training. Each shape is presented at various translations and scales. (b) The test set for one of the categories (Cat 2) is obtained by deforming the Basis shape (in the top-left corner) through a combination of rotation and shear operations. Here we have organised these deformations in a matrix based on their coordinate distance (measured as cosine distance) and relational distance (measured as change in 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 presented as heat-maps, assuming that accuracy is a function of coordinate distance (top), or relational distance (bottom).

Performance of CNNs

We again tested CNNs pre-trained on either `ImageNet` or `Stylized-ImageNet`. The performance of a typical CNN, here `VGG-16`, is shown in Figure 6 (we obtained a qualitatively similar pattern of results for `AlexNet`, see Appendix 1.4). For all networks, we observed 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. Thus, unlike Experiment 1, where we were able to observe only ceiling performance for both deformations, the design of Experiment 2 allowed us to compare how performance degrades for the two types of deformations. Crucially, we observed that for most categories, accuracy decreased as a function of distance to the Basis shape

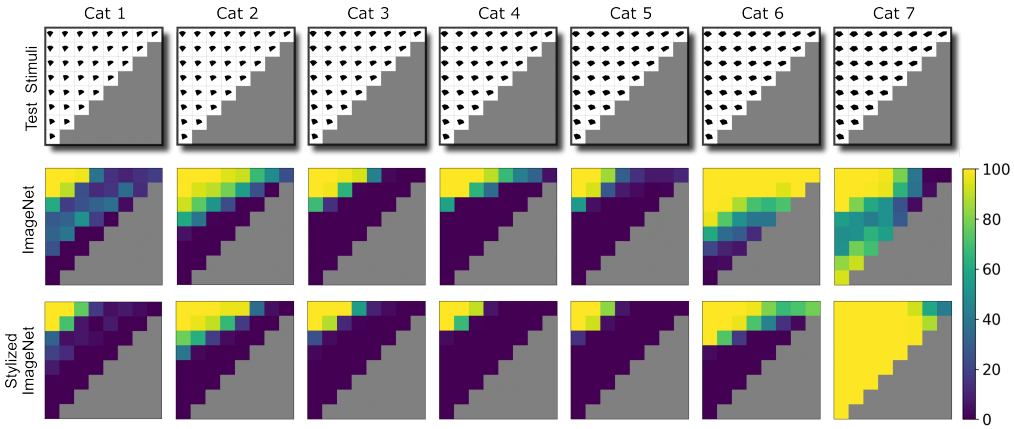


Fig 6. Test stimuli for each category (first row) and accuracy on the landscape of relational and coordinate deformations for `VGG-16` pre-trained on either `ImageNet` (middle row) or `Stylized-ImageNet` (bottom row). In each case, the network was fine-tuned on the set of seven polygons shown in Figure 5(a). 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. For most categories, accuracy is highest for small deformations (top-left corner) and decreases as a function of the coordinate distance from the basis shape (perpendicular to diagonal). The relational distance (left-to-right) has no added effect on this decrease in accuracy.

(perpendicular to the diagonals), rather than relational change (left to right). 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 on 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 Appendix 1.2 for details). Overall, these results suggest that the network does not represent the shapes in this task in a relational manner. If it did, it's performance on relational changes should have been a lot worse than it's performance on relation-preserving rotations.

But classification accuracy only provides a indirect measure of internal representations. In order to get a more insight into the network's internal representations for relational and coordinate deformations, we examined the cosine distance between internal representations for the Basis image and two test images that were equidistant from it. An example of these images is highlighted (dashed red squares) in Figure 5(b). These cosine similarities for two `VGG-16` networks are plotted in Figure 7 (we again obtained qualitatively similar results for `AlexNet` – see Figure S11 in Appendix). At all internal layers, we observed that the average similarity between a Basis image and it's relational (shear) deformation was equal or higher than the average similarity between the Basis image and it's coordinate (rotation) deformation (compare solid (red) and dashed (blue) lines in Figure 7). In other words, relational deformation of an image was closer to the Basis image than it's coordinate deformation and pre-training on the `Stylized-ImageNet` dataset 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. It may be argued that this difference between humans and CNNs may be due to a difference in experience, with humans experiences objects in different rotations. However, even when CNNs were trained to recognise objects in different rotations, we observed a qualitatively similar pattern – performance on deformations depended on the (cosine or Euclidean) distance to the

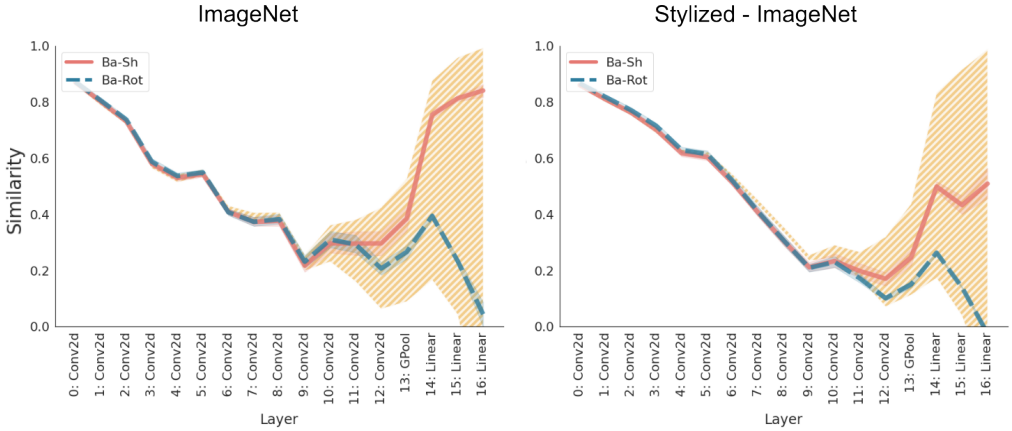


Fig 7. Cosine similarity in internal representations of `VGG-16` in Experiment 2. The solid (red) and dashed (blue) lines show the average cosine similarity between Basis images and relational (shear) and coordinate (rotation) deformations, respectively. The hatched (yellow) region shows the bounds on 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 affect on internal representations than a coordinate (rotation) deformation, one would expect the solid (red) line to be below the dashed (blue) line.

corresponding Basis images, rather than selectively on the degree of relational change (see Appendix 1.3).

Performance of human participants

The optimisation view (CNNs) and heuristic view (structural description theory) make contrasting predictions of how performance should degrade when a learned shape is deformed through rotation and shear transformations. In our next experiment, we examined which of these predictions holds for human participants.

We trained 23 participants on the same categorisation task used to train the CNNs above. Participants saw the polygons shown in Figure 5(a) and had to learn to categorise them. Once they had learned this task, they were tested on four deformations of each Basis shape – two shears and two rotations. These deformations are marked as D1 and D2 in Figure 5(b) (see Methods for details).

The average accuracy of classification on each of these deformations is shown in Figure 8. We can see that irrespective of training, `VGG-16` was more sensitive to rotation than to shear (see Figure S12 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 $t(22) < 3.48$, $p > .28$), while performance decreased significantly for each of shear deformations (both $t(22) > 14.10$, $p < .001$, $d_z > .83$). The largest shear resulted in largest decrease in performance ($M_{difference} = 25.87\%$). Thus, the behaviour of participants was in line with the prediction of structural description theories, where shape is encoded based on relations between internal parts, and in the opposite direction to the performance of the CNNs.

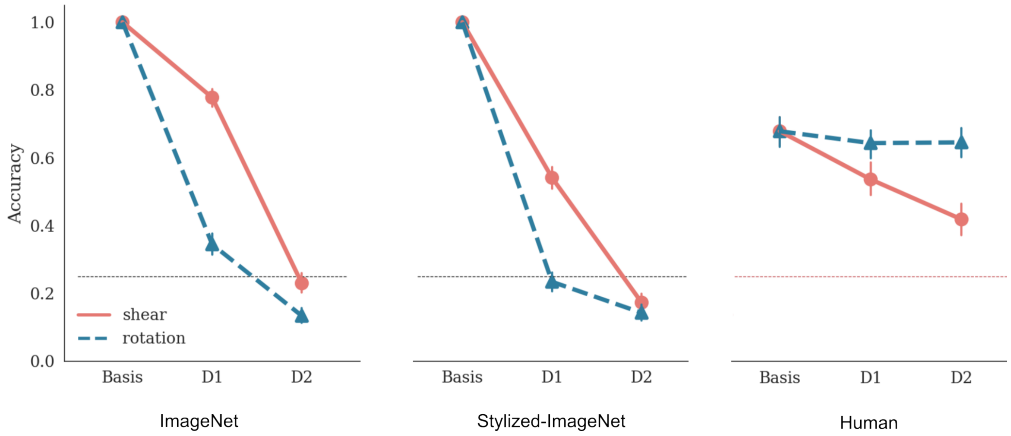


Fig 8. Comparison of how classification accuracy changes for `VGG-16` (left and middle) and human participants (right) with stimuli deformation. Each panel shows performance under three conditions: basis image, deformation D1 and deformation D2. For the shear deformation (solid, red line), D1 and D2 consists 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 95% confidence interval and dashed black line shows chance performance.

Discussion

In two sets of experiments we have shown that CNNs and humans represent shape in qualitatively different ways as reflected in the biases they show in generalisation. In Experiment 1 we compared how CNNs and humans encode multi-part objects following deformations in the categorical relation between parts (Relational Variants) and deformations that maintained the categorical relations between parts (Coordinate Variants). Whereas humans are highly sensitive to deformations in the categorical relations between parts [26, 20], we found that CNNs are entirely insensitive to these deformations, with performance only a function of the cosine distance between images in pixel space. Furthermore, we could not train CNNs to classify objects on the basis of the relations between parts. In Experiment 2, we compared CNNs and humans in the classification of single part objects when they were deformed by sheering (Relational) or by rotating (Coordinate) manipulations. Again, we found that humans are highly sensitive to relational deformations, whereas CNNs are only sensitive to coordinate distance, and once again, CNNs could not learn to be sensitive to relational manipulations.

These findings challenge the hypothesis that CNNs and humans perceive objects based on similar principles and that apparent differences arise due to "differences in the data that they see" [17]. The results of both experiments 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, Puebla and Bowers [44] have found that CNNs fail to support a simple relational judgement 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. [1] 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.

These failures may reflect a range of processes present in humans but absent in

CNNs trained to recognise objects through supervised learning, such as figure-ground segregation, completing objects behind occluders, encoding border ownership, and inferring 3D properties about the object [43]. Consistent with this hypothesis, Jacob et al. [27] and Bowers et al. [7] have recently highlighted a number of these failures in CNNs, including a failure to represent 3D structure, occlusion, and parts of objects. More broadly, these results challenge the the claim that CNNs trained to recognise objects through supervised learning are good models of the ventral visual stream of human vision (see, for example, [52, 8, 37]).

One interesting study that provides some evidence to suggest that standard CNNs have similar shape representations to humans was reported by Kubilius et al. [33]. In one of their experiments (Experiment 3), they compared the similarity of representations in various CNNs in response to a change in metric and non-accidental features of single-part objects. For instance, they compared a base object that looked like a slightly curved brick to two objects: one object that was obtained by deforming the base object into a straight brick (a non-accidental change) and a second object that was obtained by deforming the base object into a greatly curved brick (a metric change). Kubilius et al. reported that, like humans, CNNs were more sensitive to non-accidental changes. However, it is unclear whether CNNs were more sensitive to one of their manipulations because of the non-accidental change or because of other confounds accompanying these manipulations. For example, when Kubilius et al. modified some of the base shapes to non-accidental deformations, it was accompanied by a change in local features (such as properties of vertices). Recent research [1, 13] has shown that, unlike humans, CNNs are in fact highly sensitive to change in local and textural features and it is unclear whether it is these types changes that are driving the effects observed by Kubilius et al. [33]. More work is required to reconcile their findings with our own.

Of course, it is possible that training CNNs 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. generative modelling [29] or on a "self-supervised" task [16]) 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 recognising 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 [13, 17, 33]. Secondly, these models have had the largest success in predicting neural representations in human and primate visual system [51, 8, 46] and it has been argued that there is a "strong correlation between a model's categorization performance and it's ability to predict individual level IT neural unit response data" [52]. 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., 2].

It is also possible that a different Deep Learning architecture may be more successful than CNNs at encoding objects based on relations between their parts. Indeed, previous research indicates that relational reasoning may require a more powerful architecture that can explicitly and separately represent (i) parts and relations, and (ii) their bindings [e.g., to distinguish whether the brick is above the cone or vice-versa; 9, 21, 23, 24, 25]. Other Deep Learning architectures such as Capsule Networks [45], Transformers [50], LSTMs [19] or Neural Turing machines [15] may also provide the representational power necessary to represent structural descriptions, but to date this has yet to be demonstrated.

Methods

Generating training and test sets

Training and test sets for Hummel and Stankiewicz [26] We constructed six basis shapes that were identical to the shapes used by Hummel and Stankiewicz [26] in their Experiments 1–3. Each image was sized 196x196 pixels and consisted of five black line segments on a white background organised 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 8 (2x2x2) possible Basis shapes. We selected six out of these to match the six shapes used by Hummel and Stankiewicz [26]. Following Hummel and Stankiewicz [26], we constructed Rel (relational) deformations (called V1 variants by Hummel and Stankiewicz [26]) of each Basis shape by shifting the location of the top vertical segment, so that it's categorical relation to the upper horizontal segment changed from "above" to "below". Similarly, we constructed Cood (coordinate) deformations (called V2 variants by Hummel and Stankiewicz [26]) 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. Each training set contained 5000 images in each category. When no augmentation was used, all the images in each category were identical – i.e. the shape appeared at the identical location on the canvas for all images in a category. In addition, we constructed two augmented datasets, one in which the Basis image was translated to a random locations (in the range $[-50, +50]$ pixels) on the canvas and another in which it was additionally randomly scaled ($[\frac{1}{2}, 1]$) or rotated ($[-20°, +20°]$).

As described above, we generated three additional datasets for teaching CNNs to recognise relational deformations on Hummel and Stankiewicz's stimuli (see Figure 3). Each of these training sets contained five pairs of Basis shape and one unpaired shape. Each pair consisted of a shape and it's Rel deformation. The test set consisted of Rel and Cood deformations of the unpaired shape. The difference in the three datasets was the amount of overlap between the trained Rel deformations and tested deformation. In the first dataset, there were two pairs of Basis shapes where a Rel deformation was constructed by changing the same categorical relation as the one that differed between the unpaired shape and the tested Rel deformation. In the second set, there was only one such pair. And in the third set, none of the trained shapes differed in the tested deformation. Each training set again consisted of 5000 images, where each image was constructed by translating, scaling and rotating the Basis shape for that category. The test set consisted of 1000 images where each image was constructed by randomly translating, scaling and rotating the Rel and Cood deformations of the unpaired Basis shape.

Training and test sets for polygons task The training set for Experiment 2 consisted of seven symmetric filled pentagons, presented on a white canvas. Each category contained 5000 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 Basis shapes for two categories was the angles between the edges. 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, it's 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 λ was the degree of shear and Δy was the distance between y_{old} and y_{apex}, the y-coordinate of the vertex at the apex. Images could also be combination of rotations and shears. To do this, the Basis image was first sheared, then rotated. We measured the distance of a deformed image and test images were organised as shown in Figure 5, where images in each column had the same degree of shear and images along each diagonal had the same (cosine or euclidean) distance to the Basis image. We then obtained twenty exemplars of each deformed image on the grid by randomly translating and scaling the image.

CNN Simulations

We evaluated two deep convolutional neural networks, `VGG-16` [47] and `AlexNet` [32] 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 1.4. Since human participants had a lifetime experience of classifying naturalistic objects prior to the experiment, we used network implementations that had been pre-trained on a set of naturalistic images. Two types of pre-training were used: networks were either pre-trained in the standard manner on `ImageNet`, or pre-trained on a set of images where shape was made more predictive than texture by using style-transfer [12]. We used networks pre-trained by Geirhos et al. [13], who have shown that networks trained in this manner have a greater shape-bias than networks trained on `ImageNet`. Networks were either tested in the *Zero-shot* condition, where no training was given on any of our datasets and we recorded the response of the pre-trained networks to the test images, or in the *Fine-tuned* condition, where the pre-trained network was fine-tuned to classify the 5000 images per category. Each of these images were obtained from the corresponding Basis image in the manner described above. This fine-tuning was performed in the standard manner [53] by replacing the last layer of the classifier to reflect the number of target classes in each dataset. The models learnt to minimise the cross-entropy error by using the Adam optimiser [28] with a small learning rate of 10^{-5} and a weight-decay of 10^{-3}. In all simulations, learning continued till the loss function had converged. To check for overfitting, we created cross-validation sets and ensured performance on 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. For the first set of experiments (Hummel and Stankiewicz [26] stimuli), the test set consisted of 1000 Basis, Rel and Cood deformations of each category. In the second set of experiments (polygons stimuli) the test set consisted of 100 exemplars (translation and scale variants) of each test image on the grid (see Figure 5). All simulations were perfomed using the `Pytorch` framework [40] and we used `torchvision` implementation of all models.

To test the similarity of internal representations (Figures 2, 4 and 7), we obtained the embedding of an image at each convolution and fully connected layer of the CNN. For the first set of simulations (Hummel and Stankiewicz [26] stimuli), we selected one (of the six) category and randomly chose 100 pairs of images from the Basis and Rel test set. We then computed the cosine similarity between embeddings of each pair. This gives the estimated average distance in the Ba-Rel condition (solid red line in Figure 2). Similarly the cosine similarity between 100 pairs of Basis and Cood test images gives the Ba-Cood distance (dashed blue line). These distances are compared against two baseline conditions. The upper limit of similarity is given by the similarity of 100 pairs of Basis

images from the same category (upper bound of the hatched yellow area in Figure 2). The lower limit is given by the similarity of 100 pairs of Basis images from different category (in each pair, one of the images was from one category and the other from one of the other six categories). The similarity of internal reprsentations for the polygons stimuli is obtained in a similar manner. The similarity Ba-Sh (solid read line in Figure 7) is estimated by measuring the average cosine similarity between embeddings of 100 pairs images from the Basis and sheared sets of the same category. Similarly, Ba-Rot is estimated by measuring the average cosine similarity between embeddings of 100 pairs of images from Basis and rotated sets of the same category.

Behavioral experiment

Participants

Participants ($N = 37$, $M_{age} = 33$, 70% female) with normal or corrected-to-normal vision were recruited via Prolific and the experiment was conducted on the Pavlovia platform. They 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 performance during testing. The average payment was 8 GBP/hour. An written ethics approval for the study was obtained for the study from the University of Bristol Ethics board.

Stimuli

Four categories were chosen from the total data set for the behavioral study. These are 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) shear in the top row of Figure 5b. For the coordinate deformation, these were the fifth (D1) and final (D2) rotations in the left most column of Figure 5b. This made up the 5 conditions in the experiment: Basis, D1 (Shear), D2 (Shear), D1 (Rotation) and D2 (Rotation). The original stimuli were 224x224 pixels but were re-scaled for each participant to 50% of the vertical resolution to account for the variability in screen size and resolution when running the study online.

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 5 training blocks, or until they reached 85% categorization accuracy in a block. Participants who managed to reach 85% accuracy continued to the test block. The order of trials was randomised for each participant. Each trial started with a fixation cross (750 ms), then the stimulus was presented (500 ms) followed by four response buttons corresponding to the four categories (until response). 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 randomised for each participant. The procedure for each test trial was the same as in the training phase apart from the fact 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.

Code and Data

All code for generating the datasets, simulating the model as well as participant data from Experiment 2 can be downloaded from: https://github.com/gammagit/distal

References

[1] Nicholas Baker, Hongjing Lu, Gennady Erlikhman, and Philip J Kellman. Deep convolutional networks do not classify based on global object shape. *PLoS computational biology*, 14(12):e1006613, 2018.

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

[3] Irving Biederman and Eric E Cooper. Evidence for complete translational and reflectional invariance in visual object priming. *Perception*, 20(5):585–593, 1991.

[4] Irving Biederman and Eric E Cooper. Size invariance in visual object priming. *Journal of Experimental Psychology: Human Perception and Performance*, 18(1): 121, 1992.

[5] Irving Biederman and Peter C Gerhardstein. dependent mechanisms in visual object recognition: Reply to tarr and bülthoff (1995). 1995.

[6] Irving Biederman and Ginny Ju. Surface versus edge-based determinants of visual recognition. *Cognitive psychology*, 20(1):38–64, 1988.

[7] Jeffrey S Bowers, Gaurav Malhotra, Marin Dujmović, Milton Llera Montero, Christian Tsvetkov, Valerio Biscione, Guillermo Puebla, Federico G Adolfi, John Hummel, Rachel Flood Heaton, et al. Deep problems with neural network models of human vision. 2022.

[8] Charles F Cadieu, Ha Hong, Daniel LK Yamins, Nicolas Pinto, Diego Ardila, Ethan A Solomon, Najib J Majaj, and James J DiCarlo. Deep neural networks rival the representation of primate IT cortex for core visual object recognition. *PLoS Comput Biol*, 10(12):e1003963, 2014.

[9] Leonidas AA Doumas, Guillermo Puebla, Andrea E Martin, and John E Hummel. A theory of relation learning and cross-domain generalization. *Psychological Review*, in press.

[10] Willis D Ellis. *A source book of Gestalt psychology*. Routledge, 2013.

[11] Reuben Feinman and Brenden M Lake. Learning inductive biases with simple neural networks. *arXiv preprint arXiv:1802.02745*, 2018.

[12] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016.

[13] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018.

[14] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[15] Alex Graves, Greg Wayne, and Ivo Danihelka. Neural turing machines. *arXiv preprint arXiv:1410.5401*, 2014.

[16] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 33: 21271–21284, 2020.

[17] Katherine Hermann, Ting Chen, and Simon Kornblith. The origins and prevalence of texture bias in convolutional neural networks. *Advances in Neural Information Processing Systems*, 33, 2020.

[18] Julian Hochberg and Virginia Brooks. Pictorial recognition as an unlearned ability: A study of one child's performance. *The American Journal of Psychology*, 75(4): 624–628, 1962.

[19] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[20] John E Hummel. Reference frames and relations in computational models of object recognition. *Current Directions in Psychological Science*, 3(4):111–116, 1994.

[21] John E Hummel. Getting symbols out of a neural architecture. *Connection Science*, 23(2):109–118, 2011.

[22] John E Hummel. Object recognition. *Oxford handbook of cognitive psychology*, pages 32–46, 2013.

[23] John E Hummel and Irving Biederman. Dynamic binding in a neural network for shape recognition. *Psychological review*, 99(3):480, 1992.

[24] John E Hummel and Keith J Holyoak. Distributed representations of structure: A theory of analogical access and mapping. *Psychological review*, 104(3):427, 1997.

[25] John E Hummel and Keith J Holyoak. A symbolic-connectionist theory of relational inference and generalization. *Psychological review*, 110(2):220, 2003.

[26] John E Hummel and Brian J Stankiewicz. Categorical relations in shape perception. *Spatial vision*, 10(3):201–236, 1996.

[27] Georgin Jacob, RT Pramod, Harish Katti, and SP Arun. Qualitative similarities and differences in visual object representations between brains and deep networks. *Nature communications*, 12(1):1–14, 2021.

[28] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[29] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[30] David C Knill. Perception of surface contours and surface shape: from computation to psychophysics. *JOSA A*, 9(9):1449–1464, 1992.

[31] Nikolaus Kriegeskorte. Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annual review of vision science*, 1:417–446, 2015.

[32] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

[33] Jonas Kubilius, Stefania Bracci, and Hans P Op de Beeck. Deep neural networks as a computational model for human shape sensitivity. *PLoS computational biology*, 12(4):e1004896, 2016.

[34] Barbara Landau, Linda B Smith, and Susan S Jones. The importance of shape in early lexical learning. *Cognitive development*, 3(3):299–321, 1988.

[35] Gaurav Malhotra, Marin Dujmovic, and Jeffrey S Bowers. Feature blindness: a challenge for understanding and modelling visual object recognition. *bioRxiv*, 2021.

[36] Pascal Mamassian and Michael S Landy. Observer biases in the 3d interpretation of line drawings. *Vision research*, 38(18):2817–2832, 1998.

[37] Johannes Mehrer, Courtney J Spoerer, Emer C Jones, Nikolaus Kriegeskorte, and Tim C Kietzmann. An ecologically motivated image dataset for deep learning yields better models of human vision. *Proceedings of the National Academy of Sciences*, 118(8), 2021.

[38] Ken Nakayama, Zijiang J He, and Shinsuke Shimojo. Visual surface representation: A critical link between lower-level and higher-level vision. *Visual cognition: An invitation to cognitive science*, 2:1–70, 1995.

[39] Stephen Palmer. Canonical perspective and the perception of objects. *Attention and performance*, pages 135–151, 1981.

[40] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[41] Zygmunt Pizlo. Perception viewed as an inverse problem. *Vision research*, 41(24): 3145–3161, 2001.

[42] Zygmunt Pizlo and Adam K Stevenson. Shape constancy from novel views. *Perception & Psychophysics*, 61(7):1299–1307, 1999.

[43] Zygmunt Pizlo, Tadamasa Sawada, Yunfeng Li, Walter G Kropatsch, and Robert M Steinman. New approach to the perception of 3d shape based on veridicality, complexity, symmetry and volume. *Vision research*, 50(1):1–11, 2010.

[44] Guillermo Puebla and Jeffrey Bowers. Can deep convolutional neural networks support relational reasoning in the same-different task? *bioRxiv*, 2021.

[45] Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. Dynamic routing between capsules. *arXiv preprint arXiv:1710.09829*, 2017.

[46] Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J Majaj, Rishi Rajalingham, Elias B Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, et al. Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, page 407007, 2020.

[47] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[48] Kent A Stevens. The visual interpretation of surface contours. *Artificial Intelligence*, 17(1-3):47–73, 1981.

[49] Michael J Tarr and Steven Pinker. Mental rotation and orientation-dependence in shape recognition. *Cognitive psychology*, 21(2):233–282, 1989.

[50] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

[51] Daniel LK Yamins and James J DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, 19(3):356–365, 2016.

[52] Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, 111(23):8619–8624, 2014.

[53] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *arXiv preprint arXiv:1411.1792*, 2014.

Acknowledgments

This research was supported by the European Research Council Grant Generalization in Mind and Machine, ID number 741134.

Appendix

1.1 Classification performance

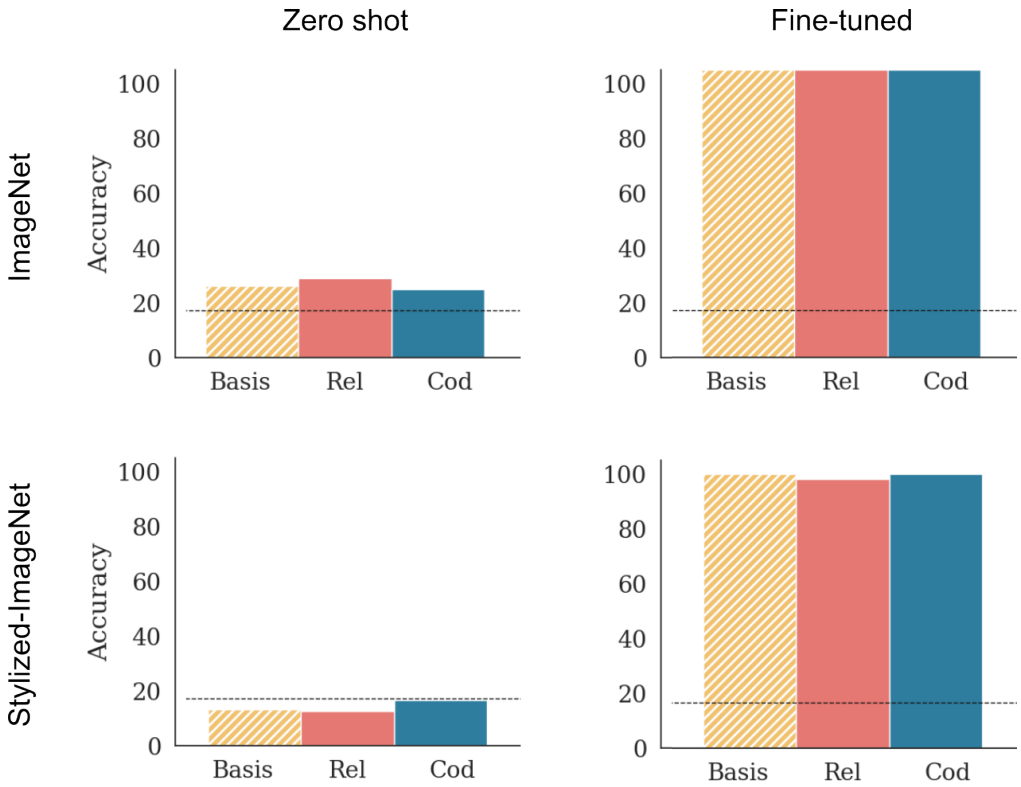


Fig S1 Accuracy on the Basis shapes as well as the two types of deformations (Rel and Cood) for a `VGG-16` network that has either been pre-trained on `ImageNet` (first row) or on `Stylized-ImageNet` (second row). Panels in the first column show classification accuracy on the pre-trained network without any further training, while panels in the second column shown test performance for a model that was fine-tuned on the set of Basis shapes. Dashed black line shows chance performance. In keeping with the cosine distance in internal representations, 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.

1.2 Examining errors in Experiment 2

In Experiment 2, 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 increase in rotation of test images, except for Category 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 S2 shows confusion matrices for two models (`VGG-16` pre-trained on `ImageNet` and `Stylized-ImageNet` respectively). Each heat

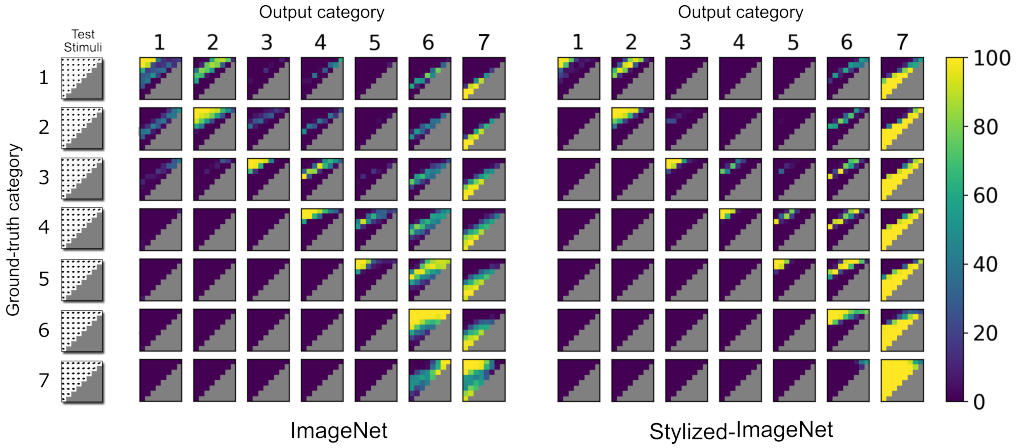


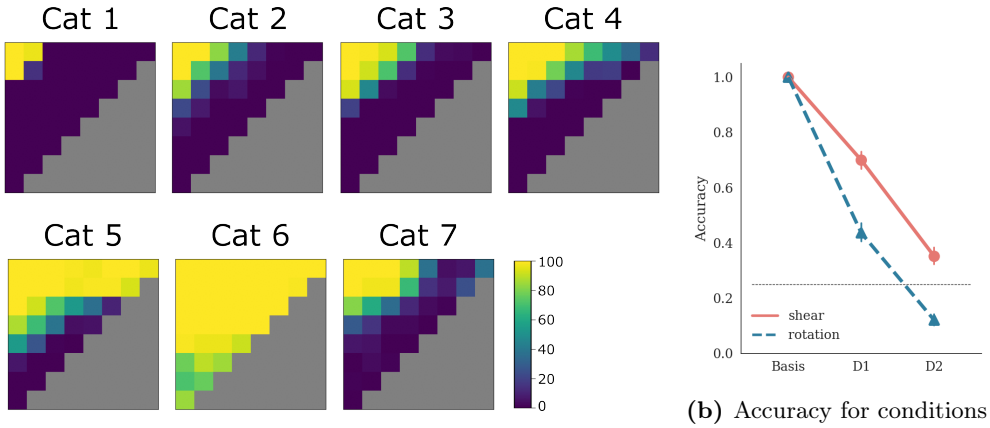
Fig S2 Confusion matrices for all deformations. For any heat map, the category label along each row shows the ground truth – i.e., all test shapes used to obtain the heat map were obtained by distorting the basis shape from that category. The category label along the column shows output category label assigned by the network. Therefore, in each row, the diagonal heat map shows the correct classifications, while the off-diagnoal heat maps show how each deformation was misclassified.

map shows the number of times an output category was chosen for all deformations of a given input category. This confusion matrix shows that the both networks were prone to mis-classify large rotations from any category as belonging to Category 7 (note large number of classifications in final column of each matrix for large rotations). These false positives (Type I errors) create a bias in the accuracy results for Category 7 in Figure 6 – that is, the high accuracy for large rotations for Category 7 category are, in fact, misleading as the networks classify large rotations for any category as Category 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 the lower triangle. In other words, the network was more likely to mis-classify images from each category as the category above rather than the category below.

1.3 Teaching relational representations for single-part objects

One response to the difference between CNNs and humans in Experiment 2 is that it arises due to the difference in experience of 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 objects in various poses). It could therefore be argued that the CNN does not learn relational representations in the polygons task because the training set did not provide an incentive for learning such a representation. Indeed, the optimisation view argues that a bias must be present in the training environment for the visual system to internalise 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 network on rotational deformations of Basis shapes, in addition to translation and scale deformations. To test whether the network generalises based on relational representations, we tested the network on the grid of deformations shown in Figure 5(b) and excluded the rotations used in that grid from the training set. Specifically, the Basis shapes were presented at random rotations in the training set in the range $[-45°, 0]$ and tested on rotations in the range $[0, +45°]$. The network performance on



(a) Accuracy landscape

(b) Accuracy for conditions in experiment

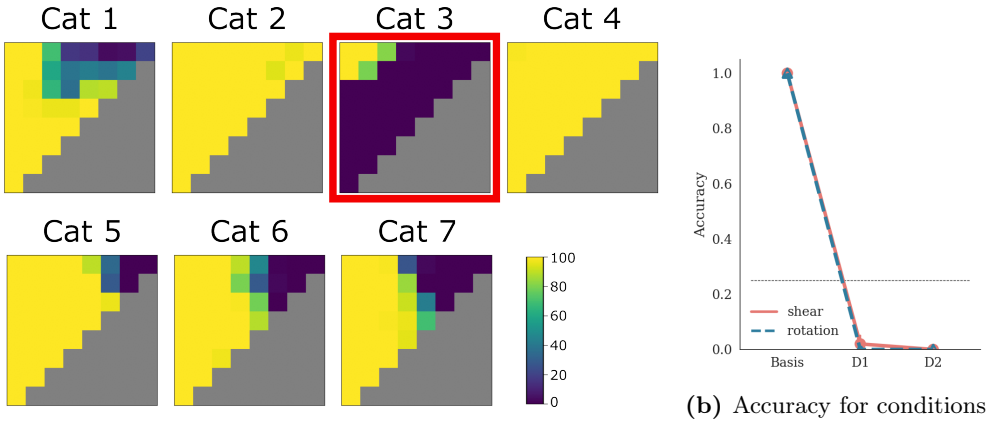
Fig S3 Performance of `VGG-16`, pre-trained on `Stylized-ImageNet` and then fine-tuned on an augmented dataset where the Basis shapes are not only translated and scaled, but also rotated randomly in the range $[-45°, 0°]$. The network is then tested in the same way as above, where the Basis shapes are deformed by shear or rotation (in the range $[0°, +45°]$) transform. (a) Accuracy (plotted) as percent correct on the landscape of deformations, and (b) accuracy for deformations used in experiment with human participants (compare with Figure 8, right-hand panel).

this test grid for each category is shown in Figure S3.

We observed that, despite being trained on this augmented dataset, results remained qualitatively similar. 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 generalising to large relational deformations (shears) than large relation-preserving deformations (rotation). The pattern was different for the final category, where the network showed good performance on large rotations. But examining the confusion matrix again revealed that the high accuracy at large rotations for these two categories was misleading as it was accompanied with large Type I errors: large rotations for shapes of any category were mis-classified as belonging to the final category. Overall, we did not find any evidence for the network learning shapes based on their internal relations.

In the second set of simulations, 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 generalised to the two types of deformations for this critical category.

The results of this simulation are shown in Figure S4. Figure S4(a) 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 – it's performance decreases across (perpendicular to) the diagonals, but increases as one moves from left-to-right along these diagonals. Figure S4(b) shows the performance on the same conditions as the human experiment (see Figure 8). Again, we see that the performance drops less rapidly across the two shear deformations (dashed line) than the two rotation deformations (solid line). This figure makes it clear that training other orientations on all rotations does not help the network generalise 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



(a) Accuracy landscape for left-out Category

(b) Accuracy for conditions in experiment

Fig S4 Performance of `VGG-16` pre-trained on `Stylized-ImageNet` and fine-tuned on an augmented dataset, where the Basis shapes are not only translated and scaled for all categories but also rotated at all angle ([0°, 360°]) for six out of seven categories. We test whether this augmented training allow the network to generalise better on the relation-preserving deformation (rotation) than the relation-changing deformation (shear) for the left-out category (Cat 3, highlighted). (b) Accuracy of network for the set of deformations D1 and D2 tested in the human experiment (compare with Figure 8, right-hand panel).

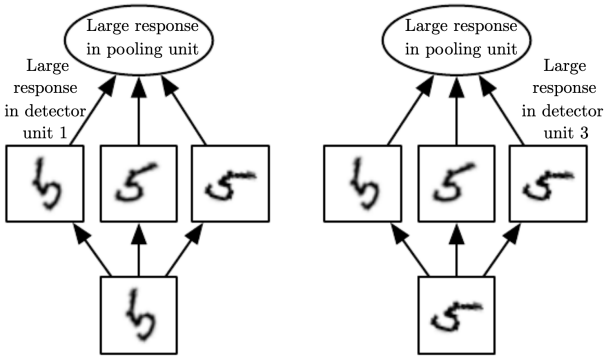


Fig S5 An example of how the operations of convolution and pooling can help a CNN achieve rotation invariance. Taken from Goodfellow et al. [14, chap. 9]

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 S5, taken from Goodfellow et al. [14, chap. 9], illustrates how a network consisting of convolution and pooling layers may learn to recognise 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 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 recognises the digit 5, irrespective of it's 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. All that is needed is to be able to recognise the internal parts of an object and check whether they are in the same relation as the learned shape. Accordingly, many psychological studies have shown that invariance, such as rotation invariance, precedes recognition [3, 4, 5, 22].

1.4 Results for AlexNet

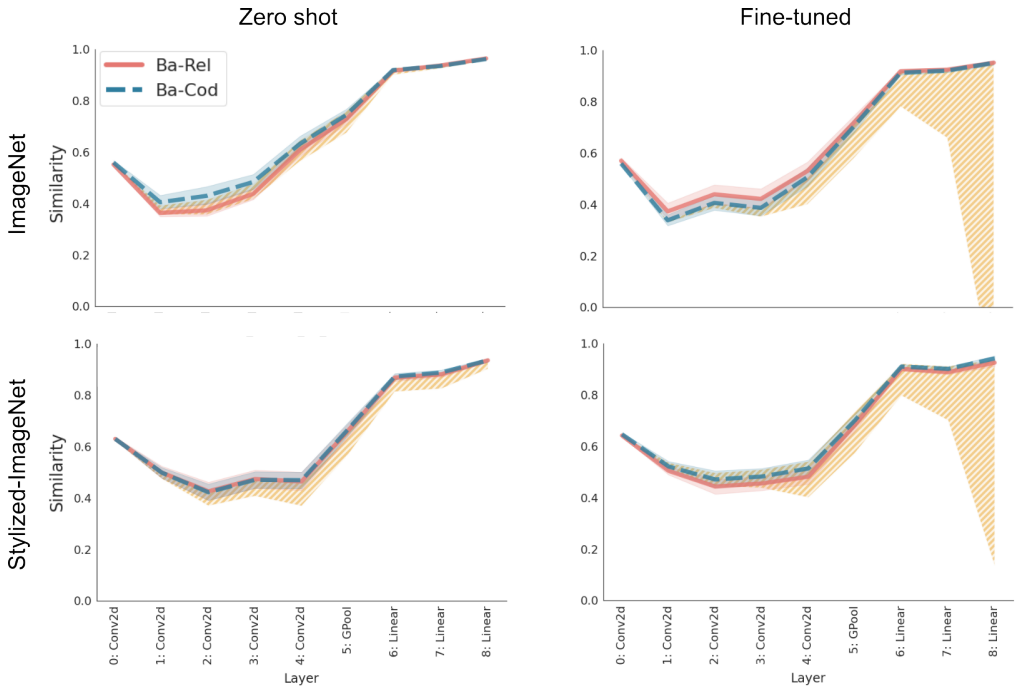


Fig S6 Cosine similarity for `AlexNet` between the internal representations of Basis images and it's Rel (relational) and Cood (coordinate) deformations. Like the results for `VGG-16` (compare with Figure 2 in the main text), the similarity between Basis images and both types of deformations is at the upper bound throughout the network, showing that the network does not distinguish the trained (Basis) image from it's Rel and Cood deformations.

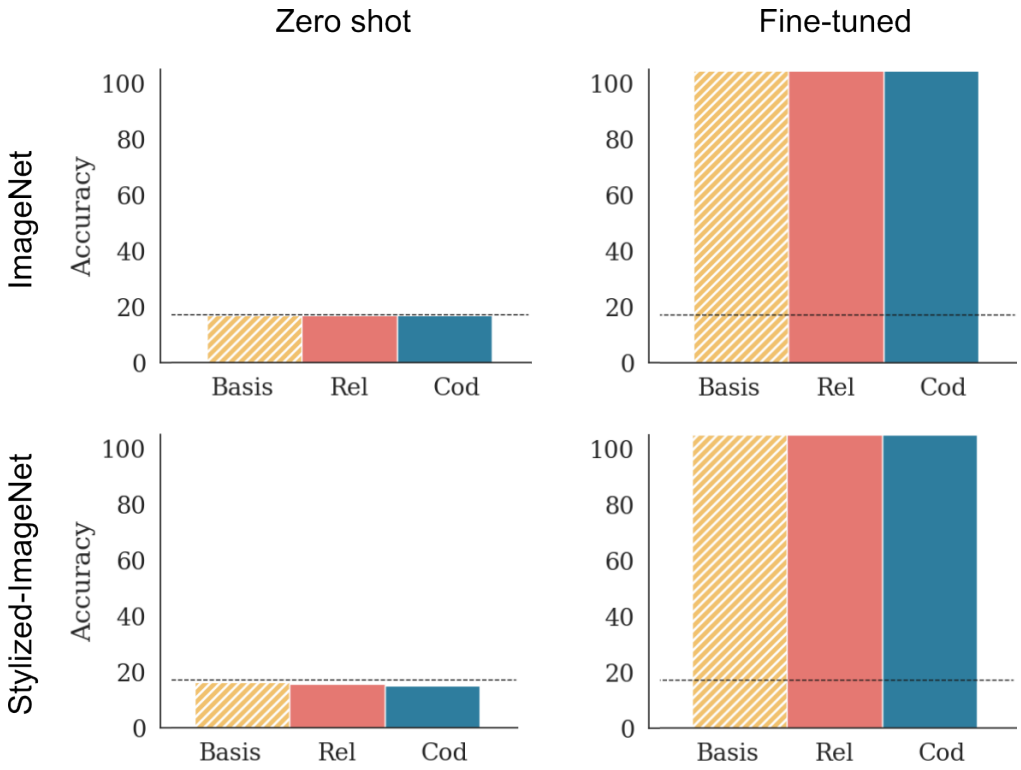


Fig S7 Performance of `AlexNet` in the test set for Experiment 1. Each panel shows accuracy on the Basis shapes as well as the two types of deformations: relational (`Rel`) which changes a categorical relation and coordinate (`Cood`), which preserves all categorical relations. Compare with performance of `VGG-16` in Figure S1.

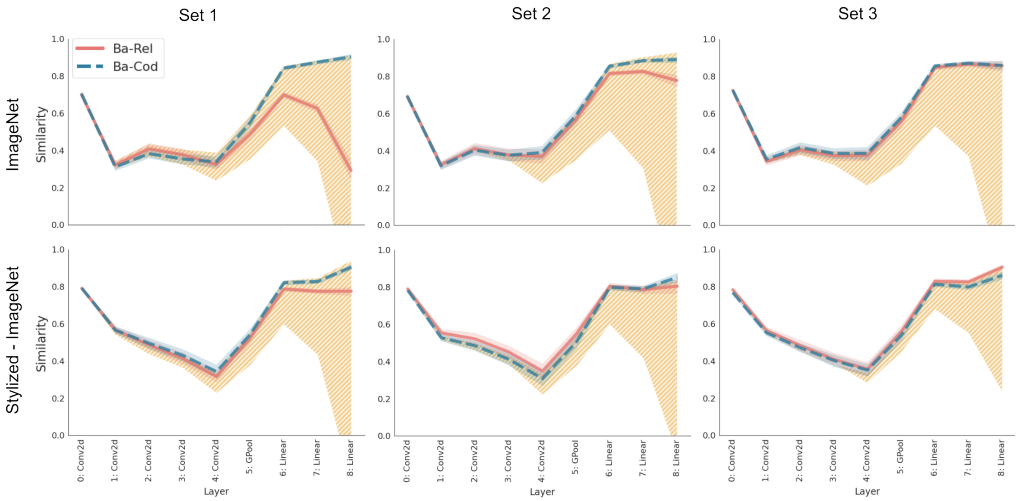


Fig S8 Cosine similarity between Basis image and two types of deformations for `AlexNet` fine-tuned on (a) Set 1, (b) Set 2, and (c) Set 3 in Figure 3. Like the results for `VGG-16` (compare with Figure 4), we see that networks learns to distinguish the Rel deformation from the Basis image for Set 1, when it has seen the specific deformation in the training set. But this sensitivity to Rel deformation diminishes in Set 2, when only one pair of trained shapes have a similar deformation and completely lost for Set 3 when the network has been trained on the Rel deformations, but the specific deformation tested is novel.

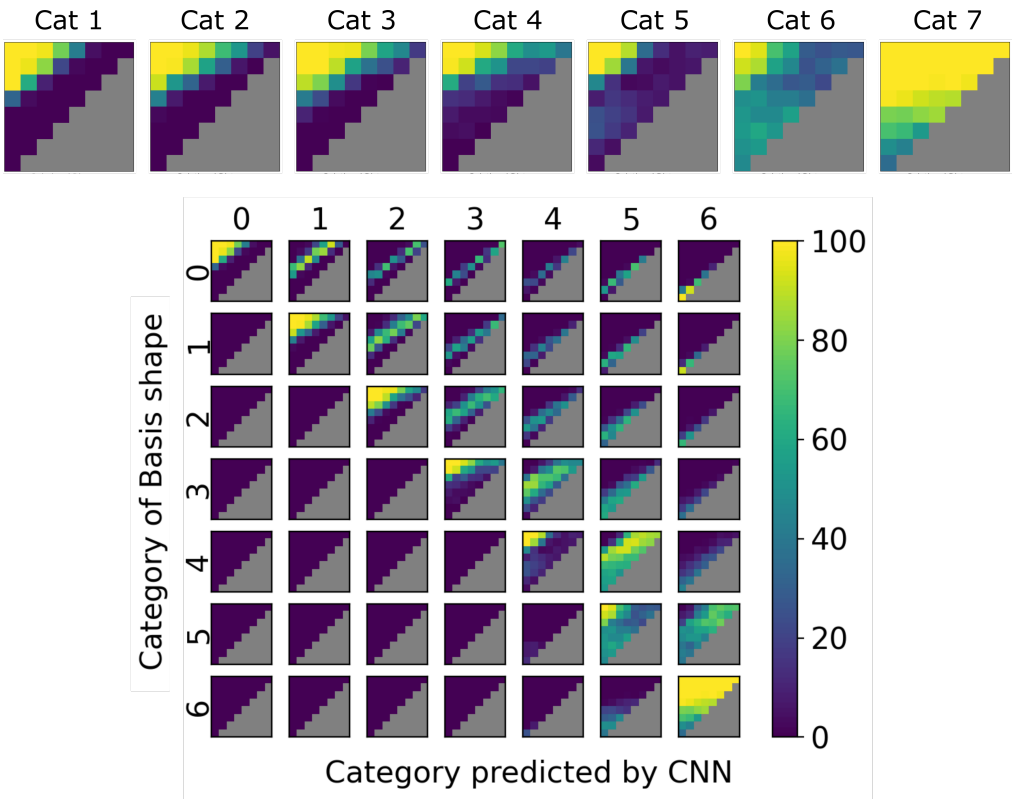


Fig S9 Each heatmap shows accuracy for on test items for a particular category for `AlexNet` pre-trained on `ImageNet` and fine-tuned on the dataset in Figure 5. Each cell in the heatmap corresponds to a deformation that is a combination of relational (shear) and coordinate (rotation) transformations of the trained Basis shapes (see Figure 5(a)). 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.

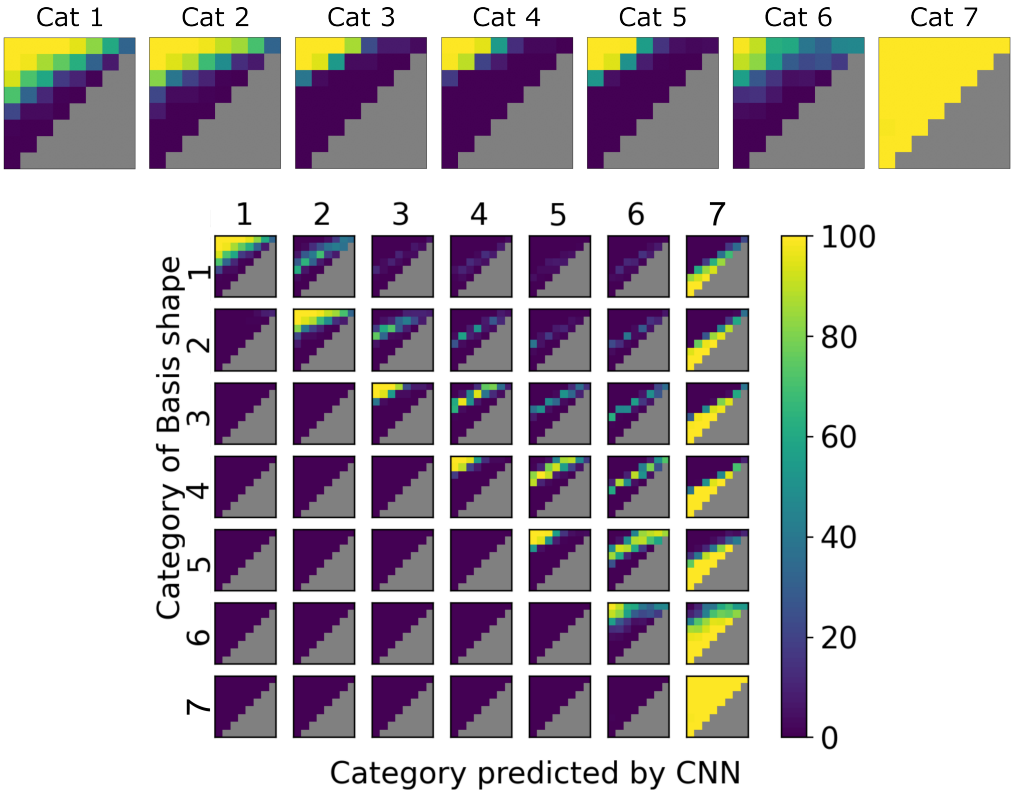


Fig S10 Each heatmap in the top row shows accuracy for on test items for a particular category for `AlexNet` pre-trained on `Stylized-ImageNet` and fine-tuned on the dataset in Figure 5. The bottom panel shows the confusion matrix. See Figure S9 for explanation.

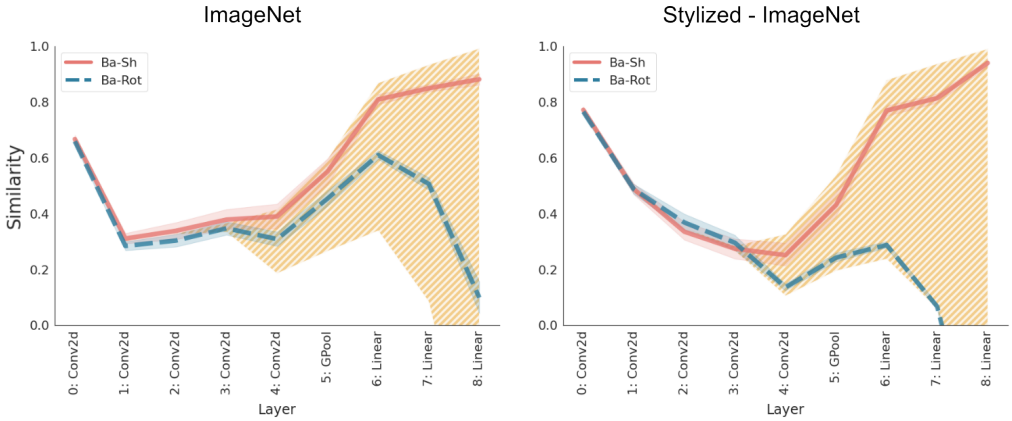


Fig S11 Cosine similarity between internal representations for the Basis shapes and two deformations of the basis shape (dashed red squares in Figure 5(b)) from the polygons dataset at each convolution and fully connected layer of `AlexNet`. Solid (red) line shows the average similarity between representations for a basis shape and its relational (shear) deformation, while dashed (blue) line shows the average similarity between a basis shape and it's 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 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.

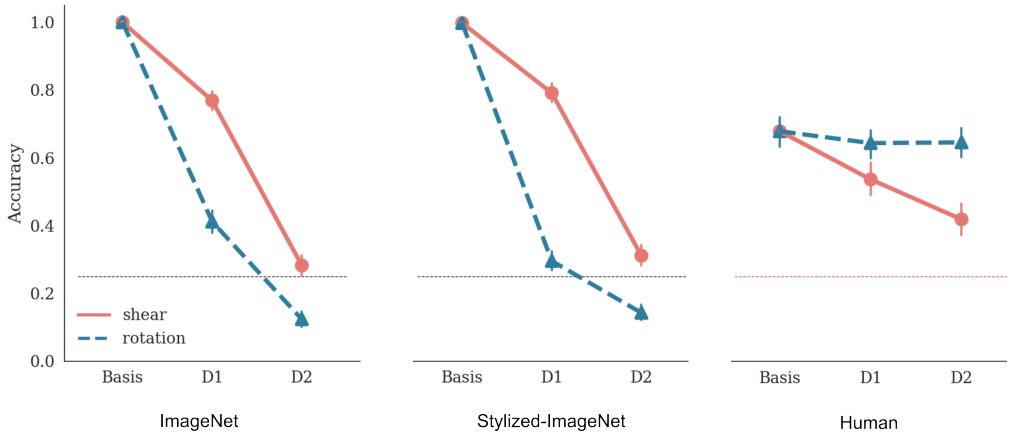
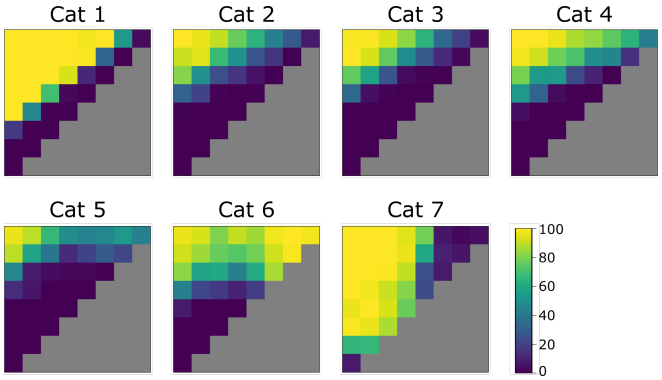
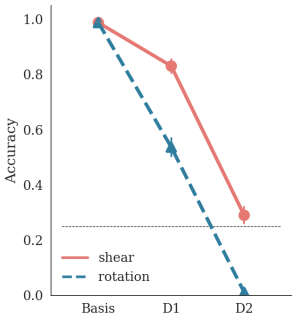


Fig S12 Comparison of change in classification accuracy with deformations for `AlexNet` (left and middle) and human participants (right). Each panel shows performance under three conditions: basis image, deformation D1 and deformation D2. For the shear deformation (solid, red line), D1 and D2 consists 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 95% confidence interval and dashed red line shows chance performance. Note that the results in right-hand panel are reproduced here for convenience but are the results of the same experiment reported in Figure 8, right-hand panel.

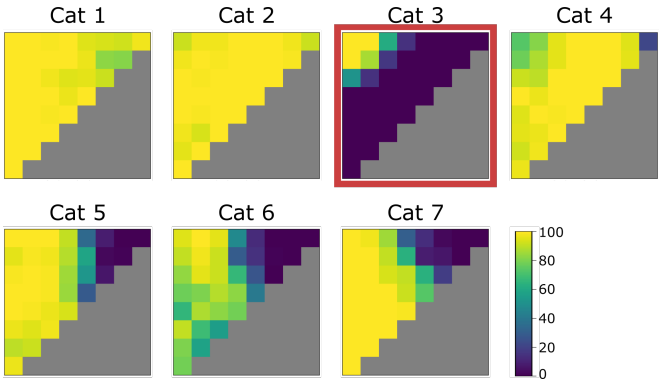


(a) Accuracy landscape

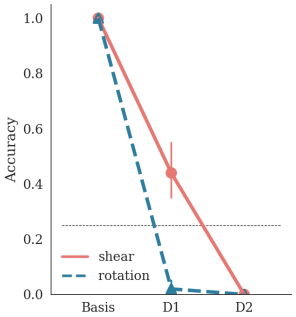


(b) Accuracy for conditions in experiment

Fig S13 Performance of `AlexNet` fine-tuned on an augmented dataset 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 on shear and rotation deformations in the range $[0°, +45°]$. Like the results for `VGG-16` (compare with Figure S3), we observed that even when the network was trained on some rotations, it's performance on untrained rotations (a coordinate transformation) was still worse than shears (a relational transformation). (b) shows accuracy for deformation level D1 and D2 used for testing human participants.



(a) Accuracy landscape for left-out Category



(b) Accuracy for conditions in experiment

Fig S14 Performance of `AlexNet` fine-tuned on an augmented dataset 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 red square in (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.