Vision Research 172 (2020) 46–61

Contents lists available at ScienceDirect

Vision Research

journal homepage: www.elsevier.com/locate/visres

Local features and global shape information in object classification by deep convolutional neural networks

Nicholas Baker[a,*], Hongjing Lu[a,b], Gennady Erlikhman[a], Philip J. Kellman[a]

[a] Department of Psychology, University of California, Los Angeles, Los Angeles, CA, United States
[b] Department of Statistics, University of California, Los Angeles, Los Angeles, CA, United States

ARTICLE INFO

Keywords:
Shape
Global and local features
Object recognition
Deep learning

ABSTRACT

Deep convolutional neural networks (DCNNs) show impressive similarities to the human visual system. Recent research, however, suggests that DCNNs have limitations in recognizing objects by their shape. We tested the hypothesis that DCNNs are sensitive to an object's local contour features but have no access to global shape information that predominates human object recognition. We employed transfer learning to assess local and global shape processing in trained networks. In Experiment 1, we used restricted and unrestricted transfer learning to retrain AlexNet, VGG-19, and ResNet-50 to classify circles and squares. We then probed these networks with stimuli with conflicting global shape and local contour information. We presented networks with overall square shapes comprised of curved elements and circles comprised of corner elements. Networks classified the test stimuli by local contour features rather than global shapes. In Experiment 2, we changed the training data to include circles and squares comprised of different elements so that the local contour features of the object were uninformative. This considerably increased the network's tendency to produce global shape responses, but deeper analyses in Experiment 3 revealed the network still showed no sensitivity to the spatial configuration of local elements. These findings demonstrate that DCNNs' performance is an inversion of human performance with respect to global and local shape processing. Whereas abstract relations of elements predominate in human perception of shape, DCNNs appear to extract only local contour fragments, with no representation of how they spatially relate to each other to form global shapes.

1. Introduction

Much of thought and behavior depends on descriptions of the world in terms of objects (Vallortigara, 2012; Spelke, 1990; Kellman & Arterberry, 2000; Kahneman, Treisman, & Gibbs, 1992). Understanding how such descriptions are obtained from sensory information is a central topic in cognitive science and neuroscience. Problems of object perception and recognition have been studied extensively in biological vision, and in recent years, object recognition has been a central focus in computer vision and artificial intelligence.

In humans, vision is primary in delivering information about objects. Research in object perception has implicated a number of computational processes as subserving human abilities to perceive and recognize objects. The visual system separates figure from ground (Rubin, 1915), distinguishes bounding contours of an object from other contours (Koffka, 1935), and encodes border ownership in contour perception (e.g., Driver & Baylis, 1996; Zhou, Friedman, & Von Der Heydt, 2000). Visible parts of objects are connected behind nearer objects that partially occlude them (Kellman & Shipley, 1991; Kellman & Spelke, 1983; Michotte, Thines, & Crabbe, 1964), and the visual system confers shape descriptions upon bounded objects, rather than, for example, the spaces between them (Koffka, 1935). Object recognition, which entails finding a match for current input in memory of a specific object or object category, is, in humans, primarily driven by these descriptions of object shape (Biederman, 1987; Elder & Velisavljević, 2009; Lloyd-Jones & Luckhurst, 2002; Marr, 1982).

In computer vision, some efforts to build artificial systems that can recognize objects have attempted to implement solutions to these computational tasks explicitly, involving information about shape (Belongie, Malik, & Puzicha, 2002; Bergevin & Levine, 1993; Rezanejad & Siddiqi, 2013), local texture patterns (Lowe, 1999), or surface feature segmentation (Shi & Malik, 2000; Shotton, Winn, Rother, & Criminisi, 2009). Algorithms of this sort have advanced in accomplishing these tasks; however, systems for object recognition based on these approaches have not attained performance levels achieved more recently by deep convolutional neural networks (DCNNs), a machine learning

* Corresponding author.
E-mail address: nbaker9@ucla.edu (N. Baker).

https://doi.org/10.1016/j.visres.2020.04.003
Received 31 October 2019; Received in revised form 13 March 2020; Accepted 6 April 2020

approach that does not recognize objects based on explicitly coded features (Krizhevsky, Sutskever, & Hinton, 2012).

Despite being trained in a purely associative fashion, remarkable similarities have been found between the human visual system and deep networks trained for object recognition. Node activity in intermediate layers of deep networks correlates with activity of cell populations in V4 (Pospisil, Pasupathy, & Bair, 2018), and some deep networks have been found to be predictive of cell populations in IT (Yamins et al., 2014). Deep networks trained for object recognition also appear to predict human behavior in judging the similarity between objects (Peterson, Abbott, & Griffiths, 2016), the memorability of objects (Dubey, Peterson, Khosla, Yang, & Ghanem, 2015), and the saliency of regions in an image (Kümmerer, Theis, & Bethge, 2014). These similarities have raised interest in using DCNNs as tools for modeling human perceptual capabilities.

Typically, DCNNs are trained for object recognition on the ImageNet database (Deng et al., 2009). A consequence of using a training set made up of millions of natural images, and of the many layers included in the architecture of DCNNs, is that it can be difficult to determine what information a DCNN is using to classify images. A photograph has many possible sources of information by which an object can be recognized: shape, luminance, surface texture, even background information. It is difficult to say how much and in what ways DCNNs use these properties in object recognition, and it is even imaginable that DCNNs access information that goes beyond or that cuts across any of these features which are used by the human visual system.

Baker, Lu, Erlikhman, and Kellman (2018) found a bias for texture over shape in DCNN classification by testing network performance for stimuli with diagnostic shape information but absent or misleading texture cues such as line drawings or glass figurines. In all cases, the absence or alteration of texture cues impaired classification performance much more than deprivation of shape cues. Geirhos et al. (2018) also compared network classification accuracy for images deprived of texture information with their accuracy for images deprived of shape information. All DCNNs did far better in the absence of shape features than in the absence of texture features, suggesting that texture plays a much larger role in classification for DCNNs. Surprisingly, Hermann and Kornblith (2019) found that networks actually begin classifying by shape more quickly than by texture, but certain data set augmentation techniques (e.g., random crop) and fine-tuned hyperparameters in training (e.g., low learning rates) appear to bias networks towards texture-based classifications.

In human perception recognition is driven by object shape more than any other cue (Erlikhman, Caplovitz, Gurariy, Medina, & Snow, 2018; Palmer, 1999). Objects simplified to line drawings can still be recognized by human perceivers and are in fact more rapidly categorized than natural images (Biederman & Ju, 1988). Developmental studies have found that 12-month-old infants represent items with different shapes as two different objects, but not objects with different sizes or colors (Xu, Carey, & Quint, 2004). The importance of shape goes beyond visual recognition to show its influence on early lexical learning: young children show a shape bias, in that they generalize word meaning to other objects on the basis of shape similarity more so than other object properties (Imai, Gentner, & Uchida, 1994; Landau, Smith, & Jones, 1988).

If DCNNs are a good model of the perceptual processes that the human visual system uses in recognition, we would expect to observe a similar preeminent role for shape in their classification performance. Kubilius, Bracci, and de Beeck (2016) tested deep network use of shape as a cue for recognition and found that DCNNs can classify image silhouettes with about 40% accuracy, well above chance performance. Moreover, Kubilius et al. (2016) found that networks are sensitive to non-accidental features of objects such as whether two edges were parallel or converging (Biederman, 1987), aspects that can be critical in distinguishing objects' identities.

Studying DCNN shape sensitivity led us to a new hypothesis regarding what shape related information is and is not encoded in DCNNs (Baker, Erlikhman, Kellman, & Lu, 2018; Baker, Lu, et al., 2018). On one hand, we suggested that DCNNs do have sensitivity to local contour features in an image. Strikingly, however, DCNNs lack the ability to represent an object's global shape and use global shapes to classify objects. We tested this *local contour feature hypothesis* by finding silhouettes of objects that the network correctly classified and altering silhouette images in the following ways. First, we scrambled the spatial relations between object parts to destroy their global shape features while preserving many of the local edge properties present in the original stimulus. Second, we preserved global shape but altered local edge features by adding serrations to bounding contours of objects. Network performance was affected very little by part-scrambled objects but was completely disrupted by the addition of a serrated edge along the contour of the object boundary, results that were the opposite from human performance in these two conditions.

Another study by Brendel and Bethge (2019) found that deep networks trained to classify based on bags of small, local features have very similar classification performance to top-performing DCNNs trained on ImageNet (Brendel & Bethge, 2019). The researchers argued that classification performance in DCNNs is not explained by sensitivity to higher-order information like the global shape of objects, but by improved learning with local features that were used in more classic machine learning algorithms like bag-of-features. Doerig, Bornet, Choung, and Herzog (2020) showed that feedforward DCNNs can also not explain configural effects of crowding, such as the uncrowding effect when a vernier is surrounded by a square instead of two unconnected vertical lines. Eckstein, Koehler, Welbourne, and Akbas (2017) showed that, in visual search, humans often miss targets that have inappropriate spatial scale relative to the scene, even when the targets are made larger and more salient (e.g., a huge toothbrush in a bathroom scene image). In contrast, DCNNs do not exhibit such performance reduction in finding mis-scaled targets, suggesting the lack of sensitivity to the global size of objects in the scene.

These findings motivated us to think more deeply about what is meant by shape. Considering that DCNNs are based on convolution operations in multiple layers, it is straightforward to understand how they may be sensitive to local oriented contrast in images, and also straightforward to understand how they might develop sensitivity to local orientation relations that characterize local features of objects (i.e., by learning concurrent activations of nearby oriented contrast detectors in early layers). These features are local constituents of shape. More global notions of shape, however, are less concerned with individual element features than with the way these features fit together into a unified whole (Koffka, 1935). These global relations drive recognition in human perception (c.f., Baker & Kellman, 2018).

Deep networks and humans may accomplish recognition in very different ways. While the human visual system encodes the global shape of object and has a high degree of flexibility about local features, deep networks appear not to encode global shape at all but are sensitive to surface information and local contour features.

In the present work, we used more direct tests to evaluate the contributions of local and global shape in object recognition for DCNNs. In Experiment 1, we tested DCNN performance on stimuli in which local and global information conflict, hypothesizing that the network would classify by local information. We also tested to see if it is possible for networks to develop sensitivity to an object's global shape given the right training data. In Experiment 2, we changed the training data to make local contour information nondiagnostic for the circle/square recognition task to test whether the network can be trained towards more global shape processing. After finding that the new training set did in fact produce more global classifications, we tested the network further in Experiment 3 to examine if it was truly classifying objects based on the configuration of its local elements rather than some statistical properties of local elements themselves.

In all the experiments, we used a *transfer learning* approach to fine-tune and probe deep networks pre-trained on a natural image dataset to a new visual task. This technique has proven to be useful for other tasks such as the classification of medical images (Esteva et al., 2017; Hoo-Chang et al., 2016), as well as in perception research to adapt trained networks for testing on psychophysical stimuli (Baker, Erlikhman, et al., 2018; Baker, Lu, et al., 2018). Transfer learning allows us to assess the kind of information captured by trained deep networks by applying the learned features to new discriminations. We used two kinds of transfer learning: restricted, where learning is limited to the last set of connection weights between the penultimate layer and the decision layer, and unrestricted, where all connection weights could be updated. Restricted learning reveals what features the network learns in typical training on ImageNet, while unrestricted learning reveals what features the network architecture could potentially learn.

2. Experiment 1

In Experiment 1, we used both restricted and unrestricted transfer learning to study the issue of local vs. global shape processing in DCNNs (specifically, AlexNet), testing whether networks classify based on the local elements composing a shape contour or more global information about how the contour is configured. In restricted learning, because fine-tuning was restricted to the last set of connection weights for decision, the retraining could only affect the weighted combination of the 4096 nodes in AlexNet's penultimate layer that drive its classification decision, not the features that are extracted earlier in network processing. Restricted learning, then, was a test on whether the capacity to use global shape information has emerged in networks trained from millions of natural images.

Unrestricted transfer learning allowed the network to learn different features in hidden layers of the network to improve performance on its new classification task. The effort is still transfer learning, as we begin with a network already trained to classify objects in natural scenes. Training was conducted in the same way as in the restricted learning condition, but all connection weights could be updated.

2.1. Experiment 1A. Training a circle vs. square classifier

We first tested AlexNet's ability to discriminate between circle and square outlines using restricted and unrestricted transfer learning. Past work suggested that deep convolutional networks' classification accuracy on outline images with object contours is very poor (Baker, Erlikhman, et al., 2018; Baker, Lu, et al., 2018), so it was an open question whether the network would be retrained to learn to detect shape differences between the two outline stimulus classes. In this phase, many training images were used, where each possessed consistent global shape and local features characteristic of a circle or a square. If this discrimination was learnable by the network, we aimed to test this newly trained version of the network on stimuli that would distinguish the roles of local curvature information vs. global shape (Experiment 1B).

2.1.1. Method

2.1.1.1. Training. We adopted AlexNet trained for recognition of natural images and then used transfer learning on this pretrained network to create a circle/square classifier, replacing AlexNet's 1000 category decision layer with a two-node layer that corresponded to circles and squares. We then trained the weights between the 4096 units in the penultimate layer of AlexNet and the final decision unit (in the restricted transfer learning condition), or all weights throughout the network (in the unrestricted transfer learning condition), by presenting it with 16,000 labeled images of circles and squares (see Fig. 1), 80% of which were used in training and 20% of which were used as the validation set. The shapes used in training spanned being little larger than a point to nearly the entire size of the image, i.e., the edge width ranged from 2 pixels to 213 pixels, and the circle radius ranged from 1 pixel to 98 pixels. Training terminated after 1030 iterations (runs through a mini-batch of size 32) in the restricted learning condition, and 1620 iterations in the unrestricted learning condition, after satisfying the criterion that error rate on the validation set increased on six consecutive iterations.

To confirm that our results were not unique consequences of AlexNet's architecture, we also conducted transfer learning on a pre-trained VGG-19 (Simonyan & Zisserman, 2014) architecture. Due to the larger hardware demands for training with VGG-19, we used half the training stimuli and mini-batch sizes of 8 instead of 32 and performed only restricted transfer learning. For these reasons, we focus our analysis on the results from AlexNet, but report the general findings from VGG-19 to show convergence across networks.

2.1.2. Results

Both restricted and unrestricted transfer learning on the pretrained DCNNs were successful. When training ended, AlexNet's error in discriminating circles and squares was 1.7% on the validation set for restricted learning and 1.0% for unrestricted learning. AlexNet was clearly able to acquire the square/circle discrimination from a wide array of examples in training, even from outlines of shapes. Transfer learning was also successful on VGG-19. The network had a 2.3% error rate on the validation set with restricted transfer learning.

2.2. Experiment 1B. Assessing the basis of classification in the circle/square classifier

The ability of a deep network to use its filters, trained for recognition of natural images, to learn the discrimination between circles and squares indicates the use of some aspects of shape information. This is true even for networks trained only on natural images, as shown by the results for the network trained with restricted transfer learning. The learning appeared to be general across circles and squares of varying sizes and positions. In this simulation, shapes were only defined by outlines, and no differing surface properties, context or background were available to support the network's discrimination.

What sort of shape information might be accessed by deep networks? There are at least two possibilities. Circles and squares differ in their global shapes; perhaps AlexNet accessed relations of contours that define (and distinguish) overall shape. A second possibility is that the network's successful discrimination performance relied on local contour features. Circles and squares differ consistently in local features. The boundaries of circles are curved in each local neighborhood, and there are no cusps or corners. For squares, edges in local neighborhoods are straight, but local contour intersections occur at each of four corners. Of course, these possibilities are not exclusive. The network could be sensitive both to global shape and local features that characterize circles and squares. Experiment 1B aimed to distinguish these possibilities.

2.2.1. Method

After the transfer learning described in Experiment 1A to enable the network to recognize squares or circles, we tested the networks on two kinds of images with different global shapes and local contour cues: squares whose boundaries were comprised of half-circle elements and circles whose boundaries were made up of half-squares. Fig. 2 shows the eight test images we constructed this way. We generated eight probe stimuli that differed both in the size of the overall global shape and in the size of the local elements composing the shape. We tested both networks' classification for each of these shapes to determine if local or global information predominated. We also measured the pattern similarity of the networks' activities at each of its eight layers for different input images to assess the contribution of local and global shape information across different stages of processing.

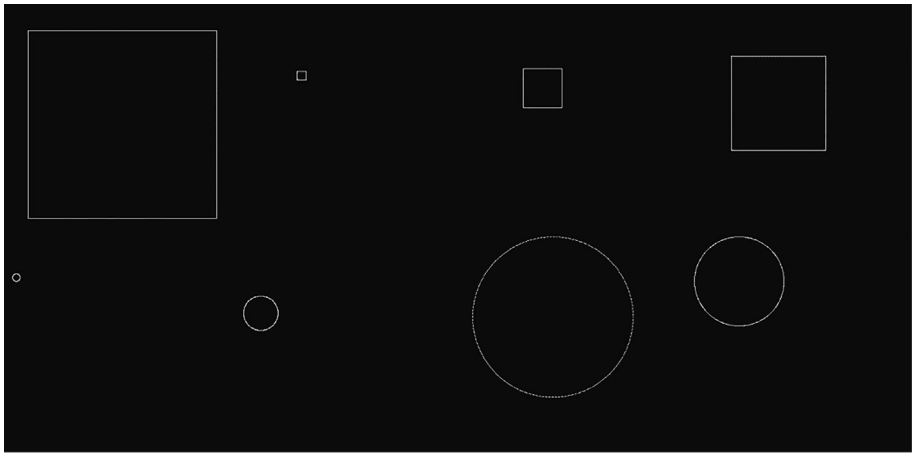


Fig. 1. Training Images in Experiment 1. The training set consisted of 16,000 white wire frame circles and squares of various sizes and spatial positions.

2.2.2. Results

The test images in Fig. 2 were used to test the circle/square classifiers. Fig. 3 shows the primary results for both training schemes. For both forms of training, the network's responses were generally inconsistent with the global shapes. For stimuli that had a square global shape comprised of local curved elements, the networks always classified the stimuli as circles. The network classified both the large circle composed of large corner elements and the large circle composed of small corner elements as squares with high confidence. Only when the corner elements were very small did the networks make the global classification. Overall, the network appeared to classify both square and circle displays according to the local elements that were constituents of its boundaries rather than the global shapes defined by these boundaries. The network made global classifications only when the local elements were extremely small, possibly because it could integrate them into a single curvature. Overall, classification performance was extremely similar between the networks trained with restricted and unrestricted transfer learning. All classifications were the same, with only small differences in confidence.

Performance for VGG-19 matched the results from AlexNet very closely. It classified the same two smaller circles globally and all other images by their local elements.

We also assessed network sensitivity to local and global shape cues by recording node activation at each layer of the DCNN for each input image. The correlation of activations across layers can be used as a measure of similarity between two input images across different stages of processing. Similar methods have been used to compare network activation at various layers with brain activity patterns in different areas of visual cortex (Cichy, Khosla, Pantazis, Torralba, & Oliva, 2016). Here, we made no direct comparison with brain activity. Rather, we examined the similarity between representations of two images that differ in local and/or global features. We input a square, circle, a square comprised of curved elements (termed a “curved square” in Fig. 4), and a circle comprised of corner elements (termed a “corner circle” in Fig. 4) to the network, and then computed the correlation between node activations for each pair of images in order to assess the representational similarity between different shapes at different layers along the network’s processing stream. All input stimuli were large, taking up nearly the whole frame for each image. For the two images where local and global cues conflicted, we selected images with large local elements since these were the stimuli for which the network showed the most local bias. We wanted to test whether that local bias was present across all layers of network processing.

As shown in Fig. 4, the pattern of correlations was very similar for networks trained with restricted and unrestricted transfer learning. There was close to zero correlation of activity in Layer 1 between any pairs of test images. However, activity patterns in layer 2 showed the highest correlation between the circle image and the corner circle image, the lowest correlation between the square image and the corner circle image, and middle-level correlations for the other pairs. Surprisingly, deeper in the network, the highest correlation was between the circle comprised of corner elements and the square comprised of circular elements. These images differ both in their local features and their global configurations, so their relatively high representational similarity on this measure is puzzling. The decision layer clearly showed a bias towards the local contour features in the final discrimination to show high correlations between square image and circle shape made of squares, and between circle image and square shape

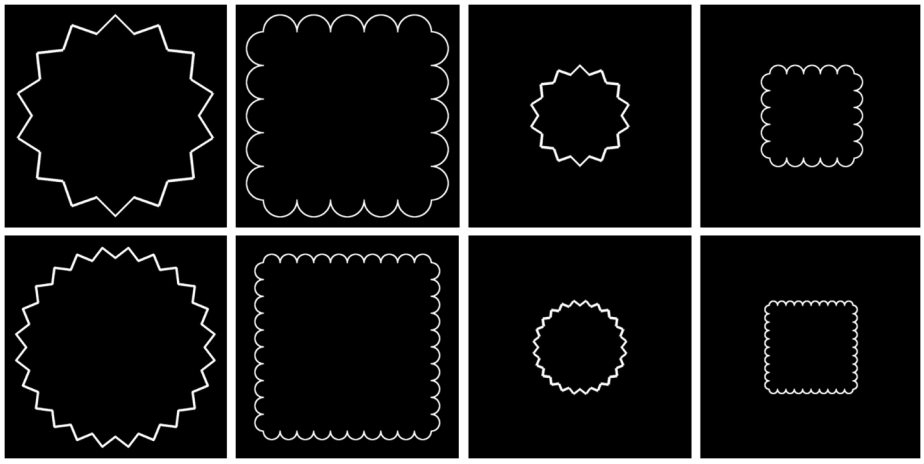


Fig. 2. Test images in Experiment 1B. A: Large circle comprised of large corner elements B: Large square comprised of large curved elements C: Small circle composed of large corner elements D: Small square comprised of large curved elements. E: Large circle comprised of small corner elements. F: Small circle comprised of small corner elements. G: Large square comprised of small curved elements H: Small square comprised of small curved element (See text).

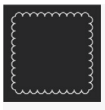
Input Image	Restricted Learning: Network Classification	Restricted Learning: Circle Probability	Restricted Learning: Square Probability	Unrestricted Learning: Network Classification	Unrestricted Learning: Circle Probability	Unrestricted Learning: Square Probability
	"Square"	0%	100%	"Square"	0%	100%
	"Circle"	100%	0%	"Circle"	99.8%	0.2%
	**"Circle"**	99.9%	0.1%	**"Circle"**	96.4%	3.6%
	"Circle"	100%	0%	"Circle"	99.9%	0.1%
	"Square"	0.6%	99.4%	"Square"	0%	100%
	"Circle"	100%	0%	"Circle"	100%	0%
	**"Circle"**	100%	0%	**"Circle"**	100%	0%
	"Circle"	100%	0%	"Circle"	99.9%	0.1%

Fig. 3. Network classification performance for the probe stimuli. Input stimulus for each condition are shown to the left. The classification responses in bold are the decisions consistent with global shapes.

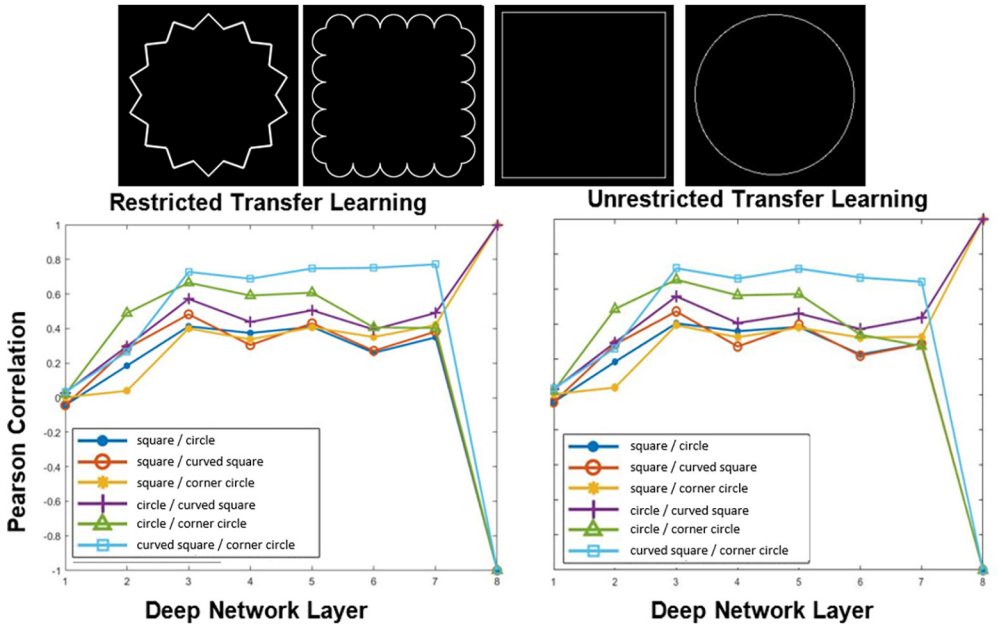


Fig. 4. Correlation of activations for different input stimulus across the layers of AlexNet. All images were large and centered in the frame. The corner circle is a global circle made from large corner elements (Fig. 2A). The curved square is a global square made up of large curved elements (Fig. 2B). Correlations were computed by transforming the activations into a vector (in which each element is a node of the network at that layer) and computing the Pearson correlation between stimulus pairs at each layer. Pairs in each correlation function are identified in the figure legend separated by a slash (e.g., "square / corner circle" indicates the correlation between the basic square pattern and a circle comprised of corner elements).

made of circles.

2.3. Experiment 1C: Are networks biased towards shape classification more sensitive to global shape?

Recent work by Geirhos et al. (2018) found that DCNNs' bias to classify by texture instead of shape can be reduced with more varied training. They employed a network that changes images to reflect different artists' styles (Gatys, Ecker, & Bethge, 2016) to augment the ImageNet database with created "stylized ImageNet" consisting of images in which texture information was less diagnostic than shape cues. When they trained a deep learning model with ResNet-50 architecture using both ImageNet and stylized ImageNet, they found an increase in classification by shape relative to texture.

It is an open question whether ResNet trained this way is more sensitive to global shape information or if the new training stimuli simply increased network sensitivity to local shape cues. Using the methods described in Experiment 1a, we used transfer learning to train the network to classify circles and squares, then tested it with the probe stimuli from Experiment 1b.

2.3.1. Method

We adopted a ResNet-50 architecture previously trained on stylized ImageNet and ImageNet, then finetuned on ImageNet (see Geirhos et al. (2018) for more details). We then performed transfer learning on the network using unrestricted transfer learning. The protocol was the same as in Experiment 1a, except the network trained for three epochs, terminating with 97.5% accuracy on the validation set.

We then tested the network on the probe stimuli used in Experiment 1b. Due to some peculiarities in network performance, we also tested it on five additional probe stimuli, shown in Fig. 5.

2.3.2. Results

Network classification for the probe stimuli is shown in Fig. 6. There appears to be a strong bias towards classifying images as circles: Aside from true squares, only the right triangle and the rectangle were given higher probability of being a square than a circle.

2.4. Discussion

In Experiment 1A, we found that AlexNet and VGG-19 were able to learn the circle/square classification to a high degree of accuracy and classify novel instances of conventional square and circle displays. The most important results, however, came from Experiment 1B, in which probe displays were presented to the network after training. Here, we used circles made of local elements that were fragments of squares, and we used squares made of local elements that were fragments of circles. The probe displays revealed that the accurate classification performance that AlexNet had achieved from the training displays rested entirely on local contour information and not global shape. These results were somewhat scale-dependent: if the jagged boundaries around a circle were small enough, the stimulus was classified as a circle. This was likely because the local contour perturbations became too small to be registered clearly by the finer filters in the network. This assessment is also supported by inspection of the classification errors of the network on regular circle and square images, which were all on small shapes (diameter < 6 pixels) and mostly on squares. Intriguingly, the network classified global squares with local curved elements as circles at all scales. We considered whether network classification for squares depended solely on the presence of local corners, but tests on straight edged squares with the corners removed revealed that the network still responded "square" with high confidence, suggesting that the straight edges of the square are also diagnostic information for the network. Below a certain size, the network appears to register small corners of a global square as part of a curve but even the smallest circular elements comprising global squares produced "circle" classifications by the network.

Our claim is not that DCNNs are inaccurate in their classifications based on local contour features. In reality, the probe stimuli we tested the network on were not simple, canonical squares or circles. However, the pattern of responses shows a clear dissociation between DCNNs and human perception. Human observers would not look at the square display made of curved elements and label it a circle, nor would anyone look at the circle display made of corner fragments and label it a square. As Gestalt psychologists pointed out long ago with similar demonstrations, global shape or configuration in human perception is a relational notion, profoundly different from a simple aggregation of parts or local elements. More modern work has shown that in visual processing the whole of an object is often extracted before or in preference to the individual parts from which it is constructed (Baker & Kellman, 2018; Elder & Zucker, 1993; Navon, 1977; Pomerantz, Sager, & Stoever, 1977). We found no evidence that sensitivity to global shape exists from the network's ultimate classification decisions; rather, the evidence clearly indicated that classification is "the sum of the parts", as it is driven by local contour characteristics.

Using unrestricted transfer learning did little to help the network learn to use global shape features to perform classification. Even with all the weights in the network free to be retrained, AlexNet still classified the images in Fig. 2 based on local contour features. The circle/square classification task could be well-served by the network learning global shape, but it does not *require* that. Accurate classification during Experiment 1A could be obtained in the training set simply by learning to look for the presence of local features such as straight lines or sharp corners. That the network relies on these features instead of global shape cues is in some ways unsurprising. Global shape is an abstract concept that requires coding of relations that remain constant across infinitely many possible physical variations (Kellman, Garrigan, & Erlikhman, 2013). In contrast, the local features that appear to drive a specific classification task are straightforward to detect using local convolution operations that lie at the heart of these networks, so it is reasonable that an associative network trained to optimize performance for this one task would emerge to become sensitive to local features rather than global relations.

Probe stimuli for both AlexNet and VGG-19 showed a strong bias for local shape cues over global shape information. In Experiment 1c, we tested whether a ResNet-50 network specifically trained to use more shape information and including more layers in the network architecture would show greater sensitivity to global shape. To the contrary,

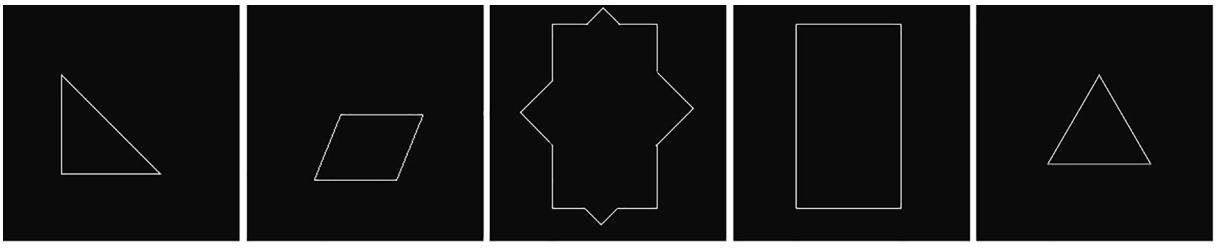


Fig. 5. Additional probe stimuli used to test ResNet-50 trained on stylized ImageNet and ImageNet.

Input Image	Classification	Circle Probability	Square Probability
	**"Circle"**	100%	0%
	"Circle"	100%	0%
	**"Circle"**	100%	0%
	"Circle"	99.6%	0.4%
	**"Circle"**	100%	0%
	"Circle"	98.6%	1.4%
	**"Circle"**	99.5%	0.5%
	"Circle"	67.5%	32.5%
	"Square"	3.0%	97.0%
	"Circle"	69.5%	30.5%
	"Circle"	100%	0%
	"Square"	0.1%	99.9%
	"Circle"	99.0%	1.0%

Fig. 6. Network classification performance for the probe stimuli. Input stimulus for each condition are shown to the left. The classification responses in bold are the decisions consistent with global shapes.

we found little evidence for global shape processing in ResNet-50 trained with stylized ImageNet and ImageNet. Instead, the network appeared to have a strong bias towards circle classifications, assigning more probability to circles for all stimuli except squares, rectangles, and right triangles. Even a rectangle with corner edges extending from the middle of its sides was classified as a circle despite being made up entirely of parts of rectangles and squares.

The ResNet-50 network we tested is both significantly deeper than AlexNet and VGG-19 and trained on stimuli that push it to use more shape cues than the typical ImageNet training scheme. Our results suggest that global shape sensitivity is not automatically developed in deep networks by increasing depth or by using stimuli in which shape is a more diagnostic cue than texture. We cannot rule out that some other training scheme might have an effect on extraction of global shape, but there is no evidence of that in the network models tested thus far.

Local superiority may not dominate at all stages of network processing. The correlations of node activities between stimulus pairs (Fig. 4) are somewhat consistent with a global shape preference at earlier layers of the network. Of the six pairs, the correlation between the circle and the circle made up of corners is the highest at layer 2 (and

correlation remained high until layer 5), consistent with global shape processing. Conversely, the circle made up of corners and the square image have the lowest correlation in layer 2 and among the lowest correlation in later layers despite having similar local elements. One explanation for this finding is that the deep network does extract features about global shape, but the global features are not strongly weighted in the network's classification decision.

This interpretation of global shape processing at earlier layers of the network, however, does little to explain the high correlation between the circle with corner elements and the square with curved elements in layers 3–7. What features make these two images representationally similar at intermediate stages of processing even though *both* their local and global cues differ? A different local feature such as the presence of many sharp points in the image or the density of local features may drive representational similarity between the stimulus pair. That the highest correlation among any pairs occurred between a circle comprised of corner elements and a square comprised of curved elements , and quite substantially so in layers 4–7, reminds us that the particular information the network is extracting is complex, difficult to specify, and not limited to the candidate properties that we may intuitively expect.

Regarding the greater correlation between globally similar shapes in early levels, another possibility is that this association might be driven by lower-level visual features such as size or the presence or absence of luminance regions or even pixels in similar regions of the image. This explanation is supported by the fact that the correlations that we might intuitively think reflect global similarity are showing up *earliest* in the network. These stages of processing are most similar to V1 and V2 in visual cortex (Zeiler & Fergus, 2014), and seem unlikely to reflect global shape similarity. When we computed the layer-wise correlation between the two large, centered probe stimuli and a circle and square that were both small and located near the top left of frame (Fig. 7), we found that the greatest correlations were between a pair of images sharing similar image size, i.e., between the two large probe stimuli and between the circle and square stimuli, despite both pairs differing in both local and global features.

The pattern of correlations across layers for unrestricted learning was extremely similar to the correlations for restricted learning. Even though all connection weights were free to change in the unrestricted condition, the network appeared to extract similar features in the pre-decision layers. This suggests that the original features from ImageNet training were sufficient for outline circle/square classification, and transfer learning mostly modifies the weights preceding the decision layer.

3. Experiment 2

In Experiment 2, we tested whether the network could be guided into using global configuration as its primary basis for recognition by curating the training data. We generated circles and squares with four different kinds of local elements and trained the network to classify by labels of global shape, irrespective of the features of the local elements. We hoped to deprioritize local contour cues by making them non-diagnostic for classification decisions in order to see if more global shape information would be extracted to drive classification. As in Experiment 1, we then tested the network on probe stimuli where local and global information conflicted to examine if our new training regimen resulted in use of global information.

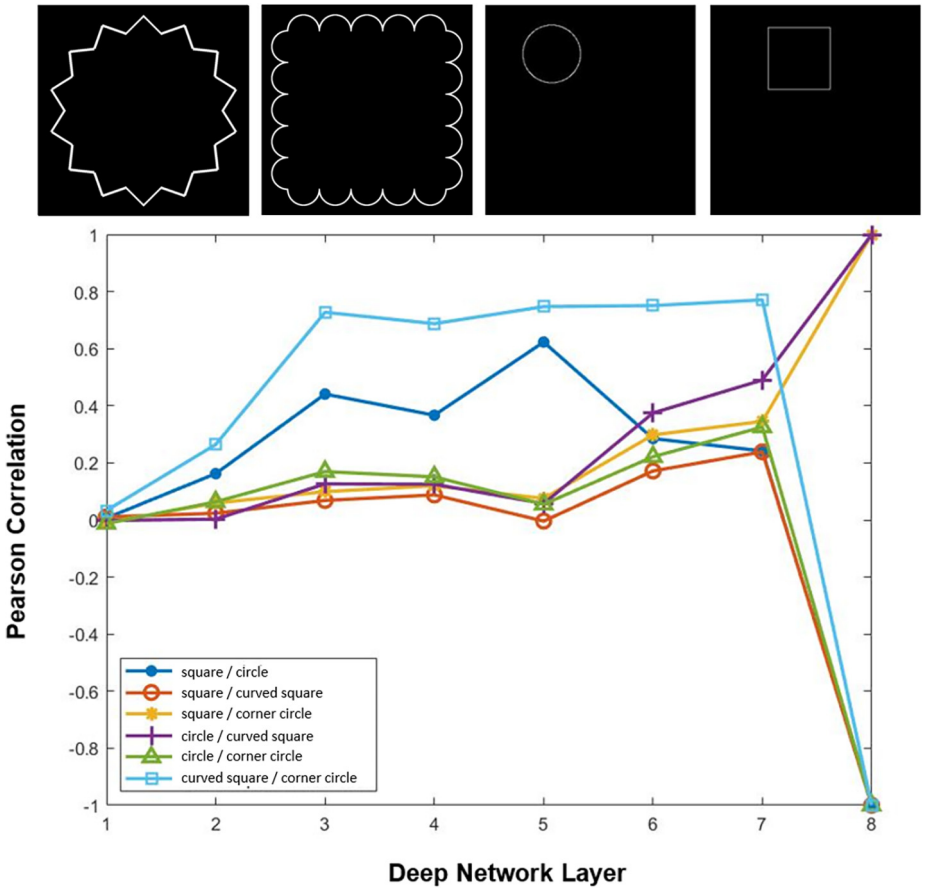


Fig. 7. Layerwise correlation with small, off-center circles and squares and large, on-center squares comprised of curved elements and circles comprised of corner elements. See Fig. 4 for details.

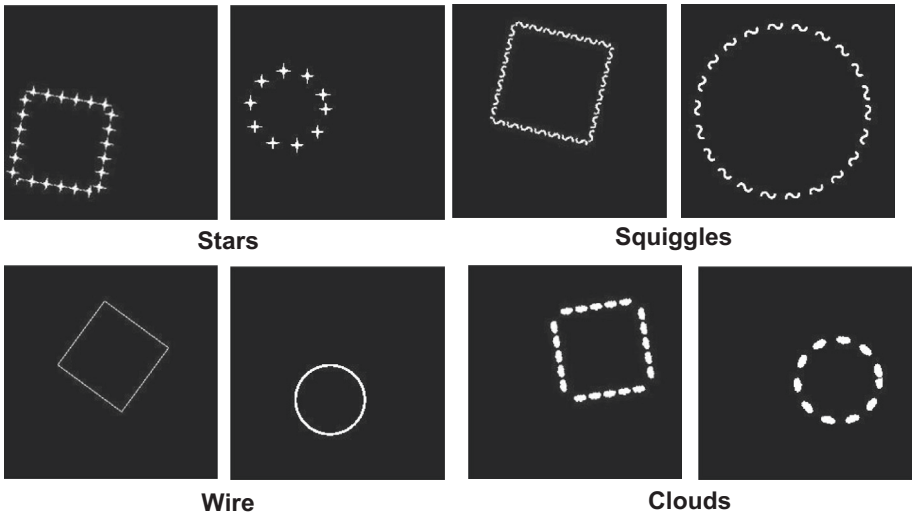


Fig. 8. Training stimuli used in Experiment 2. Training consisted of 1369 squares and 1369 circles with each local element. Size, position, and orientation were all randomly varied in the training data.

3.1. Method

We generated circles and squares with four kinds of local elements: straight edges, squiggles, clouds, and stars. For each kind of local element, there were 1369 square images and 1369 circle images that varied in size, position, and orientation (see Fig. 8). Labels were given to each image depending on their global configuration, not their composing elements. We trained two classifiers, one with restricted transfer learning and one with unrestricted transfer learning, to make the circle/square discrimination. We used 80% of the images in training and the other 20% as a test set. Other than the data used in training, methods for transfer learning were identical to Experiment 1. As in Experiment 1, we also trained VGG-19 using restricted transfer learning on the same set of training data.

Once transfer learning was complete, we tested the network on the same set of probe stimuli used in Experiment 1, where local and global shape cues are in conflict. We also computed the layer-wise correlation between the square images, circle images, circles made up of corners, and squares made up of curved segments.

3.2. Results

Learning was successful in AlexNet for both restricted transfer (criterion reached after 1850 iterations) and unrestricted transfer (criterion reached after 1070 iterations). Accuracy on the validation (withheld) set was very high, 99.95% for restricted transfer learning and 100% for unrestricted. As in Experiment 1, performance was very similar for the two kinds of transfer learning. For brevity, we report performance on AlexNet for the unrestricted transfer learning network since there were virtually no performance differences between the two networks. VGG-19 achieved 100% accuracy on the validation set.

Testing on probe stimuli revealed more classifications consistent with the object's global shape. Classification performance on the same set of probe stimuli as used in earlier studies (Fig. 2) is shown in Fig. 9. Whereas in Experiment 1, two of the eight stimuli received classifications consistent with global shape processing, in Experiment 2, six of the eight stimuli were classified consistently with global shape.

For VGG-19, there was slightly less improvement in the number of global shape classifications. While AlexNet increased from two to six global classifications, VGG-19 improved from two to four out of eight possible classifications. The main difference in performance was for large circles made of corner elements. Whereas AlexNet shifted to classifying these as circles, VGG-19 continued to classify them as squares.

The layer-wise correlation of node activities between squares, circles, squares comprised of curved elements, and circles comprised of corner elements is shown in Fig. 10. The overall pattern in the seven pre-decision layers is very similar to the patterns observed in Experiment 1. The order of stimulus pairs from highest correlation to lowest was almost identical in layers 3–7 for the unrestricted transfer learning network in Experiment 2 as for the unrestricted transfer learning network in Experiment 1. In the few cases where there were ordinal differences, they were between two pairs that had very close to equal correlations in both networks, and one pair changed from slightly higher correlation to slightly lower or vice versa. However, the decision layer (layer 8) in Experiment 2 showed different patterns from the first two experiments, as more global shape responses were given by this model.

3.3. Discussion

The classification task that the network was trained to perform in Experiment 1 could be resolved equally well by using local cues as by using global shape. In Experiment 2, we explicitly trained the network with global shape labels applied to circles and squares comprised of several different kinds of local elements. Thus, the training data included stimuli that had identical local elements but belonged to different object categories. To attain high performance on the classification task, the network would need to disregard the uninformative local feature information. The network appeared to learn to classify shapes comprised of different local elements with a high degree of accuracy, and it did so as quickly as it learned the simpler shape outlines in Experiment 1, despite much more variation within each category.

More importantly, when tested on the probe stimuli in which local and global cues conflicted, the network made significantly more classifications that aligned with the object's global shape than networks trained only on shape outlines (75% global shape agreement in Exp. 2 vs. 25% global shape agreement in Exp. 1). Performance was virtually identical between the networks trained with restricted and unrestricted transfer learning in Experiment 2. Moreover, analysis of the layer-wise correlation between stimulus pairs (Fig. 10) revealed a similar pattern in the pre-decision layers to that observed in Experiments 1 (Fig. 4), despite the substantial changes in the ultimate classifications. This suggests that features that lead to more globally aligned classification results do not need to be learned through exposure to stimuli with uninformative local features. Rather, the network may already be

Input Image	Classification	Circle Probability	Square Probability
	**"Circle"**	99.2%	0.8%
	"Circle"	100%	0%
	**"Circle"**	100%	0%
	"Circle"	100%	0%
	**"Circle"**	100%	0%
	**"Square"**	0%	100%
	**"Circle"**	100%	0%
	**"Square"**	3.8%	96.2%

Fig. 9. Network classification performance for the probe stimuli in Experiment 2. Results are from the network trained with unrestricted transfer learning. The classification responses in bold are the decisions consistent with global shapes. See Fig. 2 for the input stimulus that corresponds to each condition.

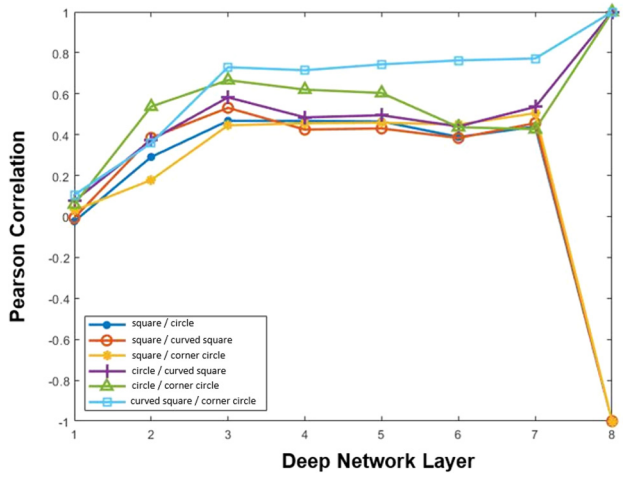


Fig. 10. Layerwise correlation between circles, squares, curved circles, and corner squares from the Experiment 2 network. The correlations reported here are from the network trained with unrestricted transfer learning. See Fig. 4 for more details.

extracting some features that are independent of local cues and in Exp. 2 gave more weight to them due to the training set used.

4. Experiment 3

Introducing images with nondiagnostic local features in transfer learning appeared to produce more network classifications consistent with an object's global shape. In humans, global shape is a complex notion that requires symbolic coding of relations between parts as well as abstraction over features inessential to the object's outline (for discussion, see Kellman et al., 2013). It would be remarkable if existing network architectures automatically develop these complex representations through convolution operations and associative learning simply by providing the right kind of training data. An alternative explanation is that in Exp. 2 the network learned some other kind of non-global visual feature that gave reliably accurate classifications on the training data and more closely aligned with global classification labels for the probe stimuli on which we tested the network. In Experiment 3, we tested the network trained in Experiment 2 on several new kinds of probe stimuli to assess more carefully whether the network had truly developed sensitivity to an object's global shape, or whether the network had exploited some other featural information in its classification performance.

4.1. Methods

All tests in Experiment 3 were conducted on the AlexNet architecture trained with unrestricted transfer learning in Experiment 2 with no new training. First, we generated 2378 new circle and square images. As in the training data for Experiment 2, these images had elements (in this case, straight lines) arranged around a virtual circle or square outline. The main difference was that instead of assigning elements orientations that were consistent with their position along the shape outline, we gave each element a random orientation (Fig. 11). Randomly assigning element orientation removed the local cue of orientation difference between adjacent elements.

Next, we took the figures that the network had classified in a way

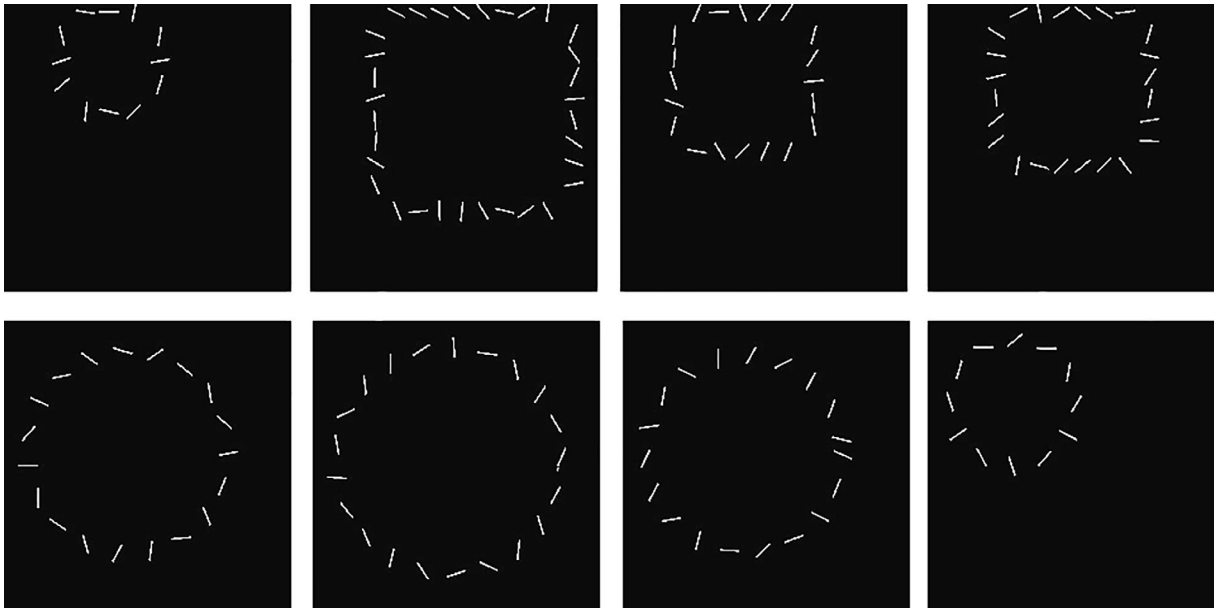


Fig. 11. Circles and squares with randomly oriented line segments from Experiment 3.

consistent with global shape processing and destroyed the global configuration by breaking them into disconnected fragments (Fig. 12). For this manipulation, we used the four figures that had previously been given a local classification but had switched to a global classification with our new training set in Exp. 2. If the network is truly classifying objects by their global shape, we predicted that this kind of perturbation should have a large detrimental effect on classification performance.

As will be described below, analyzing results from the first two sets of new probe stimuli forced us to consider a new hypothesis that, despite appearances, the network was not classifying by global configuration, but by larger, coarser local features. Features detected by operators sensitive to low spatial frequencies over larger image regions would be insensitive to the local variations along contours, which, as noted, is a necessary ingredient for seeing global shape. However, the network may still lack a way of extracting more global relations of these

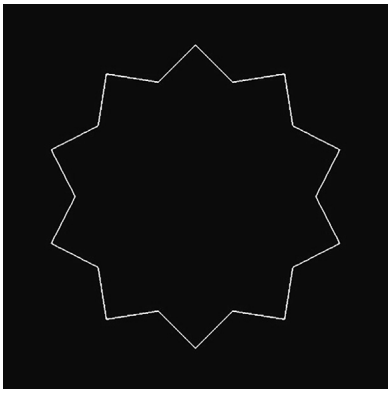


Fig. 13. Global circle with larger local corner elements.

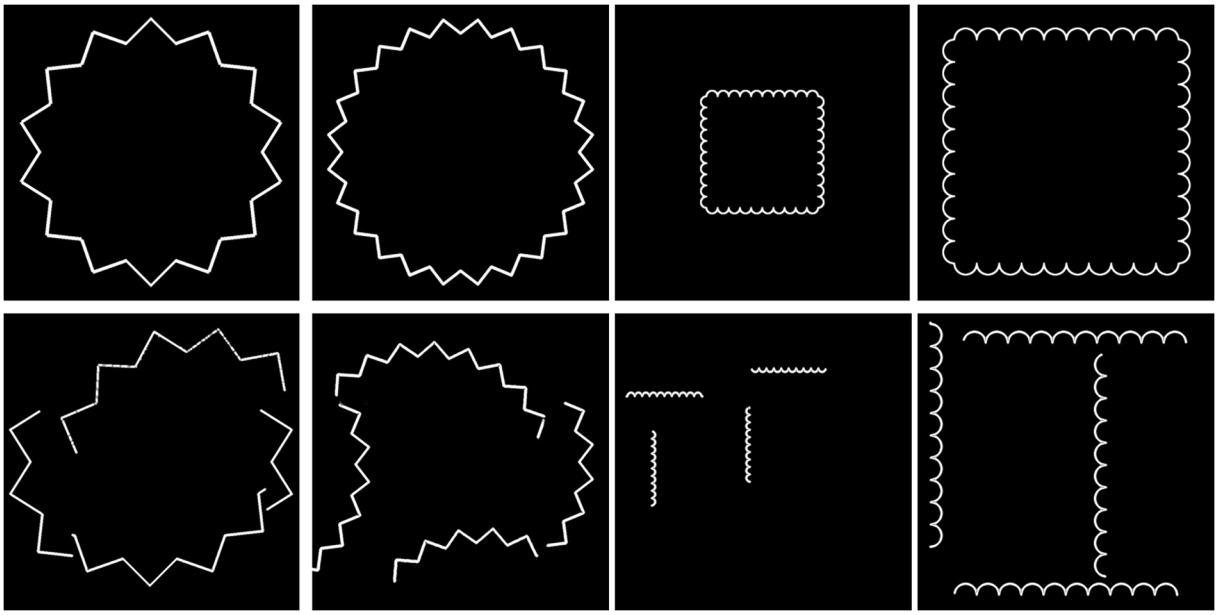


Fig. 12. Fragmented figures generated from test figures in Experiment 2. Top: Stimuli that the network classified by global shape in Exp. 2. Bottom: Figures broken into fragments in order to assess global shape processing in Exp. 3.

Input Image	Classification	Circle Probability	Square Probability
	"**Circle**"	99.2%	0.8%
	"Circle"	97.0%	3.0%
	"**Circle**"	100%	0%
	"Circle"	100%	0%
	"**Square**"	3.8%	96.2%
	"Square"	0%	100%
	"**Square**"	0%	100%
	"Square"	3.8%	96.2%

Fig. 14. Network classification results for fragmented and unfragmented images. All of the unfragmented images were classified by global shape in Experiment 2.

coarser features. Relevant to this potential explanation, we added one more display as a test. We created a new image with a circle made up of fewer, larger corner elements (Fig. 13). We wanted to test whether the network was sensitive to the radial arrangement of the corners or if it had developed large filters that allowed it to identify curved features even in our displays with large corner elements.

4.2. Results

For the 2378 circle and square images comprised of randomly oriented line segments, the network produced slightly higher than chance performance, with a mean accuracy of 59.3%. There was considerable bias in the network's response—it correctly classified 1350 of 1369 circles correctly, but only 59 of the 1369 squares correctly.

For the fragmented shapes, we compared the network's classification label and confidence level with the results we found for the unfragmented objects in Experiment 2. The results are shown in Fig. 14. In all four cases, the network gave the same classification response for the fragmented images as it did for the unfragmented images with high confidence. There was no reliable difference between the network's confidence for the fragmented images and the unfragmented images ($t(6) = -1.01$, $p = .39$). We also compared the activations at the layer immediately before the network's confidence is converted to a probability score. Activations at this layer are unbounded and do not need to add up to the same value for different input images. There was also no difference in network confidence for the fragmented vs. unfragmented images at this layer, ($t(6) = -0.64$, $p = .55$).

The network classified the image with a circle shape comprised of fewer and larger corner elements shown in Fig. 13 as a square with 84.6% confidence. In Experiment 2, all images with a global circle shape comprised of corner elements were classified as circles. This result showed that the classification decision depends on large fragments of the shape image.

4.3. Discussion

The purpose of Experiment 3 was to investigate whether a deep network trained with nondiagnostic local element cues truly uses global shape to classify objects. To clarify the findings of Experiment 3, it might be helpful to distinguish between two processes that take place in shape representation in the human visual system. One process is encoding the configuration of parts within an object. For example, we know that the head of a dog is at one end of its body and the tail is at the other end. For circles and squares, configural information could include features like all sides of a square being equal length, all circles having equally long major and minor axes, and all edges of a square joining at 90-degree angles. Objects that are greatly simplified in every other way can often still be recognized with ease if the configuration of the shape is preserved. For example, Fig. 15 shows a small number of oriented rectangles that nonetheless give a strong percept of a dog. Despite having virtually no features physically in common with a dog, the arrangement of parts makes the shape look like a dog to human observers. Networks trained on ImageNet have shown no ability to classify this kind of stimuli (Baker, Erlikhman, et al., 2018; Baker, Lu, et al., 2018). Indeed, when we tested this image on AlexNet, the network assigned higher probability to 32 objects than any of its 120 dog categories. DCNNs may be accurate in reporting whether two objects are physically similar (Buckner, 2019; Zhou & Firestone, 2019), but they have no ability to predict perceived similarity based on global shape cues.

Another process in encoding shape representations depends on the detection of local contours. Objects in the world have very complicated physical boundaries, especially for deformable or articulated objects.

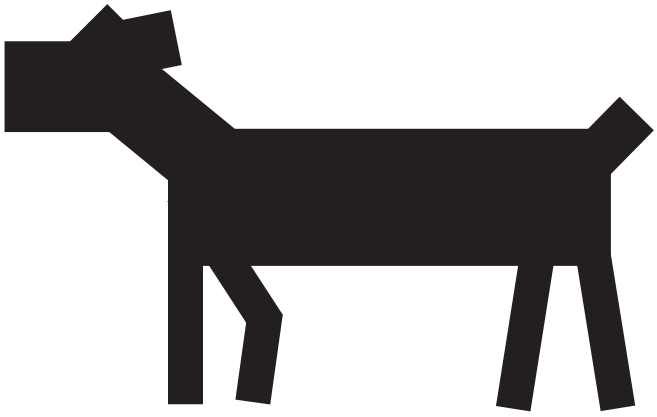


Fig. 15. A dog comprised of a number of oriented rectangles. DCNNs assigns higher probability to 32 object categories than any of the dog categories.

Abstract shape encoding in biological vision is needed because representing all the contour variation within a shape is unrealistic, as it would exceed the capacity of visual memory, and maladaptive, as it would tend to obstruct recognition between two shapes that differ in their local contour features. For example, images of two pine trees of the same species may have very different local branch and pine needle orientations, but both will have similar overall contours, allowing their categorical similarity to be abstracted.

In Experiment 3, we first tested the network to see if it could classify shapes with randomly oriented line segments placed along the virtual contour. It largely failed in this task. One reason the network might have failed to utilize global shape is that, in the training set, features of individual elements were uninformative, but changes in local feature orientations could be used to classify the object. For example, in the Experiment 2 training data (Fig. 8), the difference in orientations between pairs of adjacent elements such as a squiggle or cloud in a circle figure is always between 12 and 60 degrees. For a square, if one pair of elements is oriented along a certain direction, the next pair will always differ by either zero or 90 degrees. In the absence of these element pair relations, the network is unable to classify the objects, despite the shape's global configuration being preserved in most cases. (For some objects, particularly small ones, randomly assigning individual element orientations arguably does affect the shape's global configuration, but even where that is not the case, the network performs very poorly on the task.) Another possible explanation is that randomly oriented lines are more difficult to capture in a single low spatial frequency filter. If the network is finding local edge relations by smoothing over local elements, randomly oriented lines might make this difficult.

Our first test gave little evidence that the network was classifying based on an object's configuration, so we removed configural cues in our second test. By fragmenting the object into disconnected parts, we could measure the contribution of configural information to the network's ultimate classifications. For the DCNN, the configural contribution appears to be nonexistent: the network assigns as much probability to the global shape classification for the fragmented images in Fig. 12 as it does for the original unfragmented images, where configural information is present. This result strongly supports the idea that the network can rely on coarse feature detection but does not perform global shape extraction.

This idea is the likely explanation for the result that the network classified 75% of the probe stimuli from Experiment 2 based on their global shape similarity to squares or circles. Coarse feature detection of parts of object boundaries, and perhaps local orientation relations among such coarse feature detectors, could explain the results. This explanation is supported by the results of our fragmentation test in Exp. 3: Chopping up a global shape into large fragments made no difference in classification performance, although global circle or square shapes were no longer present. A possible counterpoint to this finding is that the network's final classification is a set of probabilities that must add up to one, so the DCNN might still assign high probability to the "square" classification for fragmented pieces of partial circles arranged along a straight line or two partial squares arranged along a curve. However, even in the layer before activation values are transformed into probability scores, there is no difference between network confidence in the fragmented image classification as compared to the unfragmented images. The activation scores at this point are unbounded scores assigned to each category, so two items can have different activations even if they are ultimately assigned similar probability.

The same idea, at a different scale, can be seen in some earlier results. In the simple training used in Experiment 1, when the network was tested on circles made up of corner elements that were sufficiently small, it classified them as circles, presumably because it was able to rely on filters coarse enough to be uninfluenced by small contour variations, allowing extraction of a curved local segment from the corner stimuli. In Experiment 2, rather than learning to classify based on configural cues, the network appears to have learned to use larger filters that respond to local curvature from larger extents along the contour.

Our data suggest that such larger, coarse filters likely already exist in trained deep networks, as evidenced by the highly similar performance between networks that were trained with restricted transfer learning and networks that were trained with unrestricted transfer learning. For typical object recognition tasks, very large filters are probably given low weighting in ultimate network classifications, as they will often blur over critical information. Training on the dataset from Experiment 2, however, would result in the giving these filters much more importance, as feedback would lead the network to downweigh smaller filters that cannot distinguish a circle made up of small clouds from a square made up of small clouds.

Finally, our hypothesis that the network is achieving more global-like performance in Experiment 2 due to more coarse, but still fragmentary, contour abstraction, rather than sensitivity to an object's part configuration, makes the prediction that if we enlarged the local elements of an object that the network classified by global shape, the model should revert to a local contour feature classification. We tested this idea with a circle made up of larger corner elements in our final test. Indeed, the DCNN classified the shape as a square, despite classifying other circles made up of smaller corner elements as circles.

5. General discussion

In this research, we used transfer learning on deep convolutional networks trained for object recognition to evaluate their sensitivity to local and global shape features. Previous results led us to hypothesize that an object's local edge properties, and some local relations of contour orientations, influence classification in DCNNs, but networks do not classify based on the global arrangement of elements composing the object's shape (Baker, Erlikhman, et al., 2018; Baker, Lu, et al., 2018). In natural images, an object's local and global shape features are generally consistent: Locally, the hook of a hammer is curved in a specific way, and globally, it is arranged behind the hammer head and at the top of the shaft. It is therefore difficult to disentangle the separate contributions of local and global shape cues to object recognition. In the present work, we carried out more direct tests by retraining the network for new classifications that probed local and global shape processing. Our work focused on three major questions. First, does a network trained for object classification with the ImageNet database learn and respond based on an object's global shape? Second, can the network learn global shape if connection weights can update in transfer learning? Finally, can sensitivity to global shape be promoted in DCNNs by curating a training set in which global shape is needed for accurate classification?

In Experiment 1, we first put local and global shape cues in

competition with each other for a network that was trained with restricted transfer learning. We adapted a DCNN to make a binary classification between circles and squares. Because connection weights from earlier layers remained unchanged, network responses in this adapted task depended on the same set of features that would be available to AlexNet for its classifications of objects. The only change to the network in our transfer learning procedures was how this set of features were weighted to drive a classification response. During training (Experiment 1A), both local and global shape properties were discriminative for the new classification task, and the network performed very well. In Experiment 1B, we separated local cues from global by taking local elements from one category and arranging them so that the global shape reflected the other category. When presented with these hybrid images, the network always classified the objects based on the local elements rather than on the organization of these elements: It detected the local curved segments of a circle but was insensitive to their arrangement into the pattern of a square (and vice versa).

Insensitivity to global shape information does not appear to be a unique artifact of AlexNet's architecture. When we trained the deeper VGG-19 with the same protocol, we found the same pattern of responses as from AlexNet. Even very deep networks explicitly trained to have a shape bias showed no sensitivity to global shape cues. Geirhos et al.'s ResNet-50 architecture trained on stylized images and natural images was greatly biased towards circle classifications, classifying even images of rhombuses and rectangles with added corners as circles even though they had no global similarity to circles and a high degree of local similarity with squares. In general, our tests of ResNet-50 revealed no evidence of sensitivity to global shape. The present results are consistent with a deep and general limitation of deep learning systems; however, it remains possible that some different architecture or training regimen would produce differing results.

Experiment 1 also tested the idea that DCNNs do not automatically develop sensitivity to global shape from ImageNet training, but might do so when exposed to training data where shape is used to define the object categories. Using the same training set from the restricted transfer learning condition, we trained the network with unrestricted transfer learning, allowing all connection weights between all layers of AlexNet to update when learning to make circle/square classifications. Training was once more successful, but tests on probe stimuli in which the local and global shape cues differed revealed no more global shape recognition in these retrained DCNNs than was observed in the DCNN that was retrained in a highly constrained manner with restricted transfer learning. Analysis of the layer-wise correlation between stimulus pairs also showed similar patterns of activation across layers in both models. These findings suggest that DCNNs learn what shape features they are capable of learning from the ImageNet database and that the critical thing for developing more shape-driven classification is adjustment of how these features are weighted in the network's decision, not adjustment of what features the network extracts. The idea that the network uses decision reweighting rather than new feature development is also supported in Experiment 2, where unrestricted and restricted transfer learning led to nearly identical response patterns even when the DCNN was trained on a database where local cues were specifically nondiagnostic.

The results of Experiment 1 suggest that deep networks trained for object recognition do not encode the shapes of objects. These networks readily classify based on local contour features, as seen in the circle/square task. If DCNNs are to be taken as models for human perception, we would expect an inverted pattern of results. As made clear by research in the Gestalt tradition and many subsequent efforts, for humans, the whole is more defining of shape and more salient than the elements composing it (Elder & Zucker, 1993; Gold, Murray, Bennett, & Sekuler, 2000; Koffka, 1935; Navon, 1977; Pomerantz et al., 1977). In human perception, local elements appear to be extracted briefly in service of encoding the configural whole, but become accessible for perceptual decision-making only later, and with greater effort than is needed to encode a global pattern (for discussion, see Baker & Kellman, 2018).

Experiments 2 and 3 examined the influence of training data in pushing the network towards more global classifications. We created a training set where elements such as clouds, stars, and squiggles were arranged along a circle or square's virtual contour and trained a deep network to do the circle/square classification. Despite the ambiguity of the local features (for the overall classification to be learned), the network learned to classify the training images very accurately. More importantly, when we tested on the probe stimuli from Experiment 1, the network trained with these new images classified 75% of the stimuli based on their global shape (as compared to 25% for the networks in Experiment 1). While these findings initially seemed promising in indicating global shape sensitivity in DCNNs with targeted training, follow-up tests in Experiment 3 revealed that global shape relations were not driving the performance of the network. When we fragmented the probe stimuli, a manipulation that destroys global shape, the network's classifications remained the same and maintained equal confidence. These results suggested that collections of responses of larger, coarser filters underlay the indications of "global" classification in Exp. 2.

This idea suggested one final test: If global shape were retained, but the grain size of features along the contour was enlarged, presumably beyond what could be "blurred" by the largest coarse filters, what would the network do? A circle made of large square elements was classified as a square. Importantly, in this test, the overall size of the circle was well within the range of global circles trained and tested in Exp. 2. This result suggests that there was no encoding of a global circular shape that the network could use in classification.

The results of Experiment 3 provide evidence that global shape configuration plays no role in DCNN object recognition, even when the training data are arranged to promote the use of configural information. In order to achieve the performance observed in Experiment 2, the network likely upweighted feature detectors that filter over large contour variations to extract a smoothed low frequency contour segments. The DCNN is able to respond to coarser image information despite uninformative local contour variation, a crucial process in visual perception (Attneave, 1954; Bell, Badcock, Wilson, & Wilkinson, 2007), but it does not encode the spatial arrangements of parts with respect to each other. The process by which the network extracts these larger contour features does not appear to be abstract. When we tested the DCNN on a circle made up of larger corner elements, it classified it as a square. We interpret this to mean that the network is simply using larger filters, not that the network has a way of registering the relations of parts into an abstract shape description.

After taking the use of larger local filters into account, the results of Experiments 2 and 3 give no evidence for the use of global shape information in network classification. Could a different training scheme force DCNNs to acquire sensitivity to the global configuration of object parts? One possibility is that training on a larger array of straight edged and curvilinear shapes could push the network towards global shape sensitivity. Our suspicion, however, is that any set of categories in which some kind of local contour information can lead to accurate classification will not produce sensitivity to spatial relations of parts in feedforward DCNNs.

If finding the correct training data cannot supply networks with global shape sensitivity, it is worth considering whether different architectures might be better suited to the task. Several studies have shown that networks' performance becomes more similar to humans when they have more nonlinear hidden layers. For example, Seijdel, Tsakmakidis, de Haan, Bohte, and Scholte (2019) showed evidence that deeper networks use more pixels from the object vs. its background in recognition tasks, although they do not claim the networks extract the object's bounding contour from the image. Kubilius et al. (2016) found that networks with more layers have greater local shape sensitivity. Our comparisons of AlexNet, VGG-19 and ResNet-50, however, showed no

improvement in sensitivity to global shape with greater depth.

Another possibility is that recurrence is needed for more global notions of shape to arise in deep networks. Recurrent networks are more biologically plausible (Kietzmann et al., 2019) and have been shown to more closely mirror human perception in tests such as the effect of configurations on crowding (Doerig, Schmittwilken, Sayim, Manassi, & Herzog, 2019). Especially intriguing for recognition from global shape information is work on horizontal gated recurrent units. Unlike feedforward networks, these horizontally connected architectures are able to detect paths from distant unconnected elements as in a Field, Hayes, and Hess (1993)-style task, potentially a key feature in representing the bounding contour of an object (Linsley, Kim, Veerabadran, Windolf, & Serre, 2018).

While recurrent architectures might be necessary to get networks to perceive global shape, we do not believe that simply building a recurrent network and training it to classify displays would produce any sensitivity to global shape. We believe there are likely two complementary problems. First, any architecture that learns through supervised learning and optimizes its weights by gradient descent, by design, converges on the simplest way to divide the high dimensional categorical space. Results from this study and from previous investigations into DCNN shape sensitivity (Baker, Erlikhman, et al., 2018; Baker, Lu, et al., 2018) may suggest that global shape is almost never the simplest way to discriminate between static images. The second problem is potentially deeper. The basic convolution operations at the heart of deep learning networks may be insufficient to extract the more abstract, relational structure that constitutes global shape. Other kinds of operations, involving abstraction, may be required (Baker & Kellman, 2018). In light of these potential requirements, it is remarkable that biological visual systems have evolved to use global shape so preferentially. For humans, objects have affordances beyond their identity that may depend on global shape far more than on texture or local edge relations. To this end, the visual system has powerful mechanisms for segmenting figure from ground (Rubin, 1915), completing objects behind occluders (Kellman & Shipley, 1991; Michotte et al., 1964), encoding an abstract description of a bounding contour (Feldman & Singh, 2006; Baker & Kellman, 2018; Baker, Garrigan, & Kellman, 2020), and inferring volumetric properties about the object (Li, Pizlo, & Steinman, 2009; Marr & Nishihara, 1978). These mechanisms are likely all important components of global shape sensitivity, but they are not required, as DCNNs' remarkable classification performance has shown, to assign classification labels to objects.

6. Conclusion

Deep convolutional networks' performance comprises an inversion of human performance with respect to global and local shape processing. While abstract relations of elements predominate both in human perception of shape and in object recognition, DCNNs appear to extract only local features, with no representation of how they relate to each other. Even when given training data specifically targeted at developing global shape sensitivity, the network simply relies on larger fragment detectors without sensitivity to the configuration of an object's parts. While other research has found interesting similarities between humans and deep networks, these findings indicate that DCNNs recognition of objects is accomplished much differently from human visual perception, in which shape is the predominant information, subserved by elaborate mechanisms that separate figure from ground, produce abstract shape descriptions for bounding contours of objects, and use these representations as the basis of object perception and recognition.

Credit authorship contribution statement

Nicholas Baker: Conceptualization, Methodology, Software, Validation, Investigation, Writing - original draft. **Hongjing Lu:** Formal analysis, Writing - review & editing, Software, Supervision. **Gennady Erlikhman:** Software, Methodology, Writing - review & editing. **Philip Kellman:** Conceptualization, Writing - review & editing, Supervision.

Acknowledgement

This research was funded by National Science Foundation grant BSC-1655300 to HL and NIH/NCI Award 1R01CA236791-01 to PK.

References

Attneave, F. (1954). Some informational aspects of visual perception. *Psychological Review, 61*(3), 183.
Baker, N., Erlikhman, G. E., Kellman, P. J., & Lu, H. (2018a). *Deep convolutional networks fail to perceive illusory contours, presented at Cognitive Science 2018, Madison, WI*.
Baker, N., & Kellman, P. J. (2018). Abstract shape representation in human visual perception. *Journal of Experimental Psychology: General, 147*(9), 1295.
Baker, N., Lu, H., Erlikhman, G., & Kellman, P. J. (2018b). Deep convolutional networks do not classify based on global object shape. *PLoS Computational Biology* [In press].
Baker, N., Garrigan, P., & Kellman, P. J. (2020). Constant curvature segments as building blocks for 2D shape. *Under Review*.
Bell, J., Badcock, D. R., Wilson, H., & Wilkinson, F. (2007). Detection of shape in radial frequency contours: Independence of local and global form information. *Vision Research, 47*(11), 1518–1522.
Belongie, S., Malik, J., & Puzicha, J. (2002). *Shape matching and object recognition using shape contexts.* California Univ San Diego La Jolla Dept of Computer Science and Engineering.
Bergevin, R., & Levine, M. D. (1993). Generic object recognition: Building and matching coarse descriptions from line drawings. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 15*(1), 19–36.
Biederman, I. (1987). Recognition-by-components: a theory of human image understanding. *Psychological Review, 94*(2), 115.
Biederman, I., & Ju, G. (1988). Surface versus edge-based determinants of visual recognition. *Cognitive Psychology, 20*(1), 38–64.
Brendel, W., & Bethge, M. (2019). Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. *arXiv preprint* Doi:1904.00760.
Buckner, C. (2019). *The comparative psychology of artificial intelligences*.
Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports, 6*, 27755.
Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. *Computer vision and pattern recognition, 2009. CVPR 2009. IEEE conference on* (pp. 248–255). IEEE.
Doerig, A., Bornet, A., Choung, O. H., & Herzog, M. H. (2020). Crowding reveals fundamental differences in local vs. global processing in humans and machines. *Vision Research, 167*, 39–45.
Doerig, A., Schmittwilken, L., Sayim, B., Manassi, M., & Herzog, M. H. (2019). Capsule networks as recurrent models of grouping and segmentation. *BioRxiv*747394.
Driver, J., & Baylis, G. C. (1996). Edge-assignment and figure–ground segmentation in short-term visual matching. *Cognitive Psychology, 31*(3), 248–306.
Dubey, R., Peterson, J., Khosla, A., Yang, M. H., & Ghanem, B. (2015). What makes an object memorable? *Proceedings of the IEEE international conference on computer vision* (pp. 1089–1097). .
Eckstein, M. P., Koehler, K., Welbourne, L. E., & Akbas, E. (2017). Humans, but not deep neural networks, often miss giant targets in scenes. *Current Biology, 27*(18), 2827–2832.
Elder, J. H., & Velisavljević, L. (2009). Cue dynamics underlying rapid detection of animals in natural scenes. *Journal of Vision, 9*(7) 7 7.
Elder, J., & Zucker, S. (1993). The effect of contour closure on the rapid discrimination of two-dimensional shapes. *Vision Research, 33*(7), 981–991.
Erlikhman, G., Caplovitz, G. P., Gurariy, G., Medina, J., & Snow, J. C. (2018). Towards a unified perspective of object shape and motion processing in human dorsal cortex. *Consciousness & Cognition, 64*, 106–120.
Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature, 542*(7639), 115.
Feldman, J., & Singh, M. (2006). Bayesian estimation of the shape skeleton. *Proceedings of the National Academy of Sciences, 103*(47), 18014–18019.
Field, D. J., Hayes, A., & Hess, R. F. (1993). Contour integration by the human visual system: Evidence for a local "association field". *Vision Research, 33*(2), 173–193.
Gatys, L. A., Ecker, A. S., & Bethge, M. (2016). Image style transfer using convolutional neural networks. *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2414–2423). .
Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2018). *ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness.* arXiv preprint arXiv:1811.12231.
Gold, J. M., Murray, R. F., Bennett, P. J., & Sekuler, A. B. (2000). Deriving behavioural receptive fields for visually completed contours. *Current Biology, 10*(11), 663–666.
Hermann, K. L., & Kornblith, S. (2019). *Exploring the origins and prevalence of texture bias in convolutional neural networks.* arXiv preprint arXiv:1911.09071.
Hoo-Chang, S., Roth, H. R., Gao, M., Lu, L., Xu, Z., Nogues, I., ... Summers, R. M. (2016). Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Transactions on Medical Imaging, 35*(5), 1285.

Imai, M., Gentner, D., & Uchida, N. (1994). Children's theories of word meaning: The role of shape similarity in early acquisition. *Cognitive Development, 9*(1), 45–75.
Kahneman, D., Treisman, A., & Gibbs, B. J. (1992). The reviewing of object files: Object specific integration of information. *Cognitive Psychology, 24*, 175–219.
Kellman, P. J., & Arterberry, M. E. (2000). *The cradle of knowledge: Development of perception in infancy*. MIT Press.
Kellman, P. J., Garrigan, P., & Erlikhman, G. (2013). Challenges in understanding visual shape perception and representation: Bridging subsymbolic and symbolic coding. *Shape perception in human and computer vision* (pp. 249–274). London: Springer.
Kellman, P. J., & Shipley, T. F. (1991). A theory of visual interpolation in object perception. *Cognitive Psychology, 23*(2), 141–221.
Kellman, P. J., & Spelke, E. S. (1983). Perception of partly occluded objects in infancy. *Cognitive Psychology, 15*(4), 483–524.
Kietzmann, T. C., Spoerer, C. J., Sörensen, L. K., Cichy, R. M., Hauk, O., & Kriegeskorte, N. (2019). Recurrence is required to capture the representational dynamics of the human visual system. *Proceedings of the National Academy of Sciences, 116*(43), 21854–21863.
Koffka, K. (1935). *Principles of Gestalt psychology*. Routledge.
Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* (pp. 1097–1105). .
Kubilius, J., Bracci, S., & de Beeck, H. P. O. (2016). Deep neural networks as a computational model for human shape sensitivity. *PLoS Computational Biology, 12*(4), e1004896.
Kümmerer, M., Theis, L., & Bethge, M. (2014). *Deep gaze i: Boosting saliency prediction with feature maps trained on imagenet.* arXiv preprint arXiv:1411.1045.
Landau, B., Smith, L. B., & Jones, S. S. (1988). The importance of shape in early lexical learning. *Cognitive Development, 3*(3), 299–321.
Li, Y., Pizlo, Z., & Steinman, R. M. (2009). A computational model that recovers the 3D shape of an object from a single 2D retinal representation. *Vision Research, 49*(9), 979–991.
Linsley, D., Kim, J., Veerabadran, V., Windolf, C., & Serre, T. (2018). Learning long-range spatial dependencies with horizontal gated recurrent units. *Advances in neural information processing systems* (pp. 152–164). .
Lloyd-Jones, T. J., & Luckhurst, L. (2002). Outline shape is a mediator of object recognition that is particularly important for living things. *Memory & Cognition, 30*(4), 489–498.
Lowe, D. G. (1999). Object recognition from local scale-invariant features. *Computer vision, 1999. The proceedings of the seventh IEEE international conference on (vol 2)* (pp. 1150–1157). IEEE.
Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information.* New York, NY: Henry Holt and Co. Inc 2(4.2).
Marr, D., & Nishihara, H. K. (1978). Representation and recognition of the spatial organization of three-dimensional shapes. *Proceedings of the Royal Society of London. Series B. Biological Sciences, 200*(1140), 269–294.
Michotte, A., Thines, G., & Crabbe, G. (1964). Amodal completion and perceptual organization (Tr.). *Louvein. Studia Psychologica*.
Navon, D. (1977). Forest before trees: The precedence of global features in visual perception. *Cognitive Psychology, 9*(3), 353–383.
Palmer, S. E. (1999). *Vision science: Photons to phenomenology*. MIT Press.
Peterson, J. C., Abbott, J. T., & Griffiths, T. L. (2016). *Adapting deep network features to capture psychological representations.* arXiv preprint arXiv:1608.02164.
Pomerantz, J. R., Sager, L. C., & Stoever, R. J. (1977). Perception of wholes and of their component parts: Some configural superiority effects. *Journal of Experimental Psychology: Human Perception and Performance, 3*(3), 422.
Pospisil, D. A., Pasupathy, A., & Bair, W. (2018). 'Artiphysiology'reveals V4-like shape tuning in a deep network trained for image classification. *Elife, 7*, e38242.
Rezanejad, M., & Siddiqi, K. (2013). Flux graphs for 2D shape analysis. *Shape perception in Human and computer vision* (pp. 41–54). London: Springer.
Rubin, E. (1915). *Synsoplevede figurer*.
Seijdel, N., Tsakmakidis, N., de Haan, E. H., Bohte, S. M., & Scholte, H. S. (2019). Depth in convolutional neural networks solves scene segmentation. *bioRxiv*.
Shi, J., & Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transaction on PAMI, 22*(8), 888–905.
Shotton, J., Winn, J., Rother, C., & Criminisi, A. (2009). Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *International Journal of Computer Vision, 81*(1), 2–23.
Simonyan, K., & Zisserman, A. (2014). *Very deep convolutional networks for large-scale image recognition.* arXiv preprint arXiv:1409.1556.
Spelke, E. S. (1990). Principles of object perception. *Cognitive Science, 14*(1), 29–56.
Vallortigara, G. (2012). Core knowledge of object, number, and geometry: A comparative and neural approach. *Cognitive Neuropsychology, 29*, 213–226. https://doi.org/10.1080/02643294.2012.654772.
Xu, F., Carey, S., & Quint, N. (2004). The emergence of kind-based object individuation in infancy. *Cognitive Psychology, 49*(2), 155–190.
Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences, 111*(23), 8619–8624.
Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. *European conference on computer vision* (pp. 818–833). Cham: Springer.
Zhou, Z., & Firestone, C. (2019). Humans can decipher adversarial images. *Nature Communications, 10*(1), 1–9.
Zhou, H., Friedman, H. S., & Von Der Heydt, R. (2000). Coding of border ownership in monkey visual cortex. *Journal of Neuroscience, 20*(17), 6594–6611.