

RESEARCH ARTICLE

Deep convolutional networks do not classify based on global object shape

Nicholas Baker^{1*}, Hongjing Lu¹, Gennady Erlikhman², Philip J. Kellman¹

1 Department of Psychology, University of California, Los Angeles, Los Angeles, California, United States of America, **2** University of Nevada, Reno, Nevada, United States of America

✉ Current address: Department of Psychology, University of California, Los Angeles, Los Angeles, California, United States of America

* nbaker9@ucla.edu



OPEN ACCESS

Citation: Baker N, Lu H, Erlikhman G, Kellman PJ (2018) Deep convolutional networks do not classify based on global object shape. *PLoS Comput Biol* 14(12): e1006613. <https://doi.org/10.1371/journal.pcbi.1006613>

Editor: Wolfgang Einhäuser, Technische Universitat Chemnitz, GERMANY

Received: November 3, 2017

Accepted: October 31, 2018

Published: December 7, 2018

Copyright: © 2018 Baker et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All relevant data are within the manuscript and its Supporting Information files.

Funding: This research was funded by the National Science Foundation Research Traineeship for Modeling and Understanding human behavior (MENTOR) DGE-1829071 (<http://www.math.ucla.edu/~bertozi/NRT/index.html>) to NB and the Advancing Theory and Application in Perceptual and Adaptive Learning to Improve Community College Mathematics NSF Grant ECR-1644916 (<https://www.nsf.gov/awardsearch/showAward?>

Abstract

Deep convolutional networks (DCNNs) are achieving previously unseen performance in object classification, raising questions about whether DCNNs operate similarly to human vision. In biological vision, shape is arguably the most important cue for recognition. We tested the role of shape information in DCNNs trained to recognize objects. In Experiment 1, we presented a trained DCNN with object silhouettes that preserved overall shape but were filled with surface texture taken from other objects. Shape cues appeared to play some role in the classification of artifacts, but little or none for animals. In Experiments 2–4, DCNNs showed no ability to classify glass figurines or outlines but correctly classified some silhouettes. Aspects of these results led us to hypothesize that DCNNs do not distinguish object’s bounding contours from other edges, and that DCNNs access some local shape features, but not global shape. In Experiment 5, we tested this hypothesis with displays that preserved local features but disrupted global shape, and vice versa. With disrupted global shape, which reduced human accuracy to 28%, DCNNs gave the same classification labels as with ordinary shapes. Conversely, local contour changes eliminated accurate DCNN classification but caused no difficulty for human observers. These results provide evidence that DCNNs have access to some local shape information in the form of local edge relations, but they have no access to global object shapes.

Author summary

“Deep learning” systems—specifically, deep convolutional neural networks (DCNNs)—have recently achieved near human levels of performance in object recognition tasks. It has been suggested that the processing in these systems may model or explain object perception abilities in biological vision. For humans, shape is the most important cue for recognizing objects. We tested whether deep convolutional neural networks trained to recognize objects make use of object shape. Our findings indicate that other cues, such as surface texture, play a larger role in deep network classification than in human recognition. Most crucially, we show that deep learning systems have no sensitivity to the overall shape of an object. Whereas deep learning systems can access some local shape features,

AWD_ID=1644916&HistoricalAwards=false) to PJK. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

such as local orientation relations, they are not sensitive to the arrangement of these edge features or global shape in general, and they do not appear to distinguish bounding contours of objects from other edge information. These findings show a crucial divergence between artificial visual systems and biological visual processes.

Introduction

Machine vision is one of the most challenging problems in artificial intelligence. Task-general image understanding is so difficult that it constitutes an “AI complete” problem [1], that is, a problem of sufficient difficulty and generality that it requires intelligence on a par with humans. If solved, it would be considered equivalent to the first successful completion of a Turing test [2,3]. While the general problem of image understanding is still far outside the capabilities of modern artificial systems, algorithms are beginning to reach near human capabilities on certain specialized tasks. In particular, deep convolutional neural network (DCNN) algorithms are achieving previously unseen performance on object recognition tasks.

Since their first entrance [4] in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC), deep convolutional networks have substantially outperformed other state of the art recognition algorithms (e.g., [5]), to the point of the practical extinction of the latter. Modern ILSVRC competitions (including 1.2 million images associated with 1000 categories) almost exclusively feature deep convolutional networks, and their error rates have continuously fallen with more powerful hardware and more sophisticated engineering. The current winner has a top-five error rate of less than 3% on the image classification task, meaning that it fails to include the correct category out of 1000 object categories in its top five most likely choices less than 3% of the time, which is even lower than human performance on the same task (~5.1%).

The impressive performance of DCNNs on natural image recognition tasks and certain apparent similarities between human physiology and the architecture of these networks suggest the natural question of whether these systems explain the capabilities of human perception and acquire similar representations to those used by the human visual system. As DCNNs approach human performance in object recognition tasks, we may ask whether or in what ways their architecture and processing mirror that of human vision. In this paper, we take up this question with special focus on object shape. In a series of five experiments, we probe the capabilities of DCNNs and humans to cope with object classifications with the goal of finding out whether trained networks overtly or implicitly encode object shape and use it to perform classifications.

To anticipate our results: Deep learning networks lack shape representations and processing capabilities that form the primary bases of human object classification. Deep learning networks do have access to some relations of local orientations that may be considered local shape constituents, but they do not appear to form global shape representations. We also show that deep learning networks make no special use of the bounding contours of objects, which most reliably define shape in human and biological vision.

Background

DCNNs have attracted considerable attention, and several different approaches have been used to compare their performance to human object processing. Some of the similarities begin with the basic architecture. Deep convolutional neural networks perform a series of nonlinear transformations on input data such as an image in the case of object recognition. The final transformation outputs a vector of category probability values, one for each object category.

Critically, early layers of these networks are not fully connected as in classical neural networks. Instead, they have convolutional windows that preserve spatial information in the image [6]. In modern DCNNs, early layers tend to operate on very local regions of the image, while deeper into the network, each node receives input from filters over a larger area of the image, allowing the network to access relations between more distant regions [4]. This network architecture has some obvious similarities with biological vision. Convolutional layers are analogous to receptive fields in visual cortex, which likewise consider more disparate regions together at higher levels of extrastriate cortex [7].

Do the similarities between DCNNs and biological vision go deeper than this basic architectural feature? One way to evaluate this is by comparing physiological activity of neural units in biological systems with the activity of certain nodes in an artificial network. Pospisil, Pasupathy, and Bair presented AlexNet, a groundbreaking DCNN, with shape stimuli to which V4 cells are optimally tuned [8] and found that there is some resemblance between node responses in intermediate layers of AlexNet and cell responses in V4, although the network response was quite sparse compared to biological systems, with many units responding to none or very few of the shape stimuli [9]. Other studies have looked at the correlation between a network's classification accuracy and the similarity between network representations and representations in the inferotemporal gyrus (IT). Randomly varying parameters across several networks, they found that the activity of nodes in networks that perform better on the object classification task give better predictions about the activity of clusters of neurons in IT when primates are presented with the same image [10,11].

Comparisons have also been made between a network's performance and human behavior in similar tasks. In a sense, all performance measures on image classification are a comparison to human vision, as accuracy is being measured based on labels assigned by humans [12]. However, to evaluate similarities and differences between DCNNs and human vision, it can be instructive to examine network performance on tasks for which they were not explicitly trained. Several experiments have found similarities between convolutional networks and humans in such tasks. One study used features from a convolutional network to predict the memorability of certain object segments. Features extracted from the DCNN were predictive of objects' memorability for human subjects, suggesting that humans and networks might be attending to similar features when viewing an object [13]. In another study, Peterson, Abbott, and Griffiths found a strong correlation between similarity judgments made by DCNNs with human similarity ratings [14].

Although it is interesting to observe similar performance level for object recognition between DCNNs and biological vision, it is unclear whether the systems process information for object recognition in a similar manner. The present paper focuses on the perception of shapes, the most important cue for object recognition [15]. Objects can be recognized accurately despite impoverishments across every other visual dimension provided that global shape information is preserved [16, 17]. For example, consider the image pair presented in Fig 1. In Fig 1A, the information available for recognition has been significantly reduced across several feature dimensions. The object has no texture or background context, and the information along its contour has been simplified. Still, it is far more easily recognized as a bear than the object in Fig 1B, where cues like texture and context are preserved, but object shape is interrupted. Similarly, an ordinary line drawing, or even a few well-chosen lines as in a Picasso sketch, readily allows object recognition via shape perception processes in humans.

If deep networks are to be taken as models of human perception, we would expect object shape to be a critical component of their classification decisions. Currently, it is unclear to what extent shape representations play a role in object recognition in DCNNs. Kubilius, Bracci, and Op de Beeck conducted several intriguing experiments that suggest deep networks

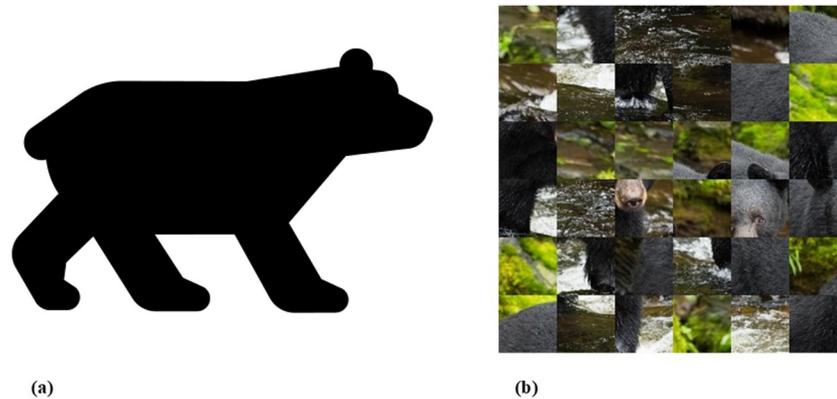


Fig 1. Demonstration of the importance of global shape in object recognition. (a) Silhouette of a bear; (b) Scrambled natural image of a bear (See text). Image URLs are in [S2 File](#).

<https://doi.org/10.1371/journal.pcbi.1006613.g001>

do have shape representations that are reasonably similar to human shape representations [18]. Networks were able to classify object silhouettes with ~40% accuracy and had some sensitivity to non-accidental features of an object, which are thought to be important for recognition in human observers [16]. They also compared the impact of shape cues on recognition performance for the networks with different architectures (e.g., different number of layers) and found some evidence that deeper networks did better on tasks where object shape was important to performance.

On the other hand, some research on DCNNs is difficult to reconcile with the claim that they utilize global shape of objects in detecting and recognizing objects. Szegedy, Zaremba, Sutskever, Bruna, Erhan, Goodfellow, and Fergus found that perturbation of a small subset of image pixels could result in consistent misclassification of the image, across multiple DCNNs, despite the changes being undetectable to human observers [19]. In the perturbed images, global shape of the object is unchanged, so the change in classification revealed that the global shape information of objects is likely not used for recognizing objects in the DCNNs. Another study used evolutionary algorithms to develop images that networks classified as certain objects with a high degree of confidence, despite a total absence of object shape in the images [20]. Zhu, Xie, and Yuille tested DCNN classification accuracy for images in which the object to be classified was removed [21]. They found that despite the object not actually being present in the image, networks performed reliably better than chance on the classification task, based purely on contextual information. These examples suggest that shape might be neither sufficient nor necessary for recognition in DCNNs.

More systematic tests are needed to understand whether or not, or in what ways, DCNNs process object shape. The fact that a network's responses are sensitive to variables other than shape (such as texture or color) may reflect a valid use of information in its training history. The supervised learning method with which DCNNs are trained is agnostic about what information to consider when classifying an image. In natural images, texture and shape information are often highly correlated, so we can learn little about what cue is most relevant to the network's classification without disentangling them. Reduction of performance by disruption of other variables may indicate that shape representations do not predominate, but such outcomes do not necessarily imply that shape information is not captured or potentially usable within the network. We term this "the latency problem". To show that shape information is not implicitly captured or usable in a system, more systematic tests are required than simply showing that variables other than shape can be decisive in classification performance. Shape

information may nevertheless be something DCNNs can in principle capture, but it may be latent in the system, overshadowed by other informational variables relevant for classification. Conversely, tests that have suggested that DCNNs do use shape information have not distinguished local shape features from more global shape characteristics, nor have they disentangled responses to local orientation relations in visible contours from the more shape-defining contributions of the bounding contours of objects.

Questions about shape processing in DCNNs can be considered at three different levels: First, do deep networks trained to recognize objects use shape features in their classification decision? Second, can deep networks be trained to use shape information in their classification decisions? Third, can formal analysis of deep networks' computational processes tell us what kinds of shape features they can and cannot extract from an image? In the current study, we focus on the first question, aiming to understand the capabilities that deep networks automatically learn through training on natural images. We consider the study of trained networks to be particularly important for two reasons. First, the attention garnered by DCNNs relates to the success of trained networks in object classification tasks. Both for theoretical understanding of these achievements and for practical applications, understanding how DCNNs achieve their high classification is important. Second, comparisons between biological vision systems and DCNNs require understanding how trained networks are operating, and require an understanding of the role of shape processing in particular. A sizable literature has recently emerged finding similarity between DCNNs trained on object recognition tasks and human perception [9–11, 13, 14, 22–26]. These efforts highlight the importance of understanding the functioning of trained DCNNs that are successful in object recognition.

In the experiments reported here, we presented DCNNs trained for object recognition (and, where appropriate, human observers) with stimuli intended to reveal information about the usability of global shape and bounding contours. The series of experiments was designed to provide multiple sources of information with regard to the latency problem in characterizing shape capabilities, in the process clarifying the roles of texture information and object shape in deep convolutional networks' classification decisions. To carry out these studies, we tested two commonly used deep convolutional networks: AlexNet [4] and VGG-19 [27]. AlexNet has eight layers and is the deep network that started the DCNN revolution for object recognition. VGG-19 is deeper, 19 layers, and approaches the state of the art in object classification. Our approach was to use a variety of systematically modified stimuli to reveal the contribution of shape information to network responses. In Experiment 1, we examined the relative importance of overall shape and texture information by using objects in which the overall shape was preserved, but different texture, from another object, was superimposed on the object's silhouette. In Experiments 2–4, we tested the networks on images with impoverished or altered texture and context information by using glass figurines, object outlines, and silhouettes. Following up the results and hypotheses emerging from these experiments, we tested the networks on images with manipulations that altered shape at a global level, while largely preserving local shape features, and vice versa (Experiment 5).

Results

Experiment 1

As noted earlier, shape is of great importance in human perception of objects, and shape information predominates in human object recognition. Prior tests of DCNNs have yielded some evidence that they classify by means of shape, whereas other work revealed examples in which images with textural similarity, but no shape in common with an object, were classified as that object with a high degree of confidence [20]. In Experiment 1, we directly compared



Fig 2. Sample stimuli used in Experiment 1. The bounding shape of an object was combined with the texture of a different object to generate each image. **a)** Shape: Teapot | Texture: Golf ball; **b)** Shape: Vase | Texture: Gong; **c)** Shape: Airplane | Texture: Otter; **d)** Shape: Obelisk | Texture: Lobster; **e)** Shape: Cannon | Texture: Pineapple; **f)** Shape: Ram | Texture: Bison; **g)** Shape: Camel | Texture: Zebra; **h)** Shape: Orca | Texture: Kimono; **i)** Shape: Otter | Texture: Sock; **j)** Shape: Elephant | Texture: Sock. The full image set is displayed in Figs 3–6.

<https://doi.org/10.1371/journal.pcbi.1006613.g002>

convolutional networks’ use of shape and texture information in their classification decision. Using object silhouettes with no surface information, we overlaid a texture from a different object on top of the black figural region. We then compared the networks’ preference for both the object whose shape is displayed and the object whose texture is displayed.

Experiment 1 Method. *Test stimuli.* Forty images of object silhouettes and 40 natural images were obtained from internet sources (for URLs of original images, see [S2 File](#)). Half of the object silhouettes were animals, and half manmade artifacts. In the ImageNet database, about 40% of categories are animals, and 50% are manmade objects (the last 10% are other inanimate objects like food). For each silhouette, a texture from one of the natural images was overlaid on the object. All shapes and textures were taken from the 1000 object categories on which the network was trained. See [Fig 2](#) for examples.

Network. Tests were conducted on a pre-trained VGG-19 network.

Experiment 1 results. For each of the displayed images, the network assigned a probability value to each of the 1000 object categories it had been trained to classify. The objects that received the five highest probability assignments for each image are shown, broken into four parts for size considerations, in [Figs 3–6](#), along with the probability assigned to the correct shape and texture label. Based on shape, the network chose as its highest probability classification the correct answer for 5 of the 40 objects. Based on texture, the network chose as its highest probability classification the correct answer for 4 of the 40 objects. In terms of including the correct answer in its top 5 possibilities, the network classified 8 of 40 objects within its top 5 choices by shape and 7 of 40 objects within its top 5 choices by texture. Overall, the assigned probability was lower than is typical for natural images for both the correct texture-object label and the correct shape-object label. For photographs of objects that include texture, shape and context, 90% or more of the total probability across 1000 object categories will ordinarily be assigned to the correct object label. In this simulation, there were a few shape-based classifications that were near natural image performance, such as the abacus and the trombone, but on average shape-based classifications were nearer to 10%. Likewise, for textures, a few of objects were assigned probabilities that were 20% or greater but average performance was quite poor. Human observers, for whom shape is predominant in object recognition, readily produce correct shape labels for all of these objects, as confirmed by pilot studies. By contrast, across the whole display set used here, the object whose shape was depicted in the displays was selected by the network, on average, as its 209th ranked choice.

Although there were indications of some use of shape information by the network for only about 20% of the displays tested in Experiment 1, an interesting pattern can be seen in the

Display Image	Shape-Object	Texture-Object	1 st Choice	2 nd Choice	3 rd Choice	4 th Choice	5 th Choice
	Otter (0%)	Odometer (0.03%)	Can Opener (12.12%)	Electric Guitar (7.66%)	Hook (3.64%)	Remote Control (3.54%)	Corkscrew (3.53%)
	Ram (31.8%)	Bison (0.06%)	Ibex (46.51%)	Ram (31.8%)	Bighorn (18.1%)	Chesapeake Bay Retriever (1.09%)	Hyena (0.48%)
	Sturgeon (0.02%)	Honeycomb (0.48%)	Starfish (15.36%)	Banded Gecko (5.6%)	Electric Ray (5.57%)	Snail (5.56%)	Kite (4.52%)
	Bee (0.04%)	Velvet (6.97%)	Stole (29.94%)	Wool (10.24%)	Velvet (6.97%)	Bonnet (4.53%)	Poncho (4.2%)
	Bison (0.01%)	Stone Wall (8.72%)	Stone Wall (8.72%)	Parachute (7.9%)	Tile Roof (5.19%)	Stole (3.6%)	Kite (3.26%)
	Dugong (0.03%)	Gorilla (0%)	Bluetick (8.53%)	German Short-Haired Pointer (7.73%)	Egyptian Cat (6.19%)	Tabby (5.79%)	Kerry Blue Terrier (5.6%)
	Flamingo (0.83%)	Python (0.05%)	Limpkin (13.48%)	Hook (12.46%)	Corkscrew (6.33%)	Bustard (3.8%)	Ostrich (3.63%)
	Gorilla (0%)	Green Mamba (1.48%)	Strawberry (19.03%)	Custard Apple (16.42%)	Granny Smith (6.81%)	Bell Pepper (5.47%)	Knot (4.5%)
	Hyena (0.15%)	Cask (0.19%)	Triceratops (19.8%)	Mortarboard (3.2%)	Pedestal (2.4%)	Pencil Sharpener (2.02%)	Kite (1.78%)
	Lobster (0.02%)	Broom (0.83%)	Wool (18.69%)	Rocking Chair (4.8%)	Knot (4.67%)	Hamper (4.41%)	Stole (3.36%)

Fig 3. Network classifications for the stimuli presented in Experiment 1 Part 1. The left most column shows the image presented. The second column in each row names the object from which the shape was sampled. The third column names the object from which the texture silhouette was obtained. Probabilities assigned to the object name in columns 2 and 3 are shown as percents below the object label. The remaining five columns show the probabilities (as percents) produced by the network for its top five classifications, ordered left to right in terms of probability. Correct shape classifications in the top five are shaded in blue and correct texture classifications are shaded in orange.

<https://doi.org/10.1371/journal.pcbi.1006613.g003>

Display Image	Shape-Object	Texture-Object	1 st Choice	2 nd Choice	3 rd Choice	4 th Choice	5 th Choice
	Ostrich (0.03%)	Corn (1.21%)	Scorpion (41.51%)	Banana (16.51%)	Panpipe (5.34%)	Rocking Chair (3.76%)	Ear (2.69%)
	Gazelle (1.95%)	Flamingo (24.44%)	Flamingo (24.44%)	Ibizan Hound (7.75%)	Tick (7.12%)	Parachute (5.22%)	Crayfish (4%)
	Bear (0.04%)	Wolf (1.3%)	Tabby (30.02%)	Egyptian Cat (14.69%)	Cardigan (7.83%)	Siamese Cat (4.16%)	Irish Wolfhound (2.71%)
	Camel (0.04%)	Zebra (3.92%)	Poncho (41.63%)	Academic Gown (29.25%)	Mortarboard (6.12%)	Zebra (3.92%)	Triceratops (2.92%)
	Elephant (0%)	Sock (3.59%)	Poncho (24.5%)	Wool (18.44%)	Muzzle (7.34%)	Parachute (7.19%)	Stole (6.55%)
	Fox (0.02%)	Tiger (1.2%)	Panpipe (16.04%)	Triceratops (9.35%)	Conch (8.81%)	Piggy Bank (3.22%)	Velvet (3.02%)
	Lion (0%)	Armadillo (0.47%)	Parachute (15.51%)	Triceratops (9.09%)	Muzzle (8.94%)	Bulletproof Vest (6.06%)	Kite (4.73%)
	Orca (0%)	Kimono (16.31%)	Pinwheel (17.14%)	Kimono (16.31%)	Flagpole (10.01%)	Mask (8.57%)	Bow Tie (3.55%)
	Hippopotamus (0.01%)	Basketball (0.16%)	Studio Couch (9.33%)	Band Aid (4.9%)	Cradle (4.25%)	Pencil Box (3.3%)	Mitten (2.17%)
	Porcupine (0.01%)	Steam Locomotive (3.19%)	Snowmobile (12.38%)	Tank (11.88%)	Warplane (10.49%)	Projectile (7.92%)	Amphibian (7.43%)

Fig 4. Network classifications for the stimuli presented in Experiment 1 Part 2.

<https://doi.org/10.1371/journal.pcbi.1006613.g004>

Display Image	Shape-Object	Texture-Object	1 st Choice	2 nd Choice	3 rd Choice	4 th Choice	5 th Choice
	Abacus (77.92%)	Broccoli (0%)	Abacus (77.92%)	Crib (4.92%)	Cradle (3.73%)	Pencil Box (2.79%)	Shopping Basket (1.93%)
	Aircraft Carrier (0.06%)	Ostrich (0.01%)	Ski (28.83%)	Warplane (5.37%)	Bassoon (5.11%)	Rifle (4.62%)	Letter Opener (3.25%)
	Airplane (0%)	Otter (0%)	Hatchet (26.16%)	Nail (14.08%)	Hook (9.24%)	Kite (5.48%)	Scabbard (4.84%)
	Airship (0.06%)	Guacamole (0.06%)	Lacewing (33.66%)	Cucumber (8.89%)	Leafhopper (6.81%)	Vine Snake (4.79%)	Nematode (2.95%)
	Apron (2.34%)	Keyboard (0.13%)	Electric Guitar (41.39%)	Muzzle (16.06%)	Bulletproof Vest (11.38%)	Holster (4.33%)	Mailbag (2.46%)
	Assault Rifle (41.01%)	Dugong (0%)	Assault Rifle (41.01%)	Rifle (20.17%)	Stretcher (14.5%)	Warplane (8.69%)	Plow (2.82%)
	Ballpoint Pen (1.64%)	Gazelle (0%)	Letter Opener (20.62%)	Spatula (12.09%)	Paintbrush (7.88%)	Scabbard (6.74%)	Hammer (5.66%)
	Banjo (1.41%)	Ram (0%)	Nail (22.43%)	Ladle (22.21%)	Spatula (10.48%)	Maraca (6.47%)	Hair Slide (5.24%)
	Binoculars (1.64%)	Soccer Ball (0.51%)	Pill Bottle (13.7%)	Dumbbell (7.56%)	Perfume (5.7%)	Whistle (5.46%)	Screw (4.46%)
	Cannon (34.48%)	Bee (0%)	Cannon (34.48%)	Tricycle (11.65%)	Plow (11.24%)	Hair Slide (7.67%)	Thresher (5.46%)
	French Horn (70.84%)	Boxer (0%)	French Horn (70.84%)	Muzzle (12.71%)	Hair Slide (8.96%)	Cornet (1.29%)	Buckle (0.78%)
	Shovel (6.59%)	Koala (0%)	Spatula (28.73%)	Broom (19.8%)	Ladle (11.35%)	Shovel (6.59%)	Letter Opener (5.82%)

Fig 5. Network classifications for the stimuli presented in Experiment 1 Part 3.

<https://doi.org/10.1371/journal.pcbi.1006613.g005>

Display Image	Shape-Object	Texture-Object	1 st Choice	2 nd Choice	3 rd Choice	4 th Choice	5 th Choice
	Teapot (0.41%)	Golf Ball (38.75%)	Golf Ball (38.75%)	Pill Bottle (5.37%)	Honeycomb (5.21%)	Ping-Pong Ball (4.71%)	Piggy Bank (3.77%)
	Torch (0.01%)	Cuirass (18.39%)	Bulletproof Vest (19.91%)	Cuirass (18.39%)	Military Uniform (11.35%)	Breastplate (5.76%)	Pickelhaube (3.43%)
	Trombone (92.64%)	Pineapple (0%)	Trombone (92.64%)	Nail (3.53%)	Cornet (1.42%)	Hook (0.4%)	Safety Pin (0.39%)
	Vase (1.79%)	Gong (22.11%)	Gong (22.11%)	Electric Fan (13.15%)	Hook (8.19%)	Shield (6.37%)	Pitcher (6.3%)
	Moped (0%)	Hyena (0.24%)	Whippet (17.99%)	Great Dane (11.96%)	Sealyham Terrier (6.86%)	Bighorn (3.86%)	Weimaraner (3.82%)
	Obelisk (24.16%)	Lobster (0%)	Candle (42.31%)	Ice Lolly (30.35%)	Obelisk (24.16%)	Matchstick (0.7%)	Plunger (0.3%)
	Guitar (0.38%)	Python (0.04%)	Corkscrew (10.13%)	Triceratops (8.31%)	Hammer (5.91%)	Can Opener (5.91%)	Hatchet (5.78%)
	Hammer (0.59%)	Ice Cream (0.03%)	Parachute (16.27%)	Kite (6.43%)	Muzzle (5.25%)	Scorpion (4.14%)	Tick (3.05%)

Fig 6. Network classifications for the stimuli presented in Experiment 1 Part 4.

<https://doi.org/10.1371/journal.pcbi.1006613.g006>

data. Use of shape information appeared to play some role in DCNN classification of artifacts but almost none for animals. The network had the object shape in its top-five classification selections for seven of the 20 artifacts, and only one of the 20 animals. The average probability assigned to the image's shape label was 10 times higher for artifacts than for animals (17.90% vs. 1.75%). Texture appears to be about equally considered for both kinds of stimuli. There are three classifications in the top 5 choices for the texture-object in the 20 artifact images, and four in the 20 animal images; the mean probability assigned to the texture object is about equal for both kinds of images (3.22% vs. 3.73%).

The data from Experiment 1 were also analyzed by directly comparing the probability value associated with the object whose shape was described by the silhouette and the object whose texture was overlaid atop the shape. Overall, texture was preferred more often than shape (23 vs. 17). However, there was a large difference between network behavior in manmade objects versus animals. The network assigned higher probability to the shape-label in 14 of the 20 manmade objects, but only three of the 20 animal images (see Figs 7 and 8).

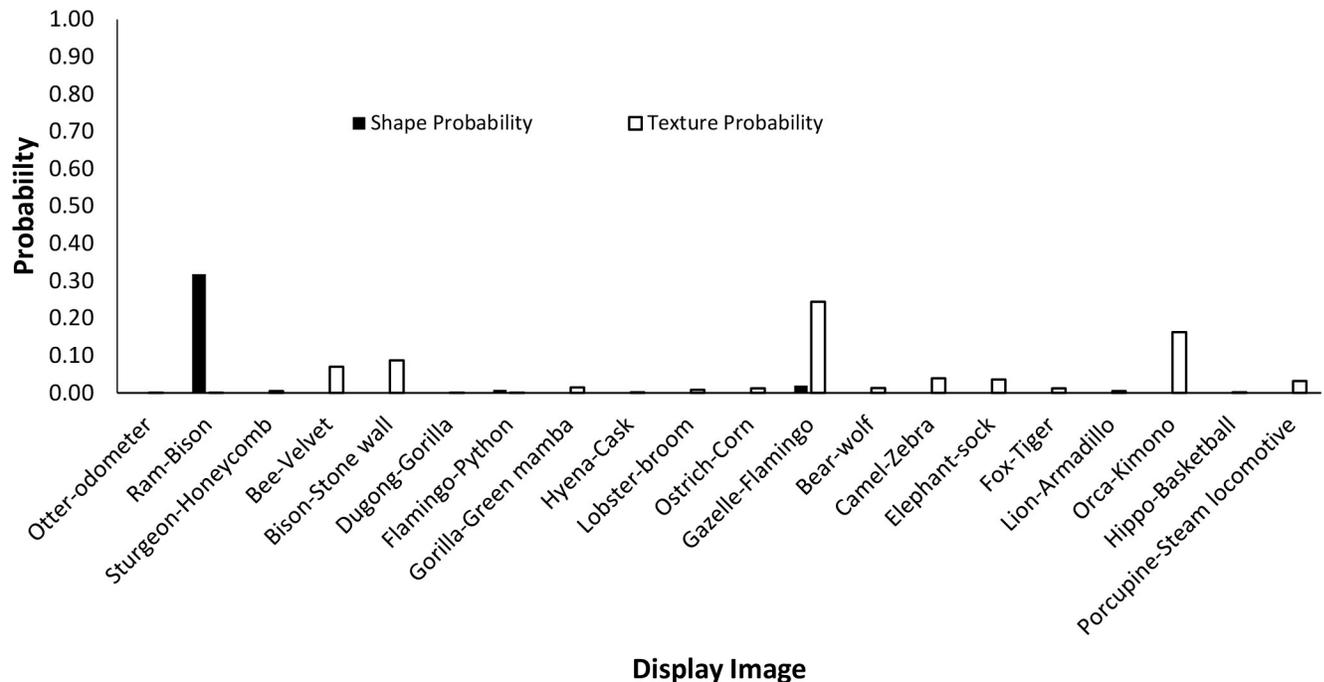


Fig 7. Comparison of probabilities assigned to image shapes and textures for animals. On the x-axis, the shape and texture of each object are given as shape-texture. Filled black bars display the probability given by the network to the correct shape. Outlined bars display the probability given by the network for the correct texture.

<https://doi.org/10.1371/journal.pcbi.1006613.g007>

Finally, we measured the contribution of shape and texture to network classification by looking at the rank order of the correct shape-object and the correct texture-object for each display used. For the correct shape response, the mean rank among network outputs was 86.70 for artifacts, and 330.50 for animals. For the correct texture response, the mean rank was 249.95 for artifacts and 65.30 for animals.

Experiment 1 Discussion. Experiment 1 suggested a major difference between human observers and DCNNs. Whereas human observers readily classify objects by shape, even in the face of uncharacteristic texture or context information, VGG-19 showed no evidence that shape information plays a primary role in DCNN classification. The correct label based on shape was chosen as the first choice classification by the network for only 5 of the 40 objects sampled, and the correct shape turned up on average as the 209th ranked choice among network outputs in object classification.

Despite the lack of a clear, general accessibility of object shape in DCNN classifications, there was some evidence suggesting use of shape information in some cases. These cases were almost entirely confined to the 20 artifacts tested, in which 5 of 20 objects selected as first-choice classifications matched on shape, with 7 out of 20 objects placing the correct shape in the top 5 choices. In contrast, no first-choice classifications were correct for animal shapes, and only one of 20 animal displays showed any shape match among the top five classifications.

Although the network appears to utilize shape for classifying artifacts but not animals, there are still inconsistent examples in the test. Some objects, even from the artifact stimuli, do not show any evidence that shape is involved in network classification. For example, the airplane with otter texture was assigned essentially zero (.000002) probability to the airplane label (where .001 would be the value obtained by randomly guessing), instead classifying the image as “hatchet”, “nail”, or “hook” as its top three choices, none of which shares any shape

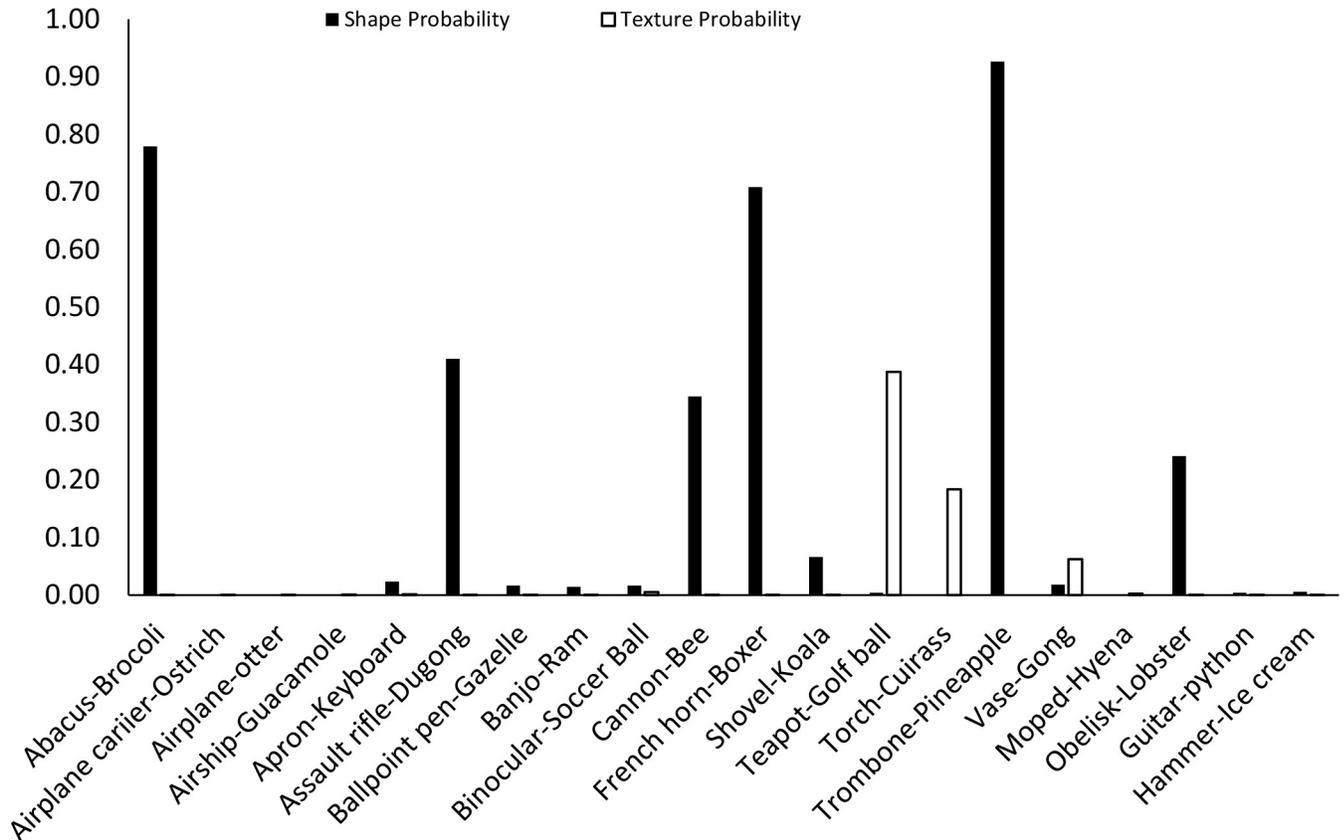


Fig 8. Comparison of probabilities assigned to image shapes and textures for artifacts. On the x-axis, the shape and texture of each object are given as shape-texture. Filled black bars display the probability given by the network to the correct shape. Outlined bars display the probability given by the network for the correct texture.

<https://doi.org/10.1371/journal.pcbi.1006613.g008>

similarity with an airplane. While this is a particularly glaring failure on the network’s part, over half of the artifacts were misclassified in all of the network’s top-five selections and the average rank order of the shape label of artifacts was 86.70. Many implausible shape misclassifications are assigned higher probability than the correct shape label. We mention this not to diminish the network’s success for some shapes on a task for which it was not explicitly trained, but to indicate that the results of Exp. 1 pose some differences in misclassification errors between the network and the human visual system. Humans have considerably less difficulty recognizing any of the objects by shape and would never consider some of the objects to which the network assigns high probability (e.g., “parachute” for hammer, or “electric guitar” for apron) as likely candidates.

Understanding why the network makes these kinds of misclassifications could be an important step to reveal the differences between network and human classification capabilities. It is possible that these erroneous responses are simply some intermediate landing point between shape and texture evidence, but classifications like “hatchet” for the airplane-otter suggest little consideration for the object’s texture in the network’s perceptual decision. It is also important to note that we tested objects separated from backgrounds and contexts. Typical tests of DCNNs include contextual information, which has been shown to be so important that networks perform reliably better than chance in classifying objects with images in which the object to be classified has been removed [21].

What could give rise to the differences we observed between artifacts and animals in networks that are not explicitly trained to recognize differences between superordinate categories like animacy? One possibility is that the network might learn to down-weight the contribution of texture cues in recognition of many artifact categories during training, since there is a more diverse range of textures and color associated with artifacts such as guitar and hammer. A sofa, for example, can be upholstered with any number of patterns, while a leopard's fur will tend to be more consistent across exemplars. Another possibility is that DCNNs attribute less importance to shape cues in natural objects due to the large variability in the bounding contours of some natural objects. Animals are non-rigid, and their bounding contours vary considerably from image to image depending on pose, so it might be maladaptive to learn shape features for natural objects during training. Yet another possibility is that DCNNs really do not encode global shape but can nevertheless make use of some local shape features, which tend to be highly diagnostic for some artifacts, but provide little discriminability between different kinds of animals. Later experiments shed light on these possibilities. After considering additional results, we return to these issues in Experiment 4 and the General Discussion.

Experiment 2

Experiment 1 showed that in displays that preserved overall object shape but altered their texture, shape was a poor predictor of network classifications. There was some indication, however, of sensitivity to shape information in certain cases. The network made several accurate classifications of objects with non-canonical surface texture. This success appeared to be largely confined to artifacts, although even artifact classifications included many implausible top selections. On the other hand, shape information appeared to be largely irrelevant for classification responses generated for animal displays. In Experiments 2–4, we developed more detailed tests to examine whether networks could classify objects only based on shape with changed or absent surface texture and context information.

It is a remarkable fact, one attesting to the primacy of shape processing in human perception, that human observers readily recognize shapes in arbitrary materials (and construct and display them, etc.). In Experiment 2, we presented two deep networks with glass figurines. All figurines were pictures of real glass objects. Since glass figurines lack the natural surface colors and textures of the objects represented, we expected that accurate classification would be difficult without a representation of the object's bounding shape. We expected that if the networks did have access to object shape, they might be able to accurately classify the glass figurines even in the absence of other, usually accompanying, cues for recognition. In other words, DCNN classifications that would resemble even a child's intuitive response the first time they see a glass elephant would furnish evidence that shape information plays a role in DCNN classification.



Fig 9. Sample stimuli used in Experiment 2.

<https://doi.org/10.1371/journal.pcbi.1006613.g009>

Experiment 2 Method. *Images.* Twenty images of glass figurines were found from the internet (see [S2 File](#) for URLs). Half of the figurines were of animals and half were of manmade objects. The images had some texture information, but the overall texture of each object was very different from a canonical instance of the represented object. The background in all but one of the 20 images was non-descriptive; either a homogeneous field or a color gradient. The one exception is the schooner figurine (see below), which is photographed on a table. See [Fig 9](#) for examples.

Network. classification was tested on two networks: AlexNet, with seven layers, and VGG-19.

Experiment 2 results. We assessed the networks as correct if they generated the names that a human observer would give to each image. (Human classification was verified in pilot work with human observers.) Neither network produced as its top choice the correct label for any of the 20 objects. Figs 10–11 show the top five classification choices for the 20 images shown for VGG-19 (the better-performing network; see below). Percentages in parentheses represent the probability assigned to each label by the network. In the absence of any evidence, the baseline probability of an object was 0.1%, as there were 1000 object categories.

Most of the top-choice responses seem bizarre for human perceivers, such as "web site" for goose, "oxygen mask" for otter, "can opener" for polar bear, and "chain" for fox. Although the stimuli here are (intentionally) different from what the networks were trained on (because they are glass figurines), the results clearly indicate that shape, if accessible at all by DCNNs, does not play the defining role in object recognition that it does for human perceivers. Comparing the networks' responses to chance level performance (0.1%), AlexNet assigned a probability below chance to the correct shape for 18 of the 20 test objects, and VGG-19 assigned a probability below chance to the correct shape for 15 of the 20 objects. Analysis of the rank order of correct labels revealed that the correct shape choice averaged, across the display set, a mean rank of 162.60.

The criterion of finding the correct answer among the top five most probable responses is often used to assess performance of DCNNs. Using this criterion, VGG-19 correctly classified two of the 20 images (the shark figurine and the grand piano figurine), while AlexNet misclassified all 20 images. VGG-19 assigned a 20.58% probability to the correct grand piano response, second only to its 53.07% probability assigned to "radio telescope." For the shark, VGG-19 selected the correct label as its 4th choice, assigning a 2.58% probability. For many of the remaining images, the objects were misidentified as glass-made or metal-made kitchen objects, such as "water jug", or "can-opener".

As another way of measuring the network's sensitivity to object shape, we compared the probability the network gave to the object in the image with the probability it gave to the nine other objects that were used in the experiment. For example, we compared the probability that the network gave to "goose" when it was shown a goose figurine to the average of the probabilities the network gave to the other nine animal labels ("otter", "peacock", and so on) for the same figurine. For this analysis, we kept glass animals and glass objects separate to ensure that higher probabilities could not be accounted for by low-level contour features like the presence or absence of a straight edge. For both animate and inanimate objects, probabilities were not higher for the correct shape-label than for the average of the other nine incorrect shape labels more often than would be expected by chance. Five of the 10 inanimate objects, and seven of the 10 animate objects were given a lower probability than the average of the other nine in their class.

Experiment 2 Discussion. The networks showed little capability of classifying glass figurines. Although glass figurines of animals remove the natural surface texture, they also introduce surface properties of their own. As in Experiment 1, surface features appear to play a

Display Image	Correct Label	1st Choice	2nd Choice	3rd Choice	4th Choice	5th Choice
	Goose (0.006%)	Web site (10.13%)	Corkscrew (6.03%)	Hook (4.23%)	Washbasin (4%)	Bathtub (3.15%)
	Otter (0.0008%)	Oxygen mask (23.6%)	Nipple (13.16%)	Piggy bank (7.49%)	Water jug (6.44%)	Pitcher (5.41%)
	Peacock (0.0000005%)	Hourglass (71.62%)	Can opener (4.51%)	Ladle (3.98%)	Oxygen mask (3.76%)	Corkscrew (2.56%)
	Elephant (0.0001%)	Egyptian cat (15.6%)	Can opener (13.76%)	Goblet (7.69%)	Mexican hairless (7.61%)	Corkscrew (7.14%)
	Polar bear (0.001%)	Can opener (13.37%)	Corkscrew (12.61%)	Hammerhead (9.08%)	Ladle (5.03%)	Breastplate (3.92%)
	Rabbit (0.01%)	Goblet (10.28%)	Piggy bank (8.04%)	Quill (7.99%)	Cup (6.34%)	Perfume (6.16%)
	Fox (0.000003%)	Chain (22.47%)	Whistle (21.69%)	Hair slide (15.05%)	Iron (5.97%)	Safety pin (5.81%)
	Robin (0.00003%)	Shower cap (25.88%)	Pinwheel (20.25%)	Bonnet (9.02%)	Hair slide (8.86%)	Piggy bank (6.14%)
	Great white shark (2.58%)	Warplane (52.57%)	Hammerhead (29.79%)	Airliner (4.02%)	Great white shark (2.58%)	Wing (2.26%)
	Tiger (1.05%)	Egyptian cat (12.8%)	Breastplate (6.72%)	Triceratops (5.43%)	Letter opener (5.08%)	Scabbard (3.33%)

Fig 10. VGG-19 classifications for glass figurines Part 1. The leftmost column shows the image presented to the VGG-19 DCNN. The second column shows the correct object label and the probability generated by the network for that label. The other five columns show probabilities for the network's top five classifications, ordered left to right from highest to lowest. Correct classifications are shaded in blue.

<https://doi.org/10.1371/journal.pcbi.1006613.g010>

major role in the network's classification decisions. For example, the peacock figurine has "water jug" and "pitcher" in its top-five objects, despite having no shape similarity to either. "Goblet", "vase", "cup", and "hourglass" are also common misclassifications made by both AlexNet and VGG-19. By contrast, the networks appear to make few misclassifications based on similarity between the shape of two objects. Aside from the piano and great white shark, only the tiger had a misclassification ("Egyptian cat") that would be consistent with use of some sort of information about object shape. It appears that shape does not play an independent, predominant role in recognition, as it does in humans [28–30].

Do the network's few successes point to a broader trend that the network is utilizing both shape and texture in its classification decisions? It is difficult to determine what is different about these three images than the other 20 images that the network fails to classify. One thing to keep in mind is that there is a chance that a plausible shape classification appears in the network's top-five selections without the network having any sensitivity to the shape of the image. The tiger seems like a likely candidate for this possibility. While Egyptian cats and tigers have some shape features in common, Egyptian cats and elephants have very few, but the network names "Egyptian cat" as its top selection for both the elephant and the tiger glass figurines. In fact, it assigns higher probability that the elephant is an Egyptian cat than that the tiger is an Egyptian cat. It is possible that some surface feature, or conjunction of surface features and local edge properties, is driving classification in both cases.

This explanation is less satisfactory for the grand piano and great white shark, whose shape label is actually in the top-five selections and does not appear in any of the other images' top selections. In the case of the piano, one possibility we considered is that the texture of the keys drove classification, but a further test showed that the network performs well even after the keys have been occluded or blurred out. We tested the network on five additional glass grand piano images (see Fig 12), and it was unable to correctly classify any of them in its top-five selections. It is unclear what information is present in the image where the network does well that is absent in the other five, but it is likely a local shape feature, as global shape is very similar across the six images. Likewise, it is likely that local contour features, not global shape, are driving the network's accurate classification of the great white shark. We discuss this hypothesis in greater detail and revisit these positive examples in Experiment 4, after we have considered more data regarding the network's sensitivity to local and global shape information.

The results of Experiment 2 clearly showed an absence of shape sensitivity for glass figurines. Human classification of such objects is affirmed by the fact that we make, display, and recognize such objects routinely. Although in natural scenes, an elephant is never made of glass, is never 4" high, and only rarely appears on anyone's desk or coffee table, human use of object shape makes recognition of a glass elephant on a desk effortless and routine. Not only is this predominance of shape not seen in DCNN performance, there is little to suggest that shape, independent of other information, is accessible at all in classification of these objects.

In this experiment, texture or surface quality information provided a stronger influence than object shape on object recognition by both AlexNet and VGG-19. We might suspect that the strength of surface texture cues pulled the networks' top-five classifications towards texture-object labels. If object shape played any role at all, however, we would expect that the correct object label would be assigned a probability that is at least greater than chance. In most

Display Image	Correct Label	1st Choice	2nd Choice	3rd Choice	4th Choice	5th Choice
	Bicycle (0.08%)	Hair Slide (32.05%)	Necklace (13.71%)	Sandal (4.69%)	Perfume (4.35%)	Sax (3.22%)
	Cannon (0.1%)	Whistle (56.76%)	Hair Slide (11.68%)	Corkscrew (5.06%)	Chain (2.73%)	Letter Opener (2.27%)
	Church (0.16%)	Thimble (16.06%)	Perfume (14.71%)	Saltshaker (8.96%)	Packet (7.65%)	Lighter (6.74%)
	Castle (0.01%)	Perfume (74.56%)	Vase (4.78%)	Nipple (3.39%)	Goblet (2.13%)	Hair Slide (1.77%)
	Grand Piano (20.68%)	Radio Telescope (53.07%)	Grand Piano (20.68%)	Space Shuttle (10.39%)	Bathtub (1.7%)	Drum (1.69%)
	Pineapple (1.02%)	Vase (14.46%)	Golf Ball (13.32%)	Pitcher (8.42%)	Perfume (6.27%)	Water Jug (5.76%)
	Schooner (0%)	Can Opener (48.97%)	Corkscrew (19.48%)	Binder (3.08%)	Table Lamp (1.92%)	Letter Opener (1.88%)
	Totem Pole (0%)	Safety Pin (25.66%)	Necklace (8.99%)	Hook (6.63%)	Hair Slide (5.73%)	Jersey (4.46%)
	Umbrella (0.02%)	Revolver (7.75%)	Can Opener (6.41%)	Corkscrew (5.43%)	Space Shuttle (4.41%)	Hook (4.13%)
	Violin (0%)	Cup (46.02%)	Teapot (8.21%)	Tray (6.64%)	Coffeepot (5.98%)	Pitcher (4.3%)

Fig 11. VGG-19 classifications for glass figurines Part 2.

<https://doi.org/10.1371/journal.pcbi.1006613.g011>



Fig 12. Five additional glass Pianos. VGG-19 incorrectly classified each of these five images despite correctly classifying the glass piano shown in Fig 11.

<https://doi.org/10.1371/journal.pcbi.1006613.g012>

cases, the correct shape label was assigned a value less than chance, and objects of similar composition but different shapes tended to receive probabilities as high as the correct shape.

Experiment 3

A remarkable fact about human object perception is that we readily extract shape from outline drawings. This ability clearly depends on shape, as outlines omit surface information completely. Object outlines have the same texture within the bounding contour as outside it, and there is no variation in texture between the outlines of two different objects. We tested outlines in Experiment 3 to extend the earlier results and specifically to remove competing texture or surface information as much as possible. If, for example, the pictures of glass figurines somehow distracted deep networks from utilizing some encoded shape information due to competing surface texture, we expected that the problem would be substantially mitigated by using outlines. On the other hand, if deep networks cannot access shape from outline drawings of objects, we expected poor performance.

Experiment 2 Method. Images. Forty images of object outlines were selected from the internet, half of which were manmade artifacts, and half of which were animals. In all images, the only contrast difference was at the object contour. Images were uniformly white at all other locations in the image. There was a degree of abstraction in the object contours, as the outlines were not boundaries of real natural objects. All were readily recognized and correctly named by human observers. Fig 13 shows examples of stimuli used in Experiment 3.

Networks: Tests were conducted using both AlexNet and VGG-19.

Experiment 2 results. None of the correct shape labels for the 40 objects were chosen as the first-choice classification by either VGG-19 or AlexNet. Two of the 40 objects were named among the top-five choices for VGG-19, and only one was given as a top-five response by AlexNet. Figs 14–17 show the full set of responses. Twenty-eight of the probabilities assigned to the object label given by humans were below the chance rate of 0.1% in VGG-19, and 35 out of 40 were below 0.1% in AlexNet. On average, the correct classification by shape was the network's 328th most preferred choice. As in Experiment 2, we compared the probability assigned to the correct shape label to the shape label of 19 other animate or inanimate objects. For animals, the probability was lower for the shape-object than for the mean of 19 other animal images in eight of the 20 trials. For artifacts, it was lower in 10 of the 20 images. These results do not differ from the chance likelihood that the target shape will be higher than the mean of the other 19 in half the trials. Of the objects that were correctly classified, all were artifacts and none were animals.

For some reason, "corkscrew" was the VGG-19's first or second choice for 10 of the 40 objects, and it appeared among the top 5 choices for 15 objects. More understandable, perhaps, is the finding that "envelope" was the first or second choice for 14 of the objects, and this response appeared in the top 5 for 19 objects. These responses are most likely due to the local similarities between training photographs of a white envelope, possibly including black letters,

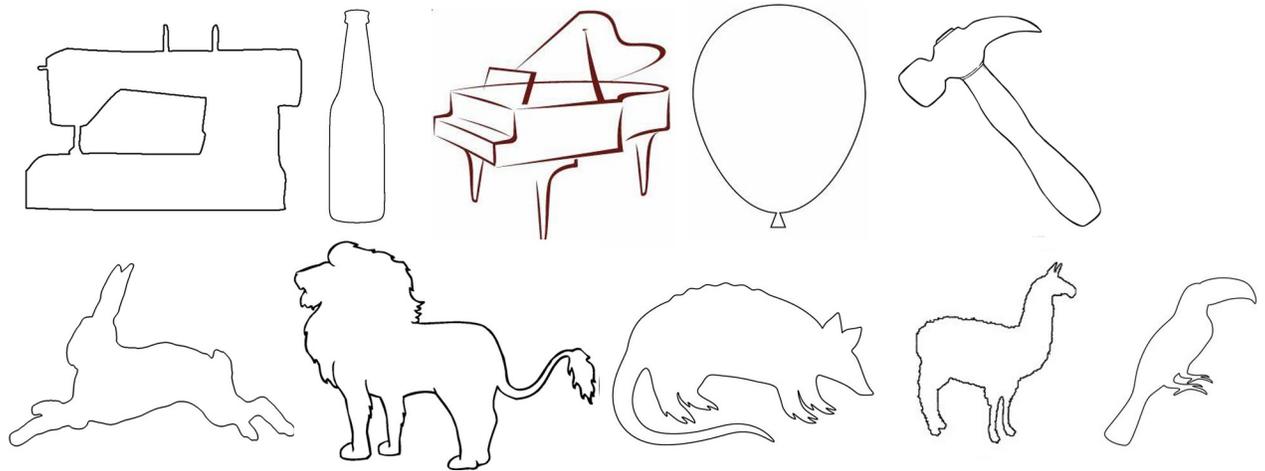


Fig 13. Sample outline stimuli used in Experiment 3.

<https://doi.org/10.1371/journal.pcbi.1006613.g013>

and the tested outline images, which are also white with thin black lines. There is no evidence that the shape of the outline images is being extracted and classified using shape features acquired from training images with similar figure boundaries.

Experiment 2 Discussion. Deep convolutional networks showed little capability to recognize objects with shapes defined by outline contours. Whereas the glass figurines had textures that were compatible with a different set of objects from their correct shape-object label, the texture in object outlines is not differentially diagnostic of any object classification, as it is identical to the surrounding background and identical throughout the object set. We hypothesized that if global shape information is extracted by the neural networks but was not used for glass figurines due to the strength of texture cues, object shape might play a larger role for outline images. The data do not support this hypothesis; classification performance was just as poor for object outlines as it was for pictures of glass objects.

The failure to classify objects correctly based on outline shape marks a clear divergence from human perception. All of the displays used in this experiment were consistently and accurately classified by human observers. We verified this with behavioral tests in which the outlines were shown together, as well as interleaved with photographs to confirm that humans have no trouble classifying object outlines, even in unexpected contexts. However, the results do not by themselves exclude some possibilities for use of shape information in DCNNs. That humans readily see shape in outline drawings is a remarkable fact in itself, and one that is not completely understood [31, 32]. It would certainly be possible to envision a shape processing system that balked at outlines and used only surface edges. As introductory drawing teachers often stress, outlines are ecologically anomalous. Within biological vision, some perceptual processes seem to treat outlines and surface edges differently, as in perceptual completion [33]. Humans can see shape given by ordinary surface edges or outlines, but if deep networks cannot utilize the latter, it does not necessarily imply that surface edges do not play some role in classification in conjunction with texture features. Humans' fluent use of outlines to see object shape indicates the strong role of shape representation in human object recognition, and human interpretation of forms in outlines probably connects naturally to some stage of perceptual representation [34]. It appears that DCNNs differ from human processors in that they have little or no linkage between shape properties embodied in outlines and the classification labels in the output layer. DCNNs' failure to use outlines does not, however, rule out the possibility that these systems may utilize some shape information in more natural cases.

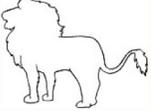
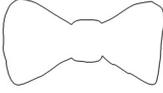
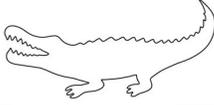
Display Image	Correct Label	1 st Choice	2 nd Choice	3 rd Choice	4 th Choice	5 th Choice
	airplane (0.05%)	walking stick (24.03%)	mantis (9.49%)	nail (4.01%)	black widow (3.3%)	harvestman (2.5%)
	armadillo (0.01%)	corkscrew (27.04%)	envelope (8.14%)	nematode (5.1%)	can opener (2.92%)	hook (2.88%)
	lion (0.01%)	corkscrew (15.73%)	jigsaw puzzle (12.33%)	envelope (6.83%)	hook (5.84%)	fire screen (3.44%)
	banana (0.03%)	hook (24.87%)	microphone (6.99%)	bow (3.55%)	table lamp (3.13%)	mouse (3.04%)
	beer bottle (0.31%)	beaker (54.21%)	red wine (4.18%)	wine bottle (3.09%)	syringe (3.08%)	pop bottle (2.96%)
	bow tie (0.11%)	harvestman (6.47%)	sunglass (6.41%)	cockroach (4.65%)	envelope (4.27%)	ant (3.24%)
	butterfly (0%)	bow (10.34%)	hook (7.14%)	microphone (6.86%)	mouse (6.01%)	cassette (5.5%)
	cello (0.91%)	hook (24.41%)	table lamp (3.8%)	microphone (3.3%)	whistle (2.81%)	syringe (2.74%)
	coffee mug (0.62%)	sunglass (12.81%)	cup (7.4%)	sunglasses (5.3%)	pick (4.66%)	safety pin (4.05%)
	crocodile (0.01%)	corkscrew (28.9%)	sea snake (7.84%)	hook (5.96%)	envelope (3.45%)	nematode (3.21%)
	dragonfly (0.1%)	hook (7.86%)	nail (5.99%)	safety pin (3.86%)	wall clock (2.66%)	tripod (2.46%)

Fig 14. VGG-19 classifications for object outlines Part 1. The leftmost column is the image presented to the DCNN. The second column from the left is the correct object label and the classification probability produced for that label. The other five columns show probabilities for the VGG-19's top five classifications, ordered left to right in terms of the probability given by the network. Correct classifications are shaded in blue.

<https://doi.org/10.1371/journal.pcbi.1006613.g014>

These issues may relate to an important factor not yet mentioned. Figure-ground assignment, or equivalently, assignment of border ownership at occluding edges, is a well-known feature of human perceptual organization [35, 36]. It appears that outlines, especially closed outlines, are interpreted in human vision as owning their borders. (The enclosed area is taken to be the bounded object.) DCNNs do not have an obvious way of representing figure vs. ground or border ownership. These seem to be more explicitly representational aspects of human perceptual processing. At least some of the problem with outlines may involve figure-ground issues. Misclassifications of objects as “hook”, “safety pin”, and “syringe”, which all have empty interior regions, suggest that DCNNs might be interpreting the actual outline as the object, rather than seeing an object as occupying the region within the outline.

On the other hand, results for a few of the images presented in Exp. 3, the beer bottle and the coffee mug, as well as a few objects like the tandem bicycle and trombone which are not correctly classified but are assigned probabilities significantly greater than chance, suggest that deep networks do not exclusively treat the black outlines as bodies of the objects themselves. If the networks were only treating the black outlines as the figure, these objects should have below-chance probability, as they do not have thin, black forms. Of course, we do not mean to imply that a DCNN employs any consistent approach or strategy; any good predictor from any of the many filters in the network, may influence the outcomes toward a correct classification. Perhaps these images have certain local features that can be extracted and further facilitate the recognition. DCNNs may not capture global shape, but may pick up some relatively local shape features, a possibility we discuss further in connection with later results.

Experiment 4

Experiments 2 and 3 found little evidence that deep convolutional networks access global shape in object recognition tasks. These results may seem surprising, as some recent reports have suggested that DCNNs do possess some shape classification abilities. Kubilius et al. [18] found that removing surface features from the Snodgrass and Vanderwart dataset of colored-in line drawings [37] did not totally destroy networks' classification performance. With the regular line-drawing images, classification performance was 80–90%. Removal of color information reduced performance to around 70%, and removal of all inner surface gradients (black silhouettes) brought performance down to about 40%. It is arguable whether 40% classification accuracy represents a success or failure in shape-based classification for DCNNs. On the one hand, this marks a divergence from human performance, which is largely unaffected by the removal of color and inner surface gradient information from most objects. On the other hand, it seems almost impossible that the network would reach even 40% accuracy without some information about object shape. In contrast, our findings about recognition for glass objects and line drawings provided almost no evidence that shape representations are used for classification in DCNNs. In Experiment 4, we tried to replicate Kubilius et al.'s findings as a first step in clarifying this apparent discrepancy.

Experiment 3 Method. Images. The same 40 object silhouettes found on the internet and used in Experiment 1 were used in Experiment 4, this time without any texture substitution. All images consisted of a single black figure on a white background. Half of the images were artifacts, and half were animals. The black figures were silhouettes of object drawings, rather than being taken from photographs of real objects with their textures removed, so some

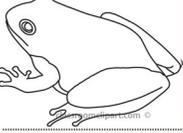
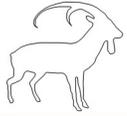
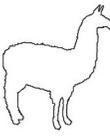
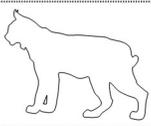
Display Image	Correct Label	1 st Choice	2 nd Choice	3 rd Choice	4 th Choice	5 th Choice
	frog (0%)	envelope (29.28%)	microphone (5.97%)	cassette (3.91%)	hook (3.05%)	cleaver (2.97%)
	hammer (0.66%)	hook (24.01%)	nematode (7.01%)	envelope (6.37%)	web site (3.29%)	mouse (2.16%)
	hare (0.02%)	jigsaw puzzle (10.88%)	corkscrew (9.11%)	nematode (5.19%)	hook (4.51%)	envelope (3.16%)
	ibex (0.05%)	envelope (16.69%)	corkscrew (3.62%)	cleaver (3.1%)	hook (2.83%)	nematode (2.12%)
	balloon (0.02%)	microphone (10.73%)	web site (6.21%)	loupe (5.01%)	necklace (4.43%)	envelope (3.02%)
	lawn mower (0.02%)	safety pin (65.9%)	corkscrew (10.89%)	can opener (5.58%)	hook (4.02%)	strainer (3.97%)
	llama (0.04%)	hook (4.38%)	corkscrew (3.2%)	envelope (2.95%)	nematode (1.86%)	rhinoceros beetle (1.7%)
	lynx (0.02%)	envelope (7.93%)	corkscrew (3.58%)	hook (2.99%)	jigsaw puzzle (2.85%)	paper towel (2.03%)
	paintbrush (0.96%)	letter opener (52.2%)	hook (5.69%)	syringe (5.22%)	safety pin (4.31%)	screwdriver (3.55%)
	peacock (0%)	envelope (14.41%)	hook (6.38%)	corkscrew (4.04%)	cleaver (3.21%)	letter opener (2.7%)
	penguin (0.04%)	web site (13.34%)	envelope (3.67%)	abaya (2.11%)	Windsor tie (1.52%)	nematode (1.21%)

Fig 15. VGG-19 classifications for object outlines Part 2.

<https://doi.org/10.1371/journal.pcbi.1006613.g015>

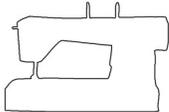
Display Image	Correct Label	1 st Choice	2 nd Choice	3 rd Choice	4 th Choice	5 th Choice
	piano (0.02%)	walking stick (9.59%)	mantis (8.01%)	hook (7.48%)	folding chair (6.01%)	nail (5.82%)
	pirate ship (0.05%)	jigsaw puzzle (5.42%)	envelope (3.81%)	laptop (3.01%)	hook (2.4%)	cup (1.93%)
	polar bear (0.12%)	envelope (22.64%)	web site (2.42%)	corkscrew (2.27%)	rubber eraser (1.7%)	hook (1.5%)
	poodle (0.01%)	hook (10.23%)	jigsaw puzzle (6.28%)	corkscrew (5%)	nematode (4.25%)	hatchet (2.82%)
	pretzel (0.06%)	envelope (41.91%)	web site (3.61%)	cleaver (2.04%)	face powder (1.6%)	binder (1.42%)
	revolver (1.67%)	hook (17.34%)	corkscrew (5.84%)	safety pin (5.48%)	can opener (3.91%)	cleaver (2.61%)
	sewing machine (0.26%)	stretcher (10.59%)	hook (8%)	studio couch (6.99%)	safety pin (4.41%)	mousetrap (3.53%)
	running shoe (0.01%)	hook (12.82%)	bow (4.89%)	microphone (2.77%)	nematode (2.59%)	bassoon (2.46%)
	snail (0%)	envelope (14.43%)	microphone (5.51%)	sunglass (3.37%)	mouse (2.64%)	cassette (2.33%)
	space shuttle (0.1%)	hook (8.86%)	envelope (7.25%)	wall clock (2.86%)	corkscrew (1.81%)	folding chair (1.59%)
	stingray (0.08%)	bow (19.45%)	hook (11.96%)	corkscrew (3.29%)	long-horned beetle (3.29%)	mouse (2.01%)

Fig 16. VGG-19 classifications for object outlines Part 3.

<https://doi.org/10.1371/journal.pcbi.1006613.g016>

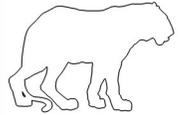
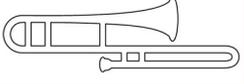
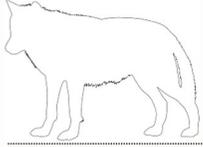
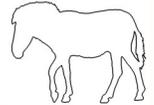
Display Image	Correct Label	1 st Choice	2 nd Choice	3 rd Choice	4 th Choice	5 th Choice
	tandem bicycle (1.03%)	hook (11.33%)	whistle (6.67%)	hair slide (4.8%)	cornet (4.17%)	safety pin (3.43%)
	teapot (0.04%)	hook (14.19%)	envelope (12.09%)	fire screen (4.42%)	sunglass (3.84%)	nematode (3.65%)
	tiger (0.01%)	envelope (19.53%)	web site (2.48%)	quill (1.63%)	cleaver (1.6%)	binder (1.58%)
	toucan (0.03%)	bow (15.72%)	corkscrew (15.5%)	hook (7.08%)	table lamp (3.25%)	nematode (3.2%)
	trombone (1.62%)	hook (26.32%)	nail (8.71%)	hatchet (3.52%)	can opener (2.91%)	microphone (2.86%)
	wolf (0.01%)	buckle (13.2%)	wallet (9.11%)	letter opener (7.36%)	hair slide (3.72%)	cleaver (3.1%)
	zebra (0.01%)	hook (10.03%)	envelope (9.23%)	corkscrew (5.86%)	nematode (3.51%)	paper towel (2.51%)

Fig 17. VGG-19 classifications for object outlines Part 4.

<https://doi.org/10.1371/journal.pcbi.1006613.g017>

contour information that would typically be in a natural instance of the object was abstracted in the silhouette images. These abstractions have no effect on the objects' recognizability to a human observer. See Fig 18 for examples. We also tested the network on the same 40 silhouettes with white figures on a black background and red figures on a white background to measure the influence of surface color on network classification performance.

Networks. As in Experiments 2 and 3, we tested AlexNet and VGG-19 on the silhouette images.

Experiment 3 results. For the 40 black silhouettes on a white background, VGG-19 and AlexNet correctly classified 20 and 15 of the 40 presented images (in their top-five classifications), respectively. Figs 19–22 shows the results for VGG-19. Performance was worse for images with white figures on black grounds, where the network classified seven of the 40 images correctly, and for images with red figures on white grounds, where the network classified nine of the 40 images correctly.

Experiment 3 Discussion. The results from Experiment 4 are largely consistent with the findings reported by Kubilius et al. In the absence of other information, deep convolutional



Fig 18. Sample stimuli used in Experiment 4.

<https://doi.org/10.1371/journal.pcbi.1006613.g018>

networks could classify object silhouettes about 50% of the time. In the better-performing VGG-19, there were 12 out of 40 correct top choices, and another eight correct choices in the network's top-five selections. Classification accuracy was once again higher for artifacts (13/20 correct) than for animals (7/20 correct).

Performance was notably worse for white-on-black and red-on-white figures than for black-on-white figures. One reason for this might be that there are more canonically black objects in the training set than white or red. For example, the cannon is correctly classified when presented as a black figure, but incorrectly classified when presented as a red figure. Instead, "fire engine" appears in the network's top-five choices, a selection obviously driven by surface properties. Another reason networks might be better at classifying black figures is that they more closely resemble photographic images that were used in network training. Objects will appear very dark or even black if they are between the camera and a bright light, as in actual silhouettes. Possibly, exposure to training examples like these makes the network more likely to accept dark figures as instances of an object, even one that is not canonically dark. The differences in network performance across these three testing sets points to the strong influence of surface information in classification. For humans, a homogenous surface texture would likely not be considered at all in recognition, as the visual system would recognize that there is not enough surface information present to be diagnostic. The network makes no such evaluation and remains highly sensitive to such cues.

Regarding shape, this experiment showed a clear contribution of contour properties in classification of object silhouettes. Within a given display set, all of the test displays shared the same coloration; therefore, all differences in classification responses from the DCNNs involved contour information. Performance at this level demands explanations that go beyond a simple conclusion that DCNNs do or do not process object shape. If DCNNs had access to global shape information, we might have expected their performance to be similar to humans, readily producing accurate classifications for all displays. Indeed, in about 50% of cases, responses seemed to be correct or close in shape to the target object. Still, the results fall far below 100%, even when scored correct if a network generated a correct classification among its top five outputs. The results also contain some rather conspicuous failures to process overall shape. For the porcupine display, for example, VGG-19 gave, as its top choice, "bald eagle". For "lion", AlexNet's top choice was "goose", with high confidence, more than twice the probability of any

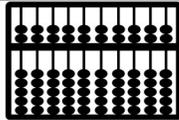
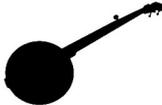
Display Image	Correct Label	1 st Choice	2 nd Choice	3 rd Choice	4 th Choice	5 th Choice
	Abacus (99.98%)	Abacus (99.98%)	Dumbbell (0.01%)	Scoreboard (0%)	Carpenters Kit (0%)	Digital Clock (0%)
	Aircraft Carrier (0.1%)	Assault Rifle (9.64%)	Rifle (6.64%)	Nail (4.69%)	Hook (4.17%)	Chain Saw (4.15%)
	Airplane (0.25%)	Plow (21.23%)	Warplane (20.04%)	Hammerhead (7.82%)	Nail (7.37%)	Wing (3.37%)
	Airship (69.75%)	Airship (69.75%)	Projectile (10.52%)	Missile (3.99%)	Warplane (3.57%)	Letter Opener (3.18%)
	Ballpoint Pen (14.36%)	Ballpoint (14.36%)	Letter Opener (14.19%)	Scabbard (8%)	Can Opener (7.21%)	Microphone (6.3%)
	Banjo (0.22%)	Ladle (13.48%)	Microphone (10.6%)	Barbell (8.17%)	Gong (5.6%)	Plunger (4.64%)
	Binoculars (9.18%)	Binoculars (9.18%)	Saltshaker (3.79%)	Megalith (3.16%)	Coffee Mug (2.48%)	Water Jug (2.22%)
	Cannon (61.38%)	Cannon (61.38%)	Thresher (8.97%)	Plow (4.48%)	Tricycle (4.48%)	Spindle (3.44%)
	Shovel (21.5%)	Spatula (35.48%)	Shovel (21.5%)	Ladle (7.58%)	Hammer (5.23%)	Plunger (3.72%)
	Teapot (52.42%)	Teapot (52.42%)	Black Grouse (16.07%)	Cup (5.95%)	Coffeepot (3.16%)	Rhinoceros Beetle (2.6%)
	Torch (0.22%)	Nail (10.84%)	Hook (10.84%)	Corkscrew (7.07%)	Albatross (6.35%)	Hammer (3.3%)
	Apron (3.75%)	Hook (10.68%)	Mailbag (8.07%)	Abaya (7.49%)	Bulletproof Vest (7.44%)	Brassiere (5.95%)

Fig 19. VGG-19 classifications for black object silhouettes Part 1. The leftmost column shows the image presented to VGG-19. The second column from the left shows the correct object label and the classification probability produced for that label. The other five columns show probabilities for the network's top five classifications, ordered left to right in terms of the probability given by the network. Correct classifications are shaded in blue.

<https://doi.org/10.1371/journal.pcbi.1006613.g019>

other response. These results contain important information regarding processing of overall shape. For humans, at least, the lion display is both recognizable as having the shape of a lion and is clearly not shaped at all like a goose. It could be pointed out that failure to give a certain label may merely indicate that the particular shape captured in silhouette may have had a particular vantage point that was uncharacteristic of examples in the training set. The implications for global shape processing here, however, hinge less on the selection of the correct name than on the incorrect answers furnished. “Bald eagle”, as well as runners up “vulture”, “ostrich”, and “buckeye” are not a close shape match for any porcupine.

These results suggest that overall shape is elusive in DCNN responses, but also that something relating to shape allows success in some of the cases tested. Perhaps a deeper analysis of what is meant by shape is needed to understand both the successes and failures of DCNNs. We consider this in successive steps below.

First, why is classification so much better for object silhouettes than for glass figures and shape outlines? We have already commented that the ability to use outlines as depicting shape, although significant in human perception, is not a necessary condition for a DCNN to be shape processor. To confirm that the difference between network performance on silhouettes and outlines was not item specific, we sampled the outline of the 40 silhouettes and tested the network on outline images of the stimuli in Experiment 4. The results closely matched those reported in Experiment 3—the network only classified three of the 40 images in its top-five selections.

What about the better performance of black silhouettes over glass objects? One reason may be that silhouettes successfully reduce distracting texture information, i.e., texture information that would tend to promote a classification other than the correct shape-based response. Glass objects may have contained more misleading surface information than black silhouettes. There are more objects with glass or other transparent or reflective surfaces in the 1000 categories on which the networks were trained than there are uniformly black objects. As mentioned, silhouettes also have the advantage of potentially looking similar to some photographs of the objects during network training if the photographs were taken at sunset or with a bright light behind the object. So, although we still see some misclassifications based on object color, such as “mortarboard” and “academic gown”, the network appears to be overall more robust to black surface textures.

Another important factor contributing to the differences in the previous three experiments is the involvement of figure-ground segmentation. Silhouettes are likely easier to classify than glass objects because the surface information provided contains *no internal contours*. Although figure-ground segmentation is an important part of human shape processing, it is likely that DCNNs produce object classifications with natural images without performing any explicit figure-ground segmentation. In human perception, bounding contours are defining of shape, and shape descriptions are conferred based on the bounding contours of segmented objects [38]. Without any figure-ground segmentation mechanism, all contours in training examples probably have equal status. Nothing designates a bounding contour relevant to overall shape, as opposed to contour information that may be part of surface texture, or noise, etc.

With displays stripped of all contour information except for bounding contours, the networks do better. This suggests that the networks must have used some information about the forms of objects to achieve the performance observed in Experiment 4, even though

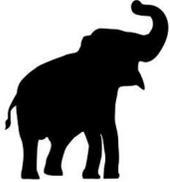
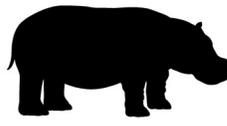
Display Image	Correct Label	1 st Choice	2 nd Choice	3 rd Choice	4 th Choice	5 th Choice
	Bear (0.05%)	Warplane (14.31%)	Stretcher (5.27%)	Cradle (4.53%)	Hammerhead (2.61%)	Studio Couch (2.53%)
	Bison (2.17%)	Capuchin (18.15%)	Ibex (12.26%)	Howler Monkey (10.32%)	Spider Monkey (2.72%)	Vulture (2.34%)
	Camel (36.83%)	Arabian Camel (36.83%)	Bighorn (8.25%)	Ibex (6.12%)	Ram (4.19%)	Harp (2.79%)
	Elephant (7.39%)	Tusker (24.3%)	African Elephant (16.99%)	Water Buffalo (8.14%)	Indian Elephant (7.39%)	Ibex (6.48%)
	Fox (0.37%)	Egyptian Cat (19.1%)	Schipperke (12.26%)	Cougar (10.33%)	Jigsaw Puzzle (6.61%)	Triceratops (3.48%)
	Fench Horn (9.09%)	Hatchet (22.82%)	French Horn (9.09%)	Mortarboard (5.96%)	Sunglass (5.02%)	Hair Slide (4.99%)
	Electric Guitar (1.82%)	Shovel (7.92%)	Hatchet (7.9%)	Hammer (7.15%)	Cleaver (6.9%)	Assault Rifle (3.35%)
	Hammer (49.92%)	Hammer (49.92%)	Hatchet (7.18%)	Nail (7.13%)	Hook (6.47%)	Shovel (3.07%)
	Hippopotamus (4.12%)	American Black Bear (22.96%)	Triceratops (8.41%)	Hippopotamus (4.12%)	Schipperke (3.7%)	Hog (2.15%)
	Lion (0%)	Scottish Deerhound (17.29%)	Giant Schnauzer (11.75%)	Flat-Coated Retriever (6.28%)	Curly-Coated Retriever (5.54%)	Kerry Blue Terrier (5.14%)

Fig 20. VGG-19 classifications for black object silhouettes Part 2.

<https://doi.org/10.1371/journal.pcbi.1006613.g020>

Display Image	Correct Label	1 st Choice	2 nd Choice	3 rd Choice	4 th Choice	5 th Choice
	Orca (1.31%)	Hammerhead (9.88%)	Jigsaw Puzzle (7.87%)	Black Grouse (5.04%)	Hook (4.17%)	Toucan (2.88%)
	Porcupine (1.64%)	Bald Eagle (19.24%)	Vulture (15.47%)	Ostrich (7.95%)	Buckeye (3.64%)	Triceratops (2.86%)
	Assault Rifle (8.43%)	Warplane (74.95%)	Assault Rifle (8.43%)	Wing (4.88%)	Rifle (3.01%)	Airliner (1.57%)
	Otter (0.17%)	Capuchin (16%)	Skunk (8.73%)	Black Stork (8.68%)	Magpie (8.43%)	Hook (4.41%)
	Ram (50.64%)	Ram (50.64%)	Bighorn (40.21%)	Ibex (8.94%)	Schipperke (0.05%)	Water Buffalo (0.04%)
	Sturgeon (0%)	Kite (65.46%)	Albatross (14.48%)	Black Stork (5.88%)	Vulture (2.55%)	Crane (1.42%)
	Bee (0.04%)	Black Grouse (10.5%)	Toucan (10.4%)	Magpie (9.67%)	Hornbill (9.04%)	Vulture (4.85%)
	Dugong (0.39%)	Hook (9.26%)	Kite (7.71%)	Triceratops (5.65%)	Vulture (4.8%)	Black Grouse (4.61%)
	Flamingo (32.2%)	Flamingo (32.2%)	Crane (20.86%)	Hook (16.95%)	Umbrella (6.17%)	Little Blue Heron (4.12%)
	Gazelle (13.16%)	Hartebeest (49.6%)	Impala (18.65%)	Gazelle (13.16%)	Ibex (7.9%)	Bighorn (1.39%)
	Gorilla (1.72%)	Triceratops (9.97%)	Chimpanzee (3.91%)	Egyptian Cat (3.53%)	American Black Bear (3.2%)	Jigsaw Puzzle (3.03%)

Fig 21. VGG-19 classifications for black object silhouettes Part 3.

<https://doi.org/10.1371/journal.pcbi.1006613.g021>

Display Image	Correct Label	1 st Choice	2 nd Choice	3 rd Choice	4 th Choice	5 th Choice
	Hyena (0.79%)	Triceratops (16.68%)	Schipperke (7.4%)	Groenendael (6.21%)	Vulture (5.81%)	Flat-Coated Retriever (5.18%)
	Lobster (1.39%)	Scorpion (34.28%)	Tick (6.47%)	Tricycle (4.33%)	Rhinoceros Beetle (3.54%)	Microphone (2.9%)
	Moped (0.39%)	Rifle (7.1%)	Assault Rifle (5.26%)	Bow Tie (4.18%)	Revolver (4.13%)	Parachute (3.93%)
	Obelisk (97.74%)	Obelisk (97.74%)	Candle (0.44%)	Pedestal (0.33%)	Missile (0.31%)	Projectile (0.22%)
	Ostrich (56.12%)	Ostrich (56.12%)	Harp (6.14%)	Arabian Camel (5.82%)	Pelican (5.33%)	Crane (5.17%)
	Trombone (13.59%)	Nail (64.34%)	Trombone (13.59%)	Hook (5.53%)	Screwdriver (2.22%)	Shovel (1.85%)
	Vase (10.54%)	Pitcher (34.46%)	Caldron (20.99%)	Water Jug (12.92%)	Vase (10.54%)	Cup (6.66%)

Fig 22. VGG-19 classifications for black object silhouettes Part 4.

<https://doi.org/10.1371/journal.pcbi.1006613.g022>

performance with silhouettes still falls far short of classification that humans readily do with shape. By comparison with the earlier experiments, it also suggests that important (bounding) contour information is more influential when no other contours are present.

What use is made of contour information? An important insight may be provided by the results for the black bear silhouette. The silhouette is substantially simplified such that key points of concavity along a bear outline are connected, mostly by straight lines. This is similar to a classic demonstration in object recognition, Attneave's cat [39]. Attneave observed that humans can robustly recognize objects whose contour has been changed significantly at a local level, provided that at the global level the spatial relationships between important points along the contour are preserved. Deep networks do not appear to have the same capabilities. We suspect that they are doing essentially the reverse of humans with regard to global and local aspects of shape. We refer to this as the *local contour feature hypothesis*. The top five responses given by the bear silhouette appears to support this conjecture. The silhouette is confused not

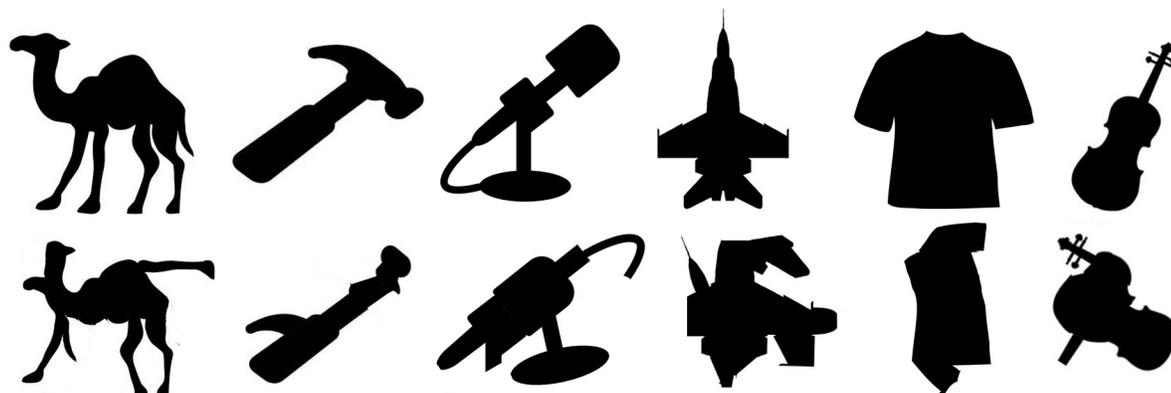


Fig 23. Stimuli used in Experiment 5a. Top row: the original silhouette images, all correctly classified by VGG-19 (appearing in top-five). Bottom row: Scrambled images on which the network was tested.

<https://doi.org/10.1371/journal.pcbi.1006613.g023>

with other quadrupedal mammals, but with manufactured objects like “warplane”, “stretcher” and “studio couch”, suggesting that the network is using shape information, but at a local, not global level. Locally, there are no parts of a bear’s contour that are straight edges, so the network does not consider “black bear” a probable response, whereas an object like a warplane, though its global shape differs completely from the presented stimulus, has a similar set of local contour segments—straight edges and a few shorter rounded edges. Likewise, the network misclassifies the electric guitar as “shovel”, “hatchet”, or “assault rifle”. No human observer would make such errors, but it is easy to see how local curvature information might produce such responses if global form is disregarded.

The local contour feature hypothesis can also help clarify some positive results from Experiments 1–3. In Experiment 1, we found that deep networks use shape much more when classifying artifacts than when classifying animals. This was also observed in Experiment 4, where 13 of the 20 correctly identified objects were manmade. Artifacts will tend to be discriminable based on local features like curvature more often than animals because they are functional and have different component parts depending on their intended purpose. The hook of a hammer, for example, is highly diagnostic in discriminating it from other shafted tools, as is the curve of a French horn, or the long, thin tubes of a trombone. On the other hand, the local features of animals tend more often to be very similar to each other, with some exceptions such as the horn of a ram, or the legs of a flamingo. Two animals will tend to be more discriminable from each other based on global features, such as the ratio of neck length to body length, or ear size to head size.

In Experiment 2, the only figurines correctly classified were the piano and the great white shark. The network’s successful classification of the piano could be attributable to recognition of local contour features. In particular, the curvature of the top board of a grand piano is highly regular and could be driving the network towards that classification. On the other hand, the curvature of the top board was visible in two of the five additional grand piano figurines we tested on, and the network failed to classify these. For the great white shark, the local contour feature hypothesis fits well with the network’s pattern of response. “Warplane”, “hammer-head”, and “airliner” were all assigned higher probability than the correct object label for the shark figurine. These objects all have some local shape features in common. The fin and tail of a shark are fairly similar to the wings and tail of a plane, but the two objects are hardly confusable to humans. This is because humans group the shape features into a unified whole, while

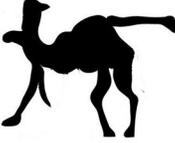
Display Image	Correct Label	1 st Choice	2 nd Choice	3 rd Choice	4 th Choice	5 th Choice
	Camel (65.51%)	Arabian camel (65.51%)	Bighorn (2.54%)	Pedestal (2.24%)	Llama (1.97%)	Ram (1.53%)
	Hammer (18.03%)	Hammer (18.03%)	Nail (12.72%)	Hook (9.67%)	Hatchet (7.96%)	Microphone (4.51%)
	Microphone (6.47%)	Muzzle (10.22%)	Scorpion (6.66%)	Microphone (6.47%)	Long-horned beetle (3.4%)	Whistle (3.03%)
	Warplane (16.1%)	Warplane (16.1%)	Mortarboard (13.16%)	Nail (6.17%)	Academic gown (4.73%)	Birdhouse (2.34%)
	Shirt (1.12%)	Academic gown (29.52%)	Mortarboard (14.68%)	Suit (9.37%)	Bulletproof vest (6.25%)	Sweatshirt (3.76%)
	Violin (6.87%)	Parachute (16.12%)	Rifle (7.03%)	Violin (6.87%)	Assault rifle (6.26%)	Tripod (6.1%)

Fig 24. VGG-19 classifications for part-scrambled silhouettes. The leftmost column shows the image presented to the DCNN. The second column shows the correct object label and the classification probability produced by the network for that label. The other five columns show probabilities for the network’s top five classifications, ordered left to right from highest to lowest. Correct classifications are shaded in blue.

<https://doi.org/10.1371/journal.pcbi.1006613.g024>

deep networks appear to be influenced by local feature aspects, but perhaps less so their relations to the whole.

In Experiment 3, the beer bottle and coffee mug were classified as “wine bottle” and “cup”. These were scored as correct, since the global shape of the objects are quite similar. Importantly, though, their local curvatures are also quite similar. In particular, the transition from the body to the neck of the beer bottle, and the handle of the coffee mug could be important local shape cues driving the network’s good performance. This seems all the more likely when one considers that “sunglass” is the network’s top classification for the coffee mug. Sunglasses share little global shape similarity with a coffee mug, but the curvature of a sunglass’s frame is often locally quite similar to a cup handle. Likewise, for the tandem bike and trombone, which were not correctly classified, but were assigned higher than chance probability, local contour

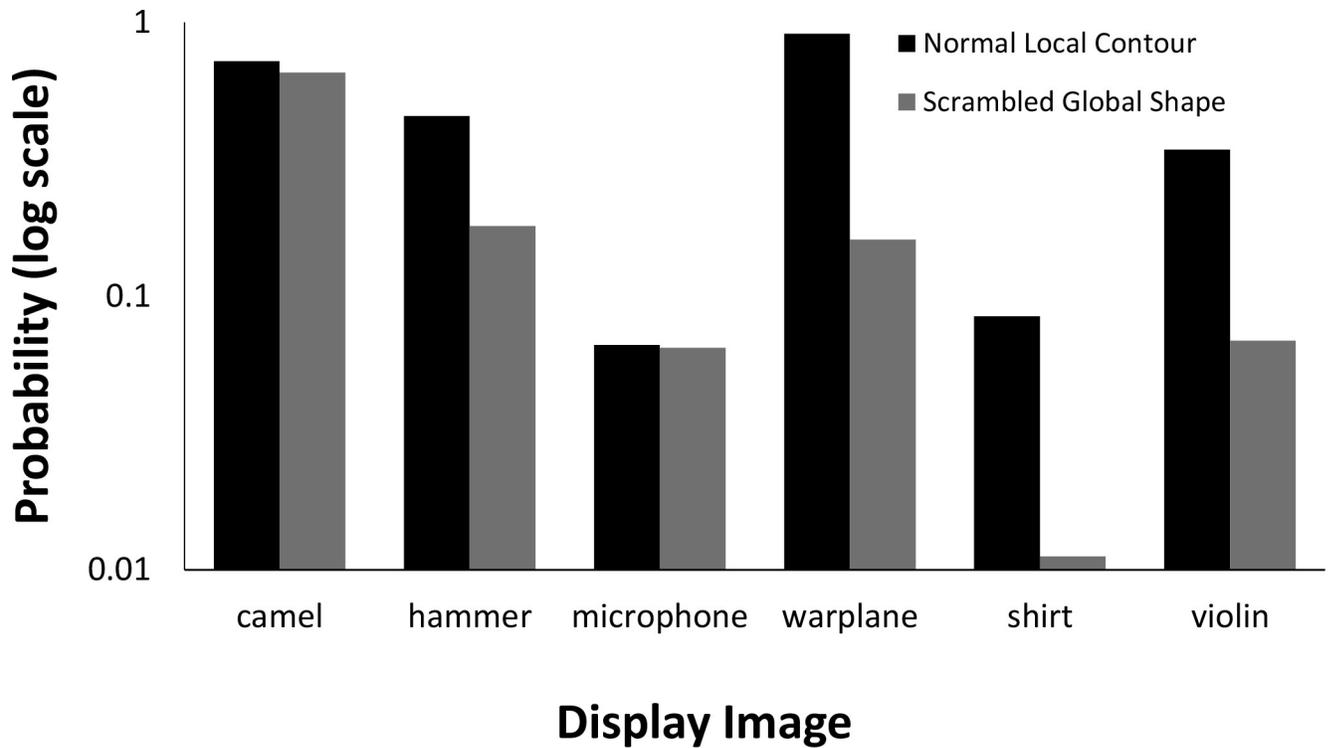


Fig 25. VGG-19 for unscrambled and part-scrambled images. Bars show probabilities for correct responses for each of the objects. Probability is plotted on a logarithmic scale to make small values visible.

<https://doi.org/10.1371/journal.pcbi.1006613.g025>

features like thin straight lines for the trombone and constant curvature in the wheels for the bike could be driving classification. We directly test this hypothesis in Experiments 5.

Experiment 5

In Experiments 2–4, only the silhouette condition provides some supportive evidence that DCNNs use shape. However, further analysis of the networks’ classification of individual items in Experiments 2–4 suggest that some accurate classification decision is likely based on local contour features, not global object shape. We tested this hypothesis in Experiment 5, by comparing the effects of changing local and global features on network classification performance.

Table 1. Human observers’ performance on individual items for part-scrambled objects.

Part-Scrambled Object	Proportion Correct Classification	
	Display Time: 1 sec	Display Time: Unlimited
Camel	40%	60%
Hammer	10%	0%
Microphone	40%	30%
Warplane	20%	20%
Shirt	20%	30%
Violin	40%	80%

<https://doi.org/10.1371/journal.pcbi.1006613.t001>



Fig 26. Stimuli used in Experiment 5b. Top row: the original silhouette images, all correctly classified by the network. Bottom row: images with local contour features disrupted.

<https://doi.org/10.1371/journal.pcbi.1006613.g026>

Experiment 5a

In Experiment 5a, we explicitly tested the hypothesis that deep networks use local shape features, such as the curvature of contour segments, but not global shape, in their classification decisions. We found new examples of shape silhouettes that could be correctly classified and tested to see if the network can still classify them despite changes to their global contour. Preserving most local curvatures, we scrambled the shapes so that the overall shape was radically

Display Image	Correct Label	1 st Choice	2 nd Choice	3 rd Choice	4 th Choice	5 th Choice
	camel (0.05%)	standard poodle (30.58%)	miniature poodle (8.06%)	Irish water spaniel (7.35%)	toy poodle (7.28%)	giant schnauzer (6.64%)
	hammer (0.13%)	parachute (18.74%)	scorpion (10.8%)	kite (5.95%)	screw (4.38%)	bald eagle (2.97%)
	microphone (2.65%)	tripod (10.69%)	whistle (5.43%)	hook (4.59%)	swing (4.03%)	muzzle (3.64%)
	warplane (0.98%)	starfish (24.1%)	space shuttle (5.79%)	Scotch terrier (3.49%)	screw (3.27%)	bell cote (2.94%)
	shirt (0.53%)	Newfoundland (10.84%)	Irish water spaniel (9.02%)	fur coat (3.44%)	curly-coated retriever (3.29%)	Labrador retriever (2.42%)
	violin (0.06%)	bald eagle (6.48%)	parachute (4.82%)	ant (1.7%)	vulture (1.63%)	teddy (1.58%)

Fig 27. VGG-19 classifications for serrated edge silhouettes. The leftmost column shows the image presented to the DCNN. The second column shows the correct object label and the classification probability produced by the network for that label. The other five columns show probabilities for the network's top five classifications, ordered left to right from highest to lowest. Correct classifications are shaded in blue.

<https://doi.org/10.1371/journal.pcbi.1006613.g027>

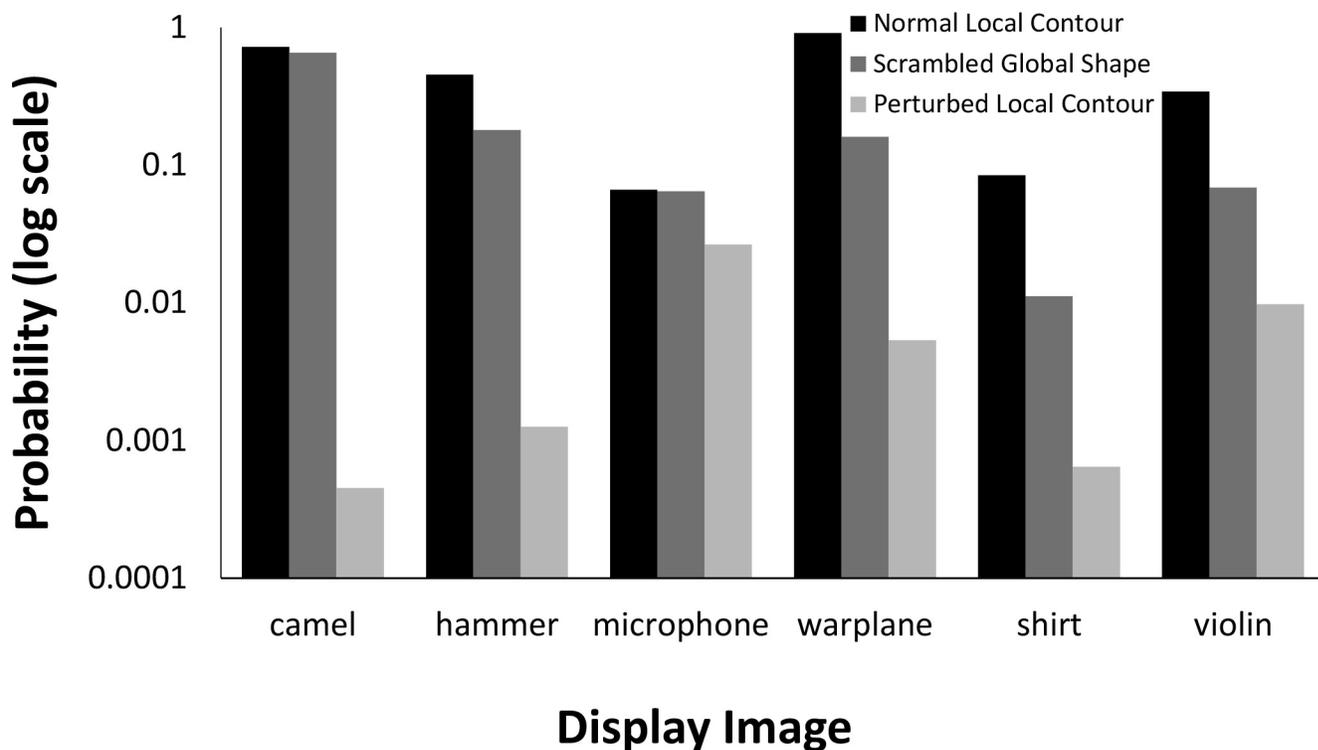


Fig 28. Comparison of VGG-19 performance for locally perturbed contours with unscrambled and part-scrambled images. Bars show probabilities for correct responses for each of the objects. Probability is plotted on a logarithmic scale to make small values visible.

<https://doi.org/10.1371/journal.pcbi.1006613.g028>

changed. If the network was robust to these alterations, that would provide evidence that the network is using local shape primitives as cues, rather than the shape’s global contour. If the network utilizes global shape, these disruptions should impede accurate classification.

Experiment 5a Method. Images. Six images were presented, two from the silhouette database that had been correctly classified, plus four new object silhouettes whose correct label appeared in the network’s top-five selections. Parts of the object were rearranged so that the global contour no longer matched the correct object label, but local edges were preserved. Fig 23 shows the stimulus objects before and after part-scrambling.

Network. The images were tested with VGG-19. We omitted simulations on AlexNet, since performance was reliably better in VGG-19 in Experiments 2–4.

Tests on human subjects. In addition to evaluating the DCNN’s performance on the six scrambled object silhouettes, we tested human subjects to see if they could recognize the objects after scrambling.

Table 2. Human observers’ performance on individual items for perturbed-contour objects.

Locally Perturbed Object	Proportion Correct Classification	
	Display Time: 1 sec	Display Time: Unlimited
Camel	100%	100%
Hammer	100%	100%
Microphone	90%	100%
Warplane	90%	90%
Shirt	100%	100%
Violin	70%	90%

<https://doi.org/10.1371/journal.pcbi.1006613.t002>

Participants. Ten human subjects (two male, eight female, $M_{\text{age}} = 19.5$) were recruited from the University of California, Los Angeles and participated in the study for course credit.

Design and procedure. Subjects completed two experiments. For the first experiment, on each trial, subjects were shown one of 30 silhouettes (24 different, unscrambled objects and the six part-scrambled objects), for one second. After presentation of each silhouette, subjects were asked to write down what they were shown on a piece of paper, after which they were free to continue to the next trial.

The second experiment was identical to the first, except that subjects' exposure duration to each silhouette image was no longer limited to one second. Subjects could view the silhouette for as much time as they wanted before writing down what they believed the object to be on a piece of paper. Once they had recorded their response, they were free to continue to the next trial. As in the first experiment, the second experiment ended when all 30 silhouettes had been presented.

Experiment 5a results. The top five responses and associated probabilities for part-scrambled objects are shown in Fig 24. Five of the six scrambled objects were "correctly" classified in the network's top-five selections. The one correct object label that did not fall in the top-five was "shirt", but "sweatshirt" did, and the misclassifications made by the network—"suit", "bulletproof vest", and "sweatshirt", for example—are clearly influenced by similar local edge features.

We also compared the probability associated with the correct response for the unscrambled image silhouettes with the probabilities for the scrambled silhouettes. Results are shown in Fig 25. On average, the correct response was given a probability 2.26 times higher for unscrambled images than for scrambled images.

Human object recognition. Among human participants, subjects correctly identified 23.33% of the part-scrambled objects and 96.67% of the unscrambled objects when their viewing time was restricted to one second. When viewing time was unrestricted, subjects correctly identified 36.67% of part-scrambled objects, and 94.8% of unscrambled objects. Table 1 shows subject accuracy for each of the individual part-scrambled objects.

Experiment 5b

Experiment 5a tested deep networks' ability to classify objects whose global shape was destroyed, with local curvature largely preserved. We hypothesized that if DCNNs did not classify based on global shape, performance would remain good, provided that enough local contour features were preserved in the scrambled objects. In Experiment 5b, we conducted a complementary study to test the hypothesis that if local contour features are disrupted, but global shape is preserved, network classification accuracy will suffer.

Experiment 5b Method. *Images.* We used the same six silhouette images as in Exp. 5a. In their original format, VGG-19 correctly classified all six images. We disrupted the local contour features in each image by adding jagged edges to the bounding contour, creating a sawtooth effect. Fig 26 shows the six original images and the sawtooth images.

Network. The perturbed contour images were tested on VGG-19.

Test on human subjects. As in Experiment 5a, we tested humans' ability to recognize images with global shape preserved and local contour information disrupted.

Participants. Ten participants (four male, six female, $M_{\text{age}} = 20.0$) were recruited from the University of California, Los Angeles and participated in the study for course credit. None of the subjects who participated in Exp. 5a participated in this experiment.

Design and procedure. The design for the study was identical to Experiment 5a in all respects except for the six disrupted stimuli that were presented. On those trials, subjects were

shown the locally disrupted sawtooth images instead of the globally disrupted scrambled images.

Experiment 5b results. The network's top-five classification selections for each of the locally disrupted images are shown in Fig 27. None of the images were correctly classified, either as the first choice or anywhere among the network's top 5 choices, and the average rank of the correct label for perturbed contour images was 96.3.

The network performed much worse on images where local contour was disrupted but global shape preserved than with either unperturbed or globally scrambled images. The average probability assigned to shape labels was 58.73 times higher for unperturbed local contour images than for perturbed local contour images, and 26.0 times higher in globally scrambled images than for perturbed local contour images. Fig 28 shows a comparison between these three conditions.

Human object recognition. Human participants recognized objects very accurately when local contours were perturbed. When viewing time was limited to one second, subjects' recognition accuracy on locally disrupted contour silhouettes was 91.67%. With unlimited viewing time, subjects' accuracy was 96.67%. Subjects' accuracy on individual items is shown in Table 2. These results did not significantly differ from subjects' performance on unperturbed object either for the brief presentation condition, $t(18) = 0.98$, $p = .341$, or for the unlimited presentation condition, $t(18) = 0.82$, $p = .425$.

Experiment 5 Discussion. The results of Experiments 5a and 5b suggest robust and interesting differences between humans and DCNNs in the use of shape information. Humans mostly failed to correctly recognize the objects presented in Experiment 5a. For one or two of the objects, they saw some recognizable parts (e.g., for the scrambled violin, one subject said "I see some pieces of a broken guitar"). Even where some parts could be recognized, every human observer pointed out that the scrambled objects are not actual versions of the label that would be correct for the unscrambled versions. In contrast, when human subjects were shown objects whose global shape was preserved, but whose local contour features were changed, they performed extremely well on the recognition task, both with short exposures and unlimited viewing time.

The DCNN results showed the reverse pattern of the human results. For VGG-19, classification for objects whose global shape had been destroyed by local scrambling remained very strong. However, in the local contour disruption condition, the network's classification accuracy fell off dramatically. Despite preservation of global shape that allowed essentially perfect human object classification for both brief exposures and unlimited viewing, VGG-19 did not select any correct label for any object in Exp. 5b among its top five choices. The influence of local contour information is clearly visible in many of the network's top-five selections from this simulation. Animals like poodles, curly-coated retrievers, and Scotch terriers are often selected despite no similarity in global shape, likely because the local features resemble a curly-furred animal. These results suggest that the network's processing of shape is restricted to relatively local features along an object's bounding contour, but not the shape as a whole. They also represent a clear difference between human and deep network recognition processes. Whereas local contour features seem to play a key role in deep network classification, global shape seems to have primacy in human perception.

These results, along with those of the earlier experiments, push us to think about what is really meant by shape. The convolution operations that form the groundwork for DCNNs are certainly capable of responding to local oriented contrast, as such filters have long been used to model orientation sensitivity [35]. For a shape feature such as the "claw" part of the hammer image, it would be sufficient for the network to encode a few orientations in proximity and in a certain spatial relation. These are, undoubtedly, aspects of "shape". And our results suggest

that they are accessible within the trained AlexNet and VGG-19 networks. There are other, more global, notions of shape, however. Larger relations of parts, and overall characteristics, such as aspect ratios, may generally be inaccessible. The network was remarkably undeterred by serious scrambling of parts, despite the fact that this scrambling undoubtedly also disrupted some local features of the sort we have noted here.

The issue of local shape features vs. global shape descriptions bears an important relation to questions of subsymbolic and symbolic representations in visual perception [36]. A number of phenomena indicate that human vision produces abstract shape representations that capture the gist of objects and support similarity relationships, despite local noise, and variations in local elements [40]. Even the perception of a clear object edge or the notion of a continuous contour are abstractions from earlier inputs of local oriented contrast in various locations and different spatial frequencies in the same location (for discussion, see [41]). Conversely, measurement of oriented contrast in two nearby locations and use of some conjunction of orientations in nearby positions can be done without conversion to a symbolic representation. These are the beginnings of shape descriptions, or perhaps even local shape descriptions. Deep networks may be accessing shape features that are conjunctions of local orientations and their relations; indeed, such representations may lie somewhere in the transition between subsymbolic and symbolic representations of an object's shape (see [42] for a related proposal regarding the representation of contour shape). Taken together with the results of Experiments 1–4, the findings from Experiment 5 are consistent with the idea that DCNNs access aspects of local edge and curvature information, but not a representation of how local parts spatially relate to each other as a whole.

Discussion

The purpose of this work was to determine the extent to which shape information is represented and used for recognition in trained deep convolutional networks. We tested deep networks trained on ImageNet to determine if shape information is relevant at all to DCNN object recognition performance, and if it is, how shape cues are weighted compared to other information, such as texture and context.

In Experiment 1, we tested VGG-19 for sensitivity to shape apart from appropriate texture and surface information, showing the network 40 object silhouettes with a different object's surface texture overlaid on each shape. Evidence for use of shape information was weak: The correct label based on shape was chosen as the first choice classification for only 5 of the 40 objects sampled, and the average rank among network outputs for the correct shape was 209. Where evidence of shape influences on classification did appear, however, they seemed to depend highly on the kind of object being classified. While classification of artifacts and rigid objects appeared to depend on shape more than on texture cues, the opposite was true for images of animals. Even for artifacts in which the network weights shape more strongly, the absence of typical surface texture greatly reduced the network's classification accuracy, resulting in many spurious classifications. The importance of texture information in object recognition was also seen in Experiment 2, where glass objects with no shape similarity to the presented stimuli were selected preferentially over the object whose shape matched the glass figurine, and in Experiment 4, where performance changed dramatically depending on whether a silhouette was black, white, or red. These results differ greatly from what is observed in research on human vision. Several studies (e.g., [43, 44, 28–33]) have found that texture plays little role in facilitating human object recognition.

In Experiments 2 and 3, we presented two networks with images whose shape matched object categories that the network had been trained to recognize, while differing in terms of

context and surface texture. Deep networks were in general unable to classify glass figurines or object outlines in the absence of other cues that are typically present in natural images. Across both datasets, only four of 80 presented images were correctly classified, and analyses of the probabilities assigned to each object label revealed that the network was assigning near-minimum probabilities to the correct objects in all but a few cases.

A possible explanation for the network's poor performance in Experiment 3 is that the outline images are flat, lacking volumetric cues that might be present in 2D photographs of real objects. If the network depends on volumetric cues to recognize objects, the flatness of the outline images might prevent it from matching these images to its trained categories. This is an intriguing hypothesis, but the network was no better for glass figurines than for object outlines, even though figurines have 3D structures that would be quite similar to those of objects the network is trained to recognize. Moreover, the network did comparatively well in Experiment 4 classifying object silhouettes, which also lack any volumetric properties.

Experiment 4 found some evidence for use of shape information in object recognition by deep networks. Networks were presented with object silhouettes and correctly classified 20 of the 40 images (based on the top-five criterion), as well as assigning non-minimum probabilities to several others. One explanation for the networks' superior performance for silhouettes is that these displays more fully eliminate competing surface texture information than do objects with glass surfaces, or objects whose texture exactly matched the background. They also remove orientation information that is not part of an object's bounding contour, thus preserving the orientation information most likely to be diagnostic of an object's shape. This selective preservation of orientation is probably especially helpful, as it seems unlikely that DCNNs trained for object classification have any differential representation of a bounding contour vs. any other contour.

Classification of object silhouettes would not be possible without some edge-based recognition capabilities, but it is unlikely that deep networks accessed global form for the task. Instead, the network appears to be to have extracted contour segments and local features based on some relations of proximate contour orientations. Indirect evidence supporting this explanation appeared in some of the network's misclassifications. For example, a great white shark figurine had warplane and airliner in its top-five labels, likely due to the local similarity between a shark's fins and a plane's wings. Likewise, a porcupine was misclassified as a bird, probably because its spines had features in common with feathers, although globally birds and porcupines differ greatly.

Experiment 5 sought direct evidence for the local contour feature hypothesis. In Experiment 5a we used silhouette images, including some of the best performing ones from Experiment 4, for which the networks produced accurate classification outcomes. These images were then scrambled in such a way that curvature of local edge segments was preserved, but the shape as a whole was radically altered. Networks performed nearly as well on these scrambled objects as they did on the original images. On the other hand, in Experiment 5b, when global shape was preserved but local contour features were changed by adding jagged edges, network classification became extremely poor. These results suggest that the features relating to shape that figure in recognition in deep networks relate to curvature or orientation relations of highly local parts of the contour. The network recognizes the object by the mere presence of these features, not by how they might globally relate to each other in space. The convolution operations that occur at the earliest levels of DCNNs are well suited to extract local oriented contrast, and conjunctions of nearby orientations could serve as a serviceable marker for local curvature, as well as local shape features. This is why rearranging existing parts of a silhouette has little effect on network performance (since local feature extraction is the same) while the

addition of a serrated border to the silhouette destroys network performance. The opposite is true in human perception.

In human vision, the transition from subsymbolic to symbolic processing depends critically on how encoded features fit together into a unified whole [36, 45]. Identifying the curvature of contour segments may be an important first step in this transition, but it cannot account for the robust capabilities of our visual system to classify objects based on their shape. For example, Attneave [39], showed that all local curvature segments in a cat can be straightened without detriment to its recognizability. We actually tested one image, a bear, that had straightened segments, much like Attneave's cat, in Experiments 1 and 4. The probabilities assigned by the network were less than chance with both inconsistent texture, in Exp. 1 (.04%), and in silhouette form, in Exp. 4 (.03%). Consistent with these ideas, the bear image in both cases was unequivocally seen as a bear by human observers. This and many other phenomena of human perception suggest that relationships between contour segments and object parts are central to human object recognition and shape perception, more so than encoding local features.

In the introduction, we mentioned three levels of questions that might be asked regarding DCNNs and the use of shape information. The present results apply to the first level: the important question of understanding the role of shape in DCNNs trained to do object recognition. We believe that our results provide a good picture of the use of shape information by trained DCNNs, specifically suggesting a profound difference between the accessibility of local shape features and more global aspects of shape. The results of this study appear to strongly suggest that such models do not classify based on an object's global shape cues. These results regarding the shape sensitivity of trained networks are important for comparisons of human perception with DCNNs, which in other respects have been argued to mirror the human visual brain and human behavior in a remarkable number of ways.

We noted earlier a second level of question—whether DCNNs can be trained from scratch to classify based on global shape information. Our present results clearly suggest that standard training on the ImageNet database does not produce such capabilities, but it leaves open the question of whether deep networks are *incapable* of classification based on global shape features. The photographs in ImageNet tend to show a single object in the foreground of the image. It is possible that training conditions in which the object is displayed in a more complete scene would produce better shape sensitivity because the network would learn to extract the spatial relationships between objects. Another way the network might be trained to encode global shape would be to include training examples with nondiagnostic texture and local shape properties. The network might have too many other rich cues for classification to develop global shape sensitivity under standard training conditions, but perhaps could encode global shape if deprived of some of these other information streams. Our suspicion is that even under training conditions specifically selected to develop global shape representations, DCNNs will not be able to classify based on global shape alone, but these hypotheses can only be confirmed through experimentation on untrained neural networks.

Conclusion

In human vision, abstract representations of shape describe how the various parts of an object spatially relate to one another. They are critical for recognition across a variety of viewing conditions, and are robust to perturbations of local contour features. Deep networks have impressive capabilities for object recognition, but they do not appear to handle the problem of recognition the same way humans do. Unlike humans, surface texture appears to be an equally strong cue for recognition as shape. Moreover, the shape information used by deep networks is highly limited: DCNNs appear to be capable of encoding local shape features

including local edge segments and relations. Sensitivity to how these local features fit together as a whole is lacking; DCNNs trained for object recognition do not appear to represent global shape at all.

Methods

Ethics statement

All research on human subjects in this and subsequent experiments was under IRB approval (IRB#11-002079-CR-00001).

Network

Tests were conducted on VGG-19 [27] and AlexNet [4]. For most experiments, the results of the better-performing VGG-19 are reported. Both networks were trained on the ImageNet database prior to testing. Tests were conducted using the Neural Networks Toolbox from Matlab 2017b.

Materials

Testing images are described in the Methods description for individual experiments. In all experiments, the correct classification for the testing images was among the 1000 object categories that the networks had been trained to classify.

Supporting information

S1 File. Results from outlines sampled from Exp. 4 silhouettes. The leftmost column shows the image presented to VGG-19. The second column from the left shows the correct object label and the classification probability produced for that label. The other five columns show probabilities for the network's top five classifications, ordered left to right in terms of the probability given by the network. Correct classifications are shaded in blue.

(DOCX)

S2 File. URLs For test images.

(DOCX)

Author Contributions

Conceptualization: Nicholas Baker, Hongjing Lu, Gennady Erlikhman, Philip J. Kellman.

Data curation: Nicholas Baker, Philip J. Kellman.

Formal analysis: Nicholas Baker.

Funding acquisition: Nicholas Baker, Philip J. Kellman.

Investigation: Nicholas Baker.

Methodology: Nicholas Baker, Hongjing Lu.

Supervision: Philip J. Kellman.

Writing – original draft: Nicholas Baker, Philip J. Kellman.

Writing – review & editing: Nicholas Baker, Hongjing Lu, Gennady Erlikhman, Philip J. Kellman.

References

1. Yampolskiy R. Turing test as a defining feature of AI-completeness. *Artificial Intelligence, Evolutionary Computing and Metaheuristics*. 2013;3–17.
2. Turing AM. Computing machinery and intelligence. *Mind*. 1950 Oct 1; 59(236):433–60.
3. Geman D, Geman S, Hallonquist N, Younes L. Visual turing test for computer vision systems. *Proceedings of the National Academy of Sciences*. 2015 Mar 24; 112(12):3618–23.
4. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: *Advances in Neural Information Processing Systems*, 2012 (pp. 1097–1105).
5. Gunji N, Higuchi T, Yasumoto K, Muraoka H, Ushiku Y, Harada T, Kuniyoshi Y. Scalable multiclass object categorization with Fisher based features. *ILSVRC 2012*, The Univ. of Tokyo.
6. LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*. 1998 Nov; 86(11):2278–324.
7. Smith AT, Singh KD, Williams AL, Greenlee MW. Estimating receptive field size from fMRI data in human striate and extrastriate visual cortex. *Cerebral Cortex*. 2001 Dec 1; 11(12):1182–90. PMID: [11709489](https://pubmed.ncbi.nlm.nih.gov/11709489/)
8. Pasupathy A, Connor CE. Shape representation in area V4: position-specific tuning for boundary conformation. *Journal of Neurophysiology*. 2001 Nov 1; 86(5):2505–19. <https://doi.org/10.1152/jn.2001.86.5.2505> PMID: [11698538](https://pubmed.ncbi.nlm.nih.gov/11698538/)
9. Pospisil D, Pasupathy A, Bair W. Comparing the brain's representation of shape to that of a deep convolutional neural network. In: *Proceedings of the 9th EAI International Conference on Bio-inspired Information and Communications Technologies (formerly BIONETICS) on 9th EAI International Conference on Bio-inspired Information and Communications Technologies (formerly BIONETICS) 2016 May 24* (pp. 516–523). ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering).
10. Yamins DL, Hong H, Cadieu CF, Solomon EA, Seibert D, DiCarlo JJ. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*. 2014 Jun 10; 111(23):8619–24.
11. Khaligh-Razavi SM, Kriegeskorte N. Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Computational Biology*. 2014 Nov 6; 10(11):e1003915. <https://doi.org/10.1371/journal.pcbi.1003915> PMID: [25375136](https://pubmed.ncbi.nlm.nih.gov/25375136/)
12. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. Imagenet: A large-scale hierarchical image database. In: *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on 2009 Jun 20* (pp. 248–255). IEEE.
13. Dubey R, Peterson J, Khosla A, Yang MH, Ghanem B. What makes an object memorable? In: *Proceedings of the IEEE International Conference on Computer Vision 2015* (pp. 1089–1097).
14. Peterson JC, Abbott JT, Griffiths TL. Adapting deep network features to capture psychological representations. *arXiv preprint arXiv:1608.02164*. 2016 Aug 6.
15. Palmer SE. *Vision science: Photons to phenomenology*. MIT Press; 1999 Apr 14.
16. Biederman I. Recognition-by-components: a theory of human image understanding. *Psychological Review*. 1987 Apr; 94(2):115. PMID: [3575582](https://pubmed.ncbi.nlm.nih.gov/3575582/)
17. Marr D, Nishihara HK. Representation and recognition of the spatial organization of three-dimensional shapes. *Proceedings of the Royal Society of London B: Biological Sciences*. 1978 Feb 23; 200(1140):269–94. PMID: [24223](https://pubmed.ncbi.nlm.nih.gov/24223/)
18. Kubilius J, Bracci S, de Beeck HP. Deep neural networks as a computational model for human shape sensitivity. *PLoS Computational Biology*. 2016 Apr 28; 12(4):e1004896. <https://doi.org/10.1371/journal.pcbi.1004896> PMID: [27124699](https://pubmed.ncbi.nlm.nih.gov/27124699/)
19. Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, Fergus R. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*. 2013 Dec 21.
20. Nguyen A, Yosinski J, Clune J. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2015* (pp. 427–436).
21. Zhu Z, Xie L, Yuille AL. Object Recognition with and without Objects. *arXiv preprint arXiv:1611.06596*. 2016 Nov 20.
22. Cichy RM, Khosla A, Pantazis D, Torralba A, Oliva A. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*. 2016 Jun 10; 6:27755. <https://doi.org/10.1038/srep27755> PMID: [27282108](https://pubmed.ncbi.nlm.nih.gov/27282108/)
23. Cadieu CF, Hong H, Yamins DL, Pinto N, Ardila D, Solomon EA, Majaj NJ, DiCarlo JJ. Deep neural networks rival the representation of primate IT cortex for core visual object recognition. *PLoS*

Computational Biology. 2014 Dec 18; 10(12):e1003963. <https://doi.org/10.1371/journal.pcbi.1003963> PMID: 25521294

24. Güçlü U, van Gerven MA. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*. 2015 Jul 8; 35(27):10005–14. <https://doi.org/10.1523/JNEUROSCI.5023-14.2015> PMID: 26157000
25. Kümmerer M, Theis L, Bethge M. Deep gaze i: Boosting saliency prediction with feature maps trained on imagenet. arXiv preprint arXiv:1411.1045. 2014 Nov 4.
26. Ithapu VK. Decoding the Deep: Exploring class hierarchies of deep representations using multiresolution matrix factorization. In: CVPR Workshop on Explainable Computer Vision and Job Candidate Screening Competition 2017 Jul 1 (Vol. 2).
27. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556. 2014 Sep 4.
28. Biederman I, Ju G. Surface versus edge-based determinants of visual recognition. *Cognitive Psychology*. 1988 Jan 31; 20(1):38–64. PMID: 3338267
29. Davidoff JB, Ostergaard AL. The role of colour in categorial judgements. *The Quarterly Journal of Experimental Psychology Section A*. 1988 Aug; 40(3):533–44.
30. Elder JH, Velisavljević L. Cue dynamics underlying rapid detection of animals in natural scenes. *Journal of Vision*. 2009 Jul 1; 9(7):7–. <https://doi.org/10.1167/9.7.7> PMID: 19761322
31. Bergevin R, Levine MD. Generic object recognition: Building and matching coarse descriptions from line drawings. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 1993 Jan; 15(1):19–36.
32. Lloyd-Jones TJ, Luckhurst L. Outline shape is a mediator of object recognition that is particularly important for living things. *Memory & Cognition*. 2002 Jun 1; 30(4):489–98.
33. Kellman PJ, Shipley TF. A theory of visual interpolation in object perception. *Cognitive Psychology*. 1991 Apr 30; 23(2):141–221. PMID: 2055000
34. Hochberg J, Brooks V. Pictorial recognition as an unlearned ability: A study of one child's performance. *The American Journal of Psychology*. 1962 Dec 1; 75(4):624–8.
35. Marr D. *Vision*. 1982. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. 1982.
36. Koffka K. *Principles of Gestalt Psychology*, International Library of Psychology, Philosophy and Scientific Method.
37. Snodgrass JG, Vanderwart M. A standardized set of 260 pictures: norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology: Human Learning and Memory*. 1980 Mar; 6(2):174.
38. Kellman PJ, Garrigan P, Erlikhman G. Challenges in Understanding Visual Shape Perception and Representation: Bridging Subsymbolic and Symbolic Coding. In *Shape Perception in Human and Computer Vision 2013* (pp. 249–274). Springer, London.
39. Attneave F. Some informational aspects of visual perception. *Psychological Review*. 1954 May; 61(3):183. PMID: 13167245
40. Baker N, Kellman PJ. Abstract shape representation in human visual perception. *Journal of Experimental Psychology: General*. 2018 Sep; 147(9):1295.
41. Kellman PJ, Massey CM. Perceptual learning, cognition, and expertise. *The Psychology of Learning and Motivation*. 2013 Jan 1; 58:117–65.
42. Kellman PJ, Garrigan P. Segmentation, grouping, and shape: some Hochbergian questions. *Perception: Essays in Honor of Julian Hochberg* Ed. Peterson MA (New York: Oxford University Press) pp. 2006:542–54.
43. Tanaka JW, Presnell LM. Color diagnosticity in object recognition. *Attention, Perception, & Psychophysics*. 1999 Aug 1; 61(6):1140–53.
44. Rossion B, Pourtois G. Revisiting Snodgrass and Vanderwart's object pictorial set: The role of surface detail in basic-level object recognition. *Perception*. 2004 Feb; 33(2):217–36. <https://doi.org/10.1068/p5117> PMID: 15109163
45. Wertheimer M. *Laws of organization in perceptual forms*. A source book of Gestalt Psychology. 1923.