Learning the 3-D structure of objects from 2-D views depends on shape, not format

Moqian Tian
Department of Psychology, Stanford University, Stanford, CA, USA

Daniel Yamins
Department of Brain and Cognitive Sciences, MIT, Cambridge, MA, USA

Kalanit Grill-Spector
Department of Psychology, Stanford University, Stanford, CA, USA
Stanford Neuroscience Institute, Stanford University, Stanford, CA, USA

Humans can learn to recognize new objects just from observing example views. However, it is unknown what structural information enables this learning. To address this question, we manipulated the amount of structural information given to subjects during unsupervised learning by varying the format of the trained views. We then tested how format affected participants' ability to discriminate similar objects across views that were rotated 90° apart. We found that, after training, participants' performance increased and generalized to new views in the same format. Surprisingly, the improvement was similar across line drawings, shape from shading, and shape from shading + stereo even though the latter two formats provide richer depth information compared to line drawings. In contrast, participants' improvement was significantly lower when training used silhouettes, suggesting that silhouettes do not have enough information to generate a robust 3-D structure. To test whether the learned object representations were format-specific or format-invariant, we examined if learning novel objects from example views transfers across formats. We found that learning objects from example line drawings transferred to shape from shading and vice versa. These results have important implications for theories of object recognition because they suggest that (a) learning the 3-D structure of objects does not require rich structural cues during training as long as shape information of internal and external features is provided and (b) learning generates shape-based object representations independent of the training format.

Introduction

Humans are able to recognize 3-D objects from the 2-D retinal input across changes in their appearance. The ability to recognize objects across views, referred to as viewpoint-invariant recognition, is particularly challenging because the shape and features of the object change drastically across different 2-D retinal views of the same object. A large body of research shows that viewpoint-invariant recognition is learned by viewing 2-D examples of an object (Bülthoff, Edelman, & Tarr, 1995; Hayward & Tarr, 1997; Tarr, Williams, Hayward, & Gauthier, 1998). It is thought that during learning people use structural information in the 2-D training images to infer the 3-D structure of the object (Nakayama, Shimojo, & Silverman, 1989). Understanding the 3-D object structure, in turn, enables participants to recognize an object across views. However, it is unknown what structural information in the visual input enables this learning.

During learning, subjects may use many sources of information about the 3-D structure of objects from both monocular cues and binocular cues. The 3-D shape of the object can be readily perceived monocularly from information provided by the external contour, internal contours, and shading. Although Marr (1982) suggested that shading information is particularly important for building an accurate 3-D representation of an object that includes surface information, empirical research shows the opposite. A large body of research indicates that the shape of the object available from both the external contour and

Citation: Tian, M., Yamins, D., & Grill-Spector, K. (2016). Learning the 3-D structure of objects from 2-D views depends on shape, not format. *Journal of Vision*, *16*(7):7, 1–17, doi:10.1167/16.7.7.

doi: 10.1167/16.7.7 Received November 18, 2015; published May 6, 2016 ISSN 1534-7362

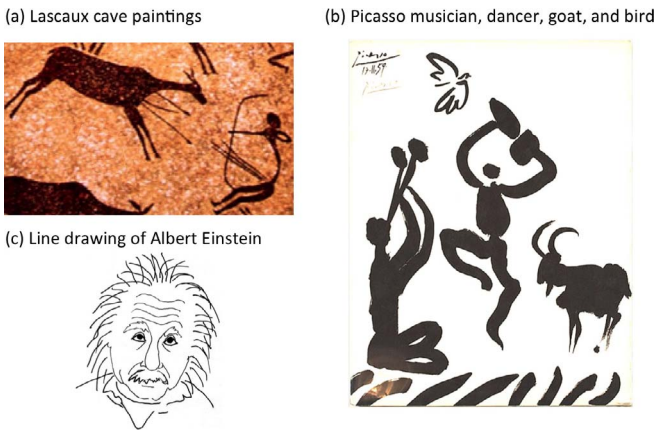

Figure 1. Example drawings demonstrating that impoverished images using silhouettes and lines can be recognized. (a) Lascaux cave paintings depicting a deer and a person with a bow. (b) Picasso's drawing of silhouettes of a musician, a dancer, a goat, and a bird. (c) A line drawing of Albert Einstein's face (copyright: imgkid.com).

internal features is the key source of information that humans use for visual recognition. For example, people can recognize objects, and particularly living things, from their external contour or their *silhouettes* (see examples in Figure 1a, b; Lloyd-Jones & Luckhurst, 2002; Mitsumatsu & Yokosawa, 2002). Nevertheless, recognition from silhouettes declines when objects appear at unusual views or when key internal parts are obscured by the external silhouette (Hayward, 1998; Lawson, 1999; Newell & Findlay, 1997). *Line drawings* provide more concrete information about the 3-D shape of objects compared to silhouettes even though they are not as realistic as photographs or objects depicted with 3-D shape-from-shading information (*shaded objects*). Indeed, people are as good at basic-level recognition of objects (e.g., discriminating a house from an apple; Rosch, 1999) from line drawings as from realistic photographs (Biederman & Ju, 1988; Rossion & Pourtois, 2004). Biederman argued that line drawings may be a privileged source of information because they provide observers with a distilled version of the critical nonaccidental properties of the object that are necessary for the basic-level recognition (Biederman, 1987; Biederman & Ju, 1988). Not only basic-level recognition, but also subordinate recognition among objects that share parts and configuration (Rosch, 1999), such as face recognition, can be performed from line drawings even when hand-drawn lines do not faithfully match any real picture of the person (e.g., Figure 1c). However, prior studies that examined learning effects from line drawings have only tested basic-level recognition, which is easier than subordinate recognition. Thus, it is possible that shading information may be critical for learning the 3-D structure of novel objects to enable subordinate recognition across views.

Another source of information about the 3-D structure of objects is available from binocular disparity in *stereoscopic* images. Stereovision provides explicit 3-D depth information. Although stereo information does not always improve object recognition ability (Liu, Ward, & Young, 2006; Pasqualotto & Hayward, 2009), several studies indicate that stereo information can be advantageous for object recognition, particularly when precise depth information needs to be recovered to perceive the 3-D structure (Bülthoff & Mallot, 1988; Burke, 2005; Edelman & Bülthoff, 1992; Humphrey & Khan, 1992; Y. L. Lee & Saunders, 2011). For example, stereo information can resolve ambiguities in the 3-D structure when objects share an external contour (Bennett & Vuong, 2006; Y. L. Lee & Saunders, 2011), lack symmetry or self-occlusion (Bennett & Vuong, 2006; Burke, 2005; Edelman & Bülthoff, 1992; Humphrey & Khan, 1992), or when their surface material has non-Lambertian properties (Nefs, 2008; Nefs & Harris, 2007). However, to date no study has systematically examined the effect of different sources of depth information—external contours, internal contours, shading, and binocular depth—on learning viewpoint-invariant recognition.

Another theoretically interesting question is whether recognition of objects from different formats, e.g., line drawings and shape from shading, is based on a common representation or this recognition is derived from different kinds of internal representations. Neuroscience studies report that neural representations in object-selective regions in the ventral occipitotemporal cortex are cue-invariant (Grill-Spector, Kushnir, Edelman, Itzchak, & Malach, 1998; Kastner, De Weerd, & Ungerleider, 2000; Kourtzi & Kanwisher, 2000; Mendola, Dale, Fischl, Liu, & Tootell, 1999; Sary, Vogels, & Orban, 1993; Vinberg & Grill-Spector, 2008). Results of these studies suggest that there is common representation of object shape across cues, such as line drawings, shading, and stereo, in the ventral visual pathway. However, two questions remain: (a) How were these object representations generated? (b) Is complete and explicit 3-D information about the object necessary during learning?

To address these gaps in knowledge, we ran psychophysical experiments varying the amount of structural information given to subjects during learning and tested how this information affected subjects' ability to recognize novel 3-D objects across views. We manipulated the amount of structural information by showing objects in four formats that contain increasing amounts of structural information (Figure 2b, c): (a) silhouettes, (b) line drawings, (c) shape from shading, and (d) shape from shading + stereo. The most impoverished format, silhouettes, provides information

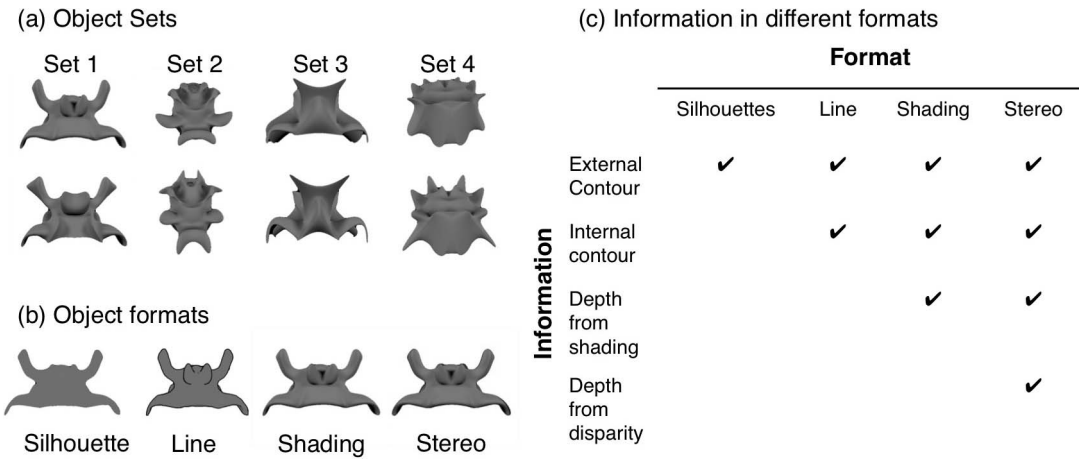

Figure 2. Examples of the stimuli used in the study. (a) Examples of the four object sets used in the study. Top row: Front view of the first object in each set. Bottom row: Front view of the second object in each set. (b) Examples of the four structural formats used in the study for an example object. Stereo used red–cyan anaglyphs. (c) Table illustrating the type of structural information available in each of the four formats. Structural information progressively increases across formats.

only about the external contour of the object. Line drawings provide additional information about the internal features as well as some perspective information. Shape from shading provides additional monocular depth information, and finally, stereo provides concrete 3-D depth from disparity. To give the reader an intuition why these formats convey increasing levels of structural information, consider the following: shading information can be extracted from the stereoscopically presented objects, but disparity information cannot be extracted from the monocular, shaded objects. Similarly, line drawings can be extracted from the shaded objects, and silhouettes can be extracted from line drawings, but the opposite information extraction is not possible.

In Studies 1 and 2, we examined whether more structural information during unsupervised learning of novel 3-D objects leads to better performance. We hypothesized that if learning depends on complete and rich structural information, then post-learning performance will steadily increase as more structural information is available to subjects during training. This hypothesis predicts that the improvement due to learning will gradually increase from silhouettes to line drawings to shaded objects to stereo objects. Alternatively, we predicted that if a certain amount of structural information is sufficient for learning the 3-D shape of objects, improvement due to learning will be equivalent across formats that contain that information. For example, if internal and external features present in line drawings are sufficient for learning the 3-D structure of objects, performance after learning from either line drawings, shaded objects, or stereo objects would be similar as all of these formats contain information about internal and external features.

In Study 3, we examined whether learning objects from different formats generates object representations that are format-specific or format-invariant. We tested these hypotheses by comparing learning within and across formats. Finding that learning transfers across formats would indicate that, even though different information was given to subjects during training, the internal representation of an object that was generated due to training was common across formats. On the other hand, finding better performance within than across formats would suggest that different formats generated different internal object representations that are format-specific.

Methods

Participants

The experimental protocol was approved by Stanford University's Human Subjects Institutional Review Board. Participants provided written informed consent before participating in the experiment. Participants received monetary compensation or course credit for their participation. Twenty-eight to fifty-two participants participated in each of the studies. These are typical sample sizes used in perceptual learning studies. Participants had normal or corrected-to-normal vision.

Stimuli

To examine the effects of different sources of structural information on unsupervised learning, we

generated sequences of novel 3-D object views that are rendered in different structural formats. We generated four sets of novel 3-D objects using 3ds Max (http://usa.autodesk.com/3ds-max/). Each set contained two objects that had the same parts and configuration but differed in the shape of features and their aspect ratio (Figure 2a). Objects were grayscale and symmetric around a vertical plane passing through their centers. All objects were rendered in the same lighting conditions with four achromatic lights, lighting the object from top right, top left, bottom right, and bottom left. Stimuli were presented on a uniform dark gray background.

Prior to these experiments, stimuli were calibrated in a pilot study. The calibration used a different group of participants and the 3-D shape-from-shading objects. Here, we modified the object sets to reach pretraining discrimination performance that was similar across object sets when presented in the shape-from-shading formats.

Images of the same 3-D object models were generated in four formats in 3ds Max (Figure 2b). These four formats contain different types of depth information summarized in Figure 2c. For each format, we generated object views by rotating the object around the vertical axis and taking a snapshot of the object. The same objects, views, and camera positions were used in all formats.

- Stereoscopic images were generated by placing two virtual cameras in 3ds Max at slightly different horizontal locations relative to the object: one from the left of the targeted view and one from the right. Images from each camera were generated in a single color (red or cyan, depending on the camera). Then, we superimposed the right (cyan) and left (red) images to generate a single image. This stereoscopic image is presented in the same view as in other formats. Subjects wore red–cyan anaglyph glasses to be able to perceive depth from the stereoscopic cues. These images contain both stereo and shading cues.
- 3-D shape-from-shading images were generated in 3ds Max by taking photos of the same views of these objects using a single camera. Images were grayscale and contained shading information as well as contours.
- Line drawings of the same objects and views were generated by extracting outlines of most features in each object image using 3ds Max. Using 3ds Max, the object's material was set to the "Ink N paint" material, which renders objects with flat shading (the paint component) with "inked" borders (the ink component). The ink and paint are two separate components with customizable settings. We adjusted the settings such that just the inked border is rendered, effectively creating line drawings. We manually edited some images to ensure that lines related to a particular feature are consistent across views.
- Silhouettes were generated by taking the outlines of each object image, using the "outline" function in 3ds Max.

Testing the information in the images across formats

We next sought to ensure that, in our stimulus sets, the amount of visual information is sufficient such that given enough training data the visual system could do the discrimination task. We chose a state-of-the-art hierarchical convolutional neural network (CNN) model (Yamins et al., 2014) to quantify the discriminative information in each format. The rationale is that if a high-performance model is able to perform the discrimination task equally well across formats after training with these images, it would indicate that images from different formats contain sufficient information from which human participants can learn. Conversely, if the CNN performs worse on one format compared to the others, it would indicate that there is insufficient information in that format from which the subjects can learn the 3-D structure of objects. In other words, the CNN serves as an ideal observer model.

CNNs are multilayer artificial neural networks that take input images and compute outputs hierarchically with units in each subsequent stage operating on responses from units in the prior stage. CNNs are characterized by a variety of parameters, including (a) discrete-valued architectural parameters, such as the number of layers in the network, the number of filter templates at each layer, the sizes of filter and pooling kernels at each layer, and the pooling exponents used at each layer and (b) continuous-valued filter weights in each layer.

The specific CNN we used in this work is similar to the network described in Yamins et al. (2014), which has been shown to approximate the neural response patterns in the primate ventral visual stream. In the present study, we used this pretrained CNN and trained an additional decision layer using as features of the responses of the CNN's top layer (the layer most similar to IT cortex). The purpose of the decision layer is to implement our behavioral testing. Thus, it determines if two images are of the same or different objects. We conducted training and testing of the CNN decision layer using images from the three monocular formats: silhouettes, line drawings, and shading. Testing and training used images from the same format.

Training

Training the decision layer used the same images as in the human behavioral training. We ran two training

regimens: one with 24 views as in Study 1 and one with seven views as in Study 2. In each training trial, the CNN received two images of random views of the novel objects shown in one format. We then computed the absolute value of the difference in the responses of the CNN's top layer to these two images. Linear classifier weights were then trained on this absolute difference feature vector, using the set of training image pairs, using a standard L2-regularized linear support vector machine from the Scikit-Learn package (Pedregosa et al., 2011). Positive labels corresponded to trials in which the two images were of the same object, and negative labels corresponded to trials in which the images showed different objects.

Testing

After training, we tested the model's ability to determine whether two images of objects rotated 90° apart were of the same object or not. Testing used new views of the objects not shown during training, taken from the human testing phase. Testing used the same object set and format as training. The model's performance was averaged across the four object sets for each format. Error bars were calculated using a bootstrap method during testing. Because results did not significantly differ across training regimens, we report the results of the model when trained with seven object views per format.

Apparatus

All experiments were run on 27-in. iMac computers using the Mac OS Mountain Lion operating system. Stimuli presentation was coded with Matlab (Mathworks.com) and Psychophysics toolbox 3 (http://psychtoolbox.org). Subjects wore red–cyan anaglyph 3-D stereo glasses (Ted Pella, Inc., Redding, CA) when stereoscopic images were presented.

Experimental protocol

Participants went through a testing phase, training phase, and a second testing phase (Figure 3a). This procedure is similar to that used in our previous study (Tian & Grill-Spector, 2015). Each participant underwent four such sequences, one for each set of novel objects (Figure 2a). Each object set was trained and tested on a single format. Each subject participated in training in each of the four structural formats. The ordering of object sets and formats was counterbalanced across subjects. All learning effects we report are within subject.

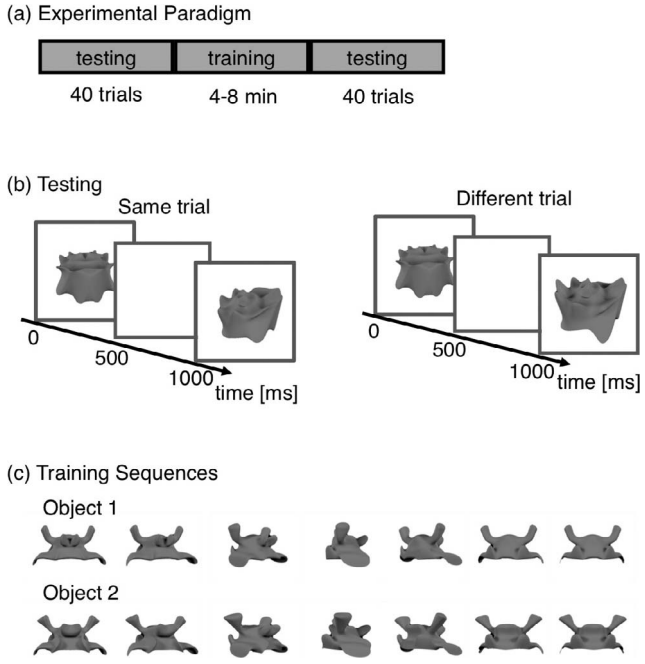

Figure 3. Experimental Procedure. (a) General procedure for each object set consisted of pretraining testing, training, and another post-training testing. (b) Illustration of discrimination testing paradigm: Subjects indicated whether two consecutive images were of the same object or of different objects. In each trial, different views of an object rotated 90° apart were shown. All experiments used rotation around the vertical axis. (c) Illustration of training paradigm: Subjects watched images of object views presented in a sequential order. The two objects in each set were shown in the same format for a given subject and appeared in different formats across subjects.

Object discrimination testing

Subjects participated in a discrimination task before and after learning. In each trial, two object images were presented sequentially, each for 500 ms, separated by a 500-ms blank screen (Figure 3b). The two images were always of objects of a single set and were shown in two views rotated 90° apart from each other (rotation around the vertical axis; images drawn from the entire range of 0°–180° views). In all studies, we tested discrimination performance on new views that are in between the trained views. In other words, we tested how learning generalizes to untrained views. In half of the trials, the two views were of the same object (Figure 3b, left), and in the other half of the trials the two views were of different objects from the same set (Figure 3b, right). The positions of the two objects in a trial were jittered randomly around fixation in a range of up to 1.3°. Object size varied randomly from $10.4° \times 10.5°$ to $12.4° \times 12.4°$. For the task, participants were instructed to indicate via key press on a keyboard if the images were two views of the same object or two views of different objects.

Participants participated in 40 trials per object set in each testing block. In all experiments, subjects' accuracy in the discrimination test was measured using d' (Green & Swets, 1974), which is the discriminability between the two objects in a set calculated by subtracting the z score of the false-alarm rate from the z score of the hit rate.

Unsupervised training

Between testing blocks, participants participated in unsupervised training. During training, subjects observed sequences of views (spanning a 180° rotation) of each of the two objects from a given set. Consecutive views of an object were presented in sequential order (Figure 3c) and were followed by consecutive views of the second object in the set. Each view was presented for 250 ms, and there was no temporal gap between successive views. Participants received four blocks of training with two blocks for each object from a given set. During each training block, the size of the object randomly decreased to 97% of its original size, three to four times. For the task, participants were instructed to watch the images and indicate when the size of the object slightly changed. This task required subjects to maintain their attention on the object. Subjects' performance on this task was not different across formats ($p > 0.05$). The format in which each object set was presented as well as the ordering of formats was counterbalanced across subjects.

Study 1

Twenty-eight participants participated in Study 1, which compared learning of novel objects across four formats: silhouettes, line drawings, shading, and stereo. Training and testing used the same format. Subjects saw each object set in one format. The assignment of format across stimulus sets was counterbalanced across subjects. Learning was done in blocks. In each block, 24 views of an object were shown. Views spanned a 180° rotation around the vertical axis, and consecutive views were 7.5° apart. During discrimination testing, subjects saw views that were 3.75° apart from the adjacent learned views (Figure 3b). During training, views were shown in succession; thus, the object appeared to rotate back and forth three times. After a training block with one object, the subject was trained in an identical manner with the second object of the set. The whole process was repeated twice. Thus, during unsupervised training, participants were shown 24 views × 8 times × 2 blocks × 2 objects = 768 trials lasting a total of about 200 s. Two objects in a given set were always learned and tested on the same structural format. We counterbalanced across subjects which object set was viewed in which condition as well as the order of conditions. All objects sets appeared in each experimental condition an equal amount of time across subjects.

Study 2

Thirty-six participants participated in Study 2, which used a paradigm similar to Study 1 in which subjects participated in learning and testing of novel object discrimination in each of the same four formats. Training and testing were done on the same format for each object set. We implemented three main changes in Study 2 relative to Study 1. First, we reduced the number of training views to seven views, spaced 30° apart. Views covered the same 180° rotation range around the vertical axis. Second, discrimination testing was done on new views that were either 7.5° or 15° away from the trained views. This enabled us to test generalization performance across a larger view range and allowed higher sensitivity to detect differences in performance across formats (Tian & Grill-Spector, 2015). Third, subjects were trained twice with the same object views across two consecutive days in the following sequence: Day 1: testing, training, and testing; Day 2: testing, training, and testing. We increased the amount of training to test whether performance on silhouettes could reach performance on other formats if more training is given. During each day's unsupervised training, participants were shown 7 views × 8 times × 2 blocks × 2 objects = 164 trials per object which is about 2 min per object.

Study 2: Control experiment

Twenty participants participated in the control experiment in Study 2. The general procedure for the control experiment is similar to Study 2. The goal of the control experiment was to test if object training produces a larger improvement in performance than that obtained just from doing the discrimination task twice. Here, we replaced the object-learning task with a baseline task in which no object training was given. Each participant was tested twice in the discrimination task for each object set, but instead of an object-training phase between the tests, the participants performed a dot task. In the dot task, participants were asked to look at a centrally presented dot and press a key when its color changed. The dot was about 3° in visual angle, and its color changed randomly in 2- to 4-s intervals. The duration of the dot task was identical to the duration of object training. No object images were presented during the dot task. Similar to Study 2, the four object sets were presented in each of four different

formats. Format presentation order and object set associated with each format were counterbalanced across subjects.

Study 3

Fifty-two subjects participated in Study 3 in which we compared learning novel objects within a format to learning across formats. Here, we examined learning objects from two formats: 3-D shape from shading and line drawings. We changed the line drawings in this study such that the foreground and background were the same gray level. The experimental paradigm is similar to Study 2. Subjects were shown seven views per object during training. Testing was done on new views that were 7.5° or 15° away from the trained views. Subjects participated in two blocks of training, which were done within the same session. Thus, the experimental sequence for each object set was as follows: discrimination testing, training, discrimination testing, training, and discrimination testing. We implemented a 2×2 design in which object sets were trained on either 3-D shape from shading or line drawings and tested on either 3-D shape from shading or line drawings. Thus, the effect of learning was tested both within format and across formats. For a given object set, the training format and testing format was held constant per subject. We counterbalanced across subjects which object set was viewed in which condition as well as the order of conditions. All object sets appeared in each experimental condition an equal amount of time across subjects.

Statistical significance and effect size

We assessed the statistical significance of learning and format for each experiment using a two-way repeated-measures analysis of variance (rmANOVA) with subjects as the repeated measure. Effect size was estimated with partial η^2, which is a measure of effect size used for ANOVA (Cohen, 1988).

Results

Study 1: Learning to discriminate novel 3-D objects across views is similar across line drawings, shape from shading, and stereo

In Study 1, we tested from which structural cues subjects can learn to recognize novel 3-D objects across views. We manipulated in different experimental conditions the amount of structural information given about the object by varying the format in which the object was presented during unsupervised training (Figure 2b, c). Specifically, during training subjects watched in an unsupervised way 24 views of objects spanning 180° presented in either silhouette, line drawings, shape-from-shading, or stereo formats. Testing was done on novel views rotated 3.75° away from the trained views. During the experiment, the order of formats and the assignment of object sets to a particular format were counterbalanced across subjects (see Methods). Because the four formats contain progressively more depth information, by comparing the effects of training across formats we can determine which structural cues are useful for learning to recognize novel 3-D objects.

Prior to training, participants were poor at discriminating whether two views of objects rotated 90° apart were of the same or different objects with an accuracy of 56.6% ± 1.1%, mean ± *SEM* (d': 0.39 ± 0.07, mean ± *SEM*; Figure 4a). After training, participants' overall accuracy in discrimination increased to 75.0% ± 1.5% (d': 1.68 ± 0.12; Figure 4a). There was a significant overall learning effect as revealed by a main effect of learning in a two-way rmANOVA of d' values with factors of training (pre-/post-training) and format (silhouette/line/shading/stereo), $F(1, 27) = 99.07$, $p < 0.001$, partial $\eta^2 = 0.31$ (Figure 4a). The learning effect (d' post − d' pre) was significant for each of the formats (all ts > 3.3, ps < 0.05; Figure 4b, diamonds). Notably, the amount of learning varied significantly across the different learning formats: learning by object format interaction, $F(3, 81) = 3.37$, $p = 0.02$, partial $\eta^2 = 0.05$. Post hoc t tests revealed that the learning effect for silhouettes was significantly lower than other formats (paired t tests, all ts > 3.27, ps < 0.01). However, there was no significant difference in the learning effect across line, shading, and stereo formats (all ts < 0.73, ps > 0.05).

To test if the lower learning in the silhouettes was a consequence of less information, we use a CNN to evaluate if there is sufficient information in each of the formats to enable successful performance in the discrimination testing (see Methods). After training, the CNN had high discriminability across formats: shaded objects, $d' = 4.17 \pm 0.13$; line drawings, $d' = 4.12 \pm 0.13$; and silhouettes, $d' = 3.96 \pm 0.02$. Importantly, a one-way ANOVA found no significant differences across formats ($p > 0.05$). These analyses indicate that there is sufficient information available to the human visual system to do the task across shading, line drawing, and silhouette formats and that the lower human performance for silhouettes is not due to insufficient visual information.

Results of Study 1 reveal a smaller improvement after unsupervised learning from silhouettes compared

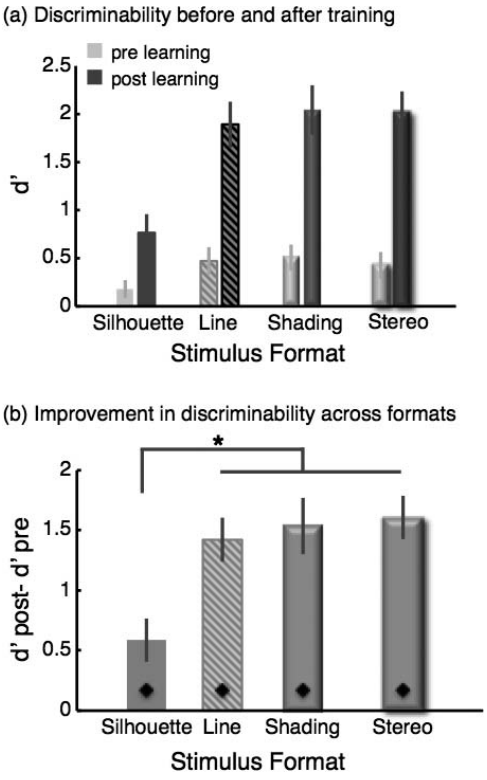

Figure 4. Study 1: Learning to discriminate novel objects from stereoscopic cues, shape from shading, and line drawings is better than learning from silhouettes. (a) Discriminability (d') performance on novel views before (prelearning, light gray) and after (postlearning, dark gray) training with 24 views using silhouettes, line drawings (line), 3-D shape from shading (shading), and stereoscopically presented objects (stereo). Results are averaged across 28 subjects. Error bars indicate standard error of the mean (*SEM*). (b) Learning-related improvement in discriminability (d') for the four object formats. *Improvement for silhouettes is significantly lower than each of the other formats, ts > 3.27, ps < 0.01. ◇Significantly greater than zero improvement, ts > 3.30, ps < 0.05.

to the other three formats. Surprisingly, learning to recognize novel 3-D objects from line drawings is as good as from 3-D shape-from-shading and stereoscopic cues. This suggests that, even though line drawings do not have as rich monocular or binocular depth information as shading or stereo objects, they are as effective for inferring the 3-D structure of an object.

Study 2: Training with fewer views replicates results of Study 1

One possible reason for the similarity in performance across line drawings, shading, and stereo formats is that training used a multitude of views (24 training views), and testing was done on views that were similar (3.75° away) to the trained views. Because training employed a large number of object views, it was perhaps more than sufficient for generating a robust representation of the object "view space," irrespective of the training format. In a prior study, we found that decreasing the number of training views increased our sensitivity to detect differences among training paradigms because less information was given during learning, and testing required generalization over bigger rotations (Tian & Grill-Spector, 2015). Thus, in Study 2, we compared the effects of training across formats when training used only seven views spaced 30° apart, spanning a 180° rotation. Testing was performed on novel views 7.5° or 15° away from the trained views. This design allowed us to better assess the generalization of learning to new views. The potential downside is that training with fewer views may reduce the overall learning effect. Thus, subjects participated in two training sessions on consecutive days to enhance learning. Each object set was shown in a single format across all testing and training sessions. Different object sets were shown in different formats.

As expected, prior to training, participants were poor at discriminating whether two views of objects rotated 90° apart were of the same or different objects with an accuracy of 56.7% ± 1.1% (d': 0.4 ± 0.06; Figure 5a). After the first training, participants' average accuracy in discrimination increased to 68.9% ± 1.7% (d': 1.29 ± 0.13; Figure 5a). There was a significant overall learning effect as shown by the main effect of learning in a two-way rmANOVA of d' during Day 1 with factors of training (pre-/post-training) and format (silhouette/line/shading/stereo), $F(1, 35) = 55.07$, $p < 0.001$, partial $\eta^2 = 0.16$ (Figure 5a). However, training with the four formats did not improve performance equally as shown by a significant interaction between training and format, $F(3, 105) = 3.98$, $p < 0.01$, partial $\eta^2 = 0.04$. There were significant learning effects for line drawings, shape-from-shading, and stereo formats (all ts > 4.11, ps < 0.05 as indicated by diamonds in Figure 5b), but there was no significant learning effect for silhouettes ($t = 1.61$, $p > 0.05$; Figure 5b). Post hoc t tests revealed that differences in learning across formats are driven by lower performance for silhouettes compared to the other three formats (all ts > 2.67, ps < 0.01). In contrast, there was no significant difference between the learning effect for stereo, shading, and lines (all ts < 1.34, ps > 0.05).

On Day 2, the average accuracy began at 68.0% ± 1.9% (d': 1.19 ± 0.14; Figure 5a), which is comparable to the performance after training on Day 1. After the second training, the average performance further increased to 75.8% ± 2.1% (d': 1.88 ± 0.17; Figure 5a). Again, there was a significant learning effect on Day 2 as shown by two-way rmANOVA of d' with factors of training and format (Figure 5a), main effect of learning: $F(1, 35) = 18.12$, $p < 0.001$, partial $\eta^2 = 0.06$.

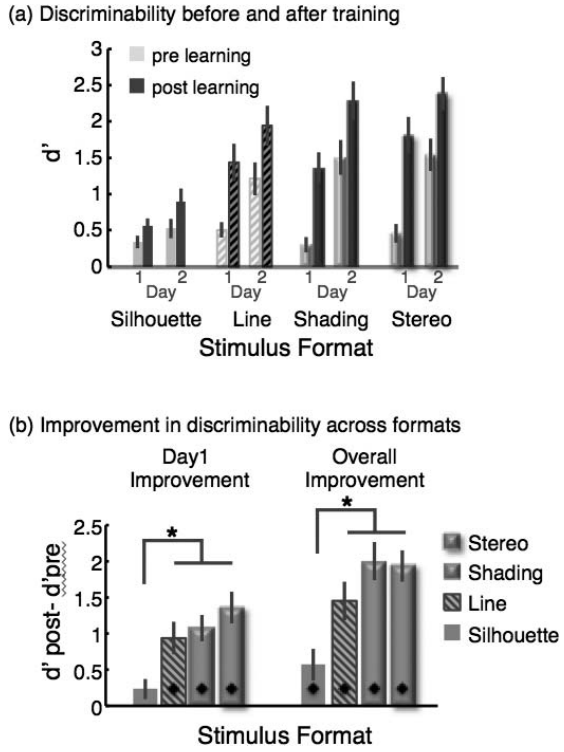

Figure 5. Study 2: Learning to discriminate novel objects from line drawings, shape from shading, and stereo is better than learning from silhouettes. (a) Discriminability (d') performance on novel views before and after training using seven object views. Light gray: Pretraining performance on each day. Dark gray: Post-training performance. Results are averaged across 36 subjects. Error bars: *SEM*. (b) Learning-related improvement in discriminability (d'). Day 1 improvement: Day 1 postlearning − Day 1 prelearning. Overall improvement: Day 2 postlearning − Day 1 prelearning. *Improvement for silhouettes is significantly lower than any of the other formats, ts > 2.57, ps < 0.05. ◇Significantly greater than zero improvement, ts > 1.61, ps < 0.05.

This learning was significant for all formats (all ts > 2.21, ps < 0.05). On Day 2, the improvement was similar across formats as there was no significant interaction between learning and format, $F(3, 105) = 0.44$, $p > 0.05$ (Figure 5a).

To test the overall learning effect, we measured the overall improvement in performance across the 2 days (Day 2 post − Day 1 pre). There was a substantial learning effect across the 2 days: A two-way rmANOVA of d' values postlearning on Day 2 and d' prelearning test on Day 1 with factors of training and format revealed a significant effect of training, $F(1, 35) = 125.03$, $p < 0.001$, partial $\eta^2 = 0.31$ (Figure 5a). We also observed a differential learning effect across formats, revealed by a significant interaction between training and format, $F(3, 105) = 6.23$, $p < 0.001$, partial $\eta^2 = 0.06$. Consistent with the results for Days 1 and 2, the overall improvement across the 2 days differed across formats, $F(3, 105) = 7.39$, $p < 0.001$, partial $\eta^2 = 0.07$ (Figure 5b). The difference was driven by lower improvement for silhouettes compared to the other three formats (post hoc paired t tests: all ts > 2.57, ps < 0.05). However, there was no significant difference in the improvement between line, shading, and stereo formats (all ts < 1.46, ps > 0.05) even though numerically there is a larger improvement for shading objects and stereo objects than line drawings. These results illustrate that learning to recognize objects across views from silhouettes is lower than from line drawings, shape-from-shading, and stereo formats.

Results of Study 2 reveal three findings. First, we found that even when we increased the difficulty of the task by reducing the number of training views compared to Study 1, learning from stereoscopic information did not significantly enhance learning compared to either shading or line drawings. Second, training with a reduced number of views replicated the result of Study 1, in which we did not find a difference between learning from shape from shading and line drawings. Third, learning from silhouettes was consistently the lowest. Extensive training over 2 days did not bring performance based on training with silhouettes even to the level of performance after just one session of training for the line, shading, or stereo formats ($p < 0.05$).

Control experiment with no training

Are the improvements in Study 2 due to the fact that participants performed the discrimination test twice on the same stimuli? To answer this question, we introduced a control experiment in which no object training was given. Participants were tested twice on the object discrimination test for each of the four object formats. In between the two tests for each object set, participants performed a dot color detection task during which no object images were shown (see Methods).

In the first test, participants were poor at discriminating whether two views of objects rotated 90° apart were of the same or different objects with an accuracy of 52.7% ± 1.2%, mean ± *SEM* (d': 0.27 ± 0.06, mean ± *SEM*; Figure 6a). In the second test, participants' overall accuracy in discrimination increased to 58.6% ± 6.3% (d': 0.57 ± 0.08; Figure 6a). There was a small but significant improvement as revealed by a main effect of test repetition in a two-way rmANOVA of d' values with factors of test repetition (first/second) and format (silhouette/line/shading/stereo), $F(1, 19) = 7.84$, $p < 0.05$ (Figure 6a). The test repetition effect (d' post − d' pre) was significant for stereo and line drawings (all ts > 2.70, ps < 0.05, Figure 6b, diamonds) but was only marginally significant for shading objects ($t = 1.81$,

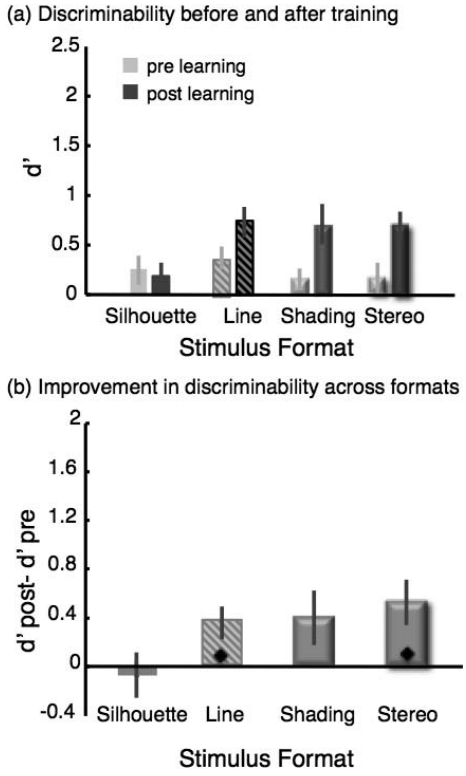

Figure 6. Study 2, control experiment: small or no improvement without object training. (a) Discriminability (d') performance on novel views before and after a dot task conducted for the same duration of object training as in Study 2. Light gray: First testing performance. Dark gray: Second testing performance. Results are averaged across 20 subjects. Error bars: *SEM*. (b) Change in discriminability (d') across the two tests. ◇Significantly greater than zero improvement, ts > 2.70, ps < 0.05.

$p > 0.05$) and not significant for silhouettes ($t = -0.38$, $p > 0.05$). However, the amount of improvement across tests did not differ significantly across learning formats: learning by object format interaction, $F(3, 57) = 0.77$, $p > 0.05$. Post hoc t tests revealed that the improvement for silhouettes was significantly lower than stereo objects (paired t tests, $t = 2.27$, $p < 0.05$) but was not significantly different between silhouettes and shading objects or silhouettes and line drawings (paired t tests, ts < 1.88, ps > 0.05). There was no significant difference in the improvement across line, shading, and stereo formats (all ts < 0.73, ps > 0.05).

The small, 6% improvement obtained with training was significantly lower compared to the 12%–20% improvement obtained with object training (Study 2, main experiment and Study 1). We performed a two-way between-subjects ANOVA with factors of experiment (Study 2 with training in the first day/control experiment without training) and format (silhouette/line/shading/stereo) on the improvement. We found a main effect of experiment, $F(1, 223) = 6.26$, $p < 0.001$, $\eta^2 = 0.03$, suggesting that learning from object images facilitates performance more than just performing the testing twice. Additionally, we found a main effect of format, $F(3, 221) = 15.65$, $p < 0.001$, $\eta^2 = 0.05$, and no interaction between format and experiment, $F(3, 221) = 0.56$, $p > 0.05$.

Results of this control study show that without training with object views there is only a slight improvement during the second test. Critically, this improvement is significantly lower than that obtained with object training. We also find a slight but not significant difference among the four formats, with which there is no improvement for silhouettes and some improvement for the other formats.

Study 3: Learning novel 3-D objects transfers across shape-from-shading and line drawings

Both Studies 1 and 2 show similar learning of novel 3-D objects when presented in shading and line drawing formats. However, it is unknown whether the learning from these two formats uses a common format-invariant representation or a format-specific representation. Therefore, in Study 3, we tested whether learning is format-specific or transfers across line drawing and shading formats by comparing learning within format to learning across formats. For the former, subjects were trained and tested on objects presented in the same format. For the latter, subjects were trained with objects presented using one format (shading or line drawing) and tested on objects presented in the other format. Here we employed a sequence of two trainings blocks, each using seven views of each object, interleaved with testing blocks with untrained views (see Methods). All training and testing were done on the same day. We considered two possible outcomes: If learning novel objects from shape from shading and line drawings generated a common object representation, then the learning effect across formats would be comparable to the learning effect within format. However, if learning novel objects depended on format-specific information, the learning effect within format would be larger than across formats.

As expected, prior to training participants were poor at discriminating whether two views of objects rotated 90° apart were of the same object or different objects as accuracy was 57.4% ± 0.8% (d': 0.4 ± 0.05; Figure 7a). After the first training block, participants' mean accuracy in discrimination increased to 67.9% ± 1.1% (d': 1.22 ± 0.10; Figure 7a), and after the second training block, participants' accuracy further increased to 73.3% ± 1.6% (d': 1.80 ± 0.15).

To examine whether learning format and testing format generate different levels of performance across the two training blocks, we performed a three-way

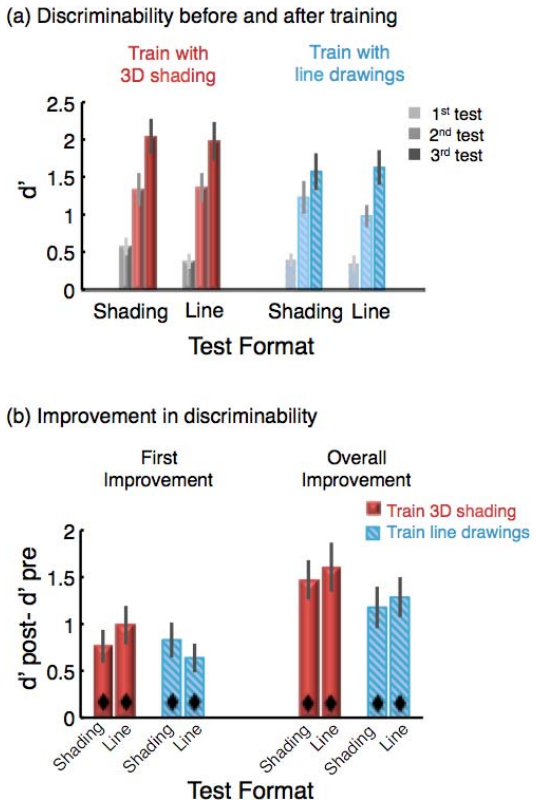

Figure 7. Study 3: Learning to discriminate novel objects transfers across shape-from-shading and line drawing formats. (a) Discriminability (d') performance for objects presented using 3-D shape from shading (shading) or line drawings (line) before and after one and two sessions of training. Red: Training with 3-D shape from shading. Blue: Training with line drawing. Results are averaged across 52 subjects. Error bars: *SEM*. (b) Learning-related improvement in d' for shaded objects is not different if training used shading or line drawings. Likewise, improvement in d' for line drawings is not different when training with line drawings or shading. ◇Significantly greater than zero improvement, ts > 4.13, ps < 0.05.

rmANOVA with factors of testing block (pre–/post–first training/post–second training), learning format (shading/line), and testing format (shading/line). First, we replicated the learning effect found in the first two studies as shown by a significant main effect of learning, $F(2, 102) = 84.27$, $p < 0.001$, $\eta^2 = 0.03$ (Figure 7a). Second, we found that learning shading or line formats generated a similar amount of learning across two learning blocks because there was no significant interaction between testing block and learning format, $F(2, 102) = 1.13$, $p > 0.05$. Third, the learning-related improvement also did not depend on which format the testing used as there was no interaction between testing block and testing format, $F(2, 102) = 1.13$, $p > 0.05$. There was also no significant interaction between learning format and testing format, $F(1, 51) = 0$, $p > 0.05$, as well as no significant interaction among all three factors, $F(2, 102) = 0.51$, $p > 0.05$. We observed a main effect of learning format, $F(1, 51) = 4.81$, $p < 0.05$, $\eta^2 = 0.0001$, as performance was higher when learning with shading compared to learning with line drawings. There was no main effect for testing format, $F(1, 51) = 0.43$, $p > 0.05$.

We further tested whether the learning effect is different across learning formats and testing formats after one learning block (Figure 7b, left panel) and after two learning blocks (Figure 7b, right panel). We again found that learning objects from shading and line drawings generated the same amount of learning after both one and two learning blocks: A two-way rmANOVA with the factors of learning format (shading/line) and testing format (shading/line) for the improvement found no main effect of learning format: after one learning block, $F(1, 51) = 0.61$, $p > 0.05$; after two learning blocks, $F(1, 51) = 1.84$, $p > 0.05$. Additionally, there was no main effect of testing format: after one learning block, $F(1, 51) = 0.01$, $p > 0.05$; after two learning blocks, $F(1, 51) = 0.39$, $p > 0.05$. Finally, there was no interaction between learning format and testing format: after one learning block, $F(1, 51) = 1.31$, $p > 0.05$; after two learning blocks, $F(1, 51) = 0$, $p > 0.05$. These analyses indicate there was similar amount of learning irrespective of which format was learned and which format was tested across both one and two learning blocks. In other words, the learning completely transferred across formats.

In Study 3, we found that learning objects from shading and line drawings generated similar improvement both within and across formats. This suggests that learning the 3-D structure of novel objects from either line drawings or shading generates a common representation of novel objects that can be used for recognition.

Discussion

We examined which structural information people use during unsupervised learning to obtain viewpoint-invariant recognition of novel objects. Our experiments used a difficult discrimination task among objects with similar parts and configuration, large rotations that affected the visible features of the object across views, and testing on new views not shown during training. Surprisingly, we found that unsupervised learning from example views of objects presented in line drawings, shading, or stereo formats generated a similar learning effect. Moreover, this learning generalized to new views that were not seen during training, suggesting that subjects learned the 3-D structure of objects. In contrast, learning with silhouettes generated a significantly lower performance across experiments, indicat-

ing that information just from the external contour is insufficient for robust learning of objects' 3-D structure. These findings suggest that formats that contain both internal and external shape information enable learning the 3-D structure of objects and, furthermore, that learning viewpoint-invariant recognition does not require complete and rich structural information during training. Strikingly, in Study 3 we found that learning was not only similar across line drawings and shading formats, but learning from one format transferred to the other format and vice versa. This suggests that the learned object representations are cue-invariant at least across monocular line and shading cues.

Three aspects of our findings are particularly notable. First, learning the 3-D structure of novel objects is possible from viewing in an unsupervised way a handful of line drawings showing the objects in different views. Second, this learning is as effective as learning from a more realistic depiction of the object containing shading cues. Third, learning completely transfers across line drawings and shading cues. In the following sections, we elaborate on these finding by (a) discussing the implications of our findings on the sources of information that are needed to learn the 3-D structure of objects and (b) considering the implications of our findings of transfer of learning across cues on the nature of object representations that are generated during training.

Which sources of information are needed to learn the 3-D structure of objects?

Shape information is sufficient for building representations of 3-D novel objects

Marr (1982) originally proposed that surface information, such as that obtained from shading or stereopsis, is key to building a complete 3-D representation of an object. However, later research (Biederman, 1987; Biederman & Ju, 1988; Rossion & Pourtois, 2004) showed that basic-level recognition from line drawings is particularly efficient and not worse than recognition from color photographs of objects. Our findings not only support prior research that demonstrated shape information rather than surface information is the key source of information for object recognition (Biederman, 1987; Marr & Nishihara, 1978; Rossion & Pourtois, 2004), but extend these finding by showing that shape information is also key for subordinate recognition of objects that have similar parts and configuration. Furthermore, we show here for the first time that learning the 3-D object structure is as efficient from impoverished examples of object views, such as line drawings, as from more realistic examples of object views that contain explicit 3-D structure from shading and stereo cues. Our data indicate that it is not necessary to have veridical examples of an object to learn its 3-D structure. Instead, an abstracted version of the object that maintains the key shape information, for example, from both external and internal contours, is sufficient.

Why is learning from silhouettes lower than from the other formats?

In contrast to the efficient learning of object structure from line drawings, our data show that learning from silhouettes is insufficient to produce a robust representation of the 3-D structure of objects. This finding is not entirely surprising because the silhouettes lack important information about the structure of objects, including most depth cues and internal features. Thus, that people can recognize objects just from silhouettes at all is amazing because in theory a single silhouette could correspond to an infinite number of possible 3-D objects. Indeed, artists have appreciated this amazing ability by depicting complex objects in silhouette format from the dawn of mankind (e.g., Lachaux cave paintings; Figure 1a) to modern abstract art (e.g., Picasso's painting; Figure 1b).

Our behavioral data is consistent with prior studies that showed reduced recognition of objects from silhouettes compared to line drawings. Previous research showed that whether objects can be recognized from their silhouettes depends on two factors: The first factor is prior knowledge. Silhouettes of familiar objects, animate objects, and canonical views are better recognized than other silhouettes (Lawson, 1999; Lloyd-Jones & Luckhurst, 2002; Mitsumatsu & Yokosawa, 2002; Newell & Findlay, 1997). The second factor is the number of features available in the silhouette. Silhouettes can be recognized if they contain a sufficient number of features that are not obscured by the external contour and those features enable an unambiguous interpretation of the silhouette (Hayward, 1998; Hayward, Wong, & Spehar, 2005). The available features and the complexity of features also depend on specific stimuli and specific views. For example, there are views in which a silhouette of a cylinder can be identical to that of a rectangular prism cube. In this case, the two objects will be indistinguishable based on their silhouettes. However, curvature information from other cues, such as stereo, shape from shading, or line drawings that depict luminance edges, will give information about the curvature of the surface and can resolve this ambiguity.

Given that in our experiments objects were nonliving and novel, familiarity did not play a role in recognition performance. To evaluate if there was sufficient feature information in our silhouettes to enable learning, we

tested the ability of a CNN that was optimized to the primate visual object recognition (Rajalingham, Schmidt, & DiCarlo, 2015; Yamins et al., 2014) to perform the same object discrimination task as our human participants across views for silhouettes, line drawings, and shading formats (see Methods). We found that the CNN performed equally well across the silhouette, line, and shading formats. Thus, in the current experiment, we generated silhouettes that had, in principle, sufficient information to enable generalization across views. However, subjects' performance after learning from silhouettes was lower than for line drawings and shape from shading. This suggests that human observers do not use this information during learning as efficiently as a computer algorithm, perhaps because they may not be naturally attuned to this source of information. It is still possible that after additional and more intensive training with silhouettes, subjects might be able to use the external contour information to recognize objects. Nevertheless, despite the lower object recognition from silhouettes in the current experiment, learning from silhouettes still improved participants' performance, which indicates that some information from the external contour contributed to learning.

Why did stereo information not improve learning of the 3-D object structure?

Theories of object perception suggest that people rely on the most informative cues to reach the optimal performance (Kersten, Mamassian, & Yuille, 2004). Intuitively, stereo seems to be a particularly useful source of information to learn the 3-D structure of a novel object because it contains explicit depth information. Indeed, several studies report a stereo benefit in object recognition (Bennett & Vuong, 2006; Burke, 2005; Humphrey & Khan, 1992; Y. L. Lee & Saunders, 2011; Norman, Todd, & Orban, 2004). However, in our study stereopsis did not improve learning the 3-D structure of objects beyond monocular cues.

We suggest that stereo provides a behavioral benefit only if it resolves ambiguity in the interpretation of the 3-D structure of objects that cannot be resolved from other sources of information. In our study, there was likely enough structural information in the shaded objects such that providing additional depth information from stereo did not add critical diagnostic information for the task. In contrast, previous reports showed a stereo benefit in object recognition in situations in which monocular information did not provide an unambiguous interpretation of the 3-D structure of the objects. Stereo benefits in object recognition have been reported for recognition across views for (a) paper clip objects that lack self-occlusion and have an ambiguous 3-D structure (Bennett & Vuong, 2006; Burke, 2005; Edelman & Bülthoff, 1992) and (b) objects with identical silhouettes in which internal features are determined from surface reflectance (Lee & Saunders, 2011; Norman et al., 2004). Thus, stereo information seems to be beneficial for 3-D object recognition when other depth cues are ambiguous or lacking. Finally, stereo information may be beneficial for other tasks that require absolute depth information, such as grasping (Melmoth, Finlay, Morgan, & Grant, 2009) or fast interactions with objects in depth (e.g., playing ping-pong; Cottereau, McKee, and Norcia, 2014).

What is the nature of representations that are generated during learning?

Learning generates cue-invariant representations

Our finding that object recognition performance after learning from line drawing completely transferred to shaded objects and vice versa suggests that the internal representation generated due to learning is cue-independent—at least across line and shading cues. However, from our behavioral measurements, we cannot infer where in the visual processing stream this cue-invariance occurs. One possibility is that learning generated object representations in the ventral occipitotemporal cortex that are shape-based and independent of the format of the visual input. This hypothesis is supported by neuroscience findings of cue-invariant representations of object shape in object-selective regions in the human lateral occipital complex (LOC; Appelbaum, Wade, Vildavski, Pettet, & Norcia, 2006; Georgieva, Todd, Peeters, & Orban, 2008; Grill-Spector et al., 1998; Kourtzi, Erb, Grodd, & Bülthoff, 2003; Kourtzi & Kanwisher, 2000, 2001; Malach et al., 1995; Vinberg & Grill-Spector, 2008) and the monkey inferotemporal cortex (Fujita, Tanaka, Ito, & Cheng, 1992; Sary et al., 1993). A second possibility is that the information arriving to the LOC is already edge-based. Because similar edges are generated from line drawings and shaded objects by processing earlier in the visual hierarchy, a common representation may be generated in the LOC. This account is supported by electrophysiological evidence that neural processing in early visual areas extracts edge information from the retinal input (Hubel & Wiesel, 1965), and this information is propagated to higher level regions in the ventral occipitotemporal cortex whereas surface and depth information is processed in parietal regions in the dorsal stream (Backus, Fleet, Parker, & Heeger, 2001; Kravitz, Saleem, Baker, Ungerleider, & Mishkin, 2012; Vin-

berg & Grill-Spector, 2008; Welchman, Deubelius, Conrad, Bülthoff, & Kourtzi, 2005).

Future experiments using electrophysiological measurements in the ventral occipitotemporal cortex as well as computational models, such as CNNs, can examine the nature of representations generated by training. For example, electrophysiological recordings of the same neural population before and after training (McMahon, Bondar, Afuwape, Ide, & Leopold, 2014) can test whether the representations learned are cue-specific or cue-invariant. Similarly, computational simulations may examine not only the output of the system before and after training, but also examine the representation in intermediate layers of the network to test if they are format-specific or format-independent. Finding that intermediate layers show similar responses across cues would suggest the cue-invariance is generated in intermediate processing stages of the ventral stream. In contrast, finding different responses to line drawings and shaded objects in intermediate layers but cue-invariant representations in output layers that are associated with LOC or IT would indicate that cue-invariant shape-based representations are built in high-level ventral stream regions.

Are the learned representations 2-D or 3-D?

A remaining question is what kinds of internal representations enable participants to discriminate objects across large rotations and novel views? One possibility is that learning generates a 3-D internal representation of objects (Kellman, Garrigan, & Shipley, 2005; Liu & Kersten, 1998; Liu, Knill, & Kersten, 1995). In particular, Kellman et al. (2005) proposed a theory of 3-D relatability that explains how local and global 3-D structures can be represented from 2-D line drawings of objects. According to this model, objects are represented by contours in the 3-D space and the connections among them. This model specifies the orientation and position of each edge in 3-D as well as each edge's connection to other edges. According to this model, generalization to new views is possible by interpolating among contours in 3-D.

Another possibility supported by psychophysical and neural investigations suggests that learning generated view-tuned units centered on the trained views with some tuning width and partial overlap to nearby views (Bülthoff & Edelman, 1992; Bülthoff et al., 1995; Logothetis & Pauls, 1995; Tarr et al., 1998). When a sufficient number of views is learned and linked via temporal or spatiotemporal proximity and the tunings of the view-tuned units overlap, the entire view space of the object is represented (Földiák, 1991; Tian & Grill-Spector, 2015; Wallis & Bülthoff, 2001). Recognition of the 3-D object is based on distributed responses across this population of view-tuned units (DiCarlo & Cox, 2007). As we found that learning transfers across line and shading formats, we hypothesize that if learning generated view-tuned units, these units will represent shape information, independent of format rather than containing a veridical "snapshot" of the trained views (see also Ullman, 1989).

Future studies parametrically varying the number of training views across formats may be useful in testing these theoretical models of object representations as view-based representations are likely to be more sensitive to number of training views than 3-D representations.

Conclusions

We investigated what structural information is used during unsupervised learning from example views to obtain a 3-D representation of an object. We found that learning objects from line drawings, shading, and stereo generated a similar improvement in object recognition across views, suggesting that the combination of internal and external shape information is sufficient for learning the 3-D structure of objects. Strikingly, not only was performance similar across line drawings and shading cues, but learning transferred across line drawings and shading cues.

In sum, these findings have advanced our understanding regarding what type of structural information is critical for learning the 3-D structure of an object from 2-D information in retinal images and has important implications for psychological and computational theories of object recognition.

Keywords: view-invariant recognition, unsupervised learning, structural cues, cue invariance

Acknowledgments

This research was supported by NEI Grant 1 R01 EY019279-01A1 to KGS, Ric Weiland Graduate Fellowship to MT, and Stanford Center for Mind Brain and Computation trainee support to MT. We thank James Manson, Thomas Rasmussen, and Darby Schumacher for their help in running behavioral experiments.

Commercial relationships: none.
Corresponding author: Moqian Tian.
Email: moqian@alumni.stanford.edu.
Address: Department of Psychology, Stanford University, Stanford, CA, USA.

References

Appelbaum, L. G., Wade, A. R., Vildavski, V. Y., Pettet, M. W., & Norcia, A. M. (2006). Cue-invariant networks for figure and background processing in human visual cortex. *Journal of Neuroscience, 26*(45), 11695–11708, doi:10.1523/JNEUROSCI.2741-06.2006.

Backus, B. T., Fleet, D. J., Parker, A. J., & Heeger, D. J. (2001). Human cortical activity correlates with stereoscopic depth perception. *Journal of Neurophysiology, 86*(4), 2054–2068.

Bennett, D. J., & Vuong, Q. C. (2006). A stereo advantage in generalizing over changes in viewpoint on object recognition tasks. *Perceptual Psychophysics, 68*(7), 1082–1093, doi:10.1167/6.6.313.

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

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

Bülthoff, H. H., & Edelman, S. (1992). Psychophysical support for a two-dimensional view interpolation theory of object recognition. *Proceedings of the National Academy of Sciences, USA, 89*(1), 60–64.

Bülthoff, H. H., Edelman, S. Y., & Tarr, M. J. (1995). How are three-dimensional objects represented in the brain? *Cerebral Cortex, 5*(3), 247–260, doi:10.1093/cercor/5.3.247.

Bülthoff, H. H., & Mallot, H. A. (1988). Integration of depth modules: Stereo and shading. *Journal of the Optical Society of America A, 5*(10), 1749–1758, doi:10.1364/JOSAA.5.001749.

Burke, D. (2005). Combining disparate views of objects: Viewpoint costs are reduced by stereopsis. *Visual Cognition, 12*(5), 705–719, doi:10.1080/13506280444000463.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. New York: Routledge Academic.

Cottereau, B. R., McKee, S. P., & Norcia, A. M. (2014). Dynamics and cortical distribution of neural responses to 2D and 3D motion in human. *Journal of Neurophysiology, 111*(3), 533–543, doi:10.1152/jn.00549.2013.

DiCarlo, J. J., & Cox, D. D. (2007). Untangling invariant object recognition. *Trends in Cognitive Sciences, 11*(8), 333–341, doi:10.1016/j.tics.2007.06.010.

Edelman, S., & Bülthoff, H. H. (1992). Orientation dependence in the recognition of familiar and novel views of three-dimensional objects. *Vision Research, 32*(12), 2385–2400, doi:10.1016/0042-6989(92)90102-O.

Földiák, P. (1991). Learning invariance from transformation sequences. *Neural Computation, 3*(2), 194–200.

Fujita, I., Tanaka, K., Ito, M., & Cheng, K. (1992, Nov 26). Columns for visual features of objects in monkey inferotemporal cortex. *Nature, 360*(6402), 343–346, doi:10.1038/360343a0.

Georgieva, S. S., Todd, J. T., Peeters, R., & Orban, G. A. (2008). The extraction of 3D shape from texture and shading in the human brain. *Cerebral Cortex, 18*(10), 2416–2438, doi:10.1093/cercor/bhn002.

Green, D. M., & Swets, J. A. (1974). *Signal detection theory and psychophysics*. Huntington, NY: R. E. Krieger Pub. Co.

Grill-Spector, K., Kushnir, T., Edelman, S., Itzchak, Y., & Malach, R. (1998). Cue-invariant activation in object-related areas of the human occipital lobe. *Neuron, 21*(1), 191–202, doi:10.1016/S0896-6273(00)80526-7.

Hayward, W. G. (1998). Effects of outline shape in object recognition. *Journal of Experimental Psychology: Human Perception and Performance, 24*(2), 427–440, doi:10.1037/0096-1523.24.2.427.

Hayward, W. G., & Tarr, M. J. (1997). Testing conditions for viewpoint invariance in object recognition. *Journal of Experimental Psychology: Human Perception and Performance, 23*(5), 1511–1521, doi:10.1037/0096-1523.23.5.1511.

Hayward, W. G., Wong, A. C., & Spehar, B. (2005). When are viewpoint costs greater for silhouettes than for shaded images? *Psychonomic Bulletin and Review, 12*(2), 321–327, doi:10.3758/BF03196379.

Hubel, D. H., & Wiesel, T. N. (1965). Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat. *Journal of Neurophysiology, 28*, 229–289.

Humphrey, G. K., & Khan, S. C. (1992). Recognizing novel views of three-dimensional objects. *Canadian Journal of Psychology, 46*(2), 170–190, doi:10.1037/h0084320.

Kastner, S., De Weerd, P., & Ungerleider, L. G. (2000). Texture segregation in the human visual cortex: A functional MRI study. *Journal of Neurophysiology, 83*(4), 2453–2457.

Kellman, P. J., Garrigan, P., & Shipley, T. F. (2005).

Object interpolation in three dimensions. *Psychological Review, 112*(3), 586–609, doi:10.1037/0033-295X.112.3.586.

Kersten, D., Mamassian, P., & Yuille, A. (2004). Object perception as Bayesian inference. *Annual Review of Psychology, 55*, 271–304, doi:10.1146/annurev.psych.55.090902.142005.

Kourtzi, Z., Erb, M., Grodd, W., & Bülthoff, H. H. (2003). Representation of the perceived 3-D object shape in the human lateral occipital complex. *Cerebral Cortex, 13*(9), 911–920, doi:10.1093/cercor/13.9.911.

Kourtzi, Z., & Kanwisher, N. (2000). Cortical regions involved in perceiving object shape. *Journal of Neuroscience, 20*(9), 3310–3318.

Kourtzi, Z., & Kanwisher, N. (2001, Aug 24). Representation of perceived object shape by the human lateral occipital complex. *Science, 293*(5534), 1506–1509, doi:10.1126/science.1061133.

Kravitz, D. J., Saleem, K. S., Baker, C. I., Ungerleider, L. G., & Mishkin, M. (2012). The ventral visual pathway: An expanded neural framework for the processing of object quality. *Trends in Cognitive Sciences, 17*(1), 26–49, doi:10.1016/j.tics.2012.10.011.

Lawson, R. (1999). The effects of view in depth on the identification of line drawings and silhouettes of familiar objects: Normality and pathology. *Visual Cognition, 6*(2), 165–195.

Lee, Y. L., & Saunders, J. A. (2011). Stereo improves 3D shape discrimination even when rich monocular shape cues are available. *Journal of Vision, 11*(9):6, 1–12, doi:10.1167/11.9.6. [PubMed] [Article]

Liu, C. H., Ward, J., & Young, A. W. (2006). Transfer between two-and three-dimensional representations of faces. *Visual Cognition, 13*(1), 51–64, doi:10.1080/13506280500143391.

Liu, Z., & Kersten, D. (1998). 2D observers for human 3D object recognition? *Vision Research, 38*(15–16), 2507–2519, doi:10.1016/S0042-6989(98)00063-7.

Liu, Z., Knill, D. C., & Kersten, D. (1995). Object classification for human and ideal observers. *Vision Research, 35*(4), 549–568, doi:10.1016/0042-6989(94)00150-K.

Lloyd-Jones, T. J., & Luckhurst, L. (2002). Outline shape is a mediator of object recognition that is particularly important for living things. *Memory & Cognition, 30*(4), 489–498, doi:10.3758/BF03194950.

Logothetis, N. K., & Pauls, J. (1995). Psychophysical and physiological evidence for viewer-centered object representations in the primate. *Cerebral Cortex, 5*(3), 270–288, doi:10.1093/cercor/5.3.270.

Malach, R., Reppas, J. B., Benson, R. R., Kwong, K. K., Jiang, H., Kennedy, W. A., & Tootell, R. B. (1995). Object-related activity revealed by functional magnetic resonance imaging in human occipital cortex. *Proceedings of the National Academy of Sciences, USA, 92*(18), 8135–8139.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. New York: Henry Holt and Co., Inc.

Marr, D., & Nishihara, H. K. (1978). Representation and recognition of the spatial organization of three-dimensional shapes. *Proceedings of the Royal Society B: Biological Sciences, 200*(1140), 269–294.

McMahon, D. B., Bondar, I. V., Afuwape, O. A., Ide, D. C., & Leopold, D. A. (2014). One month in the life of a neuron: Longitudinal single-unit electrophysiology in the monkey visual system. *Journal of Neurophysiology, 112*(7), 1748–1762, doi:10.1152/jn.00052.2014.

Melmoth, D. R., Finlay, A. L., Morgan, M. J., & Grant, S. (2009). Grasping deficits and adaptations in adults with stereo vision losses. *Investigative Visual Science & Opthalmology, 50*(8), 3711–3720. [PubMed] [Article]

Mendola, J. D., Dale, A. M., Fischl, B., Liu, A. K., & Tootell, R. B. H. (1999). The representation of illusory and real contours in human cortical visual areas revealed by functional magnetic resonance imaging. *Journal of Neuroscience, 19*(19), 8560–8572.

Mitsumatsu, H., & Yokosawa, K. (2002). How do the internal details of the object contribute to recognition? *Perception, 31*(11), 1289–1298, doi:10.1068/p3421.

Nakayama, K., Shimojo, S., & Silverman, G. H. (1989). Stereoscopic depth: Its relation to image segmentation, grouping, and the recognition of occluded objects. *Perception, 18*(1), 55–68, doi:10.1068/p180055.

Nefs, H. T. (2008). Three-dimensional object shape from shading and contour disparities. *Journal of Vision, 8*(11):11, 1–16, doi:10.1167/8.11.11. [PubMed] [Article]

Nefs, H. T., & Harris, J. M. (2007). Vergence effects on the perception of motion-in-depth. *Experimental Brain Research, 183*(3), 313–322, doi:10.1007/s00221-007-1046-5

Newell, F. N., & Findlay, J. M. (1997). The effect of depth rotation on object identification. *Perception, 26*(10), 1231–1257, doi:10.1068/p261231.

Norman, J. F., Todd, J. T., & Orban, G. A. (2004). Perception of three-dimensional shape from specular highlights, deformations of shading, and other types of visual information. *Psychological Science, 15*(8), 565–570, doi:10.1111/j.0956-7976.2004.00720.x.

Pasqualotto, A., & Hayward, W. G. (2009). A stereo disadvantage for recognizing rotated familiar objects. *Psychonomic Bulletin and Review, 16*(5), 832–838, doi:10.3758/PBR.16.5.832.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., & Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research, 12*, 2825–2830.

Rajalingham, R., Schmidt, K., & DiCarlo, J. J. (2015). Comparison of object recognition behavior in human and monkey. *Journal of Neuroscience, 35*(35), 12127–12136, doi:10.1523/JNEUROSCI.0573-15.2015.

Rosch, E. (1999). Principles of categorization. In E. Margolis & S. Laurence (Eds.), *Concepts: Core readings* (pp. 189–206). Cambridge, MA: MIT Press.

Rossion, B., & Pourtois, G. (2004). Revisiting Snodgrass and Vanderwart's object pictorial set: The role of surface detail in basic-level object recognition. *Perception, 33*(2), 217–236, doi:10.1068/p5117.

Sary, G., Vogels, R., & Orban, G. A. (1993, May 14). Cue-invariant shape selectivity of macaque inferior temporal neurons. *Science, 260*(5110), 995–997, doi:10.1126/science.8493538.

Tarr, M. J., Williams, P., Hayward, W. G., & Gauthier, I. (1998). Three-dimensional object recognition is viewpoint dependent. *Nature Neuroscience, 1*(4), 275–277, doi:10.1038/1089.

Tian, M., & Grill-Spector, K. (2015). Spatiotemporal information during unsupervised learning enhances viewpoint invariant object recognition. *Journal of Vision, 15*(6):7, 1–13, doi:10.1167/15.6.7. [PubMed] [Article]

Ullman, S. (1989). Aligning pictorial descriptions: An approach to object recognition. *Cognition, 32*(3), 193–254, doi:10.1093/cercor/5.3.270.

Vinberg, J., & Grill-Spector, K. (2008). Representation of shapes, edges, and surfaces across multiple cues in the human visual cortex. *Journal of Neurophysiology, 99*(3), 1380–1393, doi:10.1152/jn.01223.2007.

Wallis, G., & Bülthoff, H. H. (2001). Effects of temporal association on recognition memory. *Proceedings of the National Academy of Sciences, USA, 98*(8), 4800–4804, doi:10.1073/pnas.071028598.

Welchman, A. E., Deubelius, A., Conrad, V., Bülthoff, H. H., & Kourtzi, Z. (2005). 3D shape perception from combined depth cues in human visual cortex. *Nature Neuroscience, 8*(6), 820–827, doi:10.1038/nn1461.

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