# Seeing in the dark: Testing deep neural network and analysis-by-synthesis accounts of 3D shape perception with highly degraded images

**Hakan Yilmaz (hakan.yilmaz@yale.edu)**
Psychology, Yale University, New Haven, USA

**Gargi Singh (garsinhere@gmail.com)**
Indian Institute of Technology, Kanpur, Dhanbad, India

**Bernhard Egger (egger@mit.edu)**
Brain and Cognitive Science, MIT, Cambridge, Massachusetts, United States

**Joshua B. Tenenbaum (jbt@mit.edu)**
Brain and Cognitive Science, MIT, Cambridge, Massachusetts, United States

**Ilker Yildirim (ilker.yildirim@yale.edu)**
Psychology, Yale University, New Haven, USA

## Abstract

The visual system does not require extensive signal in its inputs to compute rich, three-dimensional (3D) shape percepts. Even under highly degraded stimuli conditions, we can accurately interpret images in terms of volumetric objects. What computations support such broad generalization in the visual system? To answer, we exploit two degraded image modalities – silhouettes and two-tone "Mooney" images – alongside regular shaded images. We test two distinct approaches to vision: deep networks for classification and analysis-by-synthesis for scene inference. Deep networks perform substantially sub-human even after training on 18 times more images per category compared to the existing large-scale image sets for object classification. We also present a novel analysis-by-synthesis architecture that infers 3D scenes from images via optimization in a differentiable, physically-based renderer. This model also performs substantially sub-human. Nevertheless, both approaches can explain some of the key behavioral patterns. We discuss the insights these results provide for reverse-engineering visual cognition.

**Keywords:** analysis-by-synthesis; differentiable rendering; silhouette; mooney; shape-from-x

## Introduction

Vision scientists studied many cues as possible routes to 3D object shape perception, including texture gradients, shading patterns, contour geometry, highlights, stereo disparity, and motion parallax (Bulthoff & Yuille, 1991). A striking observation across these studies is that the visual system does not require much signal in its inputs to construct rich, three-dimensional (3D) shape percepts. Even under highly degraded or atypical stimuli conditions, e.g., under dim light, behind occluders, or at unusual viewpoints, we can accurately interpret images in terms of an underlying volumetric object. A classical example of such degraded stimulus conditions is the two-tone, black and white "Mooney" images (e.g., Mooney, 1957; Moore & Cavanagh, 1998) (Fig. 1). These two-tone images lack shading, hue, or texture cues entirely, and can distort outline and contour information. Yet, most observers report a strong sense of seeing 3D objects and surfaces in such images. These observations do not just reflect curiosities about certain "corner cases", but instead they illustrate how the visual system operates at the "long-tail" of what can happen in the world. Here we ask: What computations and representations are needed to accomplish such versatile processing and broad generalization in the visual system?

To answer, we exploit two highly degraded image modalities: silhouettes and Mooney images, alongside with regular shaded images. In a behavioral experiment, we first establish that humans robustly generalize object shape information despite simultaneous viewpoint and image modality differences. We hypothesize that such broad generalization can be understood in terms of three key elements: *(i)* a hypothesis space over 3D object shapes, *(ii)* an internal model of the optical or graphics processes by which 3D scenes are projected and filtered to individual image modalities, and *(iii)* an efficient method to solve the inverse problem of inferring 3D shapes from image inputs.

We implement this hypothesis in a novel analysis-by-synthesis (AbS) architecture that builds upon and extends recent advances in computer graphics: deep implicit surface representations that are learned from an object repository to capture an expressive hypothesis space over shapes (Park, Florence, Straub, Newcombe, & Lovegrove, 2019); a physically-based differentiable renderer to capture graphics processes (Nimier-David, Vicini, Zeltner, & Jakob, 2019); and optimization through this renderer to infer shapes that best reconstruct input images (Remelli et al., 2020). In this model, perception of shape is a form of causal inference, going back from images to their underlying 3D scenes. Thus, in principle this model can generalize across viewpoints and image modalities as long as it can account for such variation using its hypothesis space. Indeed, we find that this model is able to compute consistent 3D shape percepts across both regular and degraded image modalities.

We compare this analysis-by-synthesis architecture to the current standard approaches in vision based on deep convolu-


Figure 1: Example stimuli used in the experiment showing shaded images (top row), Mooney images (middle row), and silhouettes (bottom row) from four different object categories including tables, airplanes, chairs, and cars and viewpoints (canonical and randomized).

tional neural networks (DCNNs; LeCun, Bengio, & Hinton, 2015) trained to classify (object category or identity information) through a hierarchy of non-linear transformations. Unlike our analysis-by-synthesis model, which aims to explain the variation in the image via a generative model, the functional goal in these models is to "untangle" the category (or identity) information, learning to classify objects despite variation in viewing conditions (DiCarlo, Zoccolan, & Rust, 2012). These models have not only enabled impressive engineering applications (LeCun et al., 2015), but they can also explain aspects of the variance in neural data along the visual processing hierarchy (Yamins et al., 2014; Khaligh-Razavi & Kriegeskorte, 2014). However, they are brittle to novel image modalities (e.g., line drawings, "stylized" images, etc.; Geirhos et al., 2018) and require further training (or fine-tuning) in each such domain to attain invariance. In this work, we explore how much training is needed for DCNNs to generalize across the viewpoints and image modalities we consider here.

We evaluate these models by comparing their performance to that of human observers on a shape generalization task across viewpoints and image modalities. Extending previous behavioral work (Hayward, Tarr, & Corderoy, 1999), we consider everyday object categories (including chairs, cars, airplanes, and tables) with complex geometries and test three-way generalization across not only silhouettes and shading images, but also two-tone black and white images.

We find that human performance in this task is still the golden standard. Humans substantially outperform both the best performing DCNN variant as well as the analysis-by-synthesis model (by about 10%). We find that even after training on thousands of images on each category (as high as 18 times the number of images per category in the large-scale, industry-standard ImageNet dataset (Deng et al., 2009)), the DCNNs still perform considerably below the human-level performance. The analysis-by-synthesis model performs as well as the best DCNN and thus still performs below human-level. In finer-grained comparisons between behavior and models, we find that the analysis-by-synthesis model and a variant of the DCNN model robustly explain some of the variance in human performance. These results provide insights as to the nature of computational substrate needed to understand the versatility of human shape perception and cognition.

## Task: Generalizing shape information across viewpoints and image modalities

We studied shape generalization abilities in humans, DCNNs, and in a novel analysis-by-synthesis architecture using a match-to-sample task that required matching across viewpoints and image modalities (Fig. 2).

The stimuli for this task were generated using 3D models of 60 unique objects from each of the following four everyday object categories: airplanes, cars, chairs, and tables. This resulted in a total of 240 meshes to create 120 unique mesh pairs (30 pairs per category).

The two meshes in a pair were rendered multiple times to produce 6 trials for each pair of meshes: one trial per permutation of the 3 images modalities across the target item, matching item, and distractor item. Thus, on each trial, all three image modalities were present, but their assignment to the target, matching, and distractor items were varied. This results in a total of $120 \cdot 6 = 720$ trials, uniformly distributed across categories and each trial featuring all three image modalities. The target item was always rendered at a canonical viewpoint (three-quarters view; see Fig. 2 top images). The two test items were rendered at random view points (spanning [-30°; 30°] roll, [0°; 360°] yaw, and [-30°; 30°] pitch).

### Physically-based rendering

Rendering was performed using a differentiable, physically-based renderer (Mitsuba2; Nimier-David et al., 2019). In contrast to the standard shaders such as Phong and Gouraud, physically-based rendering allows us to more realistically capture shadow patterns including ambient occlusion via ray-tracing, which is crucial in the context of producing Mooney

Figure 2: Three example trials from our match-to-sample task. The target item (the top image in each triplet) is a shaded image on the left panel, a silhouette on the middle panel, and a Mooney image on the right panel.

images that induce a strong 3D percept.

**Shading images** In rendering shading images, we use an area light (in the shape of a rectangle) that is located at x=y=4 units distance relative to the object. The camera is fixed at 45° azimuth , 15° elevation, and 3.5 units distance relative to the object; for new viewpoints, we rotate the object instead. Each object is rendered with a homogeneous texture that ensures a smooth surface appearance and equates texture-related variations.

**Mooney images** To create Mooney images, we first apply a differentiable Gaussian filter to the shading image. Then we compute a threshold using the average illumination of the blurred image ignoring the background. A final differentiable thresholding operation using Pytorch's *tanh* function (Paszke et al., 2019) effectively assigns high and low illumination to lighter and darker regions, respectively.

**Silhouettes** To create silhouette images, we remove the area light and replace it with an ambient light. We also change the texture color of the object to black. Rendering such scenes results in silhouettes with white background; to equate with other image modalities, we invert the pixel values.

## Task-specific training with DCNNs

We first explore DCNNs as one approach to solve our generalization task. Here the idea is that if we can aggregate datasets for each new generalization domain, and train or fine-tune neural networks on those specific domains, they will accomplish generalization. Here we empirically explore how much data might be needed to accomplish human-level generalization in our task. What is the order of magnitude of the required number of images for human-level generalization?

Answering this question in the general case requires a theory of DCNNs, which so far has proven elusive. Thus, we have to make decisions about the architectural details, loss function, learning rate, optimization method, and training set distribution. Here, we followed a methodology in which we first explored the following dimensions in preparatory simulations semi-systematically.

- We explored two pre-trained backbones for training including a small network VGG-11 (Simonyan & Zisserman, 2014) and a large one, Inception-v2 (Szegedy, Vanhoucke, Ioffe, Shlens, & Wojna, 2015).
- We explored several embedding sizes to use for the classification layer including 10, 20, 30, 50, and 100.
- We explored three learning rate parameters: $10^{-3}, 10^{-4}, 10^{-5}$.
- We explored two optimization methods including Stochastic Gradient Descent (SGD) and Adam (Kingma & Ba, 2017).

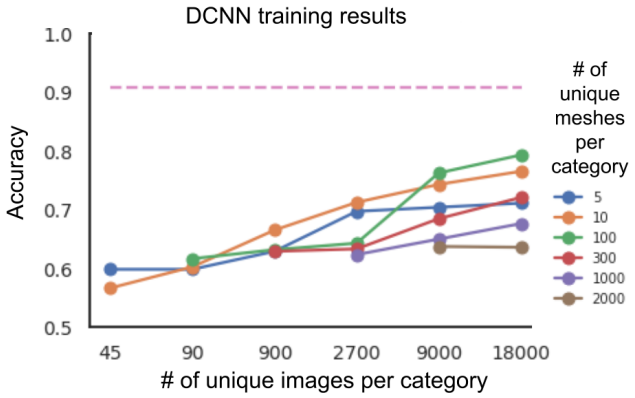


Figure 3: Accuracy plot for DCNNs trained on data generated from different numbers of unique meshes and different numbers of unique images per category. The dashed line indicates the average human performance in our task.

Based on these initial explorations, the most promising settings were as the following: VGG-11 as the pre-trained backbone, classification layer size of 20, learning rate of $10^{-3}$, and SGD optimizer with $momentum = 0.9$.

Given these settings, we directed our effort to investigating the training set sizes systematically. To that end, we considered two dimensions: the number of unique meshes and the number of unique images per category. The number of unique meshes per category could be 5, 10, 100, 300, 1000, or 2000. The number of unique images per category was 9 times (3 image modalities times 3 stimuli items) the number of mesh

pairs, resulting in the following number of unique images per category: 45, 90, 900, 2,700, 9,000, and 18,000. We considered all pairings of unique number of meshes and images where the number of images was greater than or equal to 9 times the number of unique meshes.

Given the meshes and the number of images, we constructed the training sets using the same rendering pipeline that underlies our task stimuli (Fig. 1). Using triplets from these images, we trained DCNNs using a triplet margin loss that minimizes the distance between target and matching items while maximizing the distance between target and distractor items. During training (and evaluation), if the target-matching distance was smaller than the target-distractor distance for a given triplet, that triplet was considered a "hit". We find that increasing the size of the training set in terms of the number of images is key to improve performance of the networks on this task (see Fig. 3): Highest generalization performance was achieved by selecting a moderately sized pool of unique meshes (100) and a large number of unique images rendered from them (18,000). Because these 18,000 images are obtained from 2000 mesh pairs out of 100 unique meshes, we refer to this model as DCNN-2000/100 in the following sections.

## Analysis-by-synthesis (AbS) Model

We also explored analysis-by-synthesis (AbS) as an alternative approach to accomplishing generalization in our task. AbS involves computations that are distinct from the feed-forward non-linearities learned in DCNNs for pattern classification. In AbS, shape perception amounts to inverting a generative model (or a "synthesis" function) that describes how 3D scenes form and project to images. Importantly, unlike DCNNs, AbS uses 3D scenes to explain the variation in the image ("analysis") in terms of the synthesis function. In order to perform our generalization task, we built a novel analysis-by-synthesis architecture that consists of three components (see Fig. 4): *(i)* an expressive category-specific, hypothesis space over 3D shapes for each category, *(ii)* a differentiable, physically-based renderer projecting 3D shapes to images (including degraded image modalities) using raytracing, and *(iii)* an optimization procedure that inverts this process to map images to their underlying shapes. In the following, we describe the synthesis function (components *(i)* and *(ii)*) and then the analysis function (step *(iii)*, the optimization procedure).

### The synthesis function

In order to capture the detailed shape of objects such as those present in our stimulus set, we equip the AbS model with a recently proposed learning-based model that can express high-quality shapes as continuous Signed Distance Functions (SDF) (Park et al., 2019). This model learns to map a latent vector $z$ and a query point in 3D space $x \in \mathbb{R}^3$ to the signed (indicating outside or inside) distance between $x$ and the learned shape surface via a Multi-Layer-Perceptron (MLP; this architecture is also referred to as auto-decoder; Fig. 4). The shape surface is therefore the zero-level-set of the learned implicit function that we further use to produce an explicit mesh-based surface representation using Marching Cubes (Lorensen & Cline, 1987). We train one shape model per category with 300 unique meshes from that category following the data-preparation and training procedure in Park et al. (2019).

The synthesis function also includes the same differentiable, physically-based renderer and the differentiable image modality filters for two-tone Mooney images and silhouettes (Fig. 4; see the Task section).

### The analysis function (Inference)

We infer shapes from stimuli by inverting the generative model using optimization. We initialize optimization with a random latent vector $z$, read out the corresponding shape from the generative model, produce an (optionally filtered) image given a viewpoint, and compute L2 loss between the output image and observed stimulus. Exploiting the closed-form expression of the derivative of a surface sample with respect to the underlying implicit field (Remelli et al., 2020), we preserve end-to-end differentiablity of loss with respect to $z$ despite the raytracing-based renderer. We update $z$ using Stochastic Gradient Descent (SGD) with a learning rate of 0.005 following (Remelli et al., 2020). We found 100 iterations to be sufficient to reliably infer shape for most of the stimuli. Fig. 5 showcases detailed shape reconstructions of an observed example stimulus under each of the three image modality conditions.

In our current implementation, object shape is the exclusive target of inference while lighting, viewpoint, image modality, and object category are fixed at their true values. (We note that we found image modality and, to a lesser extent, object category to be linearly decodable from the later layers of VGG-11 pre-trained on ImageNet (Deng et al., 2009) using just five meshes and a small number of images from each category. See Discussion for future work in the context of building hybrid architectures.)

### Simulation details

For each image in the stimulus set, we perform 6 randomly initialized optimization runs, collect the latent vectors associated with the smallest loss in each run, and compute the average of these 6 latent vectors. We determine "hit" for a given triplet if the linear correlation between $z_{target}$ and $z_{matching}$ is higher than the linear correlation between $z_{target}$ and $z_{distractor}$. We refer to this model as AbS in the following.

## Behavioral Experiment

### Participants

We recruited 15 subjects over Prolific, a crowdsourcing platform, each compensated $3.00. We implemented the match-to-sample task in the psiTurk framework (Gureckis et al., 2016).

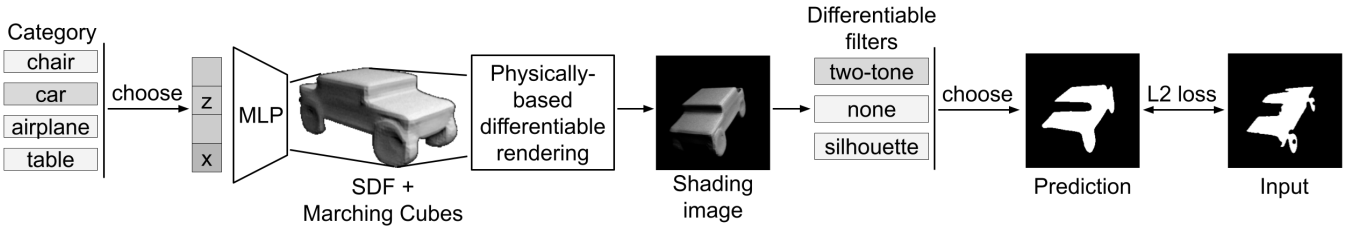


Figure 4: Analysis-by-synthesis (AbS) Model. Inference in this model is via SGD-based optimization to reduce the L2 loss with respect under a differentiable physically-based renderer and differentiable image filters.

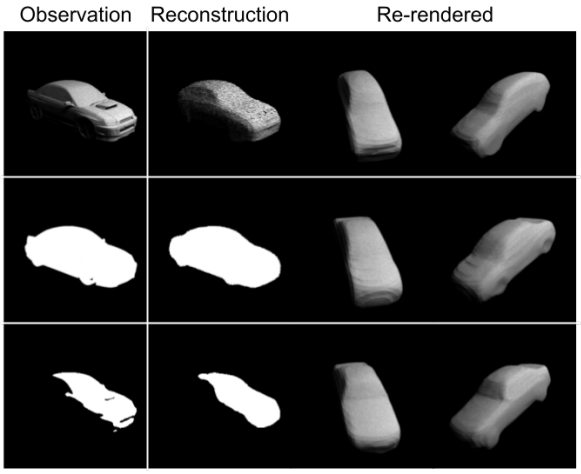


Figure 5: AbS shape inference examples. The object in the observed images is accurately reconstructed for all image modalities. Reconstruction: Same viewpoint with same rendering settings used for inference. Re-rendered: Different viewpoints at higher quality.

### Stimuli and Procedure

We used the stimuli described in the Task section. Each trial displayed all three images (the target item, matching item, and distractor item) simultaneously. Those three images were spatially organized such that one of them –the target image– was presented at the top and the other two images were presented at the bottom. The images were presented for an unlimited period of time, until the participant responded by indicating their match judgment.

The participants performed 10 training trials before proceeding to 120 test trials. They were provided feedback in the form of a running average of their performance every 10 trials. We recorded their choices as well as their response times. For each participant, the 120 test trials were drawn from the population of 720 trials, equating samples from each category and controlling for the number of permutations. The order of the trials was randomized for each participant.

### Results

Human observer accurately generalized across viewpoints and image modalities with an average performance of 91%. Observers' average performance per category is shown in Fig. 6B. We can see that the performances across all categories are all high, but not uniform: the airplane and car trials were harder than the table and chair trials. In the next section, we use models to understand this level of accuracy and its variability across categories.

## Model vs. behavior comparisons

We present the average accuracy of models and compare to behavior, in terms of both the average performance and a finer-grained trial-level comparison. Strikingly, we find that none of the models we explored reach human-level performance – in fact, humans dominate the best models (AbS and DCNN-2000/100) by more than 10% (Fig. 6A). It is not clear how many more images per category it would take to bring the DCNN models to human-level; possibly another 18,000 images per category (cf., Fig. 3).

Notice that the shape prior in the AbS model is category-specific, although this shape prior was learned in a manner agnostic to the generalization task at hand. It is possible that if we trained category-specific DCNNs (one DCNN per category), that could help bring the performance of these models to human-level, or at least improve over the category-general models. To test this possibility, we trained a separate DCNN for each category using the best-performing configuration from Fig. 3 (18,000 images, 100 unique meshes). We refer to this model as "DCNN-2000/100-spec". Surprisingly, we found that the average performance of these category-specific DCNN models was almost identical to the category-general model (80% vs. 79%; Fig. 6). This indicates that the underlying DCNN architecture had sufficient capacity for category-general training.

Despite their overall lower performance, both models captured the main pattern observed in the accuracy levels of the human observers: Their accuracy was lower for the car and airplane categories than the chair and table categories. Encouraged by this correspondence, we finally compared the models to behavior at the level of individual mesh pairs. Remember, our total of 720 task trials come from 120 unique mesh pairs. This allows for an opportunity to compare mod-

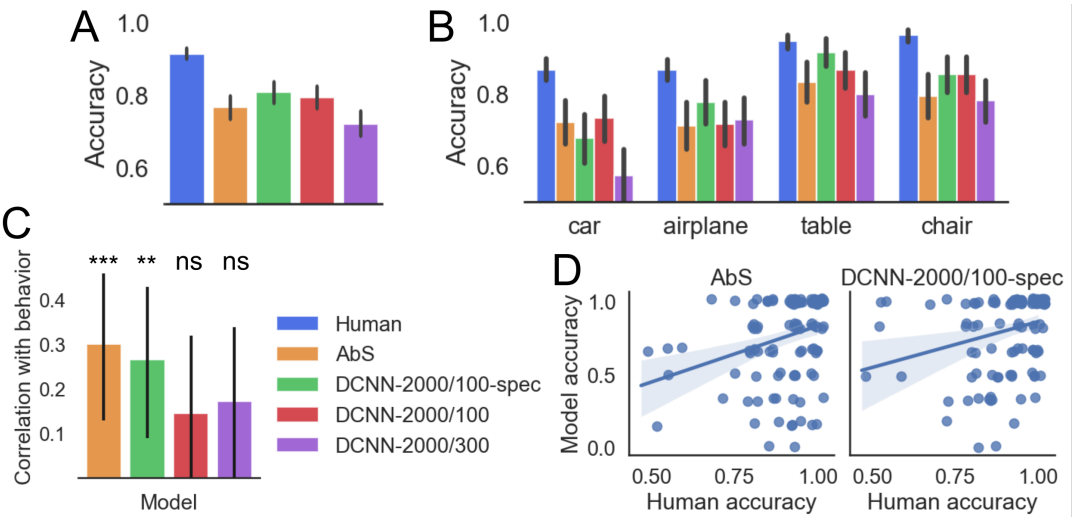


Figure 6: Comparisons between models and behavior. (A) Overall accuracy. (B) Category-wise accuracy. (C) Linear correlation between human and model accuracy based on all 120 mesh pairs. (D) Model vs. human accuracy in-depth for AbS and DCNN-2000/100-spec based on all 120 mesh pairs. Error bars/regions depict 95% CI.

els to behavior at a fine level of granularity while still leaving capacity to average accuracy across multiple trials. We would expect a small effect size for all models, but we ask if any of the models can explain any significant variance in the data. Correlations of the models to behavior are shown in Fig. 6C. We observe that only two models, the AbS and the category-specific DCNN (DCNN-2000/100-spec) show significant correlations with the data (Fig. 6C, D).

## Discussion

Our work dovetails and extends computational and behavioral studies reported in Erdogan and Jacobs (2017) and Moore and Cavanagh (1998). We extend this approach to familiar object categories (but unfamiliar individual examplars from those categories) by utilizing a flexible, learning-based prior over shapes. We go beyond their work by modeling multiple image modalities at once (not just shaded images as they do, but also Mooney images and silhouettes).

In a study using two-tone Mooney images, Moore and Cavanagh (1998) argued that perception of such scenes involves "top-down" processing where the scene illumination is factored out in the context of familiar objects retrieved from the memory. Here we provide concurring evidence and go beyond to show how these mechanisms can be implemented in a computational model. Consistent with this proposal and extending it, we find that the category-specific models (DCNN-2000/1000-spec and AbS) are the only models that significantly correlate with behavior. We wish to further explore this interesting concurrence by using the unfamiliar stimuli from Erdogan and Jacobs (2017) under our stimulus conditions to dissect the nature and impact of shape hypothesis spaces underlying our generalization abilities.

We note we are not claiming that a DCNN cannot learn to generalize in our task. However, in the settings we have explored them, such generalization would require implausibly many images, on the order of, approximately 30,000 images per category (extrapolating from Fig. 3). Still, this sample ineffiency can be improved with further research in this area. However, we are most excited about hybrid architectures that take advantage of optimization at both learning and test time, for example by integrating amortized proposals for making inferences about image modalities or categories to then drive inferences in the AbS model.

## References


Bulthoff, H. H., & Yuille, A. L. (1991). Shape-from-X: psychophysics and computation. In P. S. Schenker (Ed.), *Sensor fusion iii: 3d perception and recognition* (Vol. 1383, pp. 235 – 246). SPIE.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *Ieee conference on computer vision and pattern recognition* (pp. 248–255).

DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron*, *73*(3), 415–434.

Erdogan, G., & Jacobs, R. A. (2017). Visual shape perception as bayesian inference of 3d object-centered shape representations. *Psychological Review*.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2018). Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *CoRR*, *abs/1811.12231*. Retrieved from http://arxiv.org/abs/1811.12231

Gureckis, T. M., Martin, J., McDonnell, J., Rich, A. S., Markant, D., Coenen, A., ... Chan, P. (2016). psiturk: An open-source framework for conducting replicable behavioral experiments online. *Behavior research methods*, *48*(3), 829–842.

Hayward, W. G., Tarr, M. J., & Corderoy, A. K. (1999). Recognizing silhouettes and shaded images across depth rotation. *Perception*, *28*(10), 1197–1215.

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS Computational Biology*, *10*(11), e1003915.

Kingma, D. P., & Ba, J. (2017). *Adam: A method for stochastic optimization.*

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

Lorensen, W. E., & Cline, H. E. (1987). Marching cubes: A high resolution 3d surface construction algorithm. *ACM siggraph computer graphics*, *21*(4), 163–169.

Mooney, C. M. (1957). Age in the development of closure ability in children. *Canadian Journal of Psychology/Revue canadienne de psychologie*, *11*(4), 219.

Moore, C., & Cavanagh, P. (1998). Recovery of 3d volume from 2-tone images of novel objects. *Cognition*, *67*(1-2), 45–71.

Nimier-David, M., Vicini, D., Zeltner, T., & Jakob, W. (2019, December). Mitsuba 2: A retargetable forward and inverse renderer. *Transactions on Graphics (Proceedings of SIGGRAPH Asia)*, *38*(6). doi: 10.1145/3355089.3356498

Park, J. J., Florence, P., Straub, J., Newcombe, R., & Lovegrove, S. (2019). Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 165–174).

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems 32* (pp. 8024–8035). Curran Associates, Inc.

Remelli, E., Lukoianov, A., Richter, S. R., Guillard, B., Bagautdinov, T., Baque, P., & Fua, P. (2020). Meshsdf: Differentiable iso-surface extraction. *arXiv preprint arXiv:2006.03997*.

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

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2015). Rethinking the inception architecture for computer vision. *CoRR*, *abs/1512.00567*. Retrieved from `http://arxiv.org/abs/1512.00567`

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