Perceiving Fully Occluded Objects via Physical Simulation

Ilker Yildirim*[1] **(ilkery@mit.edu)**
BCS, MIT and The Laboratory for Neural Systems, The Rockefeller University

Max H. Siegel* (maxs@mit.edu)

Joshua B. Tenenbaum (jbt@mit.edu)
BCS, MIT

Abstract

Conventional theories of visual object recognition treat objects effectively as abstract, arbitrary patterns of image features. They do not explicitly represent objects as physical entities in the world, with physical properties such as three-dimensional shape, mass, stiffness, elasticity, surface friction, and so on. However, for many purposes, an object's physical existence is central to our ability to recognize it and think about it. This is certainly true for recognition via haptic perception, i.e., perceiving objects by touch, but even in the visual domain an object's physical properties may directly determine how it looks and thereby how we recognize it. Here we show how a physical object representation can allow the solution of visual problems, like perceiving an object under a cloth, that are otherwise difficult to accomplish without extensive experience, and we provide behavioral and computational evidence that people can use such a representation.

Keywords: physical object representations; analysis-by-synthesis; object perception; occlusion; psychophysics

> *The common and almost despairing feeling . . . was that practically anything could happen in an image and furthermore that practically everything did.*
>
> David Marr, *Vision*

Introduction

Object perception is notoriously difficult, in part because the appearance of an object can vary in almost any way. The problem has been studied in neuroscience, cognitive psychology, and artificial intelligence, leading to a loose consensus that object perception can be solved by the brain (or a computer) learning to "untangle" or become "invariant to" sources of variation in the image (DiCarlo, Zoccolan, & Rust, 2012; LeCun, Bengio, & Hinton, 2015). On this account, sensory input is repeatedly transformed, ideally leading to a (biological or artificial) neural code that is diagnostic for a particular object regardless of variation in the image (Riesenhuber & Poggio, 1999).

We study an alternative solution to the object perception problem, which is enabled by a different representation for objects and a different attitude towards variation. The basic idea is to model the causal processes that lead to an observed image, explaining and reproducing image variation rather than attempting to ignore it. More specifically, we take an object to be represented (at least in part) by a set of physical attributes necessary for supporting physical interaction and image generation. We posit that objects are represented, at a minimum, by attributes including three-dimensional geometry; rigidity; mechanical material properties; and optical material properties.

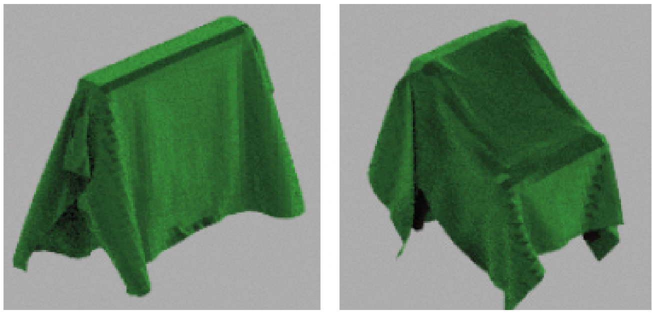

Figure 1: Two objects occluded by cloths.

Consider the covered objects displayed in Figure 1. Both objects are completely occluded, but it is easy to say which of them might be a chair. These images were generated by using a physics simulator to drop a simulated cloth onto two separate 3D models (one a chair), then using a rendering engine to produce images from the resulting scene.

This describes a process for generating the image, but the same process can also be used to interpret an image. When asked which image has a chair in it, we can simulate dropping cloths onto a chair mesh, and compare the rendered results with each candidate (this is an intuitive sketch of a procedure; it can be made precise with Bayes' theorem, which specifies how to turn a forward model into an inverse model. See also (Battaglia, Hamrick, & Tenenbaum, 2013)).

There are several notable differences between the latter approach (which we will call analysis-by-synthesis) and the "consensus" approach (which we will call the invariant features approach). First and most notably, invariant features approaches must learn each kind of scene transformation independently. We explained above how knowledge about chairs and cloths can be combined, in the analysis-by-synthesis approach, to recognize a chair underneath a cloth. By contrast, invariant features approaches cannot directly leverage existing knowledge to recognize the compound object. They must be trained, separately, to discount the cloth in order to recog-

[1] indicates equal contribution.

nize the chair[2]. Second, while invariant features approaches are mostly agnostic to the kinds of image transformations that might exist, analysis-by-synthesis implicitly handles a wide variety of transformations without being explicitly trained or taught to do so.

Analysis-by-synthesis in vision has a long history (Yuille & Kersten, 2006; Tu, Chen, Yuille, & Zhu, 2005), and has recently seen increased attention (Kulkarni, Kohli, Tenenbaum, & Mansinghka, 2015; Yildirim, Kulkarni, Freiwald, & Tenenbaum, 2015; Erdogan, Yildirim, & Jacobs, 2015). Our work focuses primarily on two less-studied aspects: First, while most work in object perception studies unoccluded or partially occluded objects, we are interested in objects that are fully occluded, so that the only perceptible effect of the object is on its occluder. Said another way, perceiving objects through cloths requires an observer to do without most or all of the visual information that one normally uses. Second, unlike most previous work, we are interested primarily in how the object representation enables this kind of perception. If objects are represented geometrically in a way that can interact with physics, then the procedure outlined above shows how to solve the cloth task without more training.

The object-under-cloth task is most interesting in the case of novel cloth-object pairs (such as an airplane under a cloth). We predict that both people and analysis-by-synthesis models will perform well on this task, but invariant features models will not without further training. We ran a large scale study to assess how well people can perform the object-under-cloth task. After describing the task and the results of the study, we will evaluate the performance of each of the candidate models, and discuss the implications of our findings.

Experiments

We performed two experiments where participants needed to generalize from a single view of 3D object shown at a canonical view to either a novel view of that object (Experiment 1) or to a fully occluded image of that object again at a novel view (Experiment 2).

Participants

58 participants were recruited from Amazon's crowdsourcing web-service Mechanical Turk. The experiment took about 20 minutes to complete. Each participant was paid $1.50.

Stimuli

The stimuli were generated using a subset of the meshes from the ShapeNet (Chang et al., 2015) database using Blender (Blender Online Community, 2015), a 3D modeling and rendering program. The meshes we used represented objects from five different categories: guns, cars/buses, bicycles, laptops and pillows.

We rendered each mesh in three ways: (1) unoccluded, with texture, and from a canonical viewpoint, (2) unoccluded, without texture, and at a random viewing angle, and (3) fully occluded with a cloth draped over it, and with the mesh randomly rotated.

For (2) and (3), the rotation was sampled from the full sphere with a slight preference around the canonical viewpoint – viewpoint of (1). More specifically, a rotation angle of ±35° of the canonical viewpoint on all three axes was 1.5 more likely than the rest of the sphere. For (3), we simulated a cotton-like cloth draped over the rotated mesh for a total of 100 simulation steps, and obtained a rendering of the very last step of the simulation.

We used a total of 197 meshes to generate 100 five-tuples of one study item of unoccluded object rendered at a canonical viewpoint with texture, and four test items consisting of two unoccluded objects without texture each rendered after randomly rotating the meshes, and two objects rendered after randomly rotating each and then occluding with a cloth. The unoccluded study items were never seen twice, but the test items were repeated multiple times, each at a different rotation or viewing angle. On 57 of the 100 tuples, the distractors were of the same category, and 43 of the 100 tuples, the distractors were of different category. Example pairs of unoccluded study items and occluded test items are shown in Figure 2a.

Procedure

Both experiments were match-to-sample tasks where both the study and the test items were presented simultaneously and all stayed on the screen until the participants responded. In Experiment 1 ($N = 27$), the study item was an unoccluded image of an object from a canonical viewpoint; the test items were images of two unoccluded objects after randomly rotating each. Participants had to choose which of the test items contained the study item (Figure 2b).

In Experiment 2 ($N = 31$), the study item was also an unoccluded image of an object; but the test items were images of two objects rotated randomly and occluded with a cloth. Again, participants had to choose which of the test images contained the study item (Figure 2b).

In each of the experiments, participants completed 10 practice trials before moving onto 90 experimental trials. Participants were provided with their running average performance at every 5th trial throughout the experiment.

Results

Results of the experiments are shown in Figure 3a. Participants performance on Experiment 1 (with the unoccluded test items) were high overall (93%).

Participants performance was surprisingly high in Experiment 2 with the occluded test items at 80% (Figure 3a), and as expected, their performance was lower when compared to Experiment 1.

The type of the distractor (whether it is of the same category as the study item or different category) introduced a much stronger decline in performance in Experiment 2 than in Experiment 1 (Figure 3b).

[2] Alternatively, models might be "retrained" or similar; these methods also require more training.

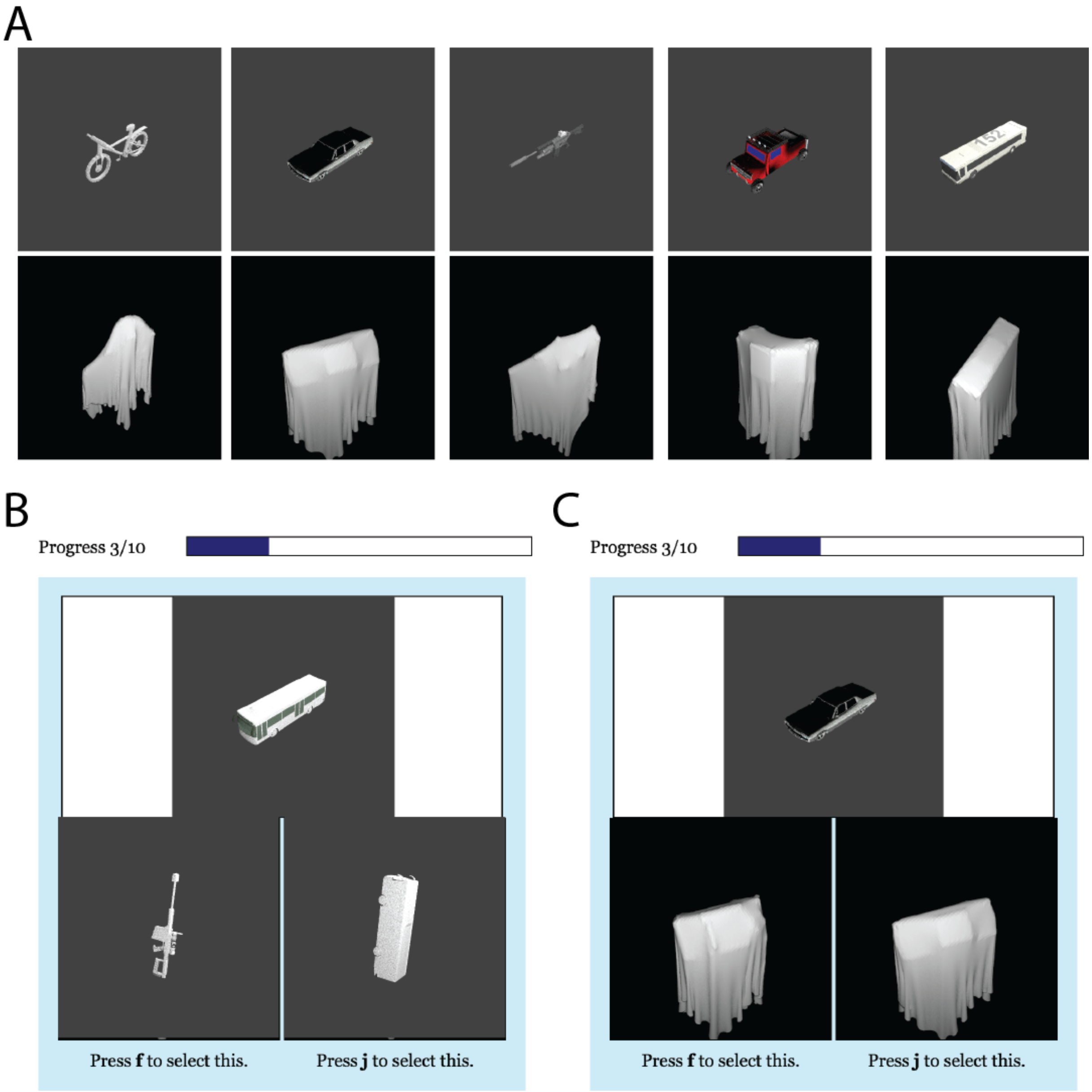

Figure 2: (A) Pairs of images of meshes and simulation results after rotating the mesh randomly and draping a cloth over it. (B) Screenshot of an example trial in the unoccluded experiment. (C) Screenshot of an example trial in the occluded experiment.

Models

We considered two models as potential accounts of our subject's performance: a physics-based analysis-by-synthesis model, and a model derived from features learned by a deep convolutional neural network trained to classify millions of *unoccluded* images.

Physics-based analysis-by-synthesis

We developed a Bayesian computational model that uses knowledge of the causal processes underlying image formation to interpret new images. Aspects of (synthetic) image formation may be divided into two categories: physics-based object factors (e.g., 3D shape, mass, friction, soft-body dynamics, soft-body and rigid-body interaction); and graphics (e.g. rotation and lighting direction). When each of these factors are specified, we end up with a likelihood function that gives the probability of an image given latent parameters, $P(I|\Psi)$.

The model maps input images to the underlying scene parameters using Bayesian inference. Bayes' rule enables us to use the (forward model) likelihood along with a prior distribution (here taken to be uniform) to get the *posterior* distribution of the parameters given the image. The posterior, which includes beliefs about the underlying mesh, is the object of

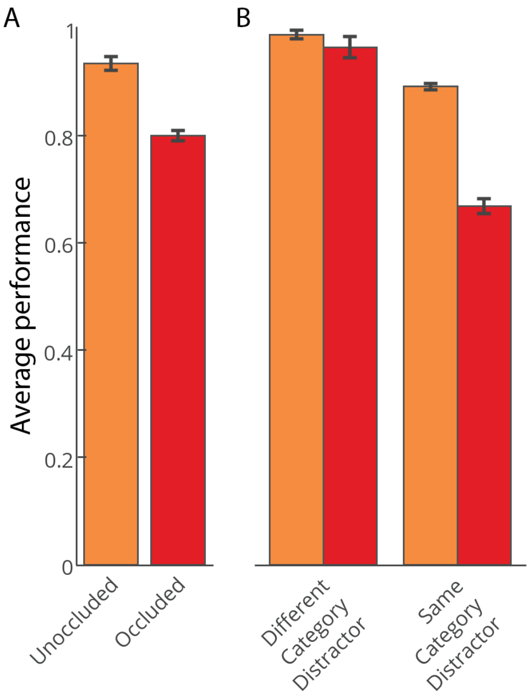

Figure 3: (A) Average performace of the participants in the two experiments. (B) Performance of the subjects divided by whether the distractor is of the same or different category as the study item. Error bars indicate standard error of the mean.

interest for inference tasks.

Bayes' rule states that the posterior is proportional to the likelihood times the prior, that is,

$$P(\Psi|I) \propto P(I|I_s,\Psi)\delta_{p(\cdot)}\delta_{g(\cdot)} \tag{1}$$

where Ψ are the latent variables (e.g. 3D shape, mass, friction, soft-body dynamics, rotation, lighting) that drive the physics and graphics engines; $\delta_{p(\cdot)}$ denotes a physics engine and $\delta_{g(\cdot)}$ denotes a graphics engine (here, we used Blender for both engines); $P(\Psi|I)$ denotes the posterior distribution over physical object representations; I_s denotes the generated image given $\Psi, g(\cdot)$, and $p(\cdot)$; and $P(I|I_s,\Psi)$ denotes the image log-likelihood for a given set of physical object parameters (we assumed a Gaussian likelihood function with $sigma = 0.01$; $N(I|I_s,\sigma)$.

In our implementation, we deterministically assigned the cloth properties such as stiffness, mass and friction to their true values, but assumed uncertainty for the exact 3d shape, S, of the study item and the rotation, R, of the test items. We approximated shape uncertainty as a uniform distribution over the most similar five shapes given a study item, I and its underlying shape, S_I, from the ShapeNet dataset. Similarity between a pair of meshes from the ShapeNet was determined by correlating the concatenated images of each mesh without any texture at four view points – the canonical viewpoint where each mesh is already aligned at in the dataset, and the three orthogonal viewpoints (top, right, front).

The uncertainty about the rotation or pose, R (a vector of length three of Euler angles), was taken to be higher for occluded test scenes than for unoccluded test scenes. For unoccluded test images, we modeled rotation as $R \sim N(R_{true}, 0.025)$, whereas for occluded test images, we modeled it as $R \sim N(R_{true}, 0.1)$, where R_{true} is the true rotation of the test items. We approximated the rotation uncertainty using five randomly chosen samples from the normal distribution.

We took a sampling-based approach to simulate participants from our experiments. For a given simulated participant and trial number, we took one joint sample of shape and rotation from a pool of 25 samples (5 possible shapes × 5 rotations). Also using the true cloth parameters for occluded trials (i.e., rest of the parameters in Ψ), the model generated a sampled image, I_s, compared it to each of the test images using correlation as its similarity metric, and returned the test image that was more similar to I_s. We simulated 40 subjects in each of the occluded and unoccluded experiments. We report the average performance of these 40 simulated participants.

Deep convolutional network

We also evaluated a pre-trained deep convolutional neural network, which has been shown to succeed at a number of challenging visual recognition tasks (Jia et al., 2014). The network, like all convolutional neural networks, is a feedforward hierarchical model that performs a series of convolutions and nonlinearities; such models can contain (as does the model we evaluate) millions of learnable parameters. The model was trained to classify objects using the Imagenet (Deng et al., 2009) dataset; it is representative of a wide class of neural network models that find application in computer vision.

Results

Figure 4 show the performance of the two models. The pre-trained network performs worse than human participants (Figure 4, left). Furthermore, unlike the human participants, the decline in the performance from by the type of the distractor category is similar for both the occluded and unoccluded stimuli sets. The network performs at chance in the most difficult condition of occluded and same-category distractor trials.

The physics-based analysis-by-synthesis model captures the average performance of the participants across the two experiments accurately (Figure 4, right). Moreover, our model captures the details of the subjects performance when broken down by the distractor type.

Discussion

Can humans recognize objects even when they are fully occluded? The behavioral study presented here indicates that the answer is yes, at least when the candidates are known (we suspect that people can often do so otherwise, as in the chair example in the introduction). What underlies our participants' performance? We think that this ability reflects the

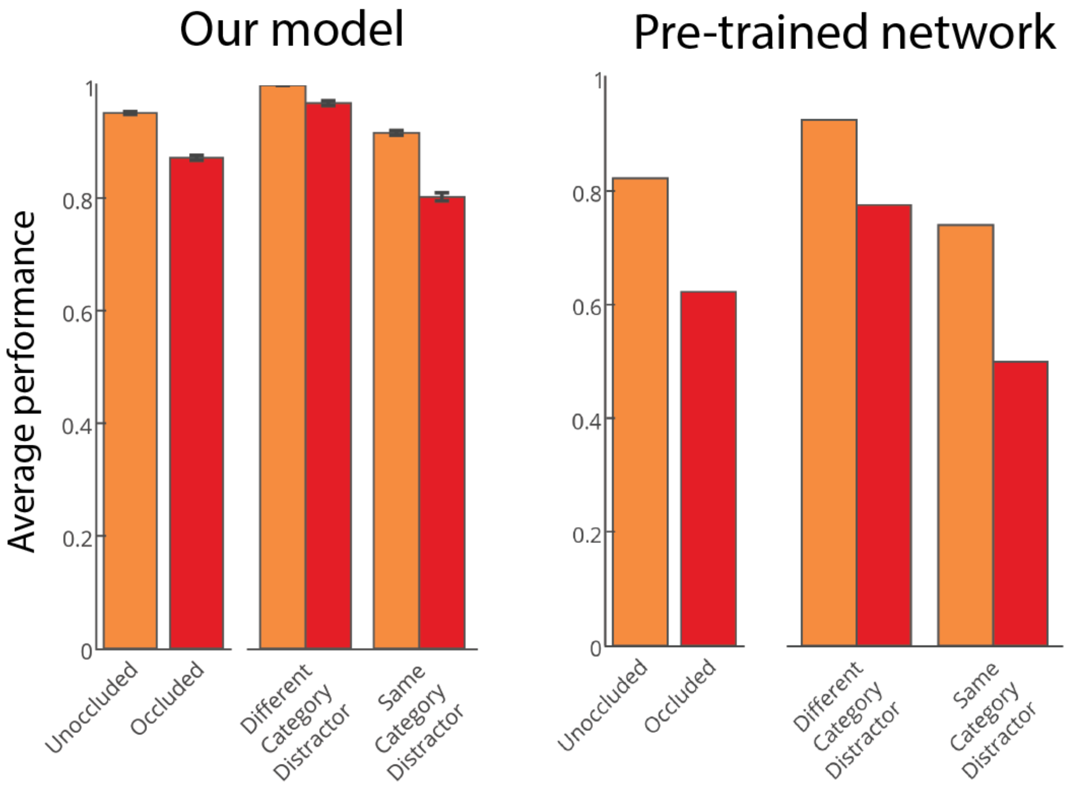

Figure 4: (Left) Average performance of the physics-based analysis-by-synthesis model on the unoccluded and occluded stimuli sets, and the breakdown of its performance by the tyep of the distractor category. Error bars indicate standard error of the mean. (Right) Average performance of the pre-trained network on the unoccluded and occluded stimuli sets, and the breakdown of its performance by the type of the distractor category. Dashed line shows the chance-level performance

presence of *causal theories* in human perception - in this case, theories about how soft and rigid objects bodies interact, and about how the resultant shape gets rendered into a visual percept. This approach is different from the standard model of visual perception. Most accounts of visual object recognition assert that the brain maps images to identity or class labels via increasingly more abstract feature hierarchies (DiCarlo et al., 2012; Krizhevsky, Sutskever, & Hinton, 2012; Riesenhuber & Poggio, 1999). In this approach, objects are not explicitly represented – they do not have physical properties such as mass and friction or 3D shape. The cost of not having explicit and causal representations of the scenes is an inability of composition – or in other words, requirement for data (lots of data) in the face of every new scene setting such as the setting presented here: objects fully occluded under cloths.

These issues are not just theoretical: we found that a state-of-the-art neural network (albeit trained only on unoccluded images) had trouble with our behavioral task. In particular, it could not approach human performance. We believe that any similar architecture (supervised learning) will have similar problems, because such approaches do not attempt to model the causal structure that gives rise to percepts.

We suggested an alternative solution. Instead of relying on the environment to teach us about variation, we proposed using basic knowledge about the world to derive and understand how the world can influence our percepts. We considered the problem of perceiving objects under cloths. Invariant features theories of perception face a problem in this domain, because cloth-draped objects differ dramatically in appearance from their unoccluded states. Treating perception as an inverse compositional process provides a solution.

There are a number of future directions that we wish to explore. First, our match-to-sample task is only one way of getting at people's abilities to perceive objects under heavy occlusion and in physical settings. We plan to build upon this paradigm for future experiments. In the Introduction we posed a question without answering - "which of these occluded objects is a chair?". This is perhaps the most interesting future direction to pursue; it requires accessing and manipulating the concept of a chair, instead of imagining object transformations as we studied here. Second, we plan to build hybrid architectures that involve fast and feedforward neural network pipelines for broad stroke comprehension of the scene and top-down physics-based architectures for in depth physical interpretation (Wu, Yildirim, Lim, Freeman, & Tenenbaum, 2015).

Acknowledgments

This research was supported by the Center for Brains Minds and Machines (CBMM), funded by NSF STC award CCF-1231216 and by ONR grant N00014-13-1-0333.

References

Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, *110*(45), 18327–18332.

Blender Online Community. (2015). Blender - a 3d modelling and rendering package [Computer software manual]. Blender Institute, Amsterdam. Retrieved from `http://www.blender.org`

Chang, A. X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., ... Yu, F. (2015). *ShapeNet: An Information-Rich 3D Model Repository* (Tech. Rep. No. arXiv:1512.03012 [cs.GR]). Stanford University — Princeton University — Toyota Technological Institute at Chicago.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *Computer vision and pattern recognition, ieee conference on* (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., Yildirim, I., & Jacobs, R. A. (2015). From sensory signals to modality-independent conceptual representations: A probabilistic language of thought approach. *PLoS Comput Biol*, *11*(11), e1004610.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., ... Darrell, T. (2014). Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the acm international conference on multimedia* (pp. 675–678).

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105).

Kulkarni, T. D., Kohli, P., Tenenbaum, J. B., & Mansinghka, V. (2015). Picture: A probabilistic programming language for scene perception. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 4390–4399).

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

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature neuroscience*, *2*(11), 1019–1025.

Tu, Z., Chen, X., Yuille, A. L., & Zhu, S.-C. (2005). Image parsing: Unifying segmentation, detection, and recognition. *International Journal of computer vision*, *63*(2), 113–140.

Wu, J., Yildirim, I., Lim, J. J., Freeman, B., & Tenenbaum, J. (2015). Galileo: Perceiving physical object properties by integrating a physics engine with deep learning. In *Advances in neural information processing systems* (pp. 127–135).

Yildirim, I., Kulkarni, T. D., Freiwald, W. A., & Tenenbaum, J. B. (2015). Efficient analysis-by-synthesis in vision: A computational framework, behavioral tests, and modeling neuronal representations. In *Thirty-seventh annual conference of the cognitive science society.*

Yuille, A., & Kersten, D. (2006). Vision as bayesian inference: analysis by synthesis? *Trends in cognitive sciences*, *10*(7), 301–308.