Humans and 3D neural field models make similar 3D shape judgements

Thomas O'Connell (tpo@mit.edu)
Brain & Cognitive Sciences, MIT, 43 Vassar St.
Cambridge, MA 02139 USA

Tyler Bonnen (tyler.ray.bonnen@gmail.com)
Psychology, Stanford University, 450 Jane Stanford Way
Stanford, CA 94305 USA

Yoni Friedman (yyf@mit.edu)
Brain & Cognitive Sciences, MIT, 43 Vassar St.
Cambridge, MA 02139 USA

Ayush Tewari (ayusht@mit.edu)
CSAIL, MIT, 32 Vassar St.
Cambridge, MA 02139 USA

Josh Tenenbaum (jbt@mit.edu)
Brain & Cognitive Sciences, MIT, 43 Vassar St.
Cambridge, MA 02139 USA

Vincent Sitzmann (sitzmann@mit.edu)
CSAIL, MIT, 32 Vassar St.
Cambridge, MA 02139 USA

Nancy Kanwisher (ngk@mit.edu)
Brain & Cognitive Sciences, MIT, 43 Vassar St.
Cambridge, MA 02139 USA

Abstract

Human visual perception captures the 3D shape of objects. While convolutional neural networks (CNNs) resemble some aspects of human visual processing, they fail to explain human shape perception. A new deep learning approach, 3D neural fields (3D-NFs), has driven remarkable recent progress in 3D graphics and computer vision. 3D-NFs encode the geometry of objects in a continuous, coordinate-based representation. Here, we investigate whether humans and 3D-NFs make similar trial-level 3D shape judgments on match-to-sample tasks with rendered stimuli. In Experiment 1, 3D-NF behavior is more similar to human behavior than standard CNNs trained on ImageNet, regardless of whether lure objects were a.) from a different category than the target, b.) the same category as the target, or c.) matched to have the most similar 3D-NF to the target as possible. In Experiment 2, to accentuate differences between humans and 3D-NFs compared to CNNs, five difficulty conditions were defined based on the performance of 25 ImageNet CNNs. Again, we find 3D-NF and human behavior are well aligned, with both showing high accuracy even for trials where CNNs fail. Overall, 3D-NFs and humans show similar patterns of 3D shape judgements, suggesting 3D-NFs as a promising framework for investigating human 3D shape perception.

Keywords: 3D shape perception; 3D neural fields; psychophysics; light field network; deep learning

Introduction

The human visual system demonstrates a remarkable capacity to perceive the three-dimensional (3D) shape of objects. Decades of research in vision science have provided insights into the mechanisms underlying human 3D shape perception (Todd, 2004). While convolutional neural networks (CNNs) have shown similarities with visual processing in the primate brain (Yamins et al., 2014; Kriegeskorte, 2015), they still perform worse than humans in 3D shape processing tasks (Kubilius, Bracci, & Op de Beeck, 2016; Geirhos et al., 2018; Bonnen, Yamins, & Wagner, 2021). What computational mechanisms does human perception leverage to recover high-fidelity representations of 3D geometry? A new deep learning technique, 3D neural fields (3D-NFs), have driven rapid recent developments in 3D graphics and computer vision. In 3D graphics, 3D-NFs are used to encode the geometry of an individual scene from 2D images from many viewpoints (Mildenhall et al., 2021). In 3D computer vision, conditional 3D-NFs can compute continuous 3D volumetric functions from images using encoder-decoder architectures (Yu, Ye, Tancik, & Kanazawa, 2021; Sitzmann, Rezchikov, Freeman, Tenenbaum, & Durand, 2021). Here, to explore potential computations underlying human 3D shape perception, we test the alignment between 3D-NFs and human behavior using 3D match-to-sample tasks.

Methods

3D Light Field Networks

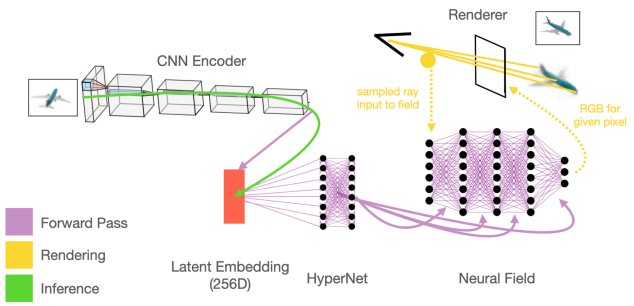

Figure 1: Schematic of the 3D-NF architecture.

The 3D-NFs used in the following experiments are variants of 3D Light Field Networks (Sitzmann et al., 2021) (Fig. 1). The essence of a neural field is a neural network that encodes a continuous function from coordinates to properties. We implement the neural field as a multi-layer perceptron (MLP) with eight layers. For light field networks, the neural field encodes the RGB value for every possible ray through a sphere surrounding each object. The neural field's input is a set of plücker coordinates, which define a ray through a scene, and the output is the RGB value of that ray.

The model pipeline follows three steps. 1. Infer a set of shape latents (256D) from an RGB input image. The image is provided as input to a resnet50 CNN encoder, and latents are regressed out of the final convolutional layer with a linear mapping. The CNN encoders used for inference are either pretrained on imagenet and finetuned when learning the 3D-NF (finetuned) or learned from scratch with the 3D-NF (learned). 2. Map from latents to weights for the 3D-NF MLP using a hypernetwork (an MLP that outputs the weights of another neural network). 3. Given a novel camera position, query rays from the 3D-NF to render an image.

The 3D-NFs are trained with a multi-view loss: given an input RGB image from one viewpoint, the model objective is to render the same object from a different viewpoint (Fig. 1, Forward Pass + Rendering). The loss is the MSE between the predicted and ground-truth novel-viewpoint image. At test time (Fig. 1, Inference), an image is provided to the trained model and latents, autodecoder intermediate units, and the weights of the final 3D-NF are extracted for 3D tasks.

3D shape judgmenets in humans and models

To quantify 3D shape judgements in humans and models, we use a match-to-sample task with rendered object stimuli. We use two datasets of objects: ShapeNet (Chang et al., 2015), a large collection of manmade shapes from 13 categories and ShapeGen (`https://github.com/jvanaken1/ShapeGen`), a shape generator that creates abstract objects without an category structure.

Human Psychophysics For a given trial (Fig. 2), three images are shown: a sample image of a rendered object is presented with two possible images to match. The target image depicts the same object as the sample from a different random viewpoint (viewpoints sampled from a 360 sphere), and the lure image shows a different object. The task for human participants is to identify the matching target image. All human data were collected online using Prolific (`https://prolific.co/`).

Model Psychophysics For models, we use a similarity-based measure to complete the task. Some set of model features (CNN unit activity, 3D-NF unit activity, 3D-NF weights) are extracted for each of the three images in a given trial. Cosine similarity is computed b/w the sample/target and sample/lure features, and if the sample/target features are more similar than the sample/lure features the trial is recorded as correct. All human and model psychophysics experiments use a variant of this task, and the primary manipulations relate to how the target-lure object pairings are selected.

Figure 2: Two example trials (L: ShapeNet, R: ShapeGen)

Results

Experiment 1: 3D-NFs and humans make similar 3D shape judgements for manmade objects

Stimuli were rendered from 13 categories of the ShapeNet dataset. Object pairs for match-to-sample trials had three conditions: 1. Target-lure objects from different categories, 2. Target-lure objects were from the same category, 3. Target-lure objects were matched to be as similar as possible in the 3D-NF weight-space. We compared human behavior (n=120) to 3D-NFs and 25 ImageNet CNNs by computing the cosine similarity across trials between each model's task performance and human accuracies.

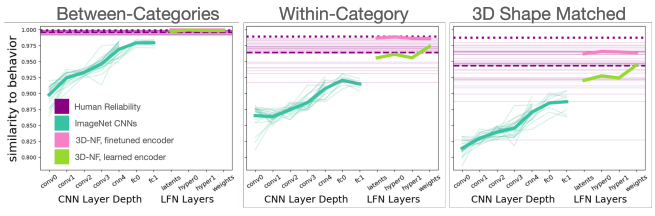

Figure 3: Results for Exp. 1. Dotted lines show human split-half reliability, dashed lines show human leave-one-subject-out reliability, and faint lines are leave-one-subject-out reliabilities for individual participants.

For between-categories trials, both humans and 3D-NFs were at ceiling performance and thus had ceiling trial-level behavioral similarity. CNN similarity to behavior improved over layers, but fell short of the noise-ceiling. For within-category trials, 3D-NFs showed high similarity to humans, exceeding the noise ceiling. For 3D-shape-matched trials, the only model to reach the noise-ceiling was the 3D-NF with a fine-tuned CNN encoder. Overall, we see that 3D-NF field models closely track human 3D shape judgements regardless of how target-lure pairs are matched. See Fig. 3 for results.

Experiment 2: 3D-NFs and humans make similar 3D shape judgements across CNN-defined difficulties

To highlight CNN failure cases relative to humans and 3D-NFs, we use the CNN model zoo to identify match-to-sample object-pairs that fall into 5 difficulty bands (ranging from an average CNN accuracy of .18 for the most difficult condition to .6 for the easiest condition). We ran two versions of the experiment with different stimuli: 1. manmade ShapeNet objects and 2. abstract ShapeGen objects. The veresion with abstract objects ensures that any similarity between human and 3D-NF behavior is not simply driven by category structure. For both manmade (n=200) and abstract (n=200) objects, we find that humans and 3D-NFs are largely unaffected by the CNN-defined conditions, displaying high performance even in cases where CNNs fail. 3D-NFs and humans again showed high trial-level similarity across conditions (Fig. 4). The 3D-NF with a fine-tuned encoder reached the stringent split-half noise-ceiling for both manmade and abstract shape tasks.

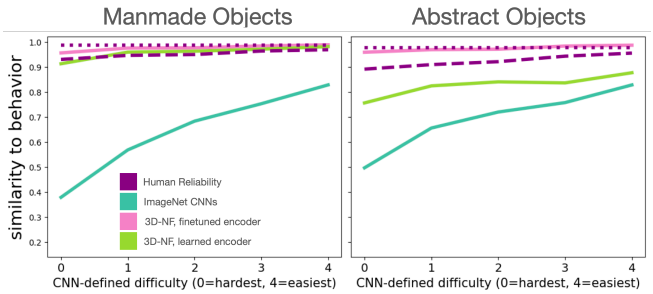

Figure 4: Results for Experiment 2. See Fig. 3 caption for human reliability explanation.

Conclusion

We find that 3D-NFs and human observers make similar 3D shape judgments regardless of target-lure similarity (Experiment 1) and for difficult CNN-defined trials (Experiment 2). In all comparisons, 3D-NFs more closely resembled human behavior than more standard ImageNet CNNs. To our knowledge, these are the first reports of models that compute continuous 3D representations from images that match human-level performance and consistency, suggesting that 3D-NF models are a promising direction for exploring behavioral and neural 3D shape processing.

Acknowledgments

This work was supported by National Institutes of Health grant DP1HD091947 and National Science Foundation grant 2124136 to Nancy Kanwisher, and the Center for Brains, Minds, and Machines (CBMM) funded by NSF STC award CC-1231216.

References

Bonnen, T., Yamins, D. L., & Wagner, A. D. (2021). When the ventral visual stream is not enough: A deep learning account of medial temporal lobe involvement in perception. *Neuron*, *109*(17), 2755–2766.

Chang, A. X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., ... others (2015). Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2018). Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*.

Kriegeskorte, N. (2015). Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annual review of vision science*, *1*, 417–446.

Kubilius, J., Bracci, S., & Op de Beeck, H. P. (2016). Deep neural networks as a computational model for human shape sensitivity. *PLoS computational biology*, *12*(4), e1004896.

Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., & Ng, R. (2021). Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, *65*(1), 99–106.

Sitzmann, V., Rezchikov, S., Freeman, B., Tenenbaum, J., & Durand, F. (2021). Light field networks: Neural scene representations with single-evaluation rendering. *Advances in Neural Information Processing Systems*, *34*, 19313–19325.

Todd, J. T. (2004). The visual perception of 3d shape. *Trends in cognitive sciences*, *8*(3), 115–121.

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.

Yu, A., Ye, V., Tancik, M., & Kanazawa, A. (2021). pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 4578–4587).