Evaluating the perceptual alignment between generative visual models and human observers on 3D shape inferences

Tyler Bonnen* Riley Peterlinz Angjoo Kanazawa Alexei Efros
Electrical Engineering and Computer Science Department, UC Berkeley
{bonnen, peterlinz, kanazawa, efros}@berkeley.edu

ABSTRACT

Humans can infer the 3D shape of objects from a single image. Computational methods in the neurosciences fail to adequately model this ability. Recently, 'generative' machine learning methods have emerged as a promising approach to modeling the geometric properties of objects. Here we develop a framework to evaluate the perceptual alignment between these generative models and humans on 3D visual tasks. Given a set of experimental images (e.g., four images within an 'oddity' trial), we use an image-conditioned generative model to infer object properties, including 3D shape. We then estimate relative viewpoints (i.e., camera positions relative to objects) across images. With these inferred object and viewpoint latents, we determine the similarity between objects within a trial, using an image generation procedure analogous to mental rotation. We evaluate how well a single instance of this generative model class, Large Reconstruction Model (LRM), predicts human behavior. We find that LRM does not achieve human-level performance on 3D visual inferences. Nonetheless, our approach provides an extensible framework to evaluate the perceptual alignment between humans and generative visual models.

Keywords: generative models; 3D perception

INTRODUCTION

Given a single image, humans can infer the shape of objects. Many theories have been proposed to account for this ability (Marr, 2010; Ullman, 1979; Koenderink, Van Doorn, & Kappers, 1992). To formalize these accounts, computational neuroscience has increasingly adopted 'stimulus-computable' models, which are able to predict both neural responses and behaviors, directly from experimental images. However, standard computational models, such as convolutional neural networks (CNNs) trained on Imagenet, systematically fail to achieve human-level performance on 3D shape inferences (Abbas & Deny, 2023; Alcorn et al., 2019). Recently, a novel class of 'generative' visual models have been used to produce images that seem consistent with human perceptual expectations, but there have not been quantitative evaluations of human-model alignment on 3D perceptual inferences. Evaluating this alignment requires novel analytic methods; while discriminative models can generate representations and behaviors conditioned on experimental stimuli, generative models often generate images, conditioned on some other input. These models can be adapted to address experimental questions (Jaini, Clark, & Geirhos, 2023), but no method has been proposed to evaluate human-model consistency in 3D shape inferences. Here we develop a framework to evaluate the perceptual alignment between generative visual models and human observers. Our approach is simple: instead of analyzing model responses to experimental images directly, we analyze the images rendered by these generative models, which are conditioned on experimental images. That is, given an experimental stimulus that contains an object, we generate images of this object rendered from novel viewpoints. To adapt this for 3D experiments, we visualize the object from each image from the viewpoints of other images, analogous to 'mental rotation' (Shepard & Metzler, 1971). We leverage this approach to evaluate a (publicly available) generative visual model: large reconstruction model (LRM, (Hong et al., 2023)). We use a dataset from Bonnen, Yamins, and Wagner (2021) where humans substantively outperform standard deep learning methods on 3D inferences; for conditions that rely on texture-level image properties, standard deep learning models (e.g., vs. DINOv2, resnet50) perform as well as humans, while they are at chance on conditions that rely on 3D shape inferences. As such, this dataset provides baselines for both model performance and human abilities (Table 1).

	Semantic 3D	Abstract 3D	Semantic Texture	Abstract Texture
Human	.85	.82	.90	.93
DINOv2	.34	.29	.89	.89

Table 1: Comparison between human and model performance on a 4-way 'oddity' task using stimuli from Bonnen et al. (2021)

RESULTS

Formulation. "Oddity" tasks are a common design used to evaluate human 3D shape perception: several images are presented simultaneously, from different viewpoints, and participants determine which image contains the object which is least like the others (see examples in Fig. 1). To evaluate generative visual models using this design, we begin with the observation that images can be decomposed into object-level properties (such as shape, texture) and camera properties (such as position and rotation). If these object and camera properties can be estimated, it is possible to determine the similarity between those objects in the following manner. First, the object from image$_j$ can be rendered using the camera properties of image$_i$ (i.e., distance, rotation, position), creating a novel image; intuitively, this can be thought of as a mental rotation; then, we can determine the similarity between the original image$_j$ and the rendered image, using a suitable met-

ric, such as l2 distance over pixels or similarity between features of a neural network. Given a set of experimental images image_{0-N} in a single trial, we can infer the underlying object and viewpoints in each image (e.g., object_0 and viewpoint_0 from image_0), render this object from the viewpoints of other images (e.g., object_0 from viewpoint_1) and compare to the reference object in that image (e.g., object_1 from viewpoint_1).

Figure 1: Example trials. Original images at top and on diagonal; renders of objects from novel viewpoints off diagonal.

Implementation. Using the formulation outlined above, we model four stimulus conditions from Bonnen et al. (2021). For each trial, we infer object-level properties of each image independently, using an open-source implementation (He & Wang, 2023) of Large Reconstruction Model (LRM; Hong et al. (2023)). We estimate the relative viewpoints between pairs of images via an exhaustive search: given the object latents inferred from image_0 using LRM, we render images of object_0 sampled uniformly from a sphere of fixed radius, searching for camera coordinates that minimize the distance between this rendered image and ground truth (e.g., between object_0 from viewpoint_1 and object_1 from viewpoint_1). We refine our grid search over 3 iterations, such that each iteration more densely samples from the region which contained the viewpoint with the highest similarity between the rendered image and the reference image. We determine image similarity using several metrics: L2 over pixels, cosine similarity of deep features from CLIP and DINOv2, and LPIPS (Zhang, Isola, Efros, Shechtman, & Wang, 2018). This procedure with each trial yields a 4x4 matrix of images which we visualize in Fig. 1: each row corresponds to a model inferred from one of four reference images, while each column corresponds to the viewpoint inferred from those same reference images. The diagonal contains a model rendered from the original image it was inferred from (e.g., model_0 inferred from image_0), while the off-diagonal contains images of each model rendered from the viewpoint of other images (e.g., model_0 inferred from image_1).

Evaluation. We compare the accuracy of LRM to human performance and a suitable computational baseline. Given that the base encoder to LRM is a powerful visual backbone (DINOv2), we take the baseline performance to be the accuracy of DINOv2 on the original trial images; for each trial, we pass all 4 images to the encoder, extract features from the last pool layer, compute the pairwise correlation between features from each image, and determine the oddity to be the item with the lowest mean off-diagonal correlation. For the LRM-based modeling results, we modify this analysis method, computing the correlation not between the original images, but between the rendered images and the reference images (e.g., the correlation between model_0 rendered from the viewpoint in image_1 and the reference image, which is model_1 rendered from viewpoint_1). We again determine the object with the lowest correlation to the other objects as the model-selected oddity. Again we use several perceptual metrics (l2 over images, DINOv2 and CLIP features, and lpips). For all comparisons, we average across trials within each of four conditions. As expected, for the two conditions where DINO features perform well, so too does our generative modeling approach. However, these our LRM-based modeling approach does not achieve human-level performance on trials, failing exactly where the original DINO features do (Fig. 2, left). More concretely, our LRM-based modeling approach was able to achieve human-level performance on conditions that rely on texture-level properties of objects, the same stimuli where DINO features perform well, but not on those conditions that require 3D shape inferences. Notably, we find that these results are consistent across the different perceptual metrics used to estimate the relative viewpoints and determine the oddity in each trial (Fig. 2, right). These data suggest that objects generated using LRM are not well aligned with human inferences of 3D shape.

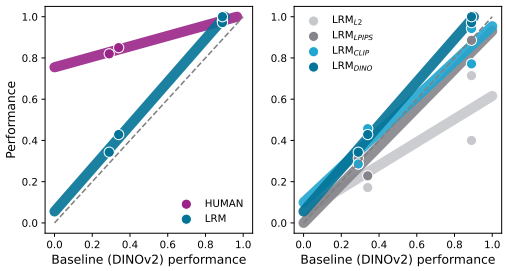

Figure 2: A generative model, LRM, (green) is not able to match human 3D visual inferences (purple; left), regardless of the perceptual metric (L2, lpips, and DINOv2, CLIP; right).

CONCLUSION

We developed a framework to evaluate the perceptual alignment between humans and generative visual models. An image-conditioned generative model, LRM, is not able to achieve human-level performance on 3D shape inferences. Our extensible approach will be useful for evaluating the align-

ment between humans and future generative visual models.

References

Abbas, A., & Deny, S. (2023). Progress and limitations of deep networks to recognize objects in unusual poses. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 37, pp. 160–168).

Alcorn, M. A., Li, Q., Gong, Z., Wang, C., Mai, L., Ku, W.-S., & Nguyen, A. (2019). Strike (with) a pose: Neural networks are easily fooled by strange poses of familiar objects. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 4845–4854).

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.

He, Z., & Wang, T. (2023). *Openlrm: Open-source large reconstruction models.* `https://github.com/3DTopia/OpenLRM`.

Hong, Y., Zhang, K., Gu, J., Bi, S., Zhou, Y., Liu, D., ... Tan, H. (2023). Lrm: Large reconstruction model for single image to 3d. *arXiv preprint arXiv:2311.04400*.

Jaini, P., Clark, K., & Geirhos, R. (2023). Intriguing properties of generative classifiers. *arXiv preprint arXiv:2309.16779*.

Koenderink, J. J., Van Doorn, A. J., & Kappers, A. M. (1992). Surface perception in pictures. *Perception & psychophysics*, *52*, 487–496.

Marr, D. (2010). *Vision: A computational investigation into the human representation and processing of visual information*. MIT press.

Shepard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science*, *171*(3972), 701–703.

Ullman, S. (1979). The interpretation of structure from motion. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, *203*(1153), 405–426.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. In *Cvpr.*